POLYTECHNIQUE MONTRÉAL
affiliée à l'Université de Montréal

Computational Profile Likelihood for fairness evaluation and correction of deep neural network classifiers

BENJAMIN PROSPER PAUL DJIAN
Département de génie informatique et génie logiciel

Mémoire présenté en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*
Génie informatique

Avril 2025

POLYTECHNIQUE MONTRÉAL
affiliée à l'Université de Montréal

Ce mémoire intitulé :

Computational Profile Likelihood for fairness evaluation and correction of deep neural network classifiers

présenté par **Benjamin Prosper Paul DJIAN**
en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*
a été dûment accepté par le jury d'examen constitué de :

Mohammad HAMDAQA, président
Ettore MERLO, membre et directeur de recherche
Sebastien GAMBS, membre et codirecteur de recherche
Heng LI, membre

DEDICATION

All we have to decide is what to do with the time that is given us

ACKNOWLEDGEMENTS

I could never have completed this work alone, and I am very fortunate to be surrounded by people who encouraged me throughout my journey.

I wish to express my gratitude to my research director Ettore Merlo for his guidance, pedagogy, and constant availability despite his busy schedule. I've made much progress under his supervision. I am also grateful for the valuable feedback and expertise provided by my co-director Sébastien Gambs, without whom this work could not have been done.

I want to acknowledge the DEEL (DEpendable & Explainable Learning) project and all those involved, as well as the CRIAQ (Consortium de Recherche et d'Innovation en Aérospatiale au Québec) and the NSERC (Natural Sciences and Engineering Research Council of Canada).

I thank Professors Mohammad Hamdaqa and Heng LI for accepting to evaluate my work.

To my parents, I offer heartfelt appreciation for their unwavering belief in me and their continued support despite the distance. Thank you, Nathan, Raphaëlle for being here, you remind me how precious a family is.

I want to express my gratitude to Alexane, who has always been there for me, both in good times and challenging moments.

I extend my gratitude to Guillaume and Catherine for their hospitality and their constant kindness.

Finally, I express my gratitude to my friends, this job would have been very different without them.

* * *

Je n'aurais jamais pu terminer ce travail seul, et je suis très chanceux d'être entouré de personnes qui m'ont encouragé tout au long de ce parcours.

J'aimerais exprimer ma reconnaissance envers mon directeur de recherche Ettore Merlo, pour son encadrement, sa pédagogie, et sa disponibilité constante malgré son agenda chargé. J'ai beaucoup progressé sous ta supervision. Je remercie également mon co-directeur Sébatien Gambs pour ses retours précieux et son expertise, et sans qui ce travail n'aurait pas pu voir le jour.

Je voudrais remercier le projet DEEL (DEpendable & Explainable Learning) et toutes les personnes qui en font partie, ainsi que le CRIAQ (Consortium de Recherche et d'Innovation

en Aérospatiale au Québec) et le NSERC (Natural Sciences and Engineering Research Council of Canada).

J'adresse mes chaleureux remerciements aux professeurs Mohammad Adnan Hamdaqa et Heng LI pour avoir accepté d'évaluer mon travail.

Je remercie mes parents qui ont toujours cru en moi et m'ont toujours supporté, spécialement ces dernières années, malgré la distance. Merci Nathan, Raphaëlle d'être là, vous me rappelez à quel point une famille est précieuse.

Je tiens à remercier Alexane, qui n'a jamais hésité une seconde à me supporter, dans les bons comme dans les mauvais moments.

Merci, Guillaume et Catherine pour votre hospitalité et votre gentillesse constante.

Enfin, je remercie mes amis, ce travail aurait été très différent sans vous.

RÉSUMÉ

L'apprentissage machine a gagné en popularité ces dernières années, et se retrouve appliqué dans de plus en plus de secteurs, notamment dans des domaines critiques, comme la défense, la santé ou les transports. L'entraînement de modèles d'apprentissage requiert de grandes quantités de données, dont certaines peuvent présenter des biais propres à l'humain et aux sociétés. En conséquence, il est crucial d'assurer que ces modèles ne reproduisent ni n'amplifient ces schémas néfastes.

L'objectif de ce travail est d'étudier en détail la structure des réseaux de neurones profonds pour tenter de quantifier la discrimination des prédictions associées. Plus précisément, dans le cadre de la classification binaire, nous étudions la distribution statistique des niveaux d'activation des neurones d'un classificateur, et nous les comparons entre individus possédant des caractéristiques sensibles différentes (par exemple entre les hommes et les femmes).

Pour cela nous utilisons une mesure, appelée *Computational Profile Likelihood* (CPL), qui est initialement définie dans le cadre de la détection de données situées hors des distributions rencontrées lors de l'entraînement de modèles. Nous exploitons cet outil pour évaluer la vraisemblance qu'un individu appartienne à une classe de sortie donnée, en fonction des niveaux d'activation internes du réseau. Ainsi, cette mesure permet d'avoir une intuition sur le "raisonnement" interne du classificateur, et fournit des informations essentielles pour comprendre les biais qui pourraient exister dans ce modèle. Contrairement à de nombreuses méthodes de détection de biais, CPL ne nécessite pas l'entraînement d'un second modèle, et repose plutôt sur des estimations non-paramétriques de densités. CPL peut s'appliquer à n'importe quel modèle de réseaux de neurones profonds. En ce sens, ce travail propose de compléter la littérature de l'équité algorithmique avec une stratégie robuste d'évaluation et de correction de biais

Dans un premier axe, nous comparons les distributions de CPL entre différents groupes démographiques de trois bases de données différentes. Nous validons expérimentalement cette approche sur des groupes définis par un attribut sensible comme l'ethnie, l'âge ou le genre. Pour cela, on entraîne un réseau de neurones sur chaque base de données, et on trace les distributions de CPL. Les résultats expérimentaux montrent qu'il existe effectivement la plupart du temps de larges différences entre les groupes sensibles, ce qui suggère l'existence d'un biais dans le réseau. Ensuite, nous appliquons aux données des techniques de mitigation de biais avant d'étudier les distributions de CPL. En particulier nous appliquons séparément deux méthodes : un "suppresseur de corrélation" et un "suppresseur d'impact disparate".

Les différences entre groupes sensibles existent toujours, mais sont beaucoup moins présentes qu’avant. Ces résultats montrent que ces techniques sont assurément efficaces pour réduire un biais négatif existant, sans toutefois le supprimer complètement.

Dans un second axe, nous utilisons le CPL comme un indicateur pour corriger les prédictions biaisées d’un modèle. En se basant sur trois définitions de l’équité dans les problèmes de classification tirés de la littérature, nous créons une nouvelle méthode de post-traitement capable d’appliquer ces définitions à n’importe quels réseaux de neurones profonds, tout en préservant une performance acceptable. Des expériences menées sur trois bases de données comparent cette méthode avec deux méthodes de post-traitement : la “classification avec option de rejet” et l’“optimiseur de seuil”. Les résultats expérimentaux montrent que cet algorithme est meilleur que ces deux méthodes concurrentes sur deux des trois bases de données, à la fois en termes de précision de classification et d’équité. Plus généralement, ces constats nous invitent à considérer les régions où les différentes méthodes s’accordent pour atteindre une correction robuste des biais.

En résumé, ce document présente deux nouvelles manières d’appliquer le CPL au champ de l’équité algorithmique. Il démontre que les classificateurs possèdent des biais relatifs à des caractéristiques sensibles des individus sur des données réelles. Il prouve que l’application de certaines méthodes de prétraitement de données permet de réduire les différences de CPL, et ainsi de diminuer ces biais. Enfin, il présente une nouvelle méthode de post-traitement des décisions d’un modèle afin de faire respecter des contraintes d’équité définies par la littérature de l’équité algorithmique. Ces résultats prometteurs encouragent à appliquer CPL au champ de l’équité algorithmique dans de nouvelles situations, comme dans l’équité intersectionnelle, ou les problèmes de classification multi-classes.

ABSTRACT

Machine learning has gained significant popularity in recent years in a growing number of sectors, including critical areas such as defense, health and transport. The training of these models use large amounts of data, some of which may present human and societal biases. Therefore, it is crucial to ensure that these tools do not perpetuate or exacerbate these harmful patterns.

This work aims to study in detail the structure of deep neural networks to detect and address injustices in the decision-making process. Specifically, in the context of binary classification, we analyze the statistical distribution of the levels of activation of the neurons of a classifier, and we compare them between individuals with different sensitive characteristics (for example, between men and women).

We employ a measure called *Computational Profile Likelihood* (CPL), initially designed to detect out-of-distribution inputs. We harness this tool to assess the likelihood that an individual belongs to a specific output class, based on internal activation levels of the network. This tool offers insights into the "reasoning" of the classifier's decisions, providing valuable context for understanding potential biases. Unlike many bias detection methods, CPL does not require additional model training, relies on nonparametric density estimations, and can be implemented in any deep neural network model. From this perspective, this work aims to complete the algorithmic fairness literature with a robust bias evaluation and correction strategy.

In the first axis, we compare CPL distributions between different demographic groups of three databases. We validate this approach experimentally on groups defined by a sensitive attribute such as ethnicity, age, or gender. For this, neural networks are trained on each database, and the CPL distributions are plotted. The findings reveal that significant disparities exist between sensitive groups most of the time, which suggests the existence of a bias in the network. Following this observation, we apply bias mitigation techniques to the data before re-examining CPL distributions. Specifically, we employ two methods: correlation remover and disparate impact remover. Differences between sensitive groups still exist, but are much less present than before. These results show that these techniques are certainly effective in reducing an existing bias, but not eliminating it.

In a second axis, we exploit the CPL as an indicator for correcting model bias. Based on three definitions of fairness in classification problems from the literature, we create a new post-processing bias mitigation approach, capable of applying these definitions to any deep neural network, while maintaining acceptable performance. Experiments conducted across three databases compare this method with two other post-processing approaches: *Reject Option Classification* and *Threshold Optimizer*. Experimental results show that this strategy is better regarding precision and fairness on two of the three databases. More generally, these findings invite us to consider areas where multiple methods agree to achieve robust correction of biases.

In summary, this study presents two new ways to apply CPL to algorithmic fairness. It demonstrates with real examples that classifiers can harbor biases related to sensitive characteristics of individuals. It shows the effectiveness of certain data preprocessing methods for reducing CPL differences and thus decreases these biases. Finally, it presents a new method of post-processing the decisions of a model to enforce fairness constraints defined by the literature. These promising results encourage the application of CPL to the field of algorithmic fairness in new situations, such as intersectional fairness or multi-class classification problems.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ACRONYMS

AO	Average Odds
CP	Computational Profile
CPD	Computational Profile Distance
CPL	Computational Profile Likelihood
DP	Demographic Parity
EO	Equality of Opportunity
LR	Low Revenue
HR	High Revenue
ML	Machine Learning
NN	Neural Networks
OOD	Out-Of-Distribution
ROC	Reject Option Classification
TO	Threshold Optimizer

LIST OF APPENDICES

CHAPTER 1 INTRODUCTION

1.1 Motivation

Recently, machine learning techniques have been increasingly adopted in various fields, ranging from healthcare [1] and defense [2] to transportation [3]. However, the integration of machine learning models into sensitive sectors calls for thoughtful consideration. Among these concerns lies the imperative to ensure algorithmic fairness. Easy to grasp but hard to define, fairness is most of the time defined by the absence of unfairness or negative bias. A fair outcome would be determined by disregarding any bias regarding an individual's characteristics that are not pertinent to the specific decision-making context considered [4]. For instance, a study from 2015 [5] revealed alarming disparities regarding the exposure of online advertisements between genders: female profiles were less exposed to ads for high-paying jobs than male profiles.

Machine learning algorithms are characterized by their extensive reliance on voluminous datasets for training and optimizing predictive performance. However, the historical data used in these processes may unwittingly encapsulate societal and human value biases, manifesting as discrimination based on gender, age, or ethnicity. These historical biases pose a significant risk of perpetuating or exacerbating discrimination through algorithmic decision-making, thereby jeopardizing the very objectives these technologies aim to achieve. Moreover, recent research published in the journal Scientific Reports [6] reveals that individuals exhibit a heightened trust towards recommendations generated by algorithms as opposed to human advice. This phenomenon further underscores the necessity of addressing and mitigating biases within machine learning applications.

Over the past few years, substantial progress has been made in defining, identifying, preventing and rectifying the unfairness of such machine learning systems [7–9]. Analyzing the activation patterns in a neural network can be an effective approach to detecting and correcting potential biases. Although some studies have explored this direction [10–12], none have proposed a robust *post hoc* method that does not necessitate additional model training. This gap in research serves as the primary motivation behind the present work.

1.2 General Approach

The objective of this research is aligned with the broader goal of creating fairer and more robust machine learning algorithms. Specifically, it focuses on identifying and mitigating classifier biases. To that end, our study aims to take a closer look at the internal structure of neural networks, specifically examining how neuron activation patterns differ between sensitive demographic groups.

This work builds on the concept of *Computational Profile Likelihood* (CPL), first introduced by Merlo, Marhaba, Khomh, Braiek, and Antoniol [13]. Initially designed as a robust approach for identifying Out-Of-Distribution (OOD) inputs in the context of adversarial attacks, this method aggregates neuron activation levels, denoted as *Computational Profile* (CP), into histograms [14]. The *likelihood* of an input belonging to a particular output class is assessed by comparing its CP to the histograms of the output class.

Our work extends the applicability of CPL to the domain of algorithmic fairness. In particular, we present two applications of CPL in the domain of fair machine learning. Our approach's main idea is to construct CP histograms for each output class and scrutinize the differences between various demographic groups. For instance, in the context of a loan approval classifier that evaluates the probability that individuals get their bank loan *accepted* or *refused*, we compare separately men and women to the histograms of class *accepted* and class *refused*. Notable disparities between these groups indicate that the model's internal activation pattern differs between men and women. This suggests that the network processes individuals differently based on their gender, raising a serious question of fairness with respect to its prediction.

Our work is divided into two sections, which correspond to two methods to use CPL to detect the unfairness of fully connected neural networks:

- Section 4 concerns fairness evaluation, more precisely how to assert that a model has a negative bias and to what extent. In addition, we investigate the impact of two bias mitigation methods on CPL and evaluate how much these techniques can reduce differences in CPL distributions to achieve a higher fairness between demographic groups.
- Section 5 extents CPL to correction of unwanted bias. It introduces a novel way to handle discrimination with this tool. More precisely, we present a post-processing method to correct the negative bias of an already-trained model. This strategy is compared to existing thresholding techniques in this domain.

This document aims to present an in-depth analysis of CPL's effectiveness in addressing negative biases across various real-world fairness datasets.

1.3 Contributions

This work gathers several contributions. First, we introduce a novel method based on CPL to evaluate the fairness of deep neural networks. The effectiveness of this approach is evaluated through experimental analyses conducted on synthetic datasets, as well as on three widely studied datasets in fairness literature. Then, we experimentally compare two pre-processing bias mitigation methods: Correlation Remover and Disparate Impact Remover. Next, we introduce an innovative post-processing method based on CPL for group fairness, followed by a comparative evaluation with multiple post-processing algorithms across various datasets.

1.4 Outline

The outline of the paper is as follows. Chapter 2 provides an introductory overview of machine learning, and details the concerns regarding algorithmic equity. It also presents Computational Profile Likelihood. Chapter 3 discusses its application in the context of algorithmic fairness, and presents preliminary results of various experiments. Chapter 4 is a paper submitted to the special issue “Responsible and Reproducible Machine Learning” of Computational Intelligence. It presents an approach using CPL to detect discrimination with respect to model predictions and to measure the effectiveness of pre-processing bias mitigation strategies. Next, Chapter 5 presents a new algorithm, CPD Thresholding, to correct model bias towards more group fairness. Finally, Chapter 6 recaps our work, underlines its limitations, and discusses potential future research avenues.

CHAPTER 2 BACKGROUND

This section aims to provide introductory definitions as well as the contextualization of our work. Basic notions of machine learning, deep learning and algorithmic fairness are presented. Then, we give an overview of Computational Profile Likelihood and the intuition behind it.

2.1 Neural Networks and Deep Learning

Among all technologies deployed in Artificial Intelligence, Machine Learning (ML) algorithms are one of the most promising domains, due to their impressive performance in multiple fields, such as image classification or text generation. These algorithms operate by learning patterns from historical data, and generalize on unseen data effectively, without explicit human intervention. In the context of "supervised" learning (as opposed to "unsupervised" learning), each training instance consists of a couple composed of a vector representing an observation and its desired outcome. The objective is to identify a mapping, called *model* or *predictor*, associating an observation (an *input*) to its corresponding outcome (an *output*), while minimizing a *loss function.* When discussing statistical data related to people, we employ the term *individual* instead of observation.

As an example, in the application of predictive property valuations, the input parameters encompass attributes such as the land area, garden presence, and location information. The output corresponds to the estimated price of the house, and the loss function can be defined by the mean squared difference between forecasted prices and actual market values.

This approach relies on the hypothesis that the training data is representative of all the existing observations. The desired outcome can be either quantitative, in which case it is termed *regression*, or qualitative, designated as *classification.* In the following, we will focus exclusively on classification, and we use the term *classifier* to designate a predictor. We differentiate between multi-class scenarios and binary classification contexts, in which only two outcomes are considered. Considering binary classification, the two classes can be separated between the *positive* and the *negative classes.* The term "positive" refers to the output that the model seeks to identify or forecast, denoting the occurrence of a condition or an attribute while the "negative" class is complementary.

Neural networks (NN) are specific ML algorithms composed of interconnected units called *neurons*, which are arranged in numerous layers [15]. The initial layer is the *input layer*, where each neuron corresponds directly to a data feature. Subsequent layers are called *hidden layers.*

Deep Neural Networks (DNNs) are NNs with multiple hidden layers. The terminal layer is the *output layer*, which returns the final calculation performed by the network.

The primary operations executed by NNs involve *forward propagation* and *backpropagation* [15]. Forward propagation refers to the process where input data flows through the network, traversing from the input layer through one or more hidden layers to reach the output layer, yielding a numerical result based on the predictive task. Each neuron receives one or more inputs that are multiplied by the weights of the connections. These values are summed with an additional bias term. Subsequently, an *activation function* is applied to generate the *activation level* of the neuron.

Most of the time, the activation function is non-linear, permitting the NN to tackle complex tasks such as non-linear separable problems. The sigmoid function, hyperbolic tangent, and Rectified Linear Unit (ReLU) are some of the most used functions for this purpose [16]. The Leaky ReLU function is a variation of ReLU, defined by a slope parameter p, typically much smaller than 1. Compared to ReLU, this function mitigates the issue of "dying neurons", *i.e.* neurons that cease contributing towards the network's performance by having systematically null values [16]. In the following of this work, we will focus mainly on NNs with Leaky ReLU as activation function. Equation 2.1 details the definition of LeakyReLU :

$$LeakyReLU_p(x) = \begin{cases} px & x < 0 \\ x & else \end{cases} \tag{2.1}$$

The succession of multiple layers allows NNs to extract abstract features from inputs and effectively process high-dimensional data, such as images or extensive tables. A fully connected NN refers to a network in which each neuron of each layer is connected to every neuron of the preceding and subsequent layers. A visual representation of a DNN with two fully connected hidden layers appears in Figure 2.1. Three-dimensional input data are fed to the network in the input layer (left of the graph) and computations through the layers are done in the reading sense. The network returns two outputs after the processing of the output layer on the right of the graph.

The second operation performed by the NNs is the backpropagation, designed to improve model performance through iterative processing known as training [15]. Backpropagation consists of three distinct steps. Following each forward propagation on training data, the loss function assesses the network's performance by evaluating disparities between actual outcomes and predictions. The objective of training is to minimize this loss. Next, the gradient of the loss function is computed for every model parameter. Finally, an optimization

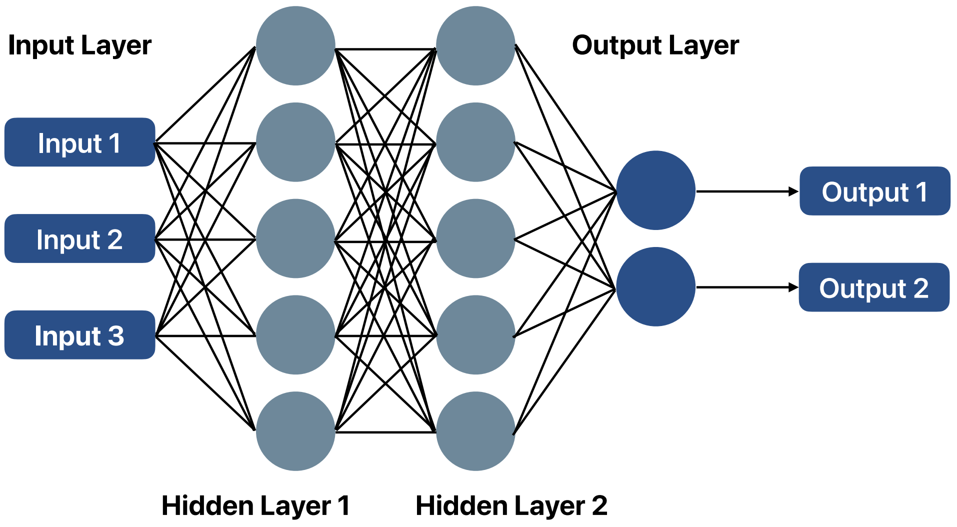


Figure 2.1 Representation of a Deep Neural Network architecture

algorithm such as stochastic gradient descent is employed to adjust each weight and bias in an opposite direction of their gradients by an amount specified by the *learning rate*.

For supervised learning tasks, such as classification, NNs are particularly well-suited due to their capacity to accept constant-sized inputs and generate fixed-sized outputs, such as a probability assigned to each possible class. For instance, in a five-class classification problem, the NN's anticipated outcome is one of the classes and an output layer of five neurons would return the probability that the associated input belongs to each of the classes.

2.2 Algorithmic Fairness for Machine Learning

Over the past decade, advancements in computational power, coupled with its widespread usage, have fueled the adoption and application of DNNs across diverse domains [17]. Concurrently, significant growth in storage capacity and cost reduction have resulted in unprecedented availability of massive datasets [18]. A new paradigm, called *Datafication* [19], describes the transformation of multiple domains into data, gathering and sharing larger and larger amounts of it. The technologies based on data, especially DNNs composed of huge numbers of parameters, have particularly profited from this trend.

However, the increasing complexity of such algorithms raises several concerns for two primary reasons: their opaque nature ("black-box" model) makes them challenging to understand, and they are increasingly being employed in a wide range of areas. To tackle issues related to these aspects, there has been a surge in research activity on emerging issues such as robustness (maintaining performance under perturbations and adversarial inputs) [20, 21], explainability (understanding the functioning of an algorithm) [22] and fairness [8, 9]. Our study aims to provide tools and methods to address discrimination that NNs may perpetrate.

During the life-cycle of AI models, unfairness can manifest at multiple steps, as identified by a recent review [23]. These steps include production bias, data bias, learning bias, and deployment bias. The first type of bias is also known as human bias and encompasses all the biases that humans are sensitive to, including cognitive and behavioral biases. Taking the example of hiring candidates for a software development role, an instance of human bias would be the preference towards male applicants over female ones. Data bias refers to the methods employed for collecting data as well as the selection of features. In this same context, a data bias could involve collecting a dataset with a significant amount of males compared to females. Learning bias arises during the model's training process, potentially resulting in the modeling of statistical patterns that favor majorities over minority groups. For instance, the model could learn that men are more often recruited than women for the position of software developer. The selection of evaluation metrics can also inadvertently perpetuate discrimination if not carefully chosen. Lastly, deployment bias occurs when a model is deployed in an environment different from its intended purpose. In our example, deploying the predictor to evaluate candidates for management positions could introduce unintended biases, as the predictor was designed primarily for evaluating developer candidate qualifications.

The study of production biases falls under the domain of psychology, whereas deployment biases are addressed through meticulous case-by-case analysis and within legal frameworks. Therefore, in our investigation into the NNs' functioning, our primary focus lies on the data and learning biases; however, we recognize the importance of acknowledging both production and deployment biases.

Although a mathematical and rigorous description of fairness is still a subject of discussion (with multiple characterizations available [24–26], some of which are discussed later), it is possible to give a consensual definition of this principle: "the absence of any prejudice or favoritism toward an individual or group based on their inherent or acquired characteristics" [27]. For instance, taking into account physical characteristics to formulate decisions about intellectual competence can be considered unfair.

Removing sensitive information of the individuals before applying the algorithm's prediction process is a simple yet naive approach for improving fairness. This idea, called "fairness through unawareness" [28] ignores the capacity of the algorithms to identify hidden patterns, often in unexpected ways. For instance, removing the attribute “race” of the individuals but keeping the location information, such as the ZIP code, overlooks the fact that some neighborhoods may be inhabited by a higher number of individuals of a specific race than other neighborhoods, thus hinting the model on individuals’ race [29]. In this situation, the “ZIP code” is called a *proxy* and thus, the existence of proxies attributes calls towards more refined strategies. On the contrary, deleting all attributes, including those relevant to classification, is not a viable strategy either. Research in this area [30–32] has expressed the fairness-utility trade-off as the idea that if fairness is enforced, then “utility” (*e.g.*, accuracy) decreases.

For simplification purposes, we encapsulate the sensitive information within a single binary attribute, known as the *sensitive attribute* [33] or the *protected attribute*. This attribute can be created from the combination of multiple sensitive attributes or the thresholding of a continuous variable. Groups of individuals defined by a sensitive attribute are called sensitive groups. In the context of an undesirable bias, we differentiate between the *privileged group*, which benefits from the bias, and the *unprivileged group*, consisting of individuals suffering from the bias.

The notions of “biased model” or “biased data”, while seemingly straightforward, may lead to misinterpretation if not explicitly defined. A recent study of bias in neural networks [34] proposes to split the concept of bias into two separate definitions. The first type of bias occurs when uneven performance is observed across different groups. The second type of bias arises when the NN relies on sensitive information to formulate a prediction. This classification has the advantage of grouping many fairness metrics into two distinct categories, that include one of the most used fairness paradigms called *group fairness*. In the literature, several approaches have been proposed to improve fairness. In addition to group fairness, *individual fairness* and *causal fairness* are two other fairness paradigms.

2.2.1 Individual Fairness

Individual fairness considers that individuals with comparable characteristics should receive equivalent treatment from algorithmic predictions [34]. Concretely, this notion evaluates similarities between profiles using metrics, ensuring consistent decisions for alike individuals. In mathematical terms, individual fairness can be defined such that the distance between the mapping of any couple of individuals is smaller than the distance between these individuals, in terms of specified distance metrics. This property of the mapping, called the Lipschitz property, is the one proposed in the work of Dwork, Hardt, Pitassi, Reingold, and Zemel [24]. In other words, people who look alike (“close” to each other) should get very similar treatments by the mapping.

Causal discrimination [35, 36] is another approach used in individual fairness testing: identical individuals that only differ by a sensitive attribute should receive identical outcomes. To tackle the fairness-utility trade-off, Li, Wu, and Su [37] extend the definition of causal discrimination and propose *accurate fairness*. A classifier is considered as accurately fair if the distance between the ground truth of an input x and the prediction of any inputs differing from x at most only by a sensitive attribute is lower than the distance between these two inputs, multiplied by a constant K. As an illustrative example, consider an accurately fair model designed to forecast whether a bank customer possesses the ability to repay a loan. In this context, the algorithm should yield extremely similar outcomes for male and female clients exhibiting identical characteristics. Additionally, for any given individual, the difference between the prediction output and the actual outcome is precisely zero, signifying that the predictor accurately predicts outcomes. This definition not only adheres to the Lipschitz property of individual fairness but also prioritizes prediction accuracy, as the outcome for input should match its ground truth.

Even if individual fairness definitions promise strong guarantees, implementing one of these in practice remains challenging, in particular, to build relevant metrics [38], which often requires domain experts. Thus, the majority of studies about bias in DNNs revolve around group fairness [39], which presents criteria easy to verify and enforce.

2.2.2 Group Fairness

In contrast to individual fairness, group fairness considers that populations defined by sensitive information should benefit to the same extent from the model predictions. Statistics are computed for each group, then are compared across all the groups [39].

Group Fairness Metrics

Enhancing fairness in NN models is a complex endeavor, as multiple interpretations exist. Indeed, fairness is not a clear and easy concept to grasp as it depends on sociocultural norms and existing legislation. Consequently, various metrics have been defined. Among these, *Demographic Parity* (DP), also called *statistical parity* is a commonly used metric in group fairness [40], which requires independence between model predictions and sensitive attributes. We consider the context of binary classification of two output classes $Y \in \{0, 1\}$ (designated by *favored* and *unfavored* class). The sensitive attribute Z can take values in $\{0, 1\}$, thus defining a privileged group and an unprivileged group. The formal definition of Demographic

Parity is :

$$P(\hat{Y}=1|Z=1)=P(\hat{Y}=1|Z=0) \tag{2.2}$$

In the context of a biased algorithm, a disproportionately larger number of inputs belonging to the privileged group are assigned positive outcomes compared to those from the unprivileged group. Remark that a perfect classifier only achieves the DP criterion if the outcome classes are perfectly balanced between protected and non-protected groups.

This definition of group fairness neglects any potential correlations between the protected attribute and the prediction task. As a result, satisfying DP is achievable through operations that can be perceived as unfair. Specifically, choosing suitable individuals from a specific protected group and selecting random individuals from the other group ensures compliance with the DP constraint while fostering discrimination between protected groups. For instance, recruiting 10% of the most qualified female applicants for a job, and recruiting a random proportion of 10% of all the male applicants for the same job is discriminating against women, while respecting the DP constraint ($P(\hat{Y}=recruited|Z=woman)=P(\hat{Y}=recruited|Z=man)=10\%$).

To remedy this issue, Hardt, Price, and Srebro [25] have introduced the notion of *Equality of Odds*, which is satisfied for any y in $\{0,1\}$, when:

$$P(\hat{Y}=y|Y=y,Z=1)=P(\hat{Y}=y|Y=y,Z=0) \tag{2.3}$$

Equality of Odds focuses on the ground truth labels of classification results across demographic groups. This condition imposes a stringent requirement by ensuring parity between true positive and negative rates for each group. Unlike DP, Equality of Odds does not allow “positive discrimination”, which grants artificial advantages to underrepresented individuals.

In the context of binary attributes, one may consider only discrimination towards a “favored” group to minimize the data alteration. This leads to a relaxed version of Equality of Odds, called *Equality of Opportunity* (EO) [25] in which there is an equality of true positive rates among privileged and unprivileged groups:

$$P(\hat{Y}=1|Y=1,Z=1)=P(\hat{Y}=1|Y=1,Z=0) \tag{2.4}$$

Finally, *Average Odds* (AO) is defined as the balance between the average of true positive and false positive rates among the privileged group and the unprivileged group. When the average odds criterion is reached, then both groups benefit to the same extent from the classification:

$$\frac{P(\hat{Y}=1|Y=1,Z=1)+P(\hat{Y}=1|Y=0,Z=1)}{2}=\frac{P(\hat{Y}=1|Y=1,Z=0)+P(\hat{Y}=1|Y=0,Z=0)}{2} \tag{2.5}$$

When considering real-world situations, a compromise between fairness and performance is often sought, and perfect fairness is very rarely reached [8]. In the following, we define fairness metrics as the gap to fill to reach perfect equality in the definitions presented above in Equations 2.2, 2.3, 2.4 and 2.5. The closer to zero the metrics are, the better the fairness. In contrast, the further away from zero the metric is (with a maximum norm of 1), the more discriminating the model is.

Bias Mitigation Strategies for Group Fairness

There exist various methods for improving fairness, which can be broadly categorized depending on where they operate during the machine learning pipeline. First, *pre-processing* methods alter the data before the model is trained in an attempt to reduce the discrimination while *in-processing* strategies adapt the training procedure itself to minimize potential biases. Finally, *post-processing* approaches change the model predictions *a posteriori* to improve the model fairness.

Several recent studies in ML fairness research have proposed novel approaches for correcting differential treatments between demographic groups [9,41,42]. Among them, the work of Celis and Keswani [43] uses Generative Adversarial Networks (GANs) to train a fair model, with a modified gradient update algorithm. Two metrics are considered to evaluate the equity of the model: the DP and the false discovery rate [44]. The authors also explore the theoretical definition of the *price of fairness* for their algorithms, which represents the smallest amount of classification loss necessary to reach perfect fairness with respect to the criterion chosen. In contrast to strategies based on GANs, our method does not require the training of an auxiliary model and is based on distribution estimates.

Other works suggest pre-processing the training data in various ways, such as by sampling [45, 46], reweighing [47] or relabeling [48, 49]. For instance, Kamiran and Calders [47] suggest changing the labels of data before model training to make the data less biased with respect to the DP difference. More precisely, four fairness-aware methods are proposed, ranging from deleting features correlated with sensitive attributes to changing the label of well-chosen inputs.

Alternative *in-processing* approaches [50–53] modify the training process instead of the data to achieve fairness. For instance, multiple works [50, 53] modify the objective function of the model, either to enforce the chosen fairness criterion or to reduce the mutual information between the sensitive attribute and the model prediction. Other studies prefer reweighting classifiers over regularization [49, 54]. In particular, Jiang and Nachum [49] describe a procedure to learn the weights of training inputs such that the resulting training corresponds to training on unbiased labels. The fairness criteria studied are the DP, EO, equality of odds, and disparate impact.

In addition to data modification and model alteration, another strategy to tackle unwanted bias is to correct the model predictions while avoiding too much alteration or performance degradation. The review of Caton and Haas [9] distinguishes various families of these post-processing methods. For instance, *calibration* aims to give the same meaning of "risk" for the privileged and unprivileged groups by ensuring that the proportion of positive predictions is equal to the proportion of positive examples. *thresholding* assumes that biased predictions are made near the decision boundary, focusing on altering these ambiguous predictions to achieve a good accuracy and fairness across all classes and groups. Finally, *transformation* learns a new representation of the data, which is fairer than the original one. However, it works mainly on numeric data, thus greatly limiting its range.

For instance, the work of Calders and Verwer [45] describes three ways to reduce the discrimination of a naïve Bayes classifier. One of them is a post-processing method that ensures that the classifier produces the same number of positive and negative outcomes as before the correction. For instance, this condition is desirable if a classifier aims at predicting whether a loan is granted, in which case the bank might want to keep the same number of loans before and after the correction. In a nutshell, this method works iteratively. While the DP difference is above zero, if the number of predicted positive labels is under the real number, the number of deprived inputs with positive outcomes is increased while the number of deprived inputs with negative outcomes is decreased. Then, if the number of predicted positive labels is above the real number, we modify similarly the privileged population, with the corrected Bayes classifier being updated accordingly. This method uses the probability distribution $P(Z|Y)$ to adjust the classifier. However unlike the CPD, internal activation patterns are not considered, which may be less reliable for individuals outside of the training distribution.

Reject Option Classification (ROC), initially proposed by Kamiran, Karim and Zhang, is another *thresholding* algorithm that defines a critical region in the input space in which model predictions are considered as ambiguous [55]. This region can reduce discrimination, by systematically assigning the privileged class to the unfavored group and the unprivileged class to the favored group. This region is defined around a well-chosen decision boundary, and its width is selected to reach the preferred fairness criterion. Outside the critical area, the standard classification rule observed by the model is followed. The idea here is to trust the model, except when its predictions are uncertain. An explanatory diagram is presented in Figure 2.2. ROC considers model scores to perform its correction and ignores internal activation information. On the contrary, our CPD correction considers activation levels of neurons to choose individuals to swap classes and has a similar computational cost.

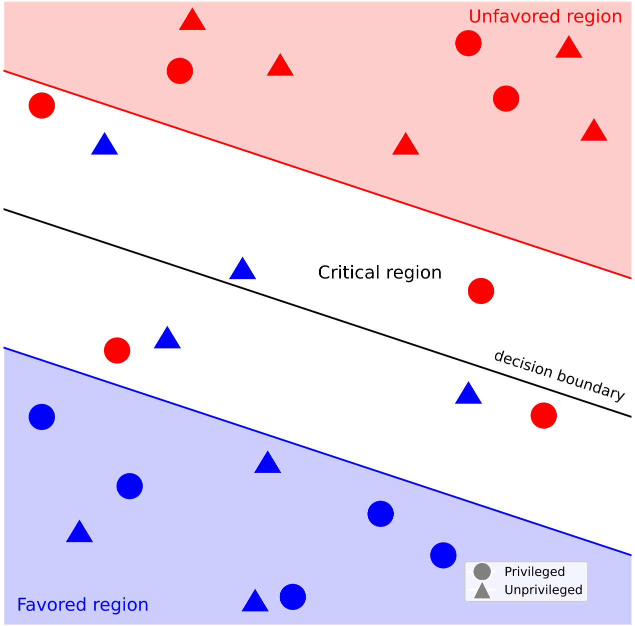


Figure 2.2 Simplified diagram of input space when ROC method operates

More generally, the predominant approach of *Thresholding* revolves around adjusting classification thresholds based on model scores towards a fairness objective. These strategies ignore that such scores can not be reliable in situations where individuals differ vastly from previously encountered distribution, and corner cases. Techniques based on CPD consider instead structural information of the classifier to adapt the thresholds classification, without needing additional computational expense, such as the training of a secondary model. Similarly to the methods presented here, the CPD correction can be applied *post hoc* to any classifiers.

ROC has a variant that applies to an ensemble of classifiers. This distinct paradigm for post-processing corrections considers not only one classifier but multiple at the same time, allegorizing an admission committee, composed of numerous members. Using the predictions of such an ensemble of classifiers could reveal which individuals are misclassified due to discrimination. Discrimination-Aware Ensemble [55] is the method exploiting this idea: if multiple classifiers agree on the output class of an input, then we consider the input correctly classified. Otherwise, the favorable class is assigned to unprivileged inputs and vice-versa.

The work of Iosifidis, Fetahu, and Ntoutsi presents the framework Fairness Aware Ensemble [46] that combines a pre-processing technique with a post-processing method. The last one consists of adjusting the classification thresholds of the unprivileged and the privileged groups separately until reaching an EO difference low enough. Misclassified individuals with positive outcomes are sorted based on a *ensemble classification score*, and thresholds for privileged and unprivileged groups are modified to reach EO with a precision chosen by the practitioner. This *ensemble classification score* is a weighted sum of the predictions returned by weak learners. These classifiers result from successive bagging and boosting. Fairness Aware Ensemble, similarly to Discrimination-Aware Ensemble, applies the paradigm of multiple classifiers for a fair classification. Unlike the post-processing method introduced in this document, this strategy does not use CPL as a criterion and favors multiple weak learners instead of a single network. The idea of ensemble classifiers is not pursued in this study, although it represents a valid direction to extend future research, considering both the CPD and the model scores as threshold-adjusting criteria.

In their work, Hardt, Price, and Srebro [56] define another *Thresholding* post-processing algorithm that learns two separate decision thresholds for unprivileged and privileged groups and apply these specific thresholds during inference. A derived predictor is learned from a biased predictor by setting the four conditional probabilities $P(\hat{Y} = 1|Y = a, Z = a')$ for $a, a' \in \{0; 1\}$. The predictor achieves perfect EO based on naive randomization of a portion of individuals. However, randomization meets limitations, such as information wastefulness or intra-group unfairness. Our CPD correction does not use a randomization technique and relies solely on CPD scores, measuring the likelihood of an individual belonging to an outcome class.

Fish, Kun, and Lelkes [57] aim to optimize the trade-off between fairness and utility by presenting a method called *Shifted Decision Boundary.* The shift introduced to the decision boundaries can be modified after model training to obtain the desired balance between accuracy and bias.

These thresholding strategies do not use structural information about the classifier and assume that discrimination predictions are made near the decision boundary. However, these approaches can overlook corner cases in which individuals differ significantly from the distributions encountered during training and could be wrongly classified in a class with high confidence in the model. Previous works on CPL have demonstrated its effectiveness in handling Out-Of-Distribution inputs. By exploiting the structural information of the network, CPL-based methods avoid the common pitfalls of such bias correction methods.

Group fairness provides fairness definitions easy to understand and implement in multiple contexts. However, a new strategy to address equity issues, called causal fairness, has emerged in response to the lack of causal explanations of model biases of previous methods.

2.2.3 Causal Fairness

More recently, a third paradigm has emerged from the algorithmic fairness literature. Although statistical approaches are easy to implement, they rely on observations and measurements of demographic groups, which can be sensible to statistical anomalies [58]. Furthermore, from a legal and philosophical setting, discrimination claims require the demonstration of causality between the decision and the sensitive attribute [59]. For these reasons, causality has been considerate appropriate to tackle unfairness issues.

As opposed to the previous approaches, causal-based fairness considers additional information about the context to understand how data is generated and to understand the causal structure of the problem. Causal reasoning can help identify which discrimination is harmful, and which discrimination is worth keeping [60]. These methods are very effective at providing explanations on how to repair a damaged dataset [61].

Causal fairness introduced the measure called *counterfactual fairness* [60], which relies on the causal discrimination idea. This metric measures to which extent is it possible to construct two identical predictions on two sensitive groups, using all the variables that are not caused by sensitive information.

Other concepts have been defined, such as *unresolved discrimination* [62], that occurs if the predicted label is dependent on at least one variable influenced by sensitive information but not considered as discriminatory. A causal model is free from *Proxy discrimination* if the target attribute is not influenced by any proxy attributes.

Using causal reasoning is useful to enhance fairness directly on the data (by removing proxy attributes for instance), before any classifier training. In particular, causal relations between attributes can be used to strategically modify the training data to reduce bias [61].

The main limitations for causal fairness are the lack of accessibility of background information needed for causal models to be defined, as well as a lack of convenience in their applications [63].

Our study applies a statistical method based on histograms of distributions to evaluate the likelihood that individuals belong to a specific class. In this context, we focus essentially on group fairness.

Algorithmic fairness is closely linked to explainability: transparent decision-making facilitates the evaluation of fairness. Nevertheless, most definitions of group fairness rely solely on the outcome of the model predictions, thereby disregarding the processes leading to these outcomes [64]. Consequently, approaches aimed at achieving fairness are restrained to result-driven mitigation, while neglecting the structural aspects of the model itself.

In this perspective of exploiting structural information about models, we provide an introductory exposition to Computational Profile Likelihood [13] in the following subsection.

2.3 Introduction of Computational Profile Likelihood

When considering fairness for NNs, many studies have proposed diverse strategies for defining, detecting, constraining, and rectifying unwanted bias. Among these strategies, examining the internal activation pattern of networks is one possible method. Following the work of Merlo, Marhaba, Khomh, Braiek, and Antoniol [13], we use the term *Computational Profiles* to denote the vectors that encapsulate neuron activation levels at different layers of a neural network. These vectors can be extracted during both the training and inference phases. The underlying principle behind Computational Profile Likelihood (CPL) relies on constructing an estimation of the distribution of the activation levels experienced by the neurons during training, followed by a comparison to the activation levels during inference. Such comparisons can tell if newly encountered inputs are Out-Of-Distribution inputs or if they are just regular inputs.

Figure 2.3 and 2.4 schematize in detail how to extract CPL from a trained DNN. First, Figure 2.3 details how estimations of the distribution of activation levels are created. The training data is fed into the network. For each input correctly classified by the model and each neuron of the last hidden layer, we record corresponding activation levels. For each neuron, we construct a histogram of the distribution of activation levels recorded.

Once these histograms are constructed, Figure 2.4 shows how to compute the CPL of an input x_{test}. The input is fed into the model, and the activation levels of each neuron of the last hidden layer are recorded. For each neuron, we record the bin of the corresponding histogram where the activation level “falls”. The ratio between the bin size and the size of all the histograms gives a probability for each neuron. We define the CPL of x_{test} as the product of all of these probabilities.

CPL focuses exclusively on the neurons of the last hidden layer for two main reasons. This layer encapsulates previous layers’ network computations, and considering only one layer can significantly reduce computation times, depending on the model architecture considered.

Details about the mathematical definition and calculations of CPL can be found in Section 4.4.

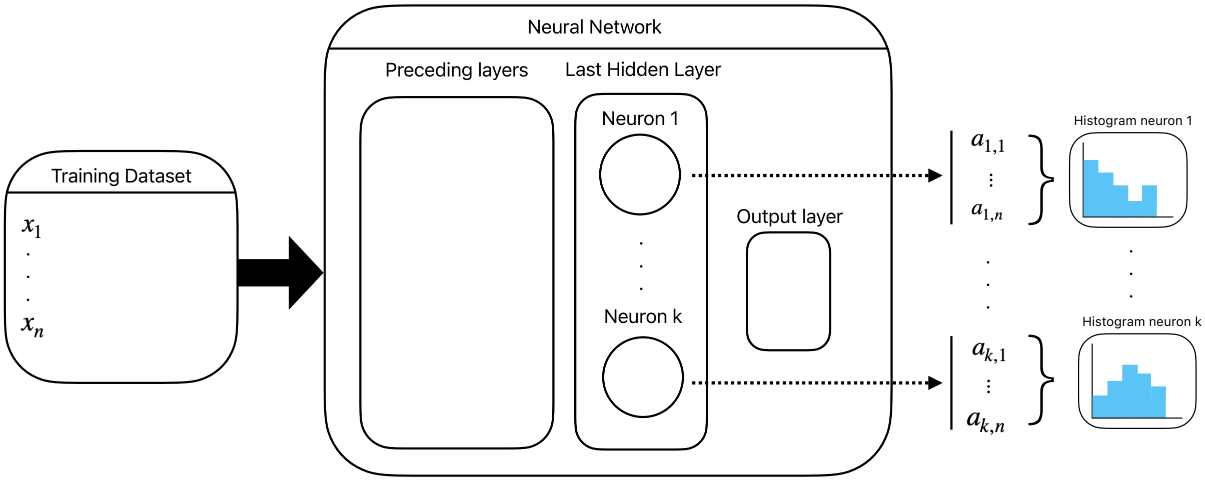


Figure 2.3 Building activation levels histogram estimations

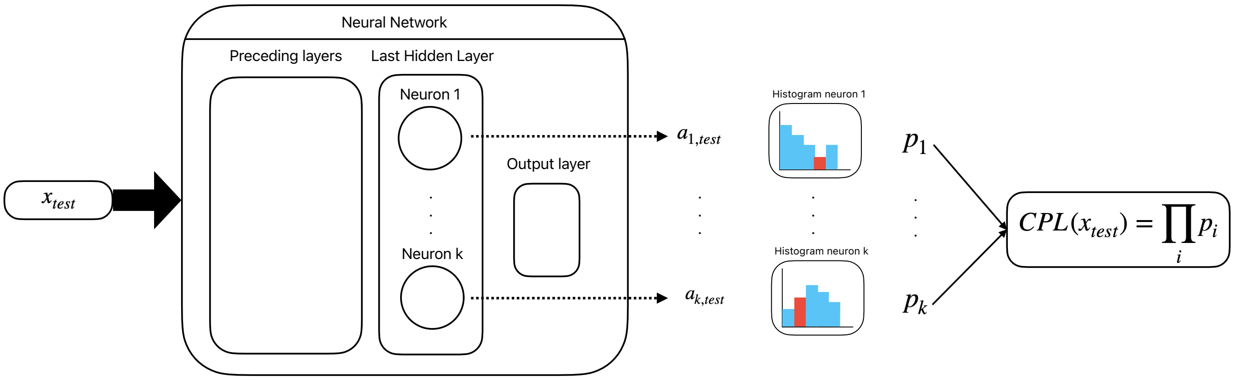


Figure 2.4 Extracting CPL

To sum up, CPL quantifies the extent to which inputs' internal pattern activation diverges from those encountered during training. The greater the disparity between an input activation pattern and the training activation pattern, the lower the probabilities associated with each neuron are, resulting in a low CPL. Conversely, inputs exhibiting activation levels that closely align with the most frequent bins of the histogram distributions are characterized by high CPL, indicating a strong likelihood of belonging to the training set.

This property of CPL was leveraged by Marhaba [65] for identifying Out-Of-Distribution (OOD) inputs. In this investigation, OODs were detected using a linear threshold on CPL: inputs with low CPL scores were likely to belong to the training distribution, whereas those with high CPL were classified as OOD. Indeed, OOD inputs tend to follow distinct "computational paths" compared to in-distribution inputs.

To address the floating-point underflow arising from the multiplication of numerous probabilities during CPL calculations, we decided instead to consider the Computational Profile Distance (CPD), which is the negative logarithm of CPL.

Consequently, high CPL means low CPD, and vice-versa. An input displaying a high CPD relative to the training distribution is less likely to be part of it. The term "distance" within Computational Profile Distance serves intuition rather than defining a distance in mathematical terms. When referring to an input with high CPD from a given distribution, we state that the input is "far" from this distribution. Conversely, when we phrase that an input is "close" to a distribution, we mean that the input has low CPD from this distribution, and thus is likely to be part of it.

Most fairness methods proven robust against OOD inputs are GAN-based techniques [43,66]. To our current knowledge, all these methods require model training or retraining. Other fairness reparation techniques, such as the Faire framework [67], leverage structural information in NN similarly to CPL but also necessitate model retraining.

CPL fills this gap by exhibiting robustness and effectiveness in managing corner-case inputs [65] without requiring model retraining. Notably, CPL can be employed as a post-hoc analysis tool for any already trained deep neural classifiers. Our research aims to propose a tool to utilize internal data from any deep neural classifiers to facilitate fairness detection and robust correction.

Next Section introduces the use of CPD in the context of algorithmic fairness and presents experimental results on synthetic datasets and on the Adult dataset.

CHAPTER 3 METHOD

This Chapter aims to expand the scope of applications for CPL to algorithmic fairness, focusing on identifying and mitigating negative biases of DNNs. Our general method is explained in Section 3.1, and preliminary experiments conducted on synthetic datasets are discussed in Section 3.2. Finally, Section 3.3 extends our findings to a widely studied dataset in fairness literature to demonstrate that CPD can indeed be used to detect negative bias.

3.1 Computational Profile Likelihood Adapted to Algorithmic Fairness

As explained in the precedent Chapter 2, CPL was primarily defined to detect OOD inputs, by comparing computational paths taken by inputs to computational paths recorded during training. This work aims to adapt this tool to algorithmic fairness.

Algorithmic fairness encompasses multiple definitions of bias and discrimination, some of which are detailed in previous Chapter 2. Fundamentally, discrimination manifests when there are unjustified disparities among demographic groups: for instance, if a classifier distinguishes the ability of individuals to be hired in a company based on their race.

Consequently, when pursuing potential bias within a classifier's prediction process, our primary focus lies on identifying whether the model treats input differently based on their sensitive attribute.

From a structural point of view, we are searching whether the internal activation patterns vary between sensitive groups during inference. If neuron activation levels remain consistent between two demographic groups, it is unlikely that the classifier bases its predictions on group membership. In particular, CPD is a good indicator of the fluctuation of internal activation patterns for any neural network classifier. In this context, we propose to assess this tool to evaluate and correct the potential bias that binary classifiers may present.

In details, we record each Computational Profiles encountered during training and draw two estimations of the distribution of activation levels, one for the positive class and the other for the negative class. These non-parametric probability density estimations have the form of histograms.

During the classification phase, Computational Profiles are extracted and the likelihood that these Profiles are produced is computed using the histograms. Thus, an individual presents a CPL for the positive class and a CPL for the negative class. These magnitudes quantify the extent to which the internal pattern activation of individuals diverges from those encountered during training. A higher CPD for a class indicates a lower likelihood that the corresponding individual is associated with that class. Conversely, a smaller CPD implies a higher likelihood of membership in that particular class. To extend the analogy with the detection of OOD, if a majority of individuals from unfavored group are "out of the distribution" of the positive class, then the model predictions suffer from negative bias towards unfavored group.

To sum up, if significant variations in CPD exist between the favored and unfavored groups, then the concerned model presents various treatments in classification depending on the sensitive attribute.

Taking the example of a sensitive attribute "gender", a classifier is responsible for sorting individuals according to their income into two classes: low revenue (*i.e.*, negative outcome) and high revenue (*i.e.*, positive outcome). During the training phase, we aggregate the Computational Profiles of the individuals predicted as high-income into a histogram and do the same for the low-income individuals. Thus, we have obtained two sets of histograms describing the distributions of activation levels of neurons, one for each output class. Subsequently, during the testing phase, we consider men and women separately, and compute their respective CPDs from both high-income and low-income classes, using the corresponding histograms. If the CPDs from low revenue class remain consistent between genders, as well as the CPDs from class high revenue, then we conclude that internal activation patterns are comparable between men and women, which indicates that the classifier's predictions are not influenced by gender.

On the other hand, if the CPDs from either low-income class or high-income class exhibit significant divergence between genders, then it implies that the activation levels of the neurons fluctuate depending on the gender of the individuals. Such a finding is an indication that a differential treatment can occur among these groups, which may result in an unfair classification. Thus, by employing CPD as an evaluation metric, we can identify instances of potential bias and work towards rectifying them to ensure fairness in the classifier prediction.

3.2 Preliminary Results on Synthetic Data

To begin our investigation into the behavior of CPD in the context of group fairness, we have conducted a series of preliminary experiments. To realize this, synthetic datasets have been generated with controlled properties, namely, imbalance and bias between privileged and unprivileged groups. We subsequently train simple NNs on these datasets and examine the CPD distributions. The objective is to study these distributions in these simple settings to assess the potential of using the CPD in situations in which undesirable biases occur in an unprivileged group. In each of the following paragraphs, we describe the objectives of each experiment, outline the anticipated outcomes, and present an analysis of the results obtained.

More precisely for the following experiments, the protected attribute will be "gender" and the privileged group refers to all the male individuals whereas the unprivileged group corresponds to their female counterparts. The prediction task consists in assigning a value of "Low revenue" (LR) - negative outcome - or "High revenue" (HR) - positive outcome - to each individual based on their attributes. An additional attribute "education" is considered, which is a binary variable representing the attributes that are not sensitive and on which the classification should rely.

Unbiased and Balanced Dataset

We consider a fair data distribution, with a perfect balance between men and women with no bias. This means that there is an identical number of males and females in each revenue class. Moreover, only educated profiles have high incomes with the same number of educated men and educated women. A summary of these proportions is provided in Table 3.1.

	High Revenue		Low Revenue	
Male	15000 educated	0 uneducated	0 educated	15000 uneducated
Female	15000 educated	0 uneducated	0 educated	15000 uneducated

Table 3.1 Proportions of individuals on the balanced and unbiased synthetic dataset

Given the characteristics of this synthetic dataset, we expect that any classifier trained on this dataset would reach a maximum accuracy since classification solely depends on the education attribute. Thus, we anticipate that CPD from the HR class would be low for educated individuals and high for uneducated individuals. The inverse trend is expected for the CPD from the LR class with minimal differences between men and women, regardless of their education level.

The results of the following experiments are represented in diagrams where the X-axis represents the CPD from the LR class while the Y-axis denotes the CPD from the HR class. An individual profile is represented by a dot on the graph, whose coordinates are its CPD from class HR and class LR. The black line separates the graph into two distinct regions:

the upper part aggregates individuals more likely to belong to HR than LR in terms of CPD, while the other region gathers individuals more inclined towards the LR class than the HR class. We differentiate genders by assigning blue coloration for male profiles and red for females. Educated profiles are symbolized by a triangle pointing up whereas uneducated profiles are represented by a triangle pointing down. All the experiments of this section are conducted with a small neuron network of three hidden fully-connected layers of five neurons each, trained during 10 epochs, with a learning parameter of 0.1. The activation function considered is Leaky-ReLU. For each experiment, the dataset considered has been divided into three random portions of equal size, to form the training, testing, and validation set.

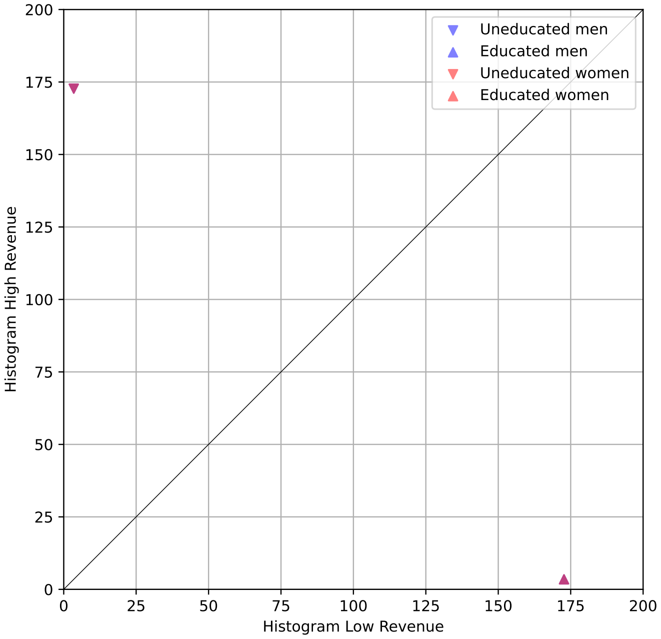


Figure 3.1 CPD distributions of individuals on the balanced and unbiased synthetic dataset

Figure 3.1 presents the result of the experiment conducted on the unbiased and balanced dataset. Notably, the accuracy reached by the network is 100%. We observe that all individuals are located at two specific coordinates: (3.4657;172.6939) and (172.6939;3.4657), which are symmetrically placed relative to the line $y = x$. Given the mathematical definition of CPD, 3.4657 is one of the lowest value possible for CPD with a network of this size, whereas 172.6939 represents the maximum achievable.

The first pair of coordinates aggregates all the uneducated individuals (blue and red markers overlapping). As anticipated, they display the lowest CPD from class LR and the highest from class HR. Conversely, the educated population is situated at the second set of coordinates, with the highest CPD from class LR and the lowest from class HR (once again, the blue and red markers overlapped). These observations match our expectations based on the dataset's design and distribution.

Unbiased and Unbalanced Dataset

In this experiment, we investigate the effects of data imbalance with respect to sensitive groups on CPD. More precisely, we generate a dataset skewed towards men, consisting of 75% males and 25% females. Despite the gender disparity, we set the same number of HR and LR among these groups, with the attribute education being equal to the income attribute.

	High Revenue		Low Revenue	
Male	22500 educated	0 uneducated	0 educated	22500 uneducated
Female	7500 educated	0 uneducated	0 educated	7500 uneducated

Table 3.2 Proportions of individuals on the unbalanced and unbiased synthetic dataset

We expect that the classifier correctly assigns all the individuals to their corresponding class. In particular, HR individuals should have greater CPD from class LR compared to their counterparts in the LR class. Conversely, LR individuals should display a higher CPD in the HR class compared to high-income individuals. Moreover, since the majority of activation levels of the histograms LR and HR correspond to men due to their higher representation in the dataset, we assume that high-revenue men have lower CPD from class HR than high-revenue women. Similarly, LR men should have lower CPD from class LR than low-revenue women.

The outcomes of the experiments indicate that the classifier achieved a perfect accuracy rate of 100% after the training phase. The repartition of CPD distributions for all the individuals is represented in Figure 3.2. In line with the previous experiment, the uneducated profiles concentrate in the region of high CPD from class HR and low CPD from class LR. Thus, uneducated individuals are more likely to belong to the LR class than the HR class. Conversely, educated men and women exhibit the opposite behavior. This is coherent with the proportions of Table 3.2 in which only educated profiles have access to high-income levels, while uneducated individuals are only present in the LR class.

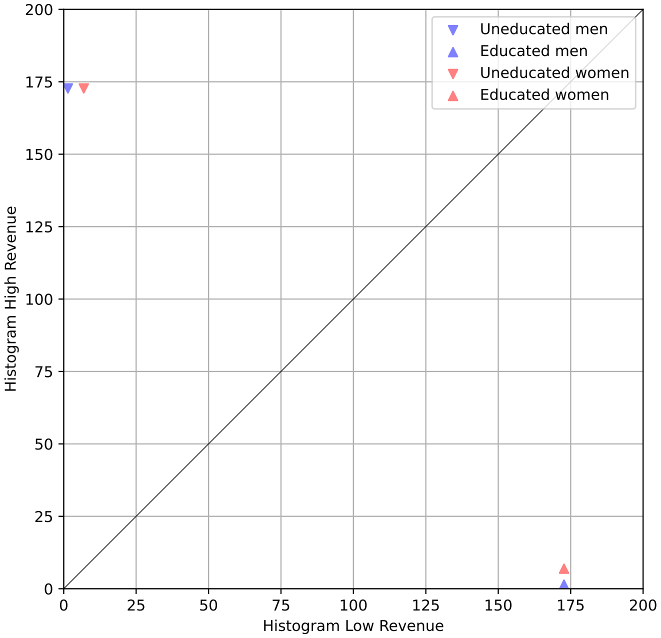


Figure 3.2 CPD distributions of individuals on the unbalanced and unbiased synthetic dataset

In addition, educated women exhibit higher CPD values from class HR than educated men. This finding aligns with our expectations: in the histogram of activation levels corresponding to the HR class, there are three times more men than women (*i.e.*, 22500 males and 7500 females).

Biased and Balanced Dataset

	High Revenue		Low Revenue	
Male	15000 educated	15000 uneducated	0 educated	0 uneducated
Female	0 educated	0 uneducated	15000 educated	15000 uneducated

Table 3.3 Proportions of individuals on the balanced and biased synthetic dataset

In addition to considering the impact of data unbalance on CPD, the following experiments aim to investigate the influence of bias on CPD. In particular, the purpose of this experiment is to study the context in which a bias exists between the balanced sensitive groups. More precisely, the generated dataset contains an equal number of men and women. The education attribute is defined such that half of the men are labeled as "educated" while the rest are not. Women get similar assignments. Thus, the dataset and each sensitive group display the same number of educated individuals and those not educated. Finally, the income is equal to the gender of the individuals. Arbitrarily, we assign all men to HR and all women to LR. In consequence, classification solely relies on gender and not on education. The sensitive attribute and the task of prediction are not independent, thus creating a positive bias towards men and a negative one towards women.

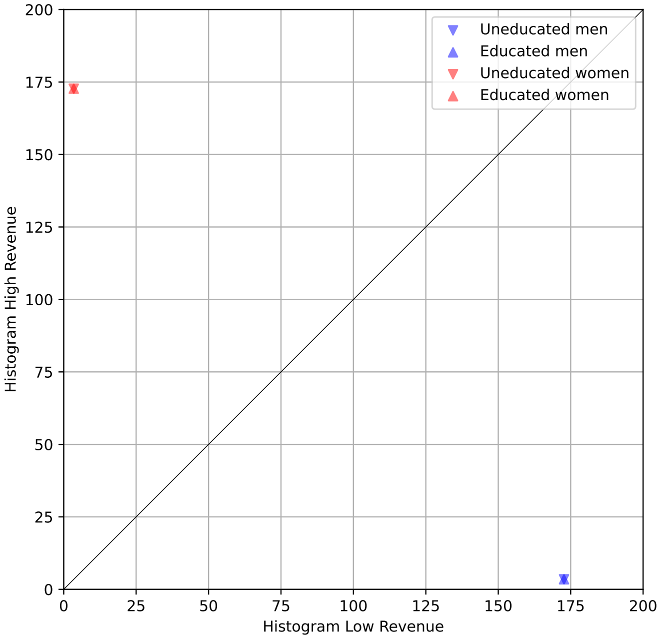


Figure 3.3 CPD distributions of individuals on the balanced and biased synthetic dataset

Given the experimental setup with a positive bias towards men and a negative bias towards women, we anticipate observing distinct differences in the CPD distributions between genders. For male individuals, CPDs from the HR class are expected to be small while the CPDs from the LR class should be very high, regardless of their education level. For women, an opposite observation is expected.

As anticipated, Figure 3.3 displays significant disparities between men's and women's CPD distributions. Unlike the previous experiments, men's and women's distributions are not symmetrical to the line of equation $y = x$. Men are gathered in the region of high CPD from class LR and low CPD from class HR, regardless of their level of education. A similar observation applies to the women's group, aggregated in the area of low CPD from class LR and high CPD from class HR, whether they are educated or not.

Biased and Unbalanced Dataset

The final experiments in this series investigate the combined effects of both bias and data imbalance on CPD calculations. The dataset is generated to ensure positive bias towards men and negative bias towards women, as well as a higher number of men compared to women. The detailed proportion of the dataset is presented in Table 3.4.

	High Revenue		Low Revenue	
Male	10000 educated	10000 uneducated	0 educated	10000 uneducated
Female	2500 educated	0 uneducated	2500 educated	5000 uneducated

Table 3.4 Proportions of individuals on the unbalanced and biased synthetic dataset

Based on the experimental design, we anticipate that CPD distributions combine the properties of the two previous experiments. Educated and uneducated men should be close to the HR class and far from the LR class, in terms of CPD. This is a consequence of the positive bias towards men that places them primarily in HR classes, regardless of their education levels. Educated women should also be in this area, but a little further away from HR class due to data unbalance. Finally, uneducated women are expected to be closely located to the LR class and far from the HR class, in terms of CPD.

After training, the classifier reaches an accuracy of 68.75% on the training set. Figure 3.4 presents the resulting CPD distributions corresponding to this experiment. As expected, uneducated women have a CPD from class HR to its maximum value and a very low CPD from class LR. Thus, uneducated women are much more likely to be classified as LR than uneducated men, who have low CPD from class HR. Educated women are likely to belong to the HR class, but less than any men, educated or not.

To summarize, the presented experimental investigations explored two aspects that data may present in unfair settings: data imbalance and data bias. CPD metric helps to measure to what extent a subgroup of individuals is likely or not to have positive outcomes, relative to other subgroups.

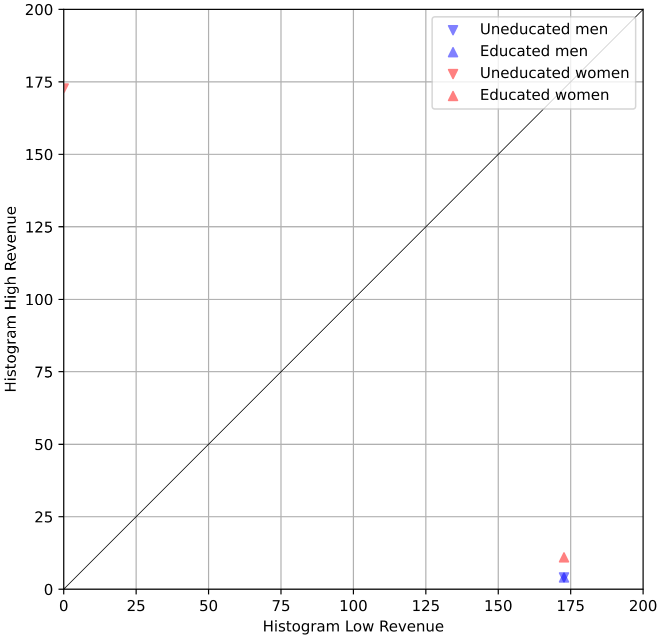


Figure 3.4 CPD distributions of individuals on the unbalanced and biased synthetic dataset

In an unbiased and balanced scenario, the CPD does not distinguish between men and women. However, in the situation of data imbalance, while CPD distributions of men and women are still symmetrical to the line of equation $y = x$, there is a notable difference: men have lower CPD from their respective outcome class than women. Conversely, in biased data scenarios, CPD distributions are not symmetrical between men and women. In this situation, the classification relies solely on gender and CPD from class HR of women is maximum whether they are educated or not. Finally, when dealing with biased and unbalanced data, CPD distributions reflect these properties on the different subgroups. For instance, educated women are statistically likely to belong to the HR class, but less than any man, regardless of their education.

These results illustrate how CPD behaves in controlled settings, highlighting how CPD and algorithmic fairness can be combined to detect potential bias. After the exploration of these synthetic datasets, the next section delves deeper into the structural analysis of a classifier trained on the Adult dataset.

3.3 Activation Levels Histograms on Real-World Data: Case Study of Adult Dataset

In the preceding section, we introduced the concept of CPD and demonstrated its application on synthetic datasets in rudimentary use cases. Nevertheless, there is no guarantee that the observed findings generalize to other datasets especially for real-world scenarios.

This section compares demographic sub-groups utilizing the Adult census income dataset [68]. This database aggregates demographic information of 45,222 individuals, including six numerical attributes and eight categorical attributes. The associated task is to predict if individuals have an annual salary higher than 50,000 USD or not. Widely studied in the algorithmic fairness literature [10, 31, 69–71], it has been established that a bias concerning the sensitive attribute *gender* exist. Women experience discrimination compared to men when considering the target attribute *income*. As defined in the last series of experiments, *income* is a binary attribute separated between HR and LR. The dataset is skewed towards high-income men, and a classifier trained to predict income level is likely to perpetuate this bias, thus disadvantaging women.

To validate this hypothesis, we selected CPD as the metric to quantify differences between men and women concerning their income level. Since CPD relies on statistical information derived from the classifier's neurons, we decided to investigate the disparities in the distribution of activation levels between genders within this dataset.

A binary classifier was consequently trained on the Adult dataset using the following architecture: three layers consisting of 128, 128, and 64 neurons respectively. The model is trained during 100 epochs, employing a learning rate of 0.001. After training, the classifier reaches an accuracy score of 84.16%. This result shows that our model effectively distinguishes between high-revenue and low-revenue individuals with a reasonable degree of precision.

In the details of the CPD definition of Section 4.4, we mention that we focus on the last hidden layer to perform the calculations of CPD. This particular layer encapsulates information from previous layers, ensuring that any differential internal activation patterns between men and women are accounted for within this layer. Furthermore, by restricting our analysis to a single layer, we allow better computational performance. Thus, in all subsequent experiments, we solely focus on the last hidden layer of the binary classifier.

The next paragraphs study the differences in distributions of activation levels, neurons by neurons. The first paragraph considers the differences between the HR and the LR profile. The second paragraph studies the disparities between low-income men and women. Finally, the last paragraph analyzes the contrast between high-revenue men and women.

Comparison Between LR and HR Histograms

To visually represent and compare the distributions of activation levels for each neuron sorted by their respective weights within our binary classifier on the Adult dataset, we opt to utilize box plots. The X-axis depicts the neuron's unique identifier. The value of the weights for each neuron is displayed by the green curve. For every neuron, two distinct distributions are displayed in the forms of blue and orange rectangles. The mean of the distributions is represented by the horizontal black line that divides each rectangle into two parts. From the mean, the rectangles extend one standard deviation above and one standard deviation below. A small rectangle denotes a distribution closely around its mean, suggesting that activation levels for this neuron are relatively consistent across the group studied. Conversely, a long rectangle implies a spread-out distribution. By examining these box plots, we aim to identify significant disparities in activation level distributions and quantify their magnitudes.

To ensure that any potential disparities in activation levels between demographic subgroups do not stem from random chance, we perform an additional analysis step. We randomly split the initial training data into two equal parts. During the training phase, for each portion, we recorded the activation levels reached by each neuron. The resulting box plots for all neurons are presented in Figure 3.5a.

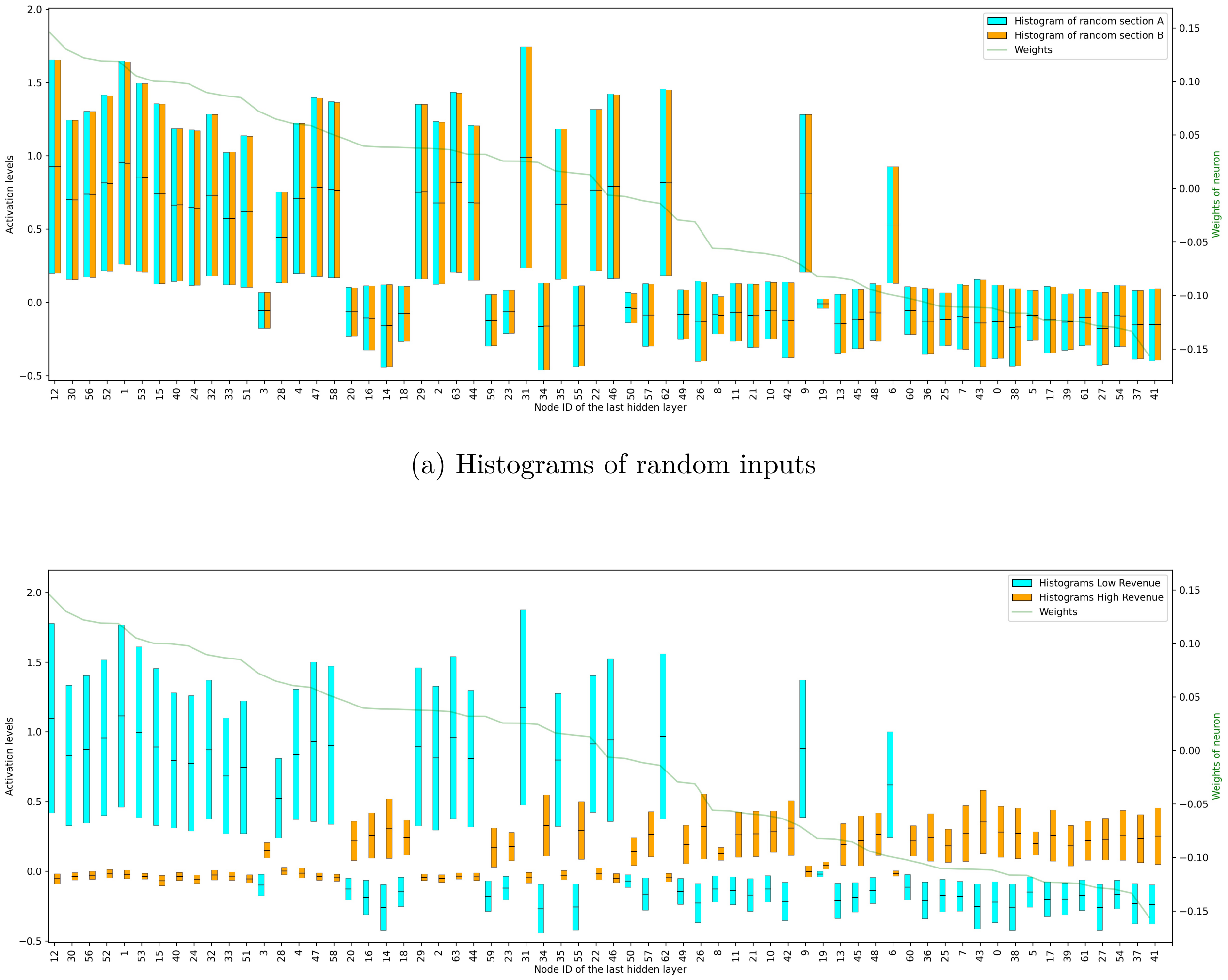
Visually, there are no or minimal disparities in the distributions of activation levels for all the involved neurons across these two random portions. The distributions are either centered around an activation level between 0.5 and 1 with a spreading of 0.6 or centered around 0 with a smaller spreading of around 0.2.

Then, we perform a similar analysis using training data split based on high-income and low-income classes rather than random portions. As shown in Figure 3.5b, distributions of activation levels are significantly distinct compared to the minimal disparities observed between our random portions. This observation is consistent with the fact that the model effectively distinguishes individuals in one of these classes with good precision.

Activation levels corresponding to neurons within low-income profiles are generally observed to be higher than those for high-income individuals when considering positive weights, whereas the opposite trend is observed for negative weights.

Comparison Between Men LR and Women LR Histograms

For a more detailed examination of disparities in activation level distributions between demographic subgroups, we investigate the specific cases of low-income men (blue) and low-income women (orange). Figure 3.6a is presented for comparison with random groups. Figure 3.6b



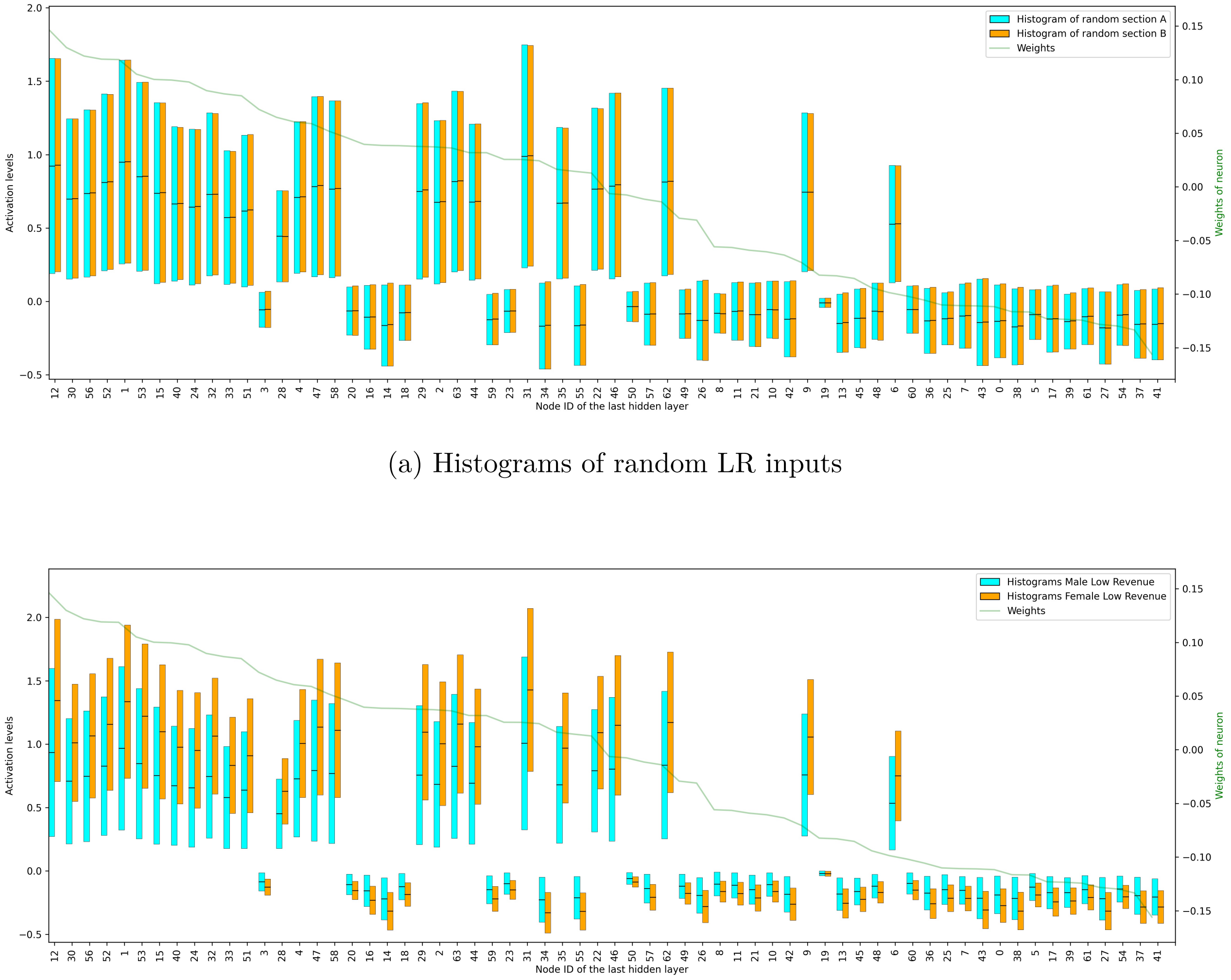
(a) Histograms of random inputs

(b) Histograms separated between LR and HR class

Figure 3.5 Histograms of activation levels for each neuron of a classifier

shows that the distributions present significant disparities between men and women compared to Figure 3.6a. On one hand, for neurons with positive weights, activation levels for women are higher than activation levels for men. On the other hand, for neurons with negative weights, levels of men are generally higher than levels of women. These two observations suggest that the classifier presents a negative bias toward women, who are more likely to belong to the low-revenue class.

The findings imply that CPD is an efficient tool to detect bias. CPD calculations rely on such histograms of distributions to evaluate the likelihood of an individual belonging to an



(a) Histograms of random LR inputs

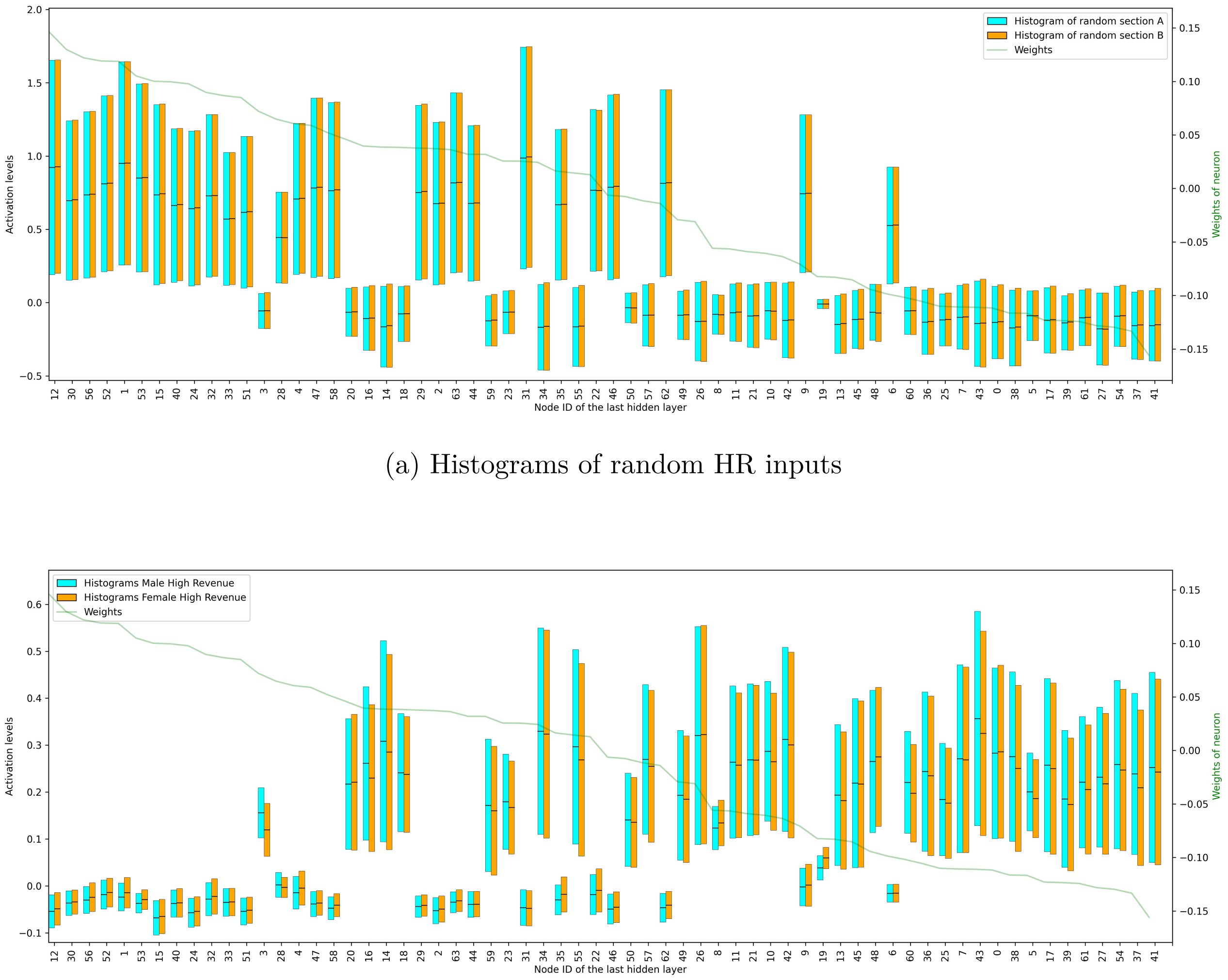
(b) Histograms of male LR and female LR classes

Figure 3.6 Histograms of activation levels corresponding to LR inputs for each neuron of a classifier

outcome class. Witnessing such disparities indicates that CPD distributions should also present disparities between male and female individuals.

Comparison Between Men HR and Women HR Histograms

The last experiment of this series considers the potential differences between high-income men and high-income women. Figure 3.7a presents a baseline with two random portions, and Figure 3.7b displays the comparison between the two groups.



(a) Histograms of random HR inputs

(b) Histograms of male HR and female HR classes

Figure 3.7 Histograms of activation levels corresponding to HR inputs for each neuron of a classifier

Although disparities are less substantial than the comparison with low-income individuals, contrast still exists between high-revenue men and high-revenue women compared to random portions of Figure 3.7a. For neurons with negative weights, activation levels seem slightly higher for male than female individuals. For neurons with the highest positive weights, distributions of activation levels are concentrated around their means, indicating that only specific activation levels correspond to classification into a HR class. This could be a consequence of the data imbalance between high-income and low-income of the Adult dataset (75% of LR and 25% of HR).

Overall, these experiments showed that the classifier trained on the Adult Census Income dataset has a distinct activation pattern when classifying individuals into HR or LR class (which is coherent with its task of prediction), but also significant disparities between men and women of the LR class, and in the lesser extent HR class. These last remarks suggest the existence of a negative bias towards women compared to men. Observing such disparities in the distributions of the internal activation patterns of a classifier promotes the idea that CPD is a convenient tool for detecting and evaluating bias.

The next Chapter delves deeper into this idea and compares CPD distributions between demographic groups defined by sensitive attributes. It takes the form of an article submitted in *Computational Intelligence*. After introducing the necessary background on fairness and CPD, it presents a new methodology based on CPD to detect unwanted bias in data. It is the natural continuation of the series of experiments conducted in this chapter and presents experimental results on three datasets (including Adult census income). Furthermore, we also propose a novel way to evaluate the efficiency of the pre-processing bias mitigation method using this tool.

CHAPTER 4 ARTICLE 1: FAIRNESS EVALUATION OF NEURAL NETWORKS THROUGH COMPUTATIONAL PROFILE LIKELIHOOD

Benjamin Djian, Ettore Merlo, Sébastien Gambs, Rosin Claude Ngueveu

Submitted to *Computational Intelligence* on 31 October 2024

4.1 Abstract

Despite high predictive performance, machine learning models can be unfair towards specific demographic subgroups characterized by sensitive attributes such as gender or race. This paper presents a novel approach using Computational Profile Likelihood (CPL) to assess potential bias in neural network decisions with respect to sensitive attributes. CPL estimates the conditional probability of a network's internal neuron excitation levels during predictions. To assess the impact of sensitive attributes on predictions, the CPL distribution of individuals sharing a particular value of a sensitive attribute and a specific outcome (*e.g.* "women" and "high income") is compared to a subgroup sharing another value of the sensitive attribute but with the same outcome (*e.g.* "men" and "high income"). The resulting disparities between distributions can be used to quantify the bias with respect to the sensitive attribute and the outcome class. We also assess the efficacy of bias reduction techniques through their influence on the resulting disparities. Experimental results on three widely used datasets indicate that the CPL of the trained models can be used to characterize significant differences between multiple protected groups, highlighting that these models display quantifiable biases. Furthermore, after applying bias mitigation methods, the gaps in CPL distributions are reduced, indicating a more similar internal representation for profiles of different protected groups.

4.2 Introduction

Machine learning (ML) has become ubiquitous due to its success across a wide range of fields. In particular, ML models significantly enhance decision-making processes, as evidenced by recent studies [72–74]. A key aspect of this success is the rise of Neural Networks (NN), which have demonstrated exceptional performance on a variety of tasks [75, 76]. However, NN models are often trained on huge quantities of personal data, which frequently reflect historically biased human decisions or social values. Thus, despite their impressive performance, ML models integrating these biases can discriminate against demographic subgroups as defined by sensitive attributes, such as gender, race, or age. Consequently, if such a model is deployed in a high-decision setting in which its predictions have an impact on individuals, it can perpetuate or even amplify this discrimination.

While there is a huge body of work on algorithmic fairness [9, 27], the fundamental question of how to detect and quantify a possible bias with respect to a sensitive attribute is still challenging. In addition, while techniques exist to mitigate bias, such as pre-processing approaches [42], quantifying the effectiveness of these approaches is also laborious. In particular, measuring bias within datasets is complex due to the multiplicity of existing metrics, which makes it difficult to choose a metric capturing the intended objectives, the context as well as the complex correlations that may exist between different attributes. In other words, a specific measure of fairness may not necessarily guarantee correct predictive behavior from the model. A particularly challenging issue is that NNs are known to extract and rely on correlations that are not intrinsically related to the problem at hand (*e.g.*, predicting enemy tank positions based on surrounding weather conditions rather than the appearance of the tanks [77]).

In this paper, we propose a novel approach to address these issues that leverage the likelihood of predictions, using computations based on activation levels of neurons of a fully connected NN, called Computational Profile Likelihood (CPL) [13]. In a nutshell, this approach relies on the comparison of distributions of CPL of distinct subgroups as defined by sensitive attributes. In most instances, it is possible to distinguish between a privileged group and a disadvantaged group, also called a protected group, within the data set. Our main contributions can be summarized as follows:

- We show that the study of CPL can provide valuable insights into how sensitive attributes impact model predictions. In particular, using it on binary classifiers trained on three widely known datasets in fairness literature, we demonstrate how the CPL profiles differ between distinct sensitive classes, for example, gender or race.
- We also demonstrate how the CPL can be used to assess the effectiveness of fairness pre-processing methods, comparing the CPL distributions with and without the use of a bias mitigation methods from the literature, namely Correlation Remover [78] and Disparate Impact Remover [69]. In addition, as observed by the results obtained, there are notable differences in the performance of these methods with respect to the effectiveness of bias removal.

The outline of the paper is as follows. First, in Section 4.3 we review the related work in the field of fairness, bias mitigation, fair representation learning, and network structural analysis for fairness. Afterwards, in Section 4.4 we introduce the necessary background on CPL necessary to the understanding of our work. Then in Section 4.5, we explain how the CPL method can be applied to perform bias detection and quantification in NN. More precisely, we rely on the CPL to verify whether profiles from privileged groups lead to different activation patterns within the model compared to profiles from protected groups. In Section 4.6, we describe the experimental setting that we have used to assess the efficiency of our approach to quantify the fairness of a network as well as the results obtained. Finally, we propose an interpretation and a discussion of their implication in Section 4.8 along with the limitations of our approach before concluding in Section 4.9.

4.3 Related work

In the context of fairness in ML, the profiles that are part of the training set of a ML model are usually assumed to be split between three types of attributes:

- The *decision attribute* C corresponds to the class that should be predicted from the profile (*e.g.*, acceptance or rejection of a loan or the income level being above a particular threshold).
- The *sensitive attribute(s)* S, from which the decisions should not be based, are known to lead to discrimination, the most common examples being gender, race, or age. In the context of a single binary sensitive attribute, it is feasible to partition the dataset into two distinct groups: the privileged group and the protected group. However, in real-life situations, two or more protected groups may exist, leading to intersectional sub-groups (*e.g.* afro-American women [79]).
- The *other attributes* correspond to all the remaining attributes that are neither prediction nor sensitive. However, these attributes may contain some proxy attributes that are correlated to sensitive ones [80] (*e.g.*, hair length), which prevents the possibility of achieving fairness simply by removing sensitive features [81, 82].

A vast body of work has led to numerous definitions of fairness for algorithms such as group fairness [9], which require the equality between statistics related to protected groups (*e.g.*, with respect to acceptance or rejection rates) and individual fairness [83], which requires that similar profiles get similar treatment. Nonetheless, a huge part of the literature focuses on group fairness [39] using notions such as demographic parity, which measures the degree of independence between classification and sensitive groups, or equalized odds [25] that assesses the equality of false positive rate and negative rate between protected groups. Another large section of the fairness literature also proposes various methods to mitigate existing biases during the learning process. We can divide these into three main families of approaches:

- *Preprocessing techniques* that aim to change the underlying dataset to satisfy fairness requirements before its use in the ML pipeline [27].
- *Inprocessing techniques* directly modify the learning algorithm by introducing fairness constraints in the objective function being optimized to enhance the resulting fairness of the trained model [27]. For instance, the method developed by Shen and co-authors [84] adds the equality of opportunity as an additional constraint during the optimization of the prediction model.
- *Postprocessing techniques* apply adjustments after the model has been trained. For instance, this can be done by measuring the model's fairness and using the observed results to calibrate accordingly the predicted outputs accordingly [27].

In our study, we focus on preprocessing methods, as in-processing and post-processing methods could potentially alter the activation levels of the network, which are fundamental for CPL calculations. Specifically, we consider two methods: Correlation Remover and Disparate Impact Remover. Correlation Remover is a classical preprocessing method, implemented in Fairlearn [78], which removes existing correlations between the sensitive attribute and other attributes. This method applies a linear transformation to the non-sensitive attributes to remove their correlation with the sensitive attribute, with a hyperparameter controlling the amount of transformation. We chose the default value for this parameter, which fully applies the transformation.

Disparate impact is a group fairness metric, which corresponds to the ratio between the protected group that received the positive outcome divided by the proportion of the privileged group that received the positive outcome. With this metric, a proportion too low is an indication of discrimination in the classification process. Disparate Impact Remover is a preprocessing method adapted from the work of Feldman and collaborators [69], whose objective is to prevent a high disparate impact by editing the values of attributes to make the different groups indistinguishable while preserving the rank of individuals inside groups. AIF360 library [85] proposes an implementation of this technique. The corresponding function comes with a parameter called “repair amount” defined between 0 and 1. Hereafter, we have applied a “full repair” to the data, thus setting the parameter to its maximum value.

A particular type of preprocessing technique called fair representation learning, aims to learn a representation independent from protected attributes [86] while preserving the decision-making property of the model. The main idea behind fair representation learning is intuitive:

if representations of different protected groups are identical, the model cannot discriminate inputs based on group membership. For instance, Creager and collaborators [87] have proposed an algorithm to learn compact and flexible fair representations for demographic subgroups. They correspond each dimension of the representation to one semantic factor of variation in the data, to eliminate the influence of sensitive attributes during inference. Other researchers have introduced FairDisCo [88] for learning fair representation independently from mutual information between prediction and sensitive attributes. This method encompasses a network comprised of three branches, each responsible for maximizing prediction quality, minimizing unfairness, and boosting classification accuracy.

To ensure desirable properties of NNs, including fairness, it is natural to look directly at the internal structure of such models. For instance, several works have investigated white-box testing methods for deep NNs to this goal such as NeuronFair [12], proposed by Zheng and co-authors, which identifies biased neurons through neuron-based analysis and generates pairs of identical inputs differing only in sensitive attributes, which generate different prediction results. Faire [67], introduced by Li and co-authors, repairs fairness issues of deep NNs by penalizing directly neurons whose outputs could be considered related to protected attributes with a conditional layer placed after each hidden layer. Another method [89] proposed by Dasu and collaborators mitigates bias through strategic dropouts during inference, balancing performance preservation and fairness enhancement. Similarly to our study, Mao and co-authors look at the last layer of deep NNs [90]. More precisely, they demonstrate that fine-tuning this small fraction of an already trained model to achieve fairness can yield a fair NN.

Adversarial methods can also be leveraged to achieve fair representation learning [91]. For example, Beutel, List, and Schweinitz [10] have trained a multi-head deep NN, with the dual objective of predicting target attribute and preventing accurate prediction of the sensitive attribute. Another recent example, GANSan [70] aims at ensuring fairness by sanitizing the data (*i.e.*, modifying it to remove correlations with the sensitive attribute) while minimizing the distortion incurred. The desired trade-off between fairness and data fidelity is controlled by a parameter α, designated by “sanitization level”, with a higher α value indicating better fairness at the expense of lower fidelity. However, as defined primarily by Goodfellow and co-authors [66], adversarial methods require a second model training to work, which can be computationally expensive.

Our approach based on CPL takes inspiration from other works in NN robustness and out-of-distribution detection [92–94]. Notably, the SADL approach [95] uses neuron activation values to calculate the level of “surprise” between training and adversarial images. Then, they rely on these surprise values to retrain a classifier to avoid the misclassification of those adversarial images. BENN is another approach that uses another DNN to estimate the bias of ML models [71], which is composed of two parts. The first part is an unsupervised DNN that indicates the degree of bias toward each feature according to the input sample while the second part processes the output of the first part to provide a final bias estimation for each feature. BENN evaluates multiple aspects of fairness using the "fairness through unawareness" principle [83] and provides an interpretation of each feature and how much they affect the model's output.

Our CPL bias assessment approach also offers detailed information on given features relative to possible discrimination. However, in contrast to BENN, CPL does not require defining a custom loss function but rather uses statistical analysis based on activation levels of inputs seen during the training of the model. In addition, while in this study we have only considered one sensitive attribute at a time, it could be easily extended to investigate several features in a multi-dimensional analysis.

4.4 Background

CPL is used in the setting of a fully connected NN, which consists of numerous layers of neurons, in which each neuron propagates activation values across layers using weights and non-linear functions [96]. CPL [13, 97] was initially designed to detect inputs that do not belong to the classifier training distribution, also called Out-Of-Distribution (OOD) inputs. Often, such inputs result in a significant performance drop of the model and are frequently used for adversarial attacks. By scrutinizing directly the internal activation pattern of the model, CPL uncovers significant variations between in-distribution examples and OOD, thus justifying its robustness and effectiveness for this task. This is particularly beneficial in sensitive and critical domains where identifying OOD inputs is challenging due to the difficulty in forecasting proper and representative samples of unknown or unexpected cases, such as aerospace, medicine, and cyber-security. The main idea behind CPL is to measure the "reasoning" likelihood of a NN prediction by assigning a probability to the activation levels of the last layer.

We will now recall the details of CPL computations initially introduced by Merlo and co-authors [13]. This approach is a non-parametric method to represent the distributions of activation levels. More precisely, the non-parametric statistics use histograms to model such distributions.

Consider a neural network with N layers, trained to assign inputs to a class of $C = \{c_1, c_2 ...\}$, in which C is the set of output classes. Let $a_{i,j}(x)$ be the activation level corresponding to an input x of the i-th neuron in the j-th layer while X denotes the set of inputs used to train such network and X' the subset of training data correctly classified. In particular, X'_c refers to all the training inputs correctly classified, belonging to the output class $c \in C$.

$\mu(i, j, c)$ denotes the mean of the distribution of the activation levels of the neuron (i, j) for all elements of X'_c:

$$\mu(i, j, c) = \frac{1}{\mid X'_c \mid} \cdot \sum_{x \in X'_c} a_{i,j}(x), \tag{4.1}$$

while $\sigma(i, j, c)$ is the standard deviation of the distribution of the $a_{i,j}(x)$ for all x in X'_c:

$$\sigma(i, j, c) = \sqrt{\frac{\sum_{x \in X'_c}(a_{i,j}(x) - \mu(i, j, c))^2}{\mid X'_c \mid}} \tag{4.2}$$

For each neuron, a specific histogram with bins proportional to its activation level's standard deviation is defined. To achieve comparable resolution across different neuron distributions, the bins used are of variable size $width(i, j, c) = \frac{1}{k} * \sigma(i, j, c)$ for each neuron (i, j), in which k is the resolution of the bins. Higher k contributes to a more accurate estimation of the distributions but engenders increased computational expenses. A lower k value enhances performance at the expense of less precise distribution estimations. In the experiments conducted afterward, we opted for $k = 1$ as a suitable compromise between precision and computational efficiency. Indeed, experiments conducted with $k = 10$ brought minor differences in the results while causing a considerable increase of the computational time.

The histogram bin frequencies $bFreq(b, i, j, c)$ are computed during training by counting the number of times inputs from $X^{'}_{c,s}$ produce an activation level that falls into bin b for neuron (i, j). More precisely, the slot probabilities are computed from slot frequencies as follows:

$$p(b, i, j, c) = \frac{bFreq(b, i, j, c)}{\mid X'_c \mid} \tag{4.3}$$

In addition, to smooth probabilities across the bins, a very low probability is assigned to all bins with null frequencies. Intuitively, during training, if a given neuron reaches often a given activation level, the corresponding slot probability will be high. Conversely, low slot probability indicates that the corresponding range of activation levels has rarely been reached during training.

To consider higher-level information about the network, we studied these slot probabilities layer-wise. An input y has a slot probability corresponding to the activation level of the neuron (i, j). If we consider all the slot probabilities $p(b, i, j, c)$ of y, we can define $CPL(y, j, c)$

as the joint probability for all neurons in a layer j:

$$CPL(y, j, c) = \prod_{i} p(b, i, j, c) \tag{4.4}$$

For practical reasons, we use the negative logarithmic of this probability:

$$dist(y, j, c) = -\sum_{i} log(p(b, i, j, c)). \tag{4.5}$$

A higher "distance" corresponds to lower CPL, although the word "distance" is used here to serve intuition rather than to define a distance in mathematical terms. For clarity, consider an input y with a great distance from a class c_{great} and a small distance from another class c_{small}. The activation levels of y were rarely reached during training on inputs of class c_{great} but were often attained when considering inputs of class c_{small}. Statistically, the internal activation pattern for y is much more similar to inputs of c_{low} than inputs of c_{great}.

Consistently with the approach proposed in the original paper [13], we focus on the penultimate $N-1$ layer, both to reduce the computational requirement and also because this layer encapsulates previous layers' network computations. In particular, we define the Computational Profile Distance (CPD) as follows:

$$CPD(y, c) = dist(y, N-1, c). \tag{4.6}$$

Finally, we extend the definitions of CPD to sets by defining:

$$CPD(Y, X'_c) = \{CPD(y, c) \mid y \in Y\} \tag{4.7}$$

in which Y is the set of inputs to study in relation with X'_c. In the following section, we discuss the application of CPL as a tool for evaluating the fairness of a NN, and what can be learned from the internal activation patterns of such model when considering protected groups.

4.5 Fairness evaluation through the lens of CPL

Overview of the approach. In this paper, we propose to apply the CPL method to assess fairness in NNs by comparing CPL distributions across groups as defined by the sensitive attribute. The CPL approach, by comparing the likelihood of a given input to a distribution of a group of inputs, aligns more closely with group fairness than individual fairness.

Instead of designating models as either “fair” or “unfair”, our approach quantifies the variance in behavior between demographic subgroups, in other words, discrimination. Notably, this approach can be used in scenarios in which a model bias is intentionally investigated by the practitioner and not restricted to instances requiring equal treatment by the model of all demographic subgroups.

In algorithmic fairness, a fundamental question to address is whether the likelihood of a prediction depends on the value of the sensitive attribute. Our objective is to answer this question by computing CPL distributions building on the intuition that profiles from the privileged group may traverse different computational paths within the model compared to profiles from protected groups. In such cases, significant differences in the likelihood of distinct groups should emerge. In particular, if such differences are observable through the lens of the CPL, the network may be biased towards a particular value of the sensitive attribute.

Another application of CPL for fairness is to assess the efficiency of bias mitigation techniques and their impact on the likelihood of predictions for potentially discriminated inputs. For example, if a bias mitigation technique is applied in a context in which the likelihood distributions of two distinct groups are initially distinguishable, one can measure these distributions *a posteriori* and compare them with the original distributions. In particular, if the new distributions are less distinguishable than the original ones, this is a strong sign that the undesired bias has been reduced.

Illustrative example. Intuition suggests that when the internal activation patterns of two sensitive groups look alike in a model, it is unlikely that the model makes distinctions in profiles based on group membership. Furthermore, bias mitigation approaches should aim to narrow down any disparities between inputs of different protected groups. In the next section, we will compare the distributions of the *CPL* of two distinct demographic groups of the testing set, relative to each output class, before and after the application of bias mitigation techniques.

For instance, considering the Adult dataset (*cf.*, next section) and *gender* as the sensitive attribute, the protected group is composed of female individuals (F), while males (M) constitute the privileged group. The output classes are *HR* (High Revenue) and *LR* (Low Revenue). We compare the distributions of $CPD(Y_F, X'_{HR})$ and $CPD(Y_M, X'_{HR})$ as well as $CPD(Y_F, X'_{LR})$ and $CPD(Y_M, X'_{LR})$ by plotting these distributions, computing the area between the curves and comparing the gap before (*i.e.*, baseline situation) and after bias mitigation methods.

In more detail, we plot the CPD distributions as presented in Figure 4.1. The Y-axis is the normalized value of CPD while the X-axis is the fraction of the respective population. We

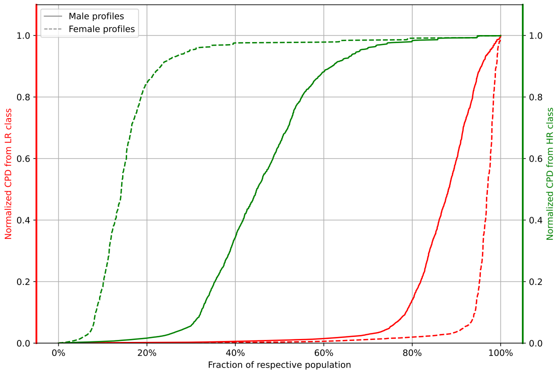


Figure 4.1 CPD distributions for attribute gender on Adult dataset.

plot the fraction of the population having a lower CPD value than the value on the Y-axis. In particular, all plots begin at $(0, 0)$ because 0% of the population has a lower normalized CPD than 0 (by definition) and end at $(100, 1)$ because at most 100% of the population has a lower normalized CPD than 1.

Afterward, the computation of the area between the curves corresponding to the CPD distributions relative to the same output class is performed. Then, the area between $CPD(Y_F, X'_{HR})$ and $CPD(Y_M, X'_{HR})$ is computed (green curves on Figure 4.1) as well as the area between $CPD(Y_F, X'_{LR})$ and $CPD(Y_M, X'_{LR})$ (red curves in Figure 4.1). The numerical values are obtained using the composite trapezoidal rule with the function *trapz* from Numpy [98]. Only the absolute values are being presented in Section 4.7. In a nutshell, high area values between distributions represent significant differences in the internal activation patterns of the model for these two distributions. In contrast, low area values imply small differences in the internal activation patterns of the model.

4.6 Experimental evaluation

4.6.1 Datasets description

We have conducted experiments on three distinct datasets that are commonly used in the fairness literature:

- The Adult dataset [68] contains the profiles of 45 222 individuals. Each profile is composed of six numerical attributes and eight categorical attributes among which the sensitive attributes usually considered are *gender*, *race*, and *age*. *income* is the decision attribute, separated in two classes: $income > 50k\$/year$ and $income \leq 50k\$/year$.

- The German Credit dataset [99] is composed of 1000 profiles of applicants for a credit loan, described by 21 attributes, including 14 categorical and 7 numerical. We chose to rely on the *age* as a sensitive attribute, after thresholding it at 25 as suggested in the work of Kamiran and Calders [33] as well as the *gender* as it has been done in previous work [100]. The binary decision attribute corresponds to a credit score (*good* or *bad*).

- The Law School dataset [29] is made of 18692 instances of law admission records described by six numerical attributes and six categorical ones. Based on a study across various law schools in the United States, the goal of prediction is to determine if a candidate would pass the bar exam (*pass* or *fail* labels). Note that the dataset is highly imbalanced towards the *pass* decision (90%) since most law students pass the bar exam successfully. While some previous works [101, 102] have considered the *race* as the sensitive attribute, others [103, 104] have relied on the *gender* of the candidate. In this work, we have considered both attributes separately.

4.6.2 Training procedure

As preprocessing steps to ensure that the data is formatted to be compatible with the training of NN, we have first performed a one-hot encoding of categorical and numerical attributes with less than 5 values, followed by a scaling between 0 and 1. For each dataset, we have trained a binary classifier for the associated task of prediction. Each of the NNs consists of an input layer, and several fully connected layers, using Leaky ReLU as the activation function, followed by the output layer of dimension two and the non-linear activation function Log Softmax. The datasets have been randomly split into three subsets: training, validation, and testing, of respective proportions 60%, 20%, and 20%. For each model, the number of neurons per layer has been chosen to maintain proximity to the number of attributes within the preprocessed dataset and to ensure enough complexity for effective learning. The number of layers assures that the models learn from the data while keeping reduced complexity. The learning rates have been chosen experimentally through iterative trials.

- Model for Adult dataset: two hidden layers of 128 neurons each, trained during 100 epochs and a learning rate of 0.001.

- Model for German Credit: two hidden layers of 80 neurons each, trained during 150 epochs and a learning rate of 0.001.

- Model for Law School: two hidden layers of 100 neurons each, trained during 500 epochs and a learning rate of 0.01.

4.7 Results

Adult dataset. The model trained on Adult achieves an accuracy of 84%. The areas between *CPD* curves are presented in Table 4.1, which summarizes the numerical values for the Adult dataset for three sensitive attributes studied:

- *Gender*: The protected groups are Male (68% of test set) and Female (32% of test set).
- *Race*: The dataset is split between profiles whose race is labeled as "White" (85% of test set) and "Non-white" (15%).
- *Age*: We created two age groups by thresholding the numerical age attribute at the median of 37 years, thus resulting in "Young" profiles representing half of the test set while "Aged" profiles represent the other half.

Sensitive attribute	Bias mitigation	Low Revenue	High Revenue
Gender	None	9.1037	29.8482
	CR	8.5171	29.6122
	DIR	9.0037	29.6873
Race	None	5.4543	17.1843
	CR	4.0613	11.9194
	DIR	5.2232	13.9488
Age	None	9.7791	25.7700
	CR	7.5391	16.7433
	DIR	N/A	N/A

Table 4.1 Area between CPD curves for Adult dataset.

Concerning the experiments with gender as the sensitive attribute, presented in Figure 4.1, one can notice that dashed curves exhibit a closer proximity to the class Low Revenue and a greater distance from the class High Revenue compared to plain curves, in terms of CPD. Thus, the CPD distributions for female inputs are "closer" to the class Low Revenue while being "further" away from class High Revenue than their male counterparts. This observation suggests that internal model representations of protected groups differ, implying a potential bias in the classifier's decision-making process. A similar trend can be observed when examining race and age as sensitive attributes. Graphical results are presented in Figure 4.2 for reference.

By comparison to the baseline, the implementation of bias mitigation methods has a high impact on CPD distribution gaps. For instance, Table 4.1 indicates that for both outcome class *Low Revenue* and *High Revenue*, CR and DIR yield substantial reduction of the disparities between curves, no matter which sensitive attributes are considered. Thus, the model's internal behavior after the preprocessing is less discriminatory between the protected and the privileged groups.

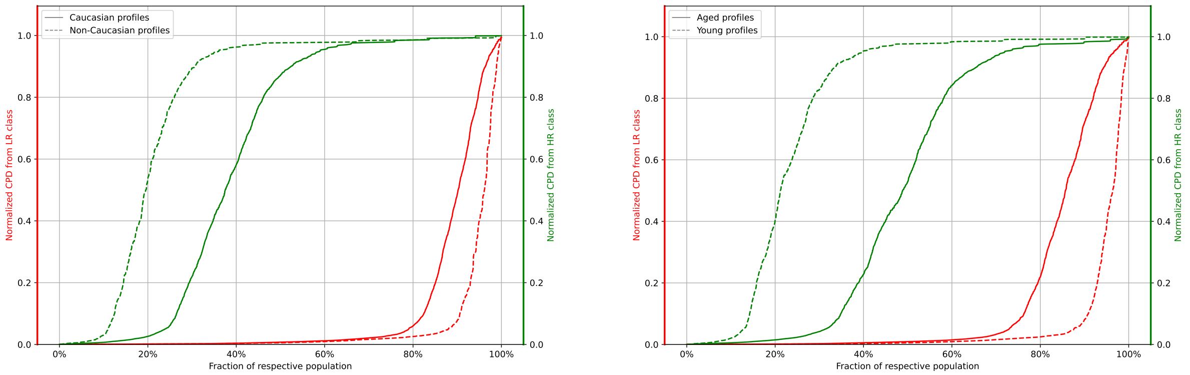


Figure 4.2 CPD distributions for sensitive attribute *race* (left) and *age* (right) on Adult dataset.

Such results are coherent with the idea that, for a model, similar internal activation patterns lead to less possible bias. Disparate Impact Remover has not been applied to the age feature in the dataset, because this method requires a binary attribute to work [85], while age is a continuous attribute. As CPD calculations depend on model structure (*cf.* Section 4.4), we have considered the same experiments with different structures. Our goal was not to conduct an extensive study on the impact of NN structure on CPD calculations, but rather explore this impact on a few other architectures. Considering the Adult dataset and gender as protected attributes, we present the results obtained on three different architectures:

1. Top left: two hidden layers of 500 neurons each, trained during 100 epochs, with a learning rate of 0.001.

2. Top right: three hidden layers of 500 neurons each, trained during 100 epochs, with a learning rate of 0.001.

3. Bottom: two hidden layers of 1000 neurons each, trained during 100 epochs, with a learning rate of 0.001.

These architectures were chosen to study an example of more complex models: more neurons with the same number of layers (1), more neurons with one additional layer (2), and a bigger number of neurons with the same number of layers (3). The number of epochs and the learning rate were chosen experimentally with repetitive trials. Information about the areas between curves is detailed in Table 4.2.

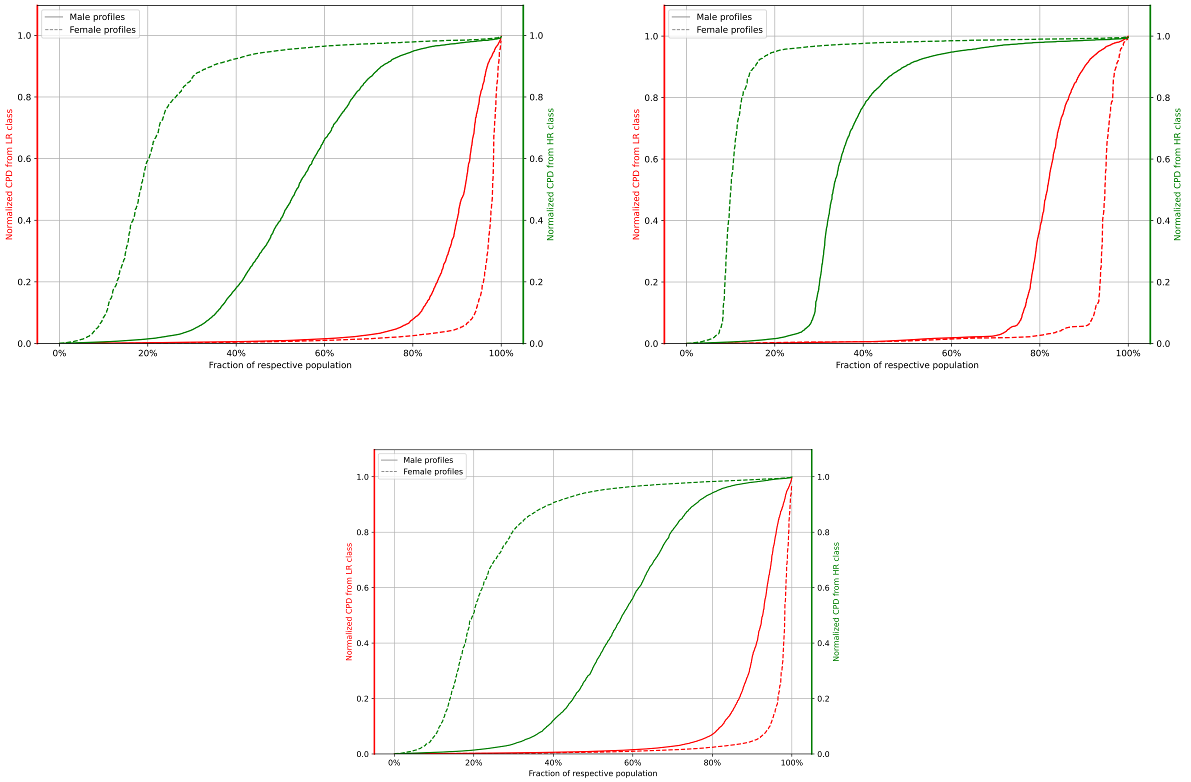


Figure 4.3 CPD distributions for sensitive attribute Gender on Adult dataset, two hidden layers of 500 neurons each (top left), three hidden layers of 500 neurons each(top right), two hidden layers of 1000 neurons each (bottom).

Results are similar to the experiment with the original structure presented in Figure 4.1. In particular, among the three architectures, the one with three layers is the most different from the others, hinting that the number of layers may be an influential factor in CPD calculations. However, it is interesting to note that the relative positioning of all four curves remains the same on the three graphics of Figure 4.3. Significant gaps between CPD distributions are also observed in these structures.

German Credit dataset. The model trained on German Credit reaches an accuracy of 77%. Previous work [33] gas found that thresholding the age at 25 years and creating *young* and

Sensitive attribute	Structure	Low Revenue	High Revenue
Gender	Two hidden layers of 128 neurons, Figure 4.1	8.7754	29.0924
	Two hidden layers of 500 neurons, Figure 4.3 (top left)	6.7357	32.2813
	Three hidden layers of 500 neurons, Figure 4.3 (top right)	11.8875	24.5813
	Two hidden layers of 1000 neurons, Figure 4.3 (bottom)	6.2292	33.6529

Table 4.2 Area between CPD curves for Adult dataset for sensitive attribute gender.

aged groups results in the maximum discrimination when considering the Demographic Parity fairness metric. More precisely, Aged people have 40% more chance to be assigned the credit class Good than Young people. Thus, we studied the attribute *age* as the sensitive attribute and considered the two groups mentioned above relative to the output class *Good* and *Bad*. The privileged group is constituted of profiles above 25 years old and represents 81% of the dataset. The dataset also contains an attribute called *Personal Status and sex*. Taking this into account, we studied the gender of the individuals, after separating male (69%) and female (31%) profiles. Similarly, numerical values for areas between CPD curves are presented in Table 4.3.

Sensitive attribute	Bias mitigation	Good	Bad
Age	None	9.4247	35.0674
	CR	0.2488	16.8462
	DIR	3.8861	21.6511
Gender	None	3.9276	10.9789
	CR	0.2681	3.8227
	DIR	3.5277	9.3884

Table 4.3 Area between CPD curves for German Credit.

It appears clearly that differences exist between the credit class *Good* and the credit class *Bad* among groups segmented by age and, to a lesser degree, gender. Namely, the activation patterns of the last hidden layer of the model demonstrate notable disparities between *young* profiles and *old* profiles, as well as between *male* and *female* individuals. A second notable observation concerns the use of bias mitigation methods: the areas between CPD distributions have been systematically reduced compared to the baseline, whether *age* or *gender* are considered. With the exception of DIR when studying gender as the sensitive attribute, reductions of 40% of the original value at least were observed. For illustration purposes, Figure 4.4 offers visual representations of the comparison between the baseline and the use of bias mitigation methods when considering *age* as sensitive.

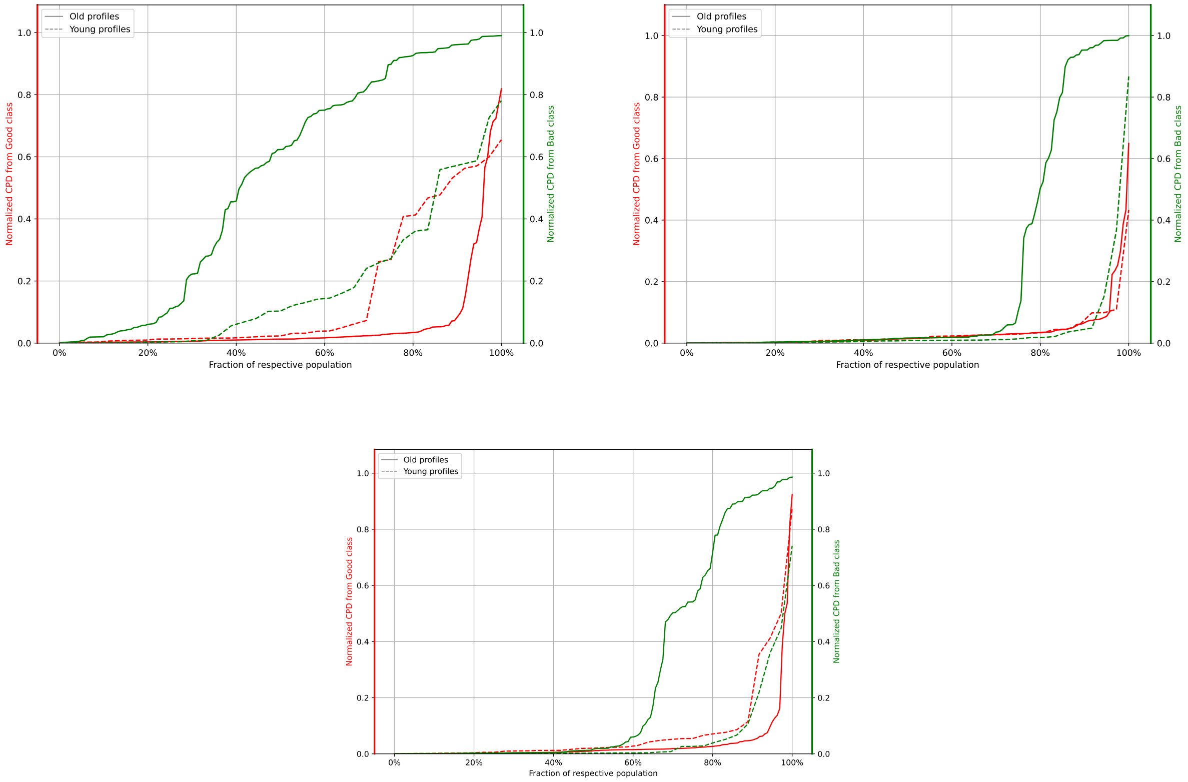


Figure 4.4 CPD curves for sensitive attribute age on German Credit, no treatment (top left), Correlation Remover (top right), Disparate Impact Remover (bottom).

Law School dataset. The model that we trained on the Law School dataset reaches an accuracy of 91%. The two sensitive attributes considered in this dataset are:

- *Race*: The dataset is split between profiles whose race is labeled as "White" (94% of test set) and the rest (6% of test set) "Non-White".
- *Gender*: the student's gender separates individuals between male (56%) and female (44%).

Numerical values for areas between CPD curves are presented in Table 4.4.

Sensitive attribute	**Bias mitigation**	**Fail**	**Pass**
Race	None	53.9589	9.0542
	CR	51.1550	8.4318
	DIR	53.3630	9.3748
Gender	None	3.2426	0.4999
	CR	2.8867	0.4307
	DIR	3.1553	0.5082

Table 4.4 Area between CPD curves for Law School.

The CPD distributions relative to the outcome classes *pass* and *fail* exhibit significant disparities when *race* is considered as the sensitive attribute. Indeed, the internal model activation patterns for individuals belonging to the protected group *non-white* display substantial differences in comparison to the privileged group composed of *white* individuals. A graphical representation of this phenomenon is depicted in Figure 4.5. In contrast, distinguishing the protected group from the privileged one proves to be more challenging. The implementation of Correlation Remover results in a decrease in the gaps between CPD distributions compared to the baseline, for both race and gender. However, this decrease does not surpass 14% of the original values and represents a smaller improvement than for Adult or German Credit. Furthermore, Disparate Impact Remover has a more mitigated effect on these CPD disparities. The impact of this bias mitigation on CPD remains tenuous and fluctuates in the range of 3% of the original value.

4.8 Discussion

Summary on findings. A natural idea coming from fair representation learning literature is that if the internal activation patterns of distinct protected groups are similar, then the

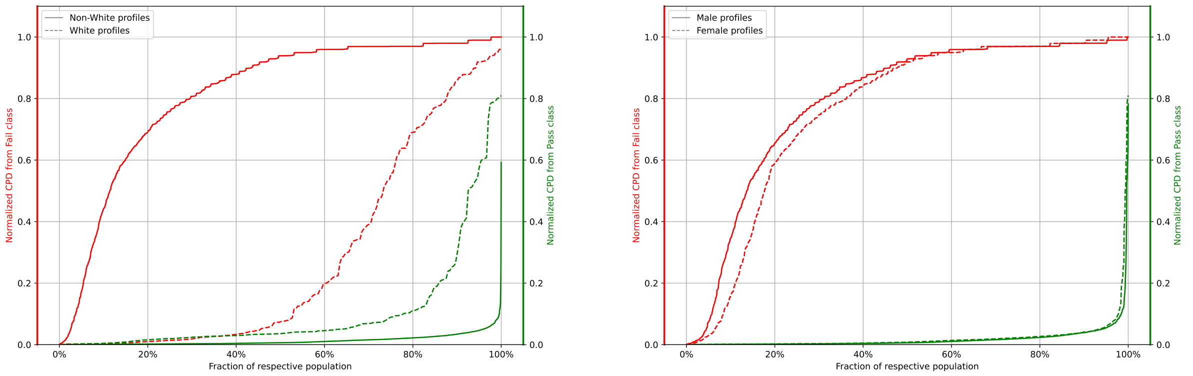


Figure 4.5 CPD curves for sensitive attribute age on Law School when considering *race* (left) and *gender* (right).

network will not discriminate inputs based on group membership (such as gender or race). By looking directly at statistical distributions of activation levels through CPL, our work aims to demonstrate that biases exist for classifiers trained on real-world datasets and that preprocessing fairness methods should aim at making such distributions more similar and less distinguishable.

The experiments presented in Section 4.7 indicate that large differences exist between CPD distributions of data points separated in privileged and protected groups, for several fairness datasets. Statistically, internal activation patterns of inputs divided according to sensitive features (such as gender or age) are different. We have observed that the models considered in the experiments treat individuals differently depending on sensitive features.

Moreover, we studied the impact of two fairness preprocessing techniques: Correlation Remover and Disparate Impact Remover, and we noticed that the disparities between CPD distributions were reduced for all the datasets studied: Adult, German Credit and Law school. In a perfectly fair model, the internal activation levels would be nearly identical between two sensitive groups, leading to negligible differences in CPD distributions.

Based on the findings, it appears that CPD is an effective approach for conducting fairness assessments. More precisely, this methodology demonstrates a substantial disparity in treatment between protected and non-protected groups before the application of bias removal treatments. With the implementation of these mitigation strategies, the separations between CPD distributions are significantly diminished, which underscores the efficacy of CPD as a tool for detecting biases and measuring the efficiency of bias reduction interventions.

On highly imbalanced datasets, such as Law School, in which the imbalance ratio for the class *fail* to *pass* is $1:9.18$, NNs may prioritize learning the frequency of output classes over relying on data features to infer predictions. This phenomenon can influence the internal model representations and CPD calculations, leading to reduced effectiveness of bias mitigation strategies compared to other, less imbalanced datasets.

Current limitations. Our work presents several limitations. First, the computation of CPL assumes that neurons are statistically independent. This initial assumption does not consider covariance and can lead to distortions. In particular, more complex and computationally expensive statistical models could be used, if required by attributes covariance values. In addition, the CPL method gives equal weights to all neurons in the computation while previous works have observed that neurons have varying importance [96]. More experiments could be performed to assess the impact of neuron weights on the overall results. As mentioned in Section 4.7, CPL measurements depend on model structure, and we did not conduct an extensive study on the impact of model dimensions on CPL. Rather, we have performed a few preliminary tests on various architectures. Finally, our study focuses exclusively on one sensitive attribute at a time, and future work will consider intersectional fairness.

4.9 Conclusion

Originally developed for OOD identification, CPL has been expanded in this study to encompass algorithmic fairness as a method for evaluating bias through an analysis of NN activation levels. Our investigation of three widely used datasets from the algorithmic fairness literature reveals considerable variation in internal activation patterns for profiles corresponding to distinct protected groups. After implementing pre-processing bias mitigation techniques, CPL typically exhibits a substantial reduction in these discrepancies. Since similar internal activation patterns result in comparable treatments by the model, our research demonstrates that CPL serves as an essential tool for assessing model bias and determining the efficacy of bias mitigation strategies.

While such results are promising, there are still many applications for CPL in fairness. Future work will investigate intersectional issues when combinations of sensitive attributes are considered, as well as the detection of proxy attributes. Other bias mitigation methods, such as in-processing and post-processing techniques, may be developed and used with CPL.

CHAPTER 5 POST-PROCESSING BIAS MITIGATION METHOD USING COMPUTATIONAL PROFILE LIKELIHOOD

The last chapter illustrated how CPL could be used to reveal model bias and evaluate pre-processing bias mitigation methods. This chapter presents another novel usage of CPL for algorithmic fairness. The following introduces an innovative post-processing method to ensure fairness while preserving performance for any fully connected DNN.

The main idea of this work is to correct the predictions of already trained algorithms. These networks may present bias toward protected subgroups, and the method that is presented in this section does not evaluate and correct the bias before prediction but rather rectifies it afterward.

5.1 Computational Profile Likelihood for Fairness Correction

A simplified version of the algorithm described in [25] is presented below. Specifically, among all possible thresholds of classification, the best thresholds for the sensitive groups are chosen to maximize a fairness criterion while avoiding too much performance degradation. We use thresholds on scores returned by the model, to swap inputs where the model is the most uncertain. In the following, we refer to this method as Threshold Optimizer (TO).

We also introduce in this section a post-processing algorithm using CPD based on the algorithm of TO, where thresholds are applied to the CPD of the inputs, instead of the model scores. We called this method CPD Thresholding. As a reminder, the details on CPD calculations are given in Section 4.4.

We present an algorithm for TO and CPD Thresholding in Algorithm 1. These methods differ in the *swap* method. *Swap* for TO is presented in Algorithm 2 and Algorithm 3 describe *swap* for CPD Thresholding.

The inputs for Algorithm 1 are the number of thresholds to test for the privileged class (called k) and for the unprivileged class (called k'), as well as the minimum and the maximum for the threshold to test (respectively min_{priv}, max_{priv} and min_{unpriv}, max_{unpriv}). Finally, parameter D represents the "degradation" allowed. For instance, if D is set at 0.02, we minimize the fairness metric in the region where the performance is at most 2% under its maximum value.

All the thresholds for the privileged and unprivileged groups are considered in lines 8 and 9 and stored respectively in arrays (lines 10 and 11). We apply the corrections of the model predictions applied with the *swap* method in line 12. Then, lines 13 and 14 store in arrays the fairness and performance metrics associated with the corrections. Once all the thresholds are explored, only the indices corresponding to the region of performance allowed by degradation are considered (line 19). Among these indices, the one that minimizes the fairness criterion is chosen (line 21). Finally, the corresponding thresholds are returned (line 24).

Algorithm 2 defines the *swap* method for TO. The variable *swapPriv* introduced at line 3 gathers all the individuals of the privileged group, that are favored (positive outcome) and with a model score for class favored below a threshold. These inputs are assigned to the unfavored outcome at line 6.

Conversely, *swapUnpriv* at line 5 aggregates all the unprivileged individuals that are unfavored (negative outcome) with a model score for class favored above a threshold, and line 7 assigns these individuals to the favored outcome.

Algorithm 3 is very similar to Algorithm 2. Instead of considering individuals whose model scores are below thresholds, individuals are selected based on their normalized CPD difference, defined in line 3. Inputs from the privileged group with a positive outcome, with a normalized CPD difference high enough are swapped to the unfavorable class. Individuals from unprivileged groups with negative outcomes, and with a normalized CPD difference below a threshold are assigned to favorable class.

In brief, TO considers model score probabilities to set the best thresholds. Our CPD correction method uses normalized CPD scores to define this region. As mentioned in Section 2.3, CPD was initially designed to detect Out-Of-Distribution inputs, i.e. inputs not belonging to the training distribution. Input with low CPL scores is not likely similar to the ones seen during training, from a statistical point of view. In the case of fairness correction, good candidates to be swapped are the ones that are not statistically similar to their output class. We aim to swap inputs (privileged into unfavored class, and unprivileged into favored class) that are not likely in the distribution of their output class until we reach the fairness criterion.

The Algorithm 3 relies on the normalized difference of CPD between favored and unfavored classes. This quantity, between 0 and 1, represents the likelihood to be similar to the unfavored class. Indeed, high values indicate high CPD for the favored class and low CPD for the unfavored class, thus a profile in this situation would be “further away” from the favored class than the unfavored class. Conversely, a quantity near 0 hints that the corresponding individuals are more similar to the favored class, for similar reasons.

Algorithm 1 Algorithm of TO and CPD Thresholding

```
Input: $data$
       $(t_{priv})_i$ ($k$ evenly spaced thresholds between $min_{priv}$ and $max_{priv}$)
       $(t_{unpriv})_i$ ($k'$ evenly spaced thresholds between $min_{unpriv}$ and $max_{unpriv}$)
       $D$ (degradation of performance allowed)

▷ Array to store fairness metric values, for instance DP difference: ◁
$fairMetricArray \leftarrow []$
▷ Array to store performance metric values, for instance the accuracy score: ◁
$perfMetricArray \leftarrow []$
$threshPrivArray \leftarrow []$
$threshUnprivArray \leftarrow []$
$cnt \leftarrow 0$

for $t_{priv} = t_{priv_0}, \ldots, t_{priv_k}$ do
    for $t_{unpriv} = t_{unpriv_0}, \ldots, t_{unpriv_k}$ do

        $threshPrivArray[cnt] \leftarrow t_{priv}$
        $threshUnprivArray[cnt] \leftarrow t_{unpriv}$

        $data \leftarrow swap(data, t_{priv}, t_{unpriv})$ ▷ Swap inputs in the region defined by thresholds

        $fairMetricArray[cnt] \leftarrow abs(computeFairMetric(data))$
        $perfMetricArray[cnt] \leftarrow computePerfMetric(data)$
        $cnt \leftarrow cnt + 1$

    end for
end for

▷ Only consider the region of performance of width D around the maximum: ◁
$bestPerfInd \leftarrow perfMetricArray \geq max(perfMetricArray) - D$

▷ Inside this region, choose the threshold that minimizes fairness criterion: ◁
$bestFairInd \leftarrow argmin(fairMetricArray[bestPerfInd])$

$thresholdPriv \leftarrow threshPrivArray[bestFairInd]$
$thresholdUnpriv \leftarrow threshUnprivArray[bestFairInd]$

return $thresholdPriv, thresholdUnpriv$
```

Algorithm 2 Swap functions for TO

```
Input: $data$
       threshPriv
       threshUnpriv

$newData \leftarrow data.copy()$

$\triangleright$ Select favored privileged inputs with model scores below threshold $\triangleleft$
$swapPriv \leftarrow data.protectedAttr = priv$ and
                    $data.predicted = fav$ and
                    $data.scores_{fav} < threshPriv$

$\triangleright$ Select unfavored unprivileged inputs with model scores above threshold $\triangleleft$
$swapUnpriv \leftarrow data.protectedAttr = unpriv$ and
                      $data.predicted = unfav$ and
                      $data.scores_{fav} > threshUnpriv$

$newData[swapPriv] \leftarrow unfav$        $\triangleright$ Assigning unfavored class to privileged inputs
$newData[swapUnpriv] \leftarrow fav$        $\triangleright$ Assigning favored class to unprivileged inputs
return $newData$
```

5.2 Experimental Setup

To conduct our experiments, we trained deep NNs on three datasets: Adult Census Income, German Credit, and Dutch Census. A description of Adult Census Income and German Credit can be found in Section 4.6.1.

Dataset Name	Favored Class	Unfavored Class	Protected Attribute Name	Privileged Group	Unprivileged Group
Adult	High Revenue	Low Revenue	Gender	Male	Female
			Age	Aged	Young
			Race	White	Non-White
German	Good	Bad	Gender	Male	Female
			Age	Aged	Young
Dutch	High	Low	Gender	Male	Female
			Age	Aged	Young

Table 5.1 Summary of all datasets studied

The Dutch Census assembled demographic information of 60,420 Dutch people. Composed of categorical 12 attributes, the dataset is considered in fairness work [105] with its binary protected attribute *gender*, with its prediction task *occupation*. The goal is to determine if an individual has a *Low Level* or *High Level* occupation, the latter being the favored class. We also binarized the attribute *age*, into two sensitive groups: “Aged” and “Young”.

Algorithm 3 Swap functions for CPD Thresholding

```
Input: $data$
       threshPriv
       threshUnpriv

$newData \leftarrow data.copy()$

▷ We consider the normalized difference of CPD of fav and unfav class: ◁
$data.CPD_{norm} \leftarrow normalize(data.CPD_{fav} - data.CPD_{unfav})$

▷ Select favored privileged inputs with CPL for the favored class above threshold ◁
$swapPriv \leftarrow data.protectedAttr = priv$ and
                    $data.predicted = fav$ and
                    $data.CPD_{norm} > threshPriv$

▷ Select unfavored unprivileged inputs with CPL for the unfavored class below threshold ◁
$swapUnpriv \leftarrow data.protectedAttr = unpriv$ and
                      $data.predicted = unfav$ and
                      $data.CPD_{norm} < threshUnPriv$

$newData[swapPriv] \leftarrow unfav$          ▷ Assigning favored class to unprivileged inputs
$newData[swapUnpriv] \leftarrow fav$          ▷ Assigning unfavored class to privileged inputs
return $newData$
```

All the datasets considered, and their protected attributes, are summarized in Table 5.1.

Before inference, the same treatment as in Section 4.6.2 is applied to the data: a one-hot encoding of categorical and numerical attributes with less than 5 values, and a scaling between 0 and 1 of the numerical features.

The models used in our experiments are the following:

- Model for Adult Census: two hidden layers of 128 neurons each, trained during 100 epochs and a learning rate of 0.001. The accuracy reached is approximately 84%.
- Model for German Credit: two hidden layers of 80 neurons each, trained during 150 epochs and a learning rate of 0.001. The accuracy reached is approximately 77%.
- Model for Dutch Census: two hidden layers of 64 neurons each, trained during 60 epochs and a learning rate of 0.01. The accuracy reached is approximately 83%.

All the additional details about training and hyper-parameters are given in Section 4.6.

Then, we apply three post-processing methods, namely ROC, TO, and the CPD Thresholding. For these methods, we use the accuracy score as the performance metric and study three fairness criteria: DP difference, AO difference, and EO difference. We allowed a degradation of 2% of the accuracy. For all three methods, we considered 100 evenly spaced thresholds between 0 and 1 for privileged and unprivileged groups. Finally, the fitting step is done on the training and the validation set, and the correction is applied to all the datasets.

It is important to acknowledge that each of the proposed techniques (ROC, TO, CPD Thresholding) presented above requires the sensitive attribute to be available during the post-processing operations.

5.3 Experimental Results

5.3.1 Post-Processing Methods Results

For all protected attributes of all three datasets, we applied the three post-processing methods and compared the fairness metric (DP difference, EO difference, and AO difference) and the accuracy score between each method and the baseline, when no method was used.

Since all three methods are *Thresholding* methods, they can achieve fairness with no remaining disparities, as long as the threshold criterion (CPD, or the model scores) of the inputs are distinct, at the expense of performance. The results presented in Table 5.2, for *gender* attribute of the Adult dataset, shows that these methods remain significantly effective even if the allowed degradation of performance is at most 2% of its maximum value. Indeed, no matter which fairness metric is studied, it is always closer to 0, in comparison to the baseline. CPD Thresholding presents competitive results with ROC and TO, in terms of fairness. Overall, the accuracy score is slightly better with CPD Thresholding than other methods.

However, fairness metrics are not systematically at their minima, because of the fairness-utility trade-off. Another reason for this phenomenon is because of the parameters of the search for thresholds: the number of thresholds between the minimum and the maximum allowed are not enough to reach the objectives. One could reach better results by increasing the thresholds to test and narrowing down the upper and lower bounds, at the expense of computational speed.

For all the protected attributes of all the datasets, CPD Thresholding, TO, and ROC achieve a significant decrease in unfairness, whether considering DP, EO, or AO criterion, while maintaining a good accuracy, with at most 2% of degradation relative to the original accuracy. As a result, fairness metrics are significantly reduced compared to the baseline. For clarity, we only present the results of *gender* attribute for each dataset. All the other results can be found in the Appendix A.

Method name	Fairness criterion	Fairness metric	Accuracy Score (%)
No Method	DP Difference	.1918	83.84
	EO Difference	.1289	83.84
	AO Difference	.1085	83.84
ROC	DP Difference	.0019	81.94
	EO Difference	.0063	83.60
	AO Difference	.0020	82.94
TO	DP Difference	.0018	81.95
	EO Difference	.0068	83.57
	AO Difference	.0005	83.41
CPD	DP Difference	.0011	82.17
	EO Difference	.0028	83.64
	AO Difference	.0007	83.41

Table 5.2 Fairness and performance achieved on Adult, with attribute gender

Table 5.2 presents the results for *gender* attribute of Adult. In this specific example, CPD Thresholding keeps a better accuracy score than TO and ROC, no matter which fairness metric is considered. In parallel, CPD Thresholding has the smallest DP Difference and EO Difference. Concerning AO Difference, TO is the best of all three methods, with a difference of .0002 with CPD Thresholding.

Table 5.3 aggregates the results for *gender* attribute of German. As mentioned above, all three methods achieve a significant decrease in DP EO and AO differences compared to the baseline, in the region of 2% of degradation performance. Once again, CPD Thresholding presents better accuracy scores. Furthermore, absolute values of fairness metrics are also closer to 0 when applying the post-processing method based on CPD. It is interesting to note that the AO difference when using this technique is below zero (-.0013): it shows that the corrections made to the model predictions are enough to advantage the unprivileged group over the privileged group, thus reverting the bias.

Finally, Table 5.4 shows performance and fairness metrics for the attribute *gender* of the Dutch dataset. Similarly to Table 5.3, CPD Thresholding shows superior results both in terms of accuracy score and bias correction, compared to the baseline, but also ROC and TO. In particular, CPD Thresholding achieves EO with a precision of four digits while the other two methods perform with at most a two-digits precision.

Method name	Fairness criterion	Fairness metric	Accuracy Score (%)
No Method	DP Difference	.1690	78.10
	EO Difference	.1039	78.10
	AO Difference	.1571	78.10
ROC	DP Difference	.0172	76.80
	EO Difference	.0343	76.70
	AO Difference	.0149	77.40
TO	DP Difference	.0162	77.30
	EO Difference	.0322	78.00
	AO Difference	.0108	77.10
CPD	DP Difference	.0069	77.80
	EO Difference	.0231	78.00
	AO Difference	-.0013	78.00

Table 5.3 Fairness and performance achieved on German, with attribute gender

Method name	Fairness criterion	Fairness metric	Accuracy Score (%)
No Method	DP Difference	.3385	83.03
	EO Difference	.0879	83.03
	AO Difference	.1521	83.03
ROC	DP Difference	.1370	81.03
	EO Difference	.0014	82.50
	AO Difference	.0015	81.74
TO	DP Difference	.1344	80.99
	EO Difference	.0011	82.04
	AO Difference	.0012	81.18
CPD	DP Difference	.1288	81.05
	EO Difference	.0000	82.59
	AO Difference	.0011	82.01

Table 5.4 Fairness and performance achieved on Dutch, with attribute gender

5.3.2 Comparison Between Methods

Since the three methods aim for the same fairness objectives using different means, it is relevant to study the profiles of the individuals affected by at least two methods.

A closer look at the scores and the CPD distributions of these affected individuals illustrates how these post-processing methods work, and details what are the profiles of individuals affected by at least two methods, and on which regions such techniques disagree. To this end, we have plotted the distributions of model scores of individuals (privileged or unprivileged) affected by CPD Thresholding and another method (either TO or ROC). We also depicted the distributions of normalized CPD difference for these inputs. Due to the high number of protected attributes studied in this work, we decided for brevity to only interpret a few examples of each dataset.

Demographic Parity for Gender of Adult Dataset

Table 5.5 summarizes the number of these individuals for each couple of methods when studying *gender* protected attribute of the Adult Census dataset, trying to reach the DP criterion.

	TO	ROC	CPD
TO	4382		
ROC	4338	4428	
CPD	3572	3591	4864

Table 5.5 Numbers of profile affected by two methods on Adult, with attribute gender, for DP

A significant portion of the affected individuals by CPD Thresholding are also affected by either TO or ROC: around 74% of all profiles changed by CPD Thresholding are also changed by ROC, and 73% of these individuals are changed by TO. Due to the similarities of TO and ROC (they use the model scores as decision criterion), a significant part of individuals affected by TO is also affected by ROC. Overall, the three methods perturb the original classification on a similar number of individuals (CPD Thresholding affects a slightly higher number of individuals than the other methods).

In this context, we present the comparison between TO and CPD Thresholding, both in terms of model scores and difference of CPD. Figure 5.1 shows the distributions of model scores of affected privileged profile (individuals whose *gender* is labeled as “male”) that have

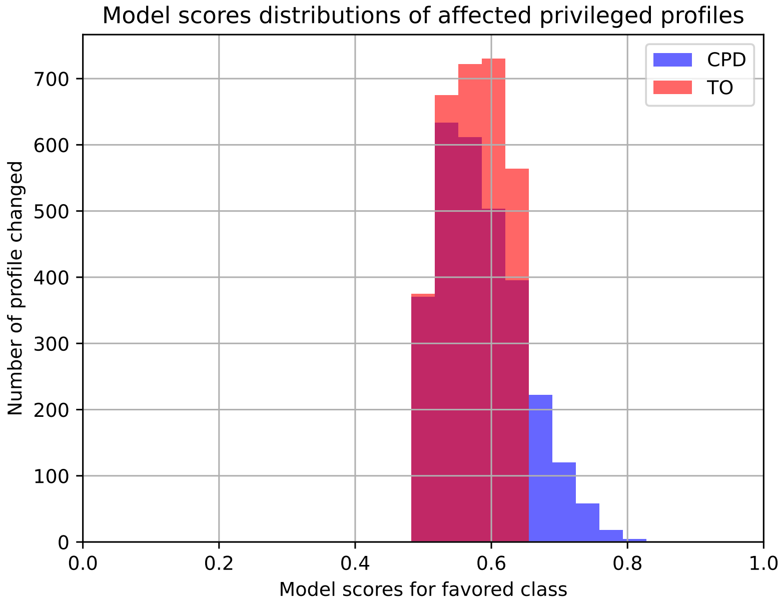


Figure 5.1 Scores distribution for the favorable class of affected privileged profiles by ROC and CPD Thresholding, considering DP and attribute *gender* of Adult

been affected by the post-processing fairness correction: in blue, by CPD Thresholding, and in red by TO. As mentioned above in Table 5.5, a significant portion of profiles is affected by both methods. They are represented as the superposition of the blue and the red histograms in the plots, i.e. in plum color in the diagrams. Along the X-axis, a greater value means that the model assigned a greater probability that the inputs are labeled as *High Revenue* (favorable class).

The red distribution corresponding to individuals changed by TO is less spread out than the blue distribution, and more concentrated on the right of the decision boundary of 0.5. This is explained by the algorithm of TO itself (Algorithm 2): only the privileged individuals with the smallest scores get affected. CPD Thresholding is not based on model scores, and considers profiles with higher scores, up to 0.8.

Figure 5.2 presents the same individuals and the same method but considers the normalized CPD difference along the X-axis. A greater value means that the CPD for the favorable class is higher than the CPD for the unfavorable class, hinting that the inputs are more likely to belong to the unfavorable class, *Low Revenue*. Conversely, a value close to 0 means that the profiles are more susceptible to belonging to the favorable class *High Revenue*.

The blue distribution is less spread out than the distribution of TO and more concentrated towards a normalized CPD difference of 0.5. As earlier, this is explained by the algorithm of

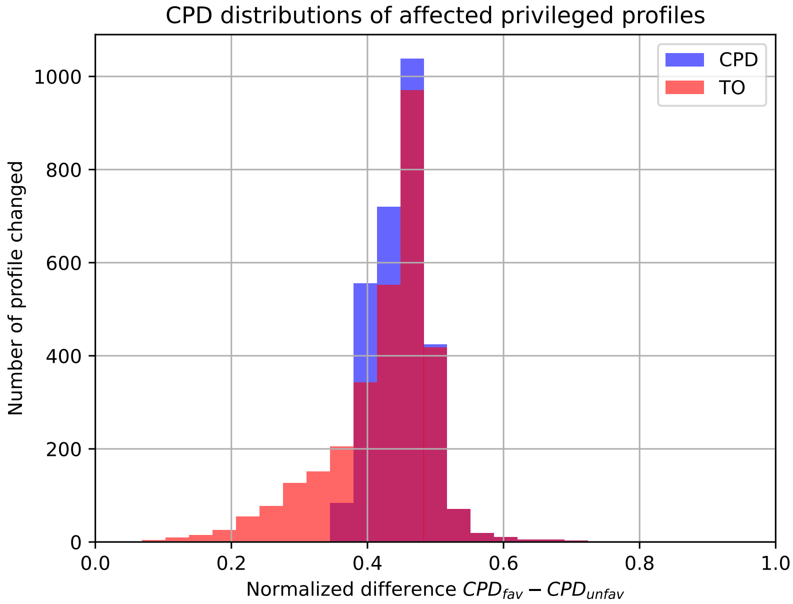


Figure 5.2 Distribution of CPD difference for favorable class of affected privileged profiles by ROC and CPD Thresholding, considering DP and attribute *gender* of Adult

CPD itself (Algorithm 3): the privileged profiles of the highest CPD difference get targeted first by the post-processing techniques. Conversely, as TO only considers model scores, corresponding affected individuals have differences of CPD more spread out, up to 0.1.

Equality of Opportunity for Race of Adult Dataset

A second example of a comparison of post-processing methods is detailed below. We consider the attribute *race* of the Adult dataset. EO is the fairness criterion considered. Table 5.6 details the number of profiles affected by each couple of methods.

	TO	ROC	CPD
TO	822		
ROC	812	1811	
CPD	17	17	110

Table 5.6 Numbers of profile affected by two methods on Adult, with attribute Race, for EO

Conversely to the study on the gender attribute, CPD Thresholding affects way fewer individuals than the other two methods, especially ROC, that modify output class for 1811 individuals. While ROC and TO agree on most individuals, CPD Thresholding impacts in majority individuals untouched by the other two methods: only 17 out of 110 inputs are shared with the other algorithms. For race attribute, CPD Thresholding shows strong disagreement.

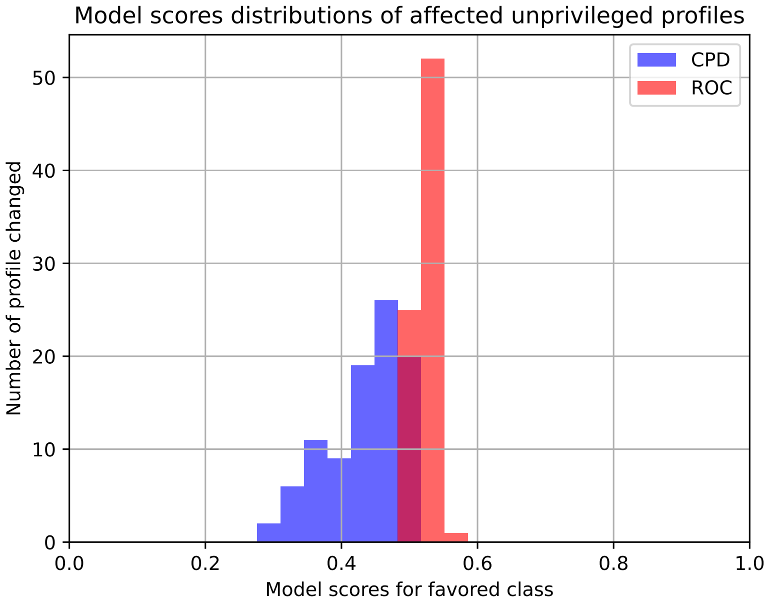


Figure 5.3 Scores distribution for the favorable class of affected unprivileged profiles by ROC and CPD Thresholding, considering EO and attribute *race* of Adult

We study in more detail the comparison between ROC and CPD Thresholding and their impact on unprivileged profiles (individuals whose *race* is labeled as “non-white”) in Figure 5.3 and Figure 5.4. The first one diagrams the distribution of model scores.

The red distribution corresponding to ROC looks like a peak of width 0.02 centered around 0.5. It is coherent with the ROC classification threshold of approximately 0.5657 and the ROC margin of 0.0132, showing that ROC only affects individuals in this critical region of the model scores. Comparatively, the blue distribution of CPD Thresholding is wider towards lower score value: since model scores are not the criterion to swap outcome class of individuals, CPD Thresholding can affect individuals with low scores, outside of the critical region of ROC. The distributions overlap around 0.5 of model scores, for 20 individuals. Figure 5.4 presents the same individuals affected by the same methods but considers the normalized CPD difference along the X-axis.

The distributions are centered around the same value of 0.5. Red distribution corresponding to ROC is more spread out towards higher CPD difference, which is expected, as CPD

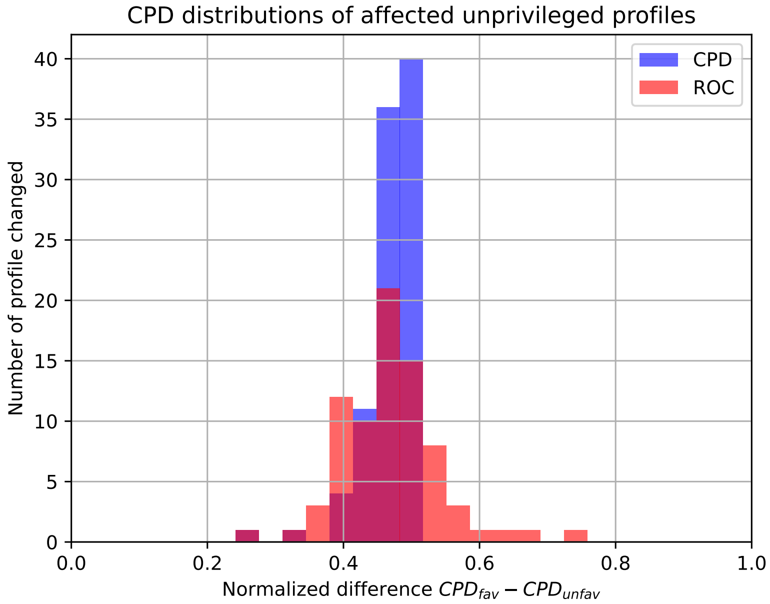


Figure 5.4 Distribution of CPD difference for the favorable class of affected unprivileged profiles by ROC and CPD Thresholding, considering EO and attribute *race* of Adult

Thresholding targets unprivileged individuals with the lowest CPD difference, i.e. the highest likelihood to belong to the favorable class. Surprisingly, the ROC distribution is also spread towards a lower CPD difference than the blue distribution. This phenomenon is explained when considering that CPD Thresholding can not assign unprivileged individuals ("non-white") to unprivileged class (*Low Revenue*), while ROC can, by adjusting the classification threshold for all individuals. In this situation, the ROC classification threshold is 0.5657, higher than the classification threshold of 0.5. Non-white individuals, outside the critical region, but above the usual classification threshold (for instance with a score of 0.53) have their outcome class changed from *High Revenue* to *Low Revenue*.

Average Odds for Age of German Dataset

In comparison to the Adult dataset, the German dataset presents a smaller sample size, and a smaller number of inputs are affected by the post-processing fairness corrections. When evaluating model corrections with the AO criterion, Table 5.7 counts the number of proposed corrections by each method and identifies those shared between at least two strategies.

In this peculiar situation, ROC and TO fully agree on the 41 corrections to be made. CPD Thresholding impacts approximately twice as many individuals and agrees with other cor-

	TO	ROC	CPD
TO	41		
ROC	41	41	
CPD	22	22	78

Table 5.7 Numbers of profile affected by two methods on German, with attribute Age, for AO

rections in 22 cases out of 78.

The following graphs illustrate in more detail the comparison between TO and CPD Thresholding by plotting the score and the CPD distributions of impacted privileged profiles. Figure 5.5 displays the scores distribution of these individuals (i.e. profiles in the "aged" category), and Figure 5.6 illustrates the normalized CPD difference distribution of the same individuals. By comparing Figure 5.5 and Figure 5.6, we can characterize which individuals are modified in priority by TO (and thus by ROC) and CPD Thresholding.

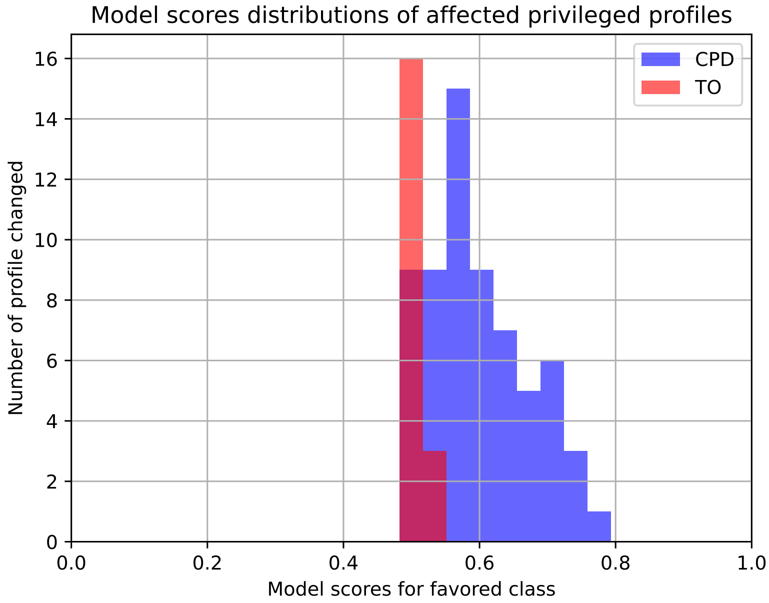


Figure 5.5 Scores distribution for the favorable class of affected unprivileged profiles by TO and CPD Thresholding, considering AO and attribute *age* of German

Figure 5.5 shows that individuals that get favored by classification with the lowest score are swapped by TO. Compared to the blue distribution, the red one is much more centered around the 0.5 frontier. CPD Thresholding does not consider model scores to swap inputs.

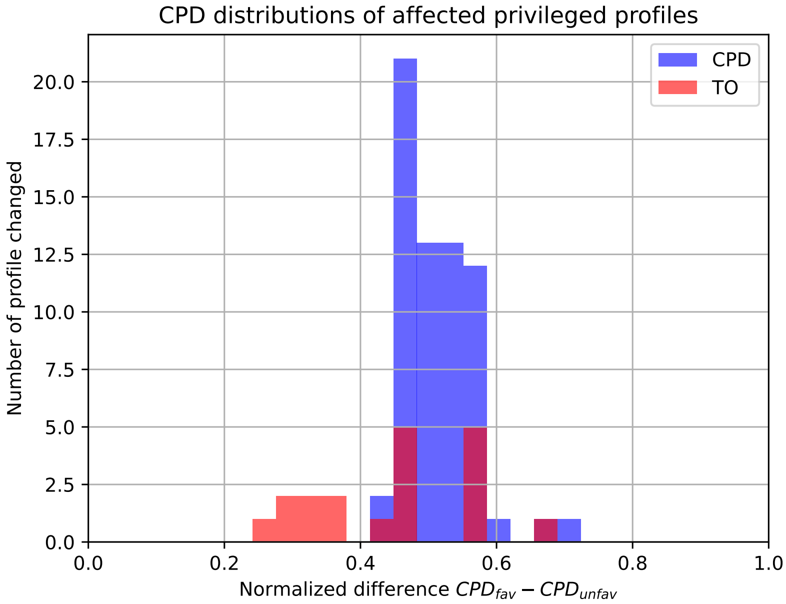


Figure 5.6 Distribution of CPD difference for favorable class of affected unprivileged profiles by TO and CPD Thresholding, considering AO and attribute *age* of German

Nevertheless, this method does not select individuals with scores higher than 0.8, demonstrating that CPD does not disagree with the model classification, and does not select irrelevant inputs.

On Figure 5.6, the blue distribution is to the right of the TO distribution: it illustrates that CPD Thresholding selects by definition the closest favored individuals from the unfavored class and the furthest from the favored class, in terms of CPD. The corrections agree on individuals with CPD differences higher than 0.4. The individuals where TO and CPD Thresholding do not agree are individuals with model scores near 0.5, near the favored class, in terms of CPD. It represents specific profiles for which the model is unsure, and requires human expertise to avoid injustice.

Demographic Parity for Gender of Dutch Dataset

Finally, we focus on the Dutch census, which is a dataset more similar to the Adult census in size and for its prediction task. The objective of prediction is to determine whether individuals have prestigious professions or not. The following analysis focuses on the *gender* attribute, separated between men and women. The fairness criterion considered for this paragraph is DP. Table 5.8 counts the number of individuals affected by each pair of methods.

	TO	ROC	CPD
TO	6155		
ROC	6008	6077	
CPD	5783	5753	6323

Table 5.8 Numbers of profile affected by two methods on Dutch, with attribute Gender, for DP

The three methods impact approximately the same number of inputs (around 6000) when trying to reach the DP. As for the precedent situations, CPD Thresholding shares fewer corrections with ROC and TO than between these two methods. As ROC and TO rely on model scores to formulate corrections, it is natural to observe this trend. Nonetheless, above 90% of all impacted profiles by CPD Thresholding are also impacted either by ROC or TO. These numbers show that for attribute gender of the Dutch dataset, the three methods reached an agreement on most of the impacted profiles.

The following Figures 5.7 and 5.8 compare affected women by ROC and affected women by CPD Thresholding. In particular, Figure 5.7 presents the model scores distribution of the affected women. As expected, swapped individuals for ROC (red distribution) are concentrated in a critical region, centered around 0.53, of width 0.18. In comparison, CPD Thresholding agrees with ROC on most profiles and affects in addition profiles with lower scores (up to 0.2).

Figure 5.8 displays the normalized CPD difference of these same inputs. ROC affects “blindly” profiles very close to the unfavored class and far from the favored class, while CPD Thresholding focuses on inputs with the lowest CPD difference. Both methods agree on most individuals to affect: profiles near the decision border, close to the favored class, and far from the unfavored class, in terms of CPD. Among women affected by only one method, we must distinguish between women affected by ROC and women affected by CPD Thresholding. Females swapped by ROC only have model scores near the border of 0.5, but are very close to the unfavored class. On the contrary, females swapped by CPD Thresholding present model scores between 0.2 and 0.3, but are much closer to the favored class, in terms of CPD.

Once again, these profiles may benefit from human expertise: they lie where the model prediction and its activation pattern are the least similar.

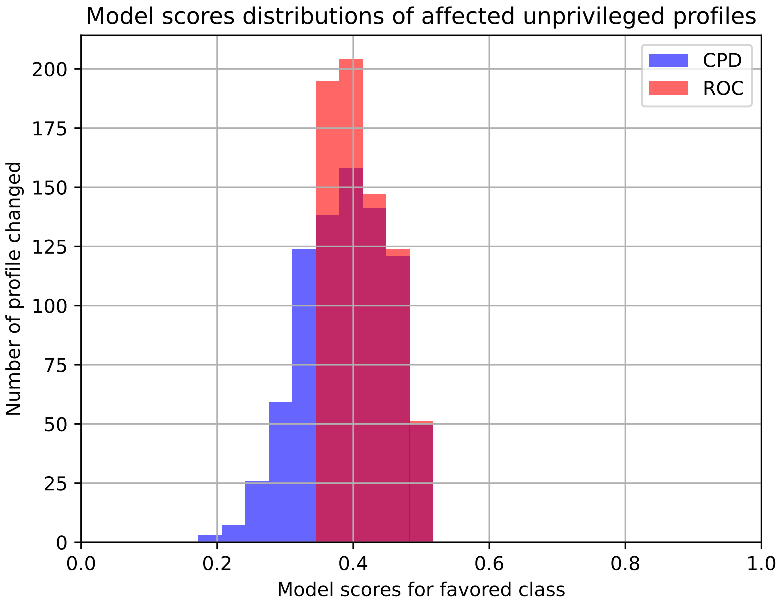


Figure 5.7 Scores distribution for the favorable class of affected unprivileged profiles by ROC and CPD Thresholding, considering DP and attribute *sex* of Dutch

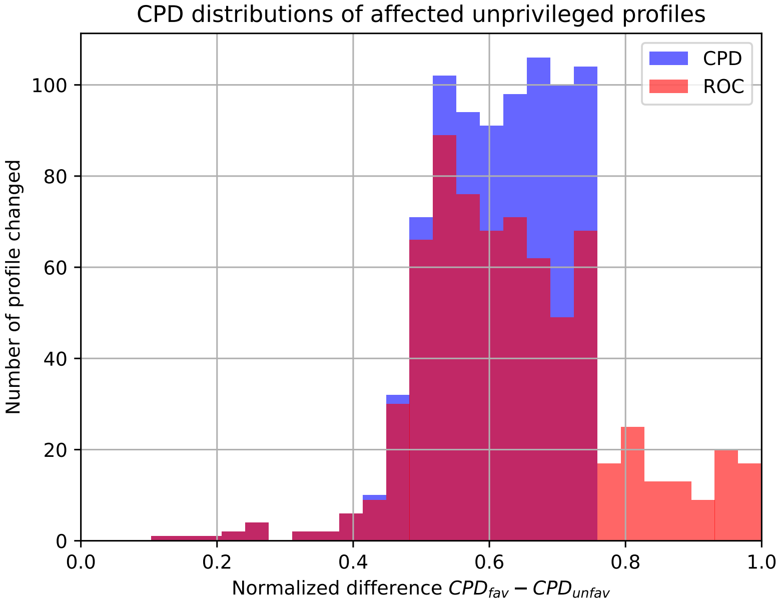


Figure 5.8 Distribution of CPD difference for the favorable class of affected unprivileged profiles by ROC and CPD Thresholding, considering DP and attribute *sex* of Dutch

5.3.3 Degradation/Fairness Trade-off

Building upon our earlier investigations into bias correction strategies comparison at a fixed degradation level, this section aims to broaden the scope of analysis by measuring performance across multiple degrees of degradation. This expanded study offers essential insights into the capabilities of each method under varying levels of performance degradation. At lower levels of degradation, bias correction techniques focus on minimizing unfairness within a restricted performance range, potentially resulting in low equity performance. Conversely, higher degradation levels allow for greater performance loss in pursuit of perfect fairness. Our comparison focuses on two research questions: Which method performs superiorly at a given level of degradation? Is the observed performance trend consistent across various degrees of degradation levels? In other words, does a strategy benefit more from the degradation/fairness trade-off than the others?

To answer these interrogations, we focus on various sensitive attributes of the three datasets presented above, and we compute the distribution of fairness metrics for a given range of degradation levels, for CPD Thresholding, ROC, and TO. We have followed the same procedure and used the same parameters as in Section 5.2, except for the introduction of varying degrees of degradation.

Demographic Parity of Gender of Adult Dataset

Figure 5.9 plots the evolution of DP difference for the attribute *gender* of the Adult dataset under varying levels of degradation. CPD Thresholding, ROC, and TO are each represented by a curve. A noticeable similarity exists among these methods: DP difference attains its maximum when no corrections are applied (zero degradation), and reaches its minimum close to 0 (DP is enforced) with degradation of 2% or higher. The curves exhibit proximity, particularly those for ROC and TO. The curve corresponding to CPD Thresholding lies slightly below the other two curves within the degradation range between 0.4% and 1.8%. This pattern conforms with the findings from the preceding section, specifically Table 5.2, revealing a marginally better fairness performance for CPD Thresholding under similar degradation levels.

In light of the dual objectives of maintaining fairness and achieving optimal performance levels, an in-depth analysis must also consider the evolution of accuracy scores for each method across varying degradation levels. Figure 5.10 showcases the accuracy score when the degradation level varies. Similar to the preceding distribution trends, these strategies present a comparable trajectory as degradation increases: they collectively tend towards

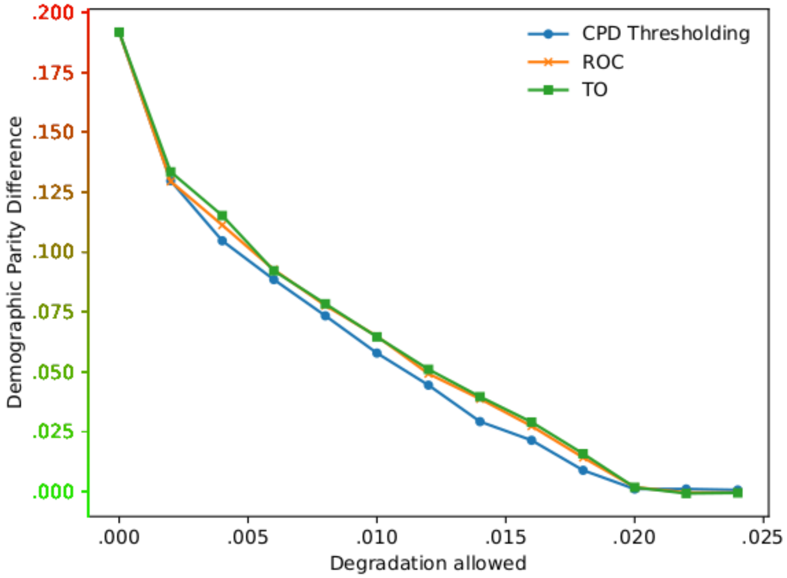


Figure 5.9 DP difference on multiple degradation levels (the lower the better)

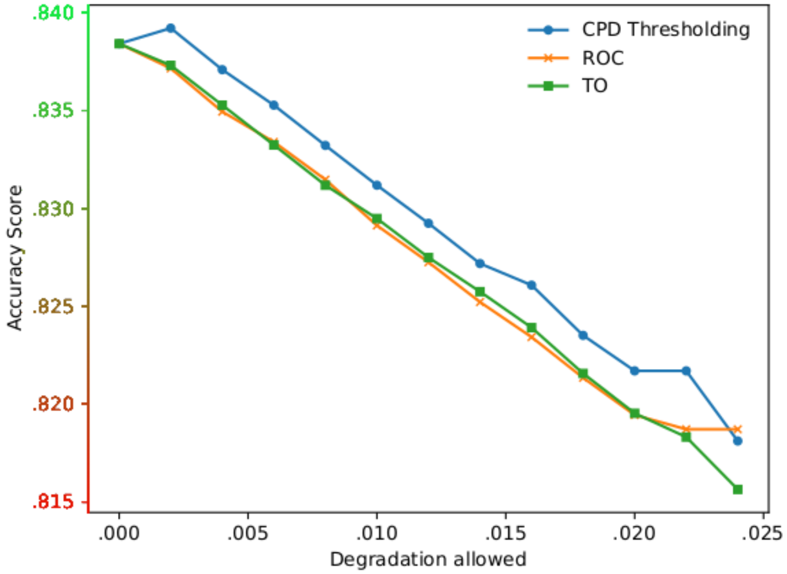


Figure 5.10 Accuracy Score on multiple degradation levels (the higher the better)

lower accuracy levels. Once again, ROC and TO are very similar. CPD Thresholding curve stands above both ROC and TO within the degradation range between 0.2% and 2.2% of degradation, exhibiting superior performance results overall. Notably, at a degradation level of 0.2%, CPD Thresholding corrections outperform the baseline by approximately 0.1% in terms of accuracy.

Equality of Opportunity of Gender of German Dataset

This analysis focuses on correcting predictions for the EO based on the attribute Gender of the German Dataset. In contrast to our earlier findings, Figure 5.11 shows that the three methods require less degradation to reach a plateau. At low degradation levels (between 0 and .002) ROC exhibits better fairness results compared to the other two methods. On high degradation settings (higher than .004), CPD Thresholding demonstrates superiority in terms of fairness performance. Overall, the three methods fail to reach EO closer than a difference of .02. Indeed, this criterion is more stringent to attain than DP.

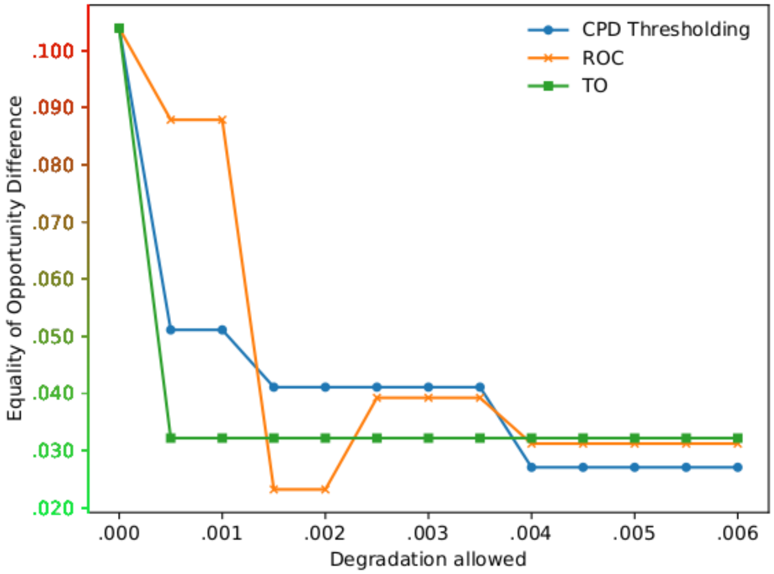


Figure 5.11 EO difference on multiple degradation levels (the lower the better)

Figure 5.12 presents the evolution of accuracy under varying degrees of degradation. Similarly to the findings from the previous Adult dataset example, CPD Thresholding outperforms the baseline for degradation levels, as well as ROC, on low degradation levels. CPD Thresholding exceeds ROC and TO for all the degradation levels considered. Throughout most of the degradation levels, TO struggles to match the performance levels of the other two strategies.

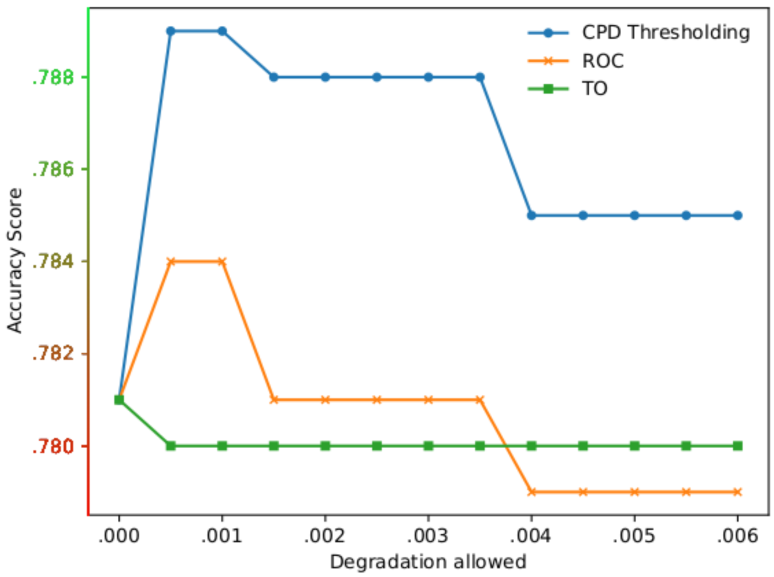


Figure 5.12 Accuracy Score on multiple degradation levels (the higher the better)

Thus, the optimal strategy varies depending on the degradation level and the relative importance assigned to the accuracy score and EO constraint. Contrary to the previous examples, CPD Thresholding does not display better performance and fairness on every degradation level, but rather in specific regions.

Average Odds of Gender of Dutch dataset

Finally, the last case study of our analysis investigates the gender attribute of the Dutch dataset, with a focus on the AO objective. Figure 5.13 showcases a comparison between CPD Thresholding, ROC, and TO as the degradation level progresses. Similarly to the first example of this Section, CPD Thresholding demonstrates a better fairness score than the other two strategies throughout the entire range of degradation levels considered, ultimately reaching a plateau at approximately 1% of accuracy loss.

Regarding the accuracy score, the three algorithms present comparable performances under the 1% degradation threshold, according to Figure 5.14. Past this threshold, CPD Thresholding maintains a constant accuracy level of around 82%, while other strategies continue to experience performance decreases as the degradation intensifies.

In this case study, CPD Thresholding stands out comparatively to ROC and TO, particularly under the 1% degradation threshold. At these levels of degradation, the algorithm achieves

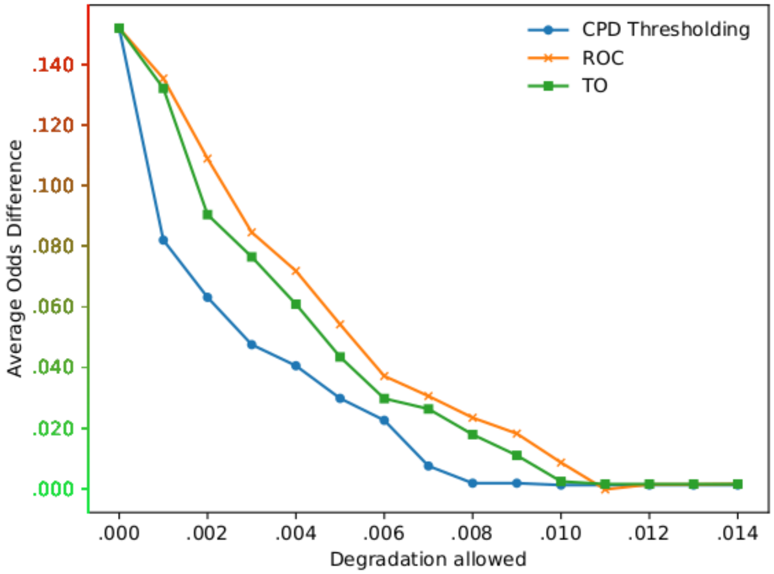


Figure 5.13 AO Difference on multiple degradation levels (the lower the better)

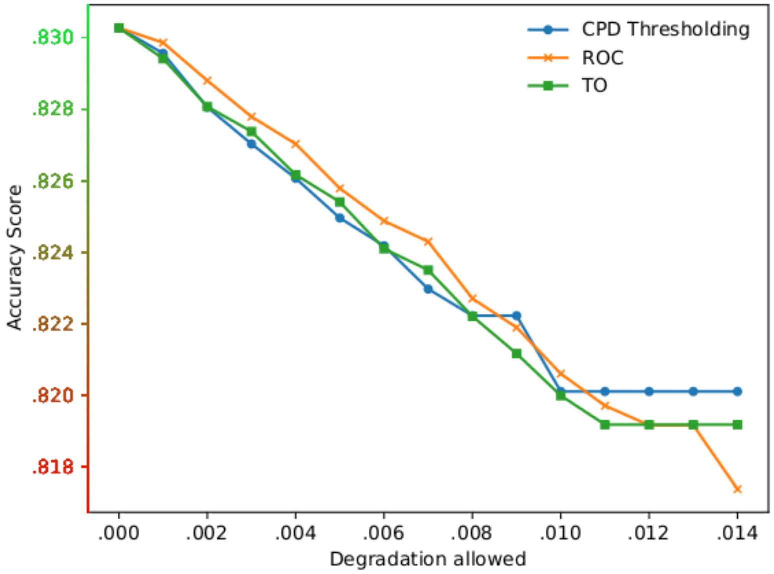


Figure 5.14 Accuracy Score on multiple degradation levels (the higher the better)

a superior fairness score for similar accuracy performance. Past reaching the plateau at 1%, performance is not negatively impacted by CPD Thresholding as opposed to ROC and TO, which continue their downward trends in performance with increasing degrees of degradation.

5.4 Discussion

5.4.1 Summary of findings

In this chapter, we have considered three post-processing methods founded on threshold use to reach fairness objectives, while preserving model performance. Two of these strategies, ROC and TO rely on the model scores to formulate predictions. The model score serves as a natural indicator for selecting inputs, reflecting the degree of certainty or confidence the model holds in its prediction. However, the reliability of these scores may be compromised if input profiles diverge significantly from the training distribution. To this end, we have introduced another criterion: Computational Profile Distance, that has shown satisfying results in detecting Out-Of-Distribution inputs. The resulting method, CPD Thresholding, aims to ensure a more robust fairness assessment than the methods based on model scores.

Experiments have shown that almost for all criteria chosen, on all datasets studied, CPD Thresholding presents better results, both in terms of fairness and performance. In one situation, CPD Thresholding exceeded the four-digit precision when trying to reach EO. Interestingly, post-processing corrections sometimes give an artificial advantage to the unprivileged group. For instance, CPD Thresholding reaches a negative AO difference on the German dataset.

In parallel, comparisons between methods demonstrate the trade-off between maintaining high accuracy levels and ensuring fairness, under varying levels of degradation. None of the three algorithms studied achieved high accuracy and fairness at the same time. However, results show that CPD Thresholding works better on large datasets (Adult and Dutch datasets) than ROC and TO, offering a better balance between fairness and performance. On smaller datasets, such as the German dataset, the performance of CPD Thresholding appears to be comparable or marginally inferior to other *Thresholding* strategies. One possible explanation of these observations comes from that CPD calculations are based on histograms (see Section 4.4 for more details), which are more adapted to a higher number of inputs. On these consequent databases, CPD Thresholding corrections are more numerous than the ones from TO and ROC but also of greater quality. On the other hand, on the small German dataset, the strategy based on CPD affects much fewer individuals than ROC and TO, for slightly less good results.

To sum up, CPD Thresholding alone seems to perform better than ROC and TO on large datasets but does rely on CPD, which is a criterion less understandable and natural than model scores. However, taken together, these three methods provide valuable corrections: most of the time a significant portion of individuals are affected by CPD Thresholding, and at least one method among TO and ROC. For these "individuals in common", a high-confidence decision can applied towards more group fairness.

The results show an encouraging tendency to correct group unfairness and suggest that CPD Thresholding is an efficient way to reach group fairness objectives for any fully connected NNs, equally to other existing methods, such as TO and ROC.

This post-processing method tends to increase the explainability of NNs, by "opening the black-box" and analyzing internal activation patterns. In addition, post-hoc methods like CPD Thresholding are easier to implement in industrial contexts, because they can be appended at the end of existing pipelines.

Despite CPD Thresholding exhibiting superior performance on large datasets, the main idea behind presenting another swapping criterion other than model scores is to illustrate that most of the time, a large portion of the individuals concerned are affected by at least two methods. For these individuals in common, a robust decision can be made towards more group fairness. A future direction of research could consider both CPD and model scores to develop an ensemble Thresholding method.

5.4.2 Limitations

The use of *Thresholding* post-processing methods only is part of the limitations of this study. Because the three methods utilize thresholds to formulate predictions, a significant portion of individuals is affected by at least two methods, but it may not be necessary the case with *Calibration* or *Transformation* methods. Another important limitation is that the three post-processing methods necessitate the sensitive attribute to be accessible during the post-processing corrections. In some cases, sensitive features are removed from the data for security and privacy reasons.

Another limitation is the absence of a guarantee that the number of positive labels (and negative) is the same before and after the post-processing corrections. In most experiments conducted, this is not the case. For an external operator that applies *post-hoc* corrections, this property may be necessary. For instance, if a company classifies applicants based on their characteristics, the fairness corrections applied to the classifiers should not reduce or increase the targeted number of applicants.

Finally, the experiments conducted present limitations. They are operated on only three datasets, and only three fairness criteria are considered. Only the accuracy score is studied as a performance metric, ignoring other additional information, such as the classification recall, or balanced accuracy score in the situation of highly imbalanced datasets.

CHAPTER 6 CONCLUSION

Machine learning algorithms, particularly neural networks, have been gaining popularity over the last few years. Often trained on personal or sensitive data, societal issues are emerging. Among these, algorithmic fairness became a priority after multiple concerns arose. Numerous techniques have been developed to detect and correct such algorithmic bias. This work proposes a new tool, Computational Profile Likelihood, to address these issues.

This approach aims to model the “reasoning” of any fully connected neural network. The non-parametric method models distributions of internal activation levels using histograms. CPL is employed to assess the extent to which the internal pattern of a model during the inference of an entry diverges from a predefined set of inputs. In this work, we used this property to measure the “distance” of individuals from a demographic group, with respect to fairness considerations. This tool does not need any model (re)training and has shown robustness against Out-Of-Distribution inputs, where unfairness is most susceptible to thrive. To our knowledge, no other work in the fairness literature aggregates all these properties.

6.1 Summary of Works

We proposed two novel ways to leverage Computational Profile Likelihood within the context of the fairness of neural networks. More particularly, we studied the CPL of demographic groups to detect and correct any differences between these. In the first axis, we discovered that internal activation patterns of neural networks are vastly different between privileged and unprivileged groups of three widely studied datasets in algorithmic fairness, suggesting potential differential treatment between populations, and thus inherent bias. Then, we applied two pre-processing methods: Correlation Remover and Disparate Impact Remover, intending to enhance fairness according to a protected attribute. In most instances, we witnessed consequential diminutions in CPL differences between the corresponding sub-groups. These findings corroborate the idea that CPL can be used to assess the unfairness and efficiency of pre-processing bias mitigation methods.

Furthermore, in the second axis, we propose an innovative post-processing algorithm based on CPL to mitigate the unfairness of neural network predictions. We conducted experiments using three classic datasets of algorithmic fairness, and we demonstrated that our method is at least as effective as other existing methods, such as Reject Option Classification and Threshold Optimizer, in attaining various fairness criteria, like Demographic Parity or

Equality of Opportunity. In every dataset, CPD Thresholding was able to get closer to every fairness criterion that we chose to study. Compared to other strategies, CPD Thresholding performs better overall, especially on large datasets. However, CPD does not exist in opposition to model scores when it comes to selection criteria. Instead, CPD Thresholding and strategies based on model scores can be regarded as compatible approaches, concurring on the correction of a substantial portion of individuals. For those individuals, a confident and robust decision can be reached to foster group fairness.

The use of CPL in algorithmic fairness is innovative, and this work demonstrated that the versatility of this method allows CPL to be used for the detection and assessment of unfairness as well as a bias mitigation method, both in pre-processing and post-processing settings.

6.2 Limitations

This work presents several drawbacks. As mentioned earlier in Chapter 4.9, this work is confined to binary classification scenarios. To mitigate this limitation, various strategies can be employed. In the context of regression tasks, it is feasible to implement thresholding on the target attribute to revert to a simple two-class prediction context. Similarly, for multi-class classification models, applying *one-vs-rest* heuristic simplifies prediction by focusing only on predicting one class against all the other classes [106]. Another approach involves converting every pair of outcome classes into binary classification problems. If k is the number of existing classes, the number of binary classification problems would be $\frac{k(k-1)}{2}$ [107]. Additionally, our work does not consider intersectional fairness or proxy attributes.

The definition of CPL also admits some limitations. CPL calculations disregard the weight of neurons in the last hidden layer. By acknowledging that neurons have varying importance in a model's prediction process, we could potentially yield finer results and insights. CPL assumes that neurons are independent, which might not always be the case.

Considering the experiments conducted in each section, a more comprehensive investigation could have been undertaken with more choice in the hyper-parameters. For instance, increasing the number of architectures of the models studied could have been interesting to examine the influence of the model architecture on CPL calculations, by modifying the number of layers, the number of neurons of each layer, or the activation function of these models. Additionally, experiments with more datasets could have led to more generalization. In the same way, more pre-processing bias mitigation methods could have been considered, for instance, sanitization methods based on Generative Adversarial Network [108], and the section 5 only consider *Thresholding* post-processing bias correction techniques.

6.3 Future Research

As outlined in Section 6.2, future research may consider increasing generalizability by examining more architectures, and a broader range of hyper-parameters, such as alternative training procedures, multiple architectures, various numbers of thresholds for Section 5, or other pre and post processing techniques. Such studies have the potential to uncover previously unobserved trends and phenomena and yield more refined analyses.

The investigation of the multi-dimensional aspect of fairness could rely on this work. For instance, we could compare the CPL distributions of the *black women* of the Adult dataset compared to the distributions of *white women*. Section 5 could be enlarged with multi-objective *Thresholding* algorithms, ensuring for instance demographic parity along the gender *and* the race of individuals. The results of this Section suggest that group fairness could benefit from the use of multiple bias correction methods, and using CPL *and* model scores at the same time for robust and precise corrections, thus getting closer to the classification ensemble paradigm, as evoked in the work of Kamiran, Karim and Zhang [55].

Other work could be dedicated to detecting proxy attributes, or sensitive groups, using CPL. Finally, the frame of this work could be extended to a broader vision than the context of binary classification, for instance, multi-class classification.

REFERENCES

[1] K. Shailaja, B. Seetharamulu, and M. A. Jabbar, “Machine learning in healthcare: A review,” in *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, 2018, pp. 910–914.

[2] E. J. Alcántara Suárez and V. Monzon Baeza, “Evaluating the role of machine learning in defense applications and industry,” *Machine Learning and Knowledge Extraction*, vol. 5, no. 4, pp. 1557–1569, 2023.

[3] T. Yuan *et al.*, “Machine learning for next-generation intelligent transportation systems: A survey,” *Transactions on Emerging Telecommunications Technologies*, vol. 33, no. 4, p. e4427, 2022.

[4] N. A. Saxena *et al.*, “How do fairness definitions fare? testing public attitudes towards three algorithmic definitions of fairness in loan allocations,” *Artificial Intelligence*, vol. 283, p. 103238, 2020.

[5] A. Datta, M. C. Tschantz, and A. Datta, “Automated experiments on ad privacy settings,” *Proceedings on Privacy Enhancing Technologies*, vol. 2015, pp. 112–92, 2014.

[6] E. Bogert, A. Schecter, and R. T. Watson, “Humans rely more on algorithms than social influence as a task becomes more difficult,” *Scientific Reports*, vol. 11, no. 1, p. 8028, apr 2021.

[7] E. O. Soremekun *et al.*, “Software fairness: An analysis and survey,” *ArXiv*, vol. abs/2205.08809, 2022.

[8] G. Alves *et al.*, “Survey on fairness notions and related tensions,” *EURO Journal on Decision Processes*, vol. 11, 2023.

[9] S. Caton and C. Haas, “Fairness in machine learning: A survey,” *ACM Computing Surveys*, vol. 56, no. 7, pp. 1–38, 2024.

[10] A. Beutel *et al.*, “Data decisions and theoretical implications when adversarially learning fair representations,” https://arxiv.org/abs/1707.00075, 2017.

[11] X. Gao *et al.*, “Fairneuron: Improving deep neural network fairness with adversary games on selective neurons,” in *2022 IEEE/ACM 44th International Conference on Software Engineering (ICSE)*, 2022, pp. 921–933.

[12] H. Zheng *et al.*, “Neuronfair: Interpretable white-box fairness testing through biased neuron identification,” in *2022 IEEE/ACM 44th International Conference on Software Engineering (ICSE)*, 2022.

[13] E. Merlo *et al.*, “Models of computational profiles to study the likelihood of dnn metamorphic test cases,” https://arxiv.org/abs/2107.13491, 2021.

[14] O. Strauss, F. Comby, and M.-J. Aldon, “Rough histograms for robust statistics,” in *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*, 2000.

[15] A. Dongare *et al.*, “Introduction to artificial neural network,” *International Journal of Engineering and Innovative Technology (IJEIT)*, vol. 2, no. 1, pp. 189–194, 2012.

[16] A. D. Rasamoelina, F. Adjailia, and P. Sinčák, “A review of activation function for artificial neural network,” in *2020 IEEE 18th World Symposium on Applied Machine Intelligence and Informatics (SAMI)*, 2020, pp. 281–286.

[17] M. I. Jordan and T. M. Mitchell, “Machine learning: Trends, perspectives, and prospects,” *Science*, vol. 349, no. 6245, pp. 255–260, 2015.

[18] I. Yaqoob *et al.*, “Big data: From beginning to future,” *International Journal of Information Management*, vol. 36, no. 6, Part B, pp. 1231–1247, 2016.

[19] S. Cossette, “Jonathan durand folco et jonathan martineau, le capital algorithmique : accumulation, pouvoir et résistance à l’ère de l’intelligence artificielle (montréal: Éditions Écosociété, 2023),” *Labour / Le Travail*, vol. 94, p. 333–335, Nov. 2024.

[20] S. H. Silva and P. Najafirad, “Opportunities and challenges in deep learning adversarial robustness: A survey,” https://arxiv.org/abs/2007.00753, 2020.

[21] H. L. França, C. Teixeira, and N. Laranjeiro, “Techniques for evaluating the robustness of deep learning systems: A preliminary review,” in *2021 10th Latin-American Symposium on Dependable Computing (LADC)*, 2021, pp. 1–5.

[22] L. H. Gilpin *et al.*, “Explaining explanations: An overview of interpretability of machine learning,” in *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, 2018, pp. 80–89.

[23] R. González-Sendino *et al.*, “A review of bias and fairness in artificial intelligence,” *International Journal of Interactive Multimedia and Artificial Intelligence*, 2023.

[24] C. Dwork *et al.*, "Fairness through awareness," in *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ser. Itcs '12. New York, NY, USA: Association for Computing Machinery, 2012, p. 214–226.

[25] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, 2016.

[26] S. Barocas, M. Hardt, and A. Narayanan, *Fairness and machine learning: Limitations and opportunities*. MIT press, 2023.

[27] N. Mehrabi *et al.*, "A survey on bias and fairness in machine learning," *ACM computing surveys (CSUR)*, vol. 54, no. 6, pp. 1–35, 2021.

[28] P. Gajane and M. Pechenizkiy, "On formalizing fairness in prediction with machine learning," https://arxiv.org/abs/1710.03184, 2018.

[29] T. Le Quy *et al.*, "A survey on datasets for fairness-aware machine learning," *WIREs Data Mining and Knowledge Discovery*, vol. 12, no. 3, p. e1452, 2022.

[30] A. K. Menon and R. C. Williamson, "The cost of fairness in classification," *ArXiv*, vol. abs/1705.09055, 2017.

[31] H. Zhao and G. J. Gordon, *Inherent tradeoffs in learning fair representations*. Curran Associates Inc., 2019.

[32] S. Dehdashtian, B. Sadeghi, and V. N. Boddeti, "Utility-fairness trade-offs and how to find them," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[33] F. Kamiran and T. Calders, "Classifying without discriminating," in *2009 2nd International Conference on Computer, Control and Communication*, 2009.

[34] J. Li *et al.*, "A critical review of predominant bias in neural networks," https://arxiv.org/abs/2502.11031, 2025.

[35] S. Galhotra, Y. Brun, and A. Meliou, "Fairness testing: testing software for discrimination," in *Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering*, ser. Esec/fse 2017. New York, NY, USA: Association for Computing Machinery, 2017, p. 498–510.

[36] W. Xie and P. Wu, “Fairness testing of machine learning models using deep reinforcement learning,” in *2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, 2020, pp. 121–128.

[37] X. Li, P. Wu, and J. Su, “Accurate fairness: Improving individual fairness without trading accuracy,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 12, pp. 14 312–14 320, Jun. 2023.

[38] C. Ilvento, “Metric Learning for Individual Fairness,” in *1st Symposium on Foundations of Responsible Computing (FORC 2020)*, ser. Leibniz International Proceedings in Informatics (LIPIcs), A. Roth, Ed., vol. 156. Dagstuhl, Germany: Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2020, pp. 2:1–2:11.

[39] M. Du *et al.*, “Fairness in deep learning: A computational perspective,” *IEEE Intelligent Systems*, vol. 36, no. 4, pp. 25–34, 2021.

[40] A. Castelnovo *et al.*, “A clarification of the nuances in the fairness metrics landscape,” *Scientific Reports*, vol. 12, 2021.

[41] D. Pessach and E. Shmueli, “A review on fairness in machine learning,” *ACM Comput. Surv.*, vol. 55, no. 3, feb 2022.

[42] Z. Zhang, S. Wang, and G. Meng, “A review on pre-processing methods for fairness in machine learning,” in *Advances in Natural Computation, Fuzzy Systems and Knowledge Discovery*, 2023.

[43] L. E. Celis and V. Keswani, “Improved adversarial learning for fair classification,” *ArXiv*, vol. abs/1901.10443, 2019.

[44] L. E. Celis *et al.*, “Classification with fairness constraints: A meta-algorithm with provable guarantees,” *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2018.

[45] T. Calders and S. Verwer, “Three naive Bayes approaches for discrimination-free classification,” *Data Mining and Knowledge Discovery*, vol. 21, no. 2, pp. 277–292, sep 2010.

[46] V. Iosifidis, B. Fetahu, and E. Ntoutsi, “Fae: A fairness-aware ensemble framework,” in *2019 IEEE International Conference on Big Data (Big Data)*, 2019, pp. 1375–1380.

[47] F. Kamiran and T. Calders, “Data preprocessing techniques for classification without discrimination,” *Knowledge and Information Systems*, vol. 33, no. 1, pp. 1–33, oct 2012.

[48] H. Wang, B. Ustun, and F. P. Calmon, “Repairing without retraining: Avoiding disparate impact with counterfactual distributions,” https://arxiv.org/abs/1901.10501, 2019.

[49] H. Jiang and O. Nachum, “Identifying and correcting label bias in machine learning,” in *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, S. Chiappa and R. Calandra, Eds., vol. 108. Pmlr, 26–28 Aug 2020, pp. 702–712.

[50] T. Kamishima *et al.*, “Fairness-aware classifier with prejudice remover regularizer,” in *Proceedings of the 2012th European Conference on Machine Learning and Knowledge Discovery in Databases - Volume Part II*, ser. Ecmlpkdd’12. Berlin, Heidelberg: Springer-Verlag, 2012, p. 35–50.

[51] M. B. Zafar *et al.*, “Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment,” in *Proceedings of the 26th International Conference on World Wide Web*, ser. Www ’17. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, 2017, p. 1171–1180.

[52] B. Woodworth *et al.*, “Learning non-discriminatory predictors,” in *Proceedings of the 2017 Conference on Learning Theory*, ser. Proceedings of Machine Learning Research, S. Kale and O. Shamir, Eds., vol. 65. Pmlr, 07–10 Jul 2017, pp. 1920–1953.

[53] Y. Bechavod and K. Ligett, “Penalizing unfairness in binary classification,” https://arxiv.org/abs/1707.00044, 2018.

[54] E. Krasanakis *et al.*, “Adaptive sensitive reweighting to mitigate bias in fairness-aware classification,” in *Proceedings of the 2018 World Wide Web Conference*, ser. Www ’18. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, 2018, p. 853–862.

[55] F. Kamiran, A. Karim, and X. Zhang, “Decision theory for discrimination-aware classification,” in *Proceedings of the 2012 IEEE 12th International Conference on Data Mining*, 2012.

[56] M. Hardt, E. Price, and N. Srebro, “Equality of opportunity in supervised learning,” in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, 2016.

[57] B. Fish, J. Kun, and Á. D. Lelkes, “A confidence-based approach for balancing fairness and accuracy,” in *Sdm*, 2016.

[58] C. Su *et al.*, “A review of causality-based fairness machine learning,” *Intelligence & Robotics*, vol. 2, no. 3, 2022.

[59] K. Makhlouf, S. Zhioua, and C. Palamidessi, “When causality meets fairness: A survey,” *Journal of Logical and Algebraic Methods in Programming*, vol. 141, p. 101000, 2024.

[60] M. J. Kusner *et al.*, “Counterfactual fairness,” https://arxiv.org/abs/1703.06856, 2018.

[61] B. Salimi, B. Howe, and D. Suciu, “Data management for causal algorithmic fairness,” *CoRR*, vol. abs/1908.07924, 2019.

[62] N. Kilbertus *et al.*, “Avoiding discrimination through causal reasoning,” in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. Nips’17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 656–666.

[63] B. Salimi *et al.*, “Interventional fairness: Causal database repair for algorithmic fairness,” in *Proceedings of the 2019 International Conference on Management of Data*, ser. Sigmod ’19. New York, NY, USA: Association for Computing Machinery, 2019, p. 793–810.

[64] Y. Zhao, Y. Wang, and T. Derr, “Fairness and explainability: Bridging the gap towards fair model explanations,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 9, pp. 11 363–11 371, Jun. 2023.

[65] M. Marhaba, “Analysis of cnn computational profile likelihood on adversarial attacks and affine transformations,” Master’s thesis, Polytechnique Montréal, April 2022. [Online]. Available: https://publications.polymtl.ca/10304/

[66] M. Krichen, “Generative adversarial networks,” in *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, 2023.

[67] T. Li *et al.*, “Faire: Repairing fairness of neural networks via neuron condition synthesis,” *ACM Trans. Softw. Eng. Methodol.*, vol. 33, no. 1, 2023.

[68] B. Becker and R. Kohavi, “Adult,” UCI Machine Learning Repository, 1996.

[69] M. Feldman *et al.*, “Certifying and removing disparate impact,” in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015.

[70] U. Aïvodji *et al.*, “Local data debiasing for fairness based on generative adversarial training,” *Algorithms*, vol. 14, no. 3, p. 87, 2021.

[71] A. Giloni *et al.*, “Benn: Bias estimation using a deep neural network,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 1, pp. 117–131, 2022.

[72] M. Leo, S. Sharma, and K. Maddulety, “Machine learning in banking risk management: A literature review,” *Risks*, vol. 7, no. 1, p. 29, 2019.

[73] J. Beutel, S. List, and G. von Schweinitz, “Does machine learning help us predict banking crises?” *Journal of Financial Stability*, vol. 45, p. 100693, 2019.

[74] E. Ramanujam, T. Perumal, and S. Padmavathi, “Human activity recognition with smartphone and wearable sensors using deep learning techniques: A review,” *IEEE Sensors Journal*, vol. 21, no. 12, pp. 13 029–13 040, 2021.

[75] H. Zhu *et al.*, “Benchmarking and analyzing deep neural network training,” in *2018 IEEE International Symposium on Workload Characterization (IISWC)*, 2018.

[76] A. JayaLakshmi and K. K. Kishore, “Performance evaluation of dnn with other machine learning techniques in a cluster using apache spark and mllib,” *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 1, pp. 1311–1319, 2022.

[77] E. Yudkowsky *et al.*, “Artificial intelligence as a positive and negative factor in global risk,” *Global catastrophic risks*, vol. 1, no. 303, p. 184, 2008.

[78] H. Weerts *et al.*, “Fairlearn: Assessing and improving fairness of ai systems,” https://arxiv.org/abs/2303.16626, 2023.

[79] J. Buolamwini and T. Gebru, “Gender shades: Intersectional accuracy disparities in commercial gender classification,” in *Fat*, 2018.

[80] G. D. Pelegrina, M. Couceiro, and L. T. Duarte, “A preprocessing shapley value-based approach to detect relevant and disparity prone features in machine learning,” in *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 2024.

[81] K. Lum and J. Johndrow, “A statistical framework for fair predictive algorithms,” https://arxiv.org/abs/1610.08077, 2016.

[82] D. Pedreshi, S. Ruggieri, and F. Turini, “Discrimination-aware data mining,” in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2008.

[83] S. Verma and J. S. Rubin, "Fairness definitions explained," *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*, pp. 1–7, 2018.

[84] A. Shen *et al.*, "Optimising equal opportunity fairness in model training," *arXiv preprint arXiv:2205.02393*, pp. 4073–4084, 2022.

[85] R. K. E. Bellamy *et al.*, "Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias," *IBM Journal of Research and Development*, vol. 63, no. 4/5, pp. 4:1–4:15, 2019.

[86] H. Zhao *et al.*, "Conditional learning of fair representations," https://arxiv.org/abs/1910.07162, 2020.

[87] E. Creager *et al.*, "Flexibly fair representation learning by disentanglement," in *Proceedings of the 36th International Conference on Machine Learning*, 2019.

[88] J. Liu *et al.*, "Fair representation learning: An alternative to mutual information," in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022.

[89] V. A. Dasu *et al.*, "Neufair: Neural network fairness repair with dropout," in *International Symposium on Software Testing and Analysis*, 2024.

[90] Y. Mao *et al.*, "Last-layer fairness fine-tuning is simple and effective for neural networks," *ArXiv*, vol. abs/2304.03935, 2023.

[91] 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*, 2018.

[92] Z. Gong, W. Wang, and W. Ku, "Adversarial and clean data are not twins," *CoRR*, vol. abs/1704.04960, 2017.

[93] K. Grosse *et al.*, "On the (statistical) detection of adversarial examples," *CoRR*, vol. abs/1702.06280, 2017.

[94] J. H. Metzen *et al.*, "On detecting adversarial perturbations," https://arxiv.org/abs/1702.04267, 2017.

[95] J. Kim, R. Feldt, and S. Yoo, "Guiding deep learning system testing using surprise adequacy," in *Proceedings of the 41st International Conference on Software Engineering*, 2019.

[96] W. Liu *et al.*, "A survey of deep neural network architectures and their applications," *Neurocomputing*, vol. 234, pp. 11–26, 2017.

[97] M. Marhaba *et al.*, "Identification of out-of-distribution cases of cnn using class-based surprise adequacy," in *2022 IEEE/ACM 1st International Conference on AI Engineering – Software Engineering for AI (CAIN)*, 2022.

[98] C. R. Harris *et al.*, "Array programming with NumPy," *Nature*, vol. 585, no. 7825, pp. 357–362, 2020.

[99] H. Hofmann, "Statlog (German Credit Data)," UCI Machine Learning Repository, 1994.

[100] J. Chakraborty *et al.*, "Fairway: a way to build fair ml software," in *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 2020.

[101] F. Yang, M. Cisse, and S. Koyejo, "Fairness with overlapping groups; a probabilistic perspective," in *Advances in Neural Information Processing Systems*, 2020.

[102] A. Ruoss *et al.*, "Learning certified individually fair representations," in *Advances in Neural Information Processing Systems*, 2020.

[103] P. Lahoti *et al.*, "Fairness without demographics through adversarially reweighted learning," in *Advances in Neural Information Processing Systems*, 2020.

[104] R. Berk *et al.*, "A convex framework for fair regression," https://arxiv.org/abs/1706.02409, 2017.

[105] A. Agarwal *et al.*, "A reductions approach to fair classification," in *Proceedings of the 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, J. Dy and A. Krause, Eds., vol. 80. Pmlr, 10–15 Jul 2018, pp. 60–69.

[106] K. P. Murphy, *Machine learning: a probabilistic perspective.* MIT press, 2012.

[107] C. M. Bishop and N. M. Nasrabadi, *Pattern recognition and machine learning.* Springer, 2006, vol. 4, no. 4.

[108] U. Aïvodji *et al.*, "Local data debiasing for fairness based on generative adversarial training," *Algorithms*, vol. 14, 2021.

APPENDIX A FAIRNESS AND PERFORMANCE OF POST-PROCESSING METHODS

Method name	Fairness criterion	Fairness metric	Accuracy Score (%)
No Method	DP Difference	.0979	83.84
	EO Difference	.0733	83.84
	AO Difference	.0562	83.84
ROC	DP Difference	.0053	82.41
	EO Difference	.0278	83.66
	AO Difference	.0168	83.81
TO	DP Difference	.0056	83.49
	EO Difference	.0308	83.85
	AO Difference	.0135	83.79
CPD	DP Difference	.0041	83.68
	EO Difference	.0259	83.86
	AO Difference	.0123	83.84

Table A.1 Fairness and performance achieved on Adult, with attribute race

Method name	Fairness criterion	Fairness metric	Accuracy Score (%)
No Method	DP Difference	.1769	83.84
	EO Difference	.0825	83.84
	AO Difference	.0751	83.84
ROC	DP Difference	.0107	81.91
	EO Difference	.0024	82.30
	AO Difference	.0000	82.72
TO	DP Difference	.0149	81.98
	EO Difference	.0008	83.55
	AO Difference	.0005	83.53
CPD	DP Difference	.0159	82.09
	EO Difference	.0027	83.80
	AO Difference	.0003	83.35

Table A.2 Fairness and performance achieved on Adult, with attribute age

Method name	Fairness criterion	Fairness metric	Accuracy Score (%)
No Method	DP Difference	.1690	78.10
	EO Difference	.1039	78.10
	AO Difference	.1571	78.10
ROC	DP Difference	.0172	76.80
	EO Difference	.0343	76.70
	AO Difference	.0149	77.40
TO	DP Difference	.0162	77.30
	EO Difference	.0322	78.00
	AO Difference	.0108	77.10
CPD	DP Difference	.0069	77.80
	EO Difference	.0231	78.00
	AO Difference	-.0013	78.00

Table A.3 Fairness and performance achieved on German, with attribute gender

Method name	Fairness criterion	Fairness metric	Accuracy Score (%)
No Method	DP Difference	.2362	78.10
	EO Difference	.1607	78.10
	AO Difference	.1815	78.10
ROC	DP Difference	.0370	76.70
	EO Difference	.0575	76.90
	AO Difference	.0273	77.60
TO	DP Difference	.0268	76.30
	EO Difference	.0586	76.70
	AO Difference	.0273	77.60
CPD	DP Difference	.0098	77.70
	EO Difference	.0516	77.90
	AO Difference	.0061	78.10

Table A.4 Fairness and performance achieved on German, with attribute age

Method name	Fairness criterion	Fairness metric	Accuracy Score (%)
No Method	DP Difference	.3385	83.03
	EO Difference	.0879	83.03
	AO Difference	.1521	83.03
ROC	DP Difference	.1370	81.03
	EO Difference	.0014	82.50
	AO Difference	.0015	81.74
TO	DP Difference	.1344	80.99
	EO Difference	.0011	82.04
	AO Difference	.0012	81.18
CPD	DP Difference	.1288	81.05
	EO Difference	.0000	82.59
	AO Difference	.0011	82.01

Table A.5 Fairness and performance achieved on Dutch, with attribute gender

Method name	Fairness criterion	Fairness metric	Accuracy Score (%)
No Method	DP Difference	.7567	83.03
	EO Difference	.9145	83.03
	AO Difference	.5958	83.03
ROC	DP Difference	-.0094	81.60
	EO Difference	-.0200	81.70
	AO Difference	.0284	82.32
TO	DP Difference	.0060	82.18
	EO Difference	-.0173	81.53
	AO Difference	.0145	82.48
CPD	DP Difference	-.0013	82.97
	EO Difference	-.0400	82.97
	AO Difference	.0204	82.97

Table A.6 Fairness and performance achieved on Dutch, with attribute age