

ROBUST LOSS AND PENALTY TO FAIR MACHINE LEARNING

Ygor de Mello Canalli

Tese de Doutorado apresentada ao Programa de Pós-graduação em Engenharia de Sistemas e Computação, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Doutor em Engenharia de Sistemas e Computação.

Orientadores: Geraldo Zimbrão da Silva
Filipe Braidão do Carmo

ROBUST LOSS AND PENALTY TO FAIR MACHINE LEARNING

Ygor de Mello Canalli

10 TESE SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO
LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA
DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS
REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE DOUTOR
EM CIÊNCIAS EM ENGENHARIA DE SISTEMAS E COMPUTAÇÃO.

Orientadores: Geraldo Zimbrão da Silva

15 Filipe Braida do Carmo

Aprovada por: Prof. Geraldo Zimbrão da Silva

Prof. Filipe Braida do Carmo

Prof. Geraldo Bonorino Xexéo

Prof. Daniel Sadoc Menasché

Prof. Carlos Eduardo Ribeiro de Mello

RIO DE JANEIRO, RJ – BRASIL

AGOSTO DE 2024

de Mello Canalli, Ygor

Robust loss and penalty to fair machine learning/Ygor de Mello Canalli. – Rio de Janeiro: UFRJ/COPPE, 2024. XI, 80 p.: il.; 29, 7cm.

Orientadores: Geraldo Zimbrão da Silva

Filipe Braidão do Carmo

Tese (doutorado) – UFRJ/COPPE/Programa de Engenharia de Sistemas e Computação, 2024.

Referências Bibliográficas: p. 68 – 80.

1. Fair machine learning. 2. Label noise robustness. 3. Multi-objective optimization. 4. Redlining effect. I. Zimbrão da Silva, Geraldo *et al.* II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia de Sistemas e Computação. III. Título.

*A alguém cujo valor é digno
desta dedicatória.*

Agradecimentos

Gostaria de agradecer a todos.

Resumo da Tese apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Doutor em Ciências (D.Sc.)

25 FUNÇÃO DE CUSTO E PENALIZAÇÃO ROBUSTAS PARA APRENDIZADO
DE MÁQUINA JUSTO

Ygor de Mello Canalli

Agosto/2024

Orientadores: Geraldo Zimbrão da Silva

Filipe Braida do Carmo

30 Programa: Engenharia de Sistemas e Computação

problema e importância do problema, proposta, metodologia experimental, resumo dos resultados, contribuições e legado no estado da arte.

Abstract of Thesis presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Doctor of Science (D.Sc.)

35

ROBUST LOSS AND PENALTY TO FAIR MACHINE LEARNING

Ygor de Mello Canalli

August/2024

Advisors: Geraldo Zimbrão da Silva

Filipe Braidão do Carmo

Department: Systems Engineering and Computer Science

40

In this work, we present ...

Contents

	List of Figures	x
	List of Tables	xi
	1 Introduction	1
45	1.1 Contextualization	2
	1.2 Objectives	2
	1.3 Contributions	2
	1.4 Results summary	2
	1.5 Thesis structure	2
50	2 Fair machine learning	3
	2.1 Sources and types of algorithmic unfairness	5
	2.2 Fairness definitions and metrics	9
	2.3 Fair classification	20
	3 Fair loss function from noise robustness	25
55	3.1 Preliminaries	25
	3.2 Fair Transition Loss	31
	3.3 Experimental setup	34
	3.4 Results and discussion	39
	4 Correlation based penalty function	53
60	4.1 Preliminaries	53
	4.2 Chatterjee Redlining Penalty	57
	4.3 Experimental setup	59
	4.4 Results and discussion	61
	5 Conclusions	66
65	5.1 Considerations on the proposal	67
	5.2 Contributions	67
	5.3 Results summary	67

5.4 Research directions	67
References	68

List of Figures

	3.1	Noise taxonomy from a statistical perspective. (a) completely random noise (NCAR), (b) random noise (NAR) and (c) non-random noise (NNAR). The arrows correspond to the statistical dependencies. For clarity, the dependency between X and Y was placed as a dashed arrow.	26
75	3.2	Sensibility analysis on optimized fitness functions within different performance and fairness metrics. Results from complete hyperparameter tuning through 100 trials with baseline model over the Adult Income dataset.	36
	3.3	Fitness values optimizing MCC and multiple fairness metrics.	42
80	3.4	Fitness values optimizing Accuracy and multiple fairness metrics.	44
	3.5	Results of false negatives and false positives within groups on protected promotion (p_1) parameter at increasing levels.	45
	3.6	Metric distribution optimizing MCC and Statistical Parity in comparison with Fair Transition Loss across multiple resample runs. Corresponding values available at Table 3.5.	47
85	3.7	Metric distribution optimizing MCC and Equal Opportunity in comparison with Fair Transition Loss across multiple resample runs. Corresponding values available at Table 3.6.	48
	3.8	Metric distribution optimizing MCC and Equalized Odds in comparison with Fair Transition Loss across multiple resample runs. Corresponding values available at Table 3.7.	49
90	3.9	Metric distribution optimizing Accuracy and Statistical Parity in comparison with Fair Transition Loss across multiple resample runs. Corresponding values available at Table 3.8.	50
	3.10	Metric distribution optimizing Accuracy and Equal Opportunity in comparison with Fair Transition Loss across multiple resample runs. Corresponding values available at Table 3.9.	51
95	3.11	Metric distribution optimizing Accuracy and Equalized Odds in comparison with Fair Transition Loss across multiple resample runs. Corresponding values available at Table 3.10.	52
100			

4.1	Multiple correlation coefficients on Anscombe's quartet.	56
4.2	Fitness values of CRP optimizing MCC and multiple fairness metrics.	61
4.3	Fitness values of CRP optimizing Accuracy and multiple fairness met- rics.	65

List of Tables

2.1	Confusion matrix of binary classification outcomes	11
3.1	Hyperparameters search ranges or options of each method.	36
3.2	Dataset details used in this work, including performance and fair- ness metrics assessed to a standard classifier without tuning, and the maximum correlation between sensitive feature and the other features.	38
3.3	Almost Stochastic Order test comparing Fair Transition Loss fitness. Values under 0.5 (in bold) mean that FTL outperforms corresponding method in such optimization scenario.	40
3.4	Fair Transition Loss hyperparameters chosen by optimizing different metrics in <i>Adult Income</i> dataset.	41
3.5	Mean and standard deviation metric values optimizing MCC and Sta- tistical Parity in comparison with Fair Transition Loss across multiple resample runs.	47
3.6	Mean and standard deviation metric values optimizing MCC and Equal Opportunity in comparison with Fair Transition Loss across multiple resample runs.	48
3.7	Mean and standard deviation metric values optimizing MCC and Equalized Odds in comparison with Fair Transition Loss across mul- tiple resample runs.	49
3.8	Mean and standard deviation metric values optimizing Accuracy and Statistical Parity in comparison with Fair Transition Loss across mul- tiple resample runs.	50
3.9	Mean and standard deviation metric values optimizing Accuracy and Equal Opportunity in comparison with Fair Transition Loss.	51
3.10	Mean and standard deviation metric values optimizing Accuracy and Equalized Odds in comparison with Fair Transition Loss across mul- tiple resample runs.	52
4.1	Anscombe's quartet	56
4.2	Hyperparameters search ranges or options of each method.	60

135	4.3	Almost Stochastic Order test comparing the use of Chatterjee Redlining Penalty on Standard MLP and MLP with Fair Transition Loss fitness. Values under 0.5 (in bold) mean that the model with CRP outperforms the same model without regularization on each optimization scenario.	61
140	4.4	Mean and standard deviation metric values optimizing MCC and Statistical Parity in comparison with Chatterjee Redlining Penalty. . . .	62
	4.5	Mean and standard deviation metric values optimizing MCC and Equal Opportunity in comparison with Chatterjee Redlining Penalty.	62
	4.6	Mean and standard deviation metric values optimizing MCC and Equalized Odds in comparison with Chatterjee Redlining Penalty. . .	63
145	4.7	Mean and standard deviation metric values optimizing Accuracy and Statistical Parity in comparison with Chatterjee Redlining Penalty. . .	63
	4.8	Mean and standard deviation metric values optimizing Accuracy and Equal Opportunity in comparison with Chatterjee Redlining Penalty.	64
150	4.9	Mean and standard deviation metric values optimizing Accuracy and Equalized Odds in comparison with Chatterjee Redlining Penalty. . .	64

Chapter 1

Introduction

The issue of fairness in machine learning has recently risen to prominence due to its implications in real-world decision-making systems (MEHRABI *et al.*, 2021; HUTCHINSON e MITCHELL, 2018). Addressing biases and discrimination is a relevant frontier in decision-making systems, as equitable outcomes across various demographic groups is both an ethical imperative and often a legal requirement. Though fairness is a multifaceted concept, it has been deeply examined within the context of machine learning. The literature presents a variety of fairness definitions, drawing concepts from political philosophy and computational techniques (HUTCHINSON e MITCHELL, 2018; CATON e HAAS, 2023). Choosing an equitable machine learning model requires the selection of a fitting definition of fairness, tailored to the specific problem at hand. Many such definitions can be precisely articulated, allowing models to be evaluated based on their predictions.

One inherent challenge in fair machine learning is the balance between fairness and accuracy. Efforts to mitigate unfairness often compromise the model’s predictive performance, a trade-off that has been well documented (MEHRABI *et al.*, 2021; CATON e HAAS, 2023). Predictors that are less biased against marginalized groups may deviate from the true class, resulting in sub-optimal performance. Also, introducing fairness considerations adds constraints to the model, further complicating the optimization process (ZAFAR *et al.*, 2017a).

In light of these challenges, we introduce the Fair Transition Loss, a novel approach to fair classification. This method estimates the influence of historical and societal biases on outcome probabilities for distinct groups within dataset. For instance, individuals from marginalized groups might have lower chances of favorable outcomes compared to their counterparts from privileged groups. Such disparate probabilities can be represented by transition matrices. Drawing inspiration from label noise robustness, we incorporate these transition matrices information into the loss function to promote fairness. The proposed method has some hyperparameters, chosen by a Multi-Objective Optimization approach combining both fairness and

model performance with a linear smooth objective. This objective is defined in such a way that it is possible to use this approach to optimize a variety of fairness and performance metrics.

185 The primary contribution of this study is the conceptualization of the Fair Transition Loss, a novel loss function influenced by label noise methodologies. In benchmark tests across common fair classification tasks, our empirical results demonstrate that this method consistently outperforms many leading in-processing fair classification techniques in a variety of scenarios. The novelty of this work lies in applying
190 label noise techniques directly within the model to mitigate unfairness. As far as we know, this is the first time that label noise techniques are directly used to address fairness in machine learning.

1.1 Contextualization

1.2 Objectives

195 1.3 Contributions

1.4 Results summary

1.5 Thesis structure

Chapter 2

Fair machine learning

200 The field of Machine Learning (ML) has experienced significant growth and is increasingly applied in various societal domains such as healthcare, finance, and criminal justice. This growth raises important ethical and operational concerns, particularly regarding the principles of Fairness, Accountability, and Transparency (FAT) (MEMARIAN e DOLECK, 2023). As ML algorithms increasingly influence a wide
205 array of societal domains, including criminal justice, healthcare, finance, and employment, the requirement to ensure these systems are designed and implemented responsibly has become paramount. This section aims to delineate the significance, scope, and prevailing challenges associated with integrating FAT principles into ML, providing a foundation for the subsequent discussion.

210 Fairness in ML concerns the equitable and just treatment of all individuals, particularly those from historically marginalized or disadvantaged groups (MEHRABI *et al.*, 2021; CATON e HAAS, 2023). It seeks to ensure that ML algorithms do not perpetuate existing biases or create new forms of discrimination. However, the multifaceted nature of fairness, encompassing various definitions and metrics, poses
215 substantial challenges in operationalizing it within algorithmic frameworks. Further in this section we will explore these complexities, examining different conceptions of fairness and the inherent trade-offs they entail.

Accountability in ML pertains to the obligation of designers, developers, and maintainers of ML systems to be answerable for the outcomes of these systems
220 (HUTCHINSON *et al.*, 2021). It involves establishing mechanisms that allow for the tracing of decisions back to the entities responsible for the deployment of the ML algorithms. Accountability also encompasses the adherence to ethical standards, legal requirements, and societal norms. This discussion frequently involves mechanisms and practices that can promote accountability in ML, like auditing,
225 documentation, and regulatory compliance.

Transparency, the third pillar, refers to the clarity and openness with which ML systems operate (BURKART e HUBER, 2021). It involves the ability of stakehold-

ers, including end-users, regulators, and the broader public, to understand how ML systems make decisions. Transparency is mandatory property of any automated decision making system to achieve trustworthiness, facilitating informed consent, and enabling the scrutiny necessary to identify and rectify biases. However, achieving transparency, particularly with complex models, presents its own set of technical and ethical challenges. This research topic includes issues as the trade-off between explainability and model performance, and discussing emerging approaches to tackle interpretability without sacrificing effectiveness.

The triad of Fairness, Accountability, and Transparency (FAT) along with data privacy forms the cornerstone of Trustworthy Artificial Intelligence (TwAI). These principles are pivotal in ensuring that AI systems are developed and deployed in a manner that respects human rights, promotes social well-being, and maintains public trust. While accountability ensures that entities behind AI systems can be held responsible for their outcomes, transparency allows stakeholders to produce and maintain environments where AI systems can be scrutinized, understood, and corrected, thereby aligning their functionality with societal norms and values.

In this context of Trustworthy AI the European Union’s High-Level Expert Group on Artificial Intelligence has outlined seven key principles that aim to ensure that AI systems are designed and used in a way that is ethically sound and trustworthy (HLEG, 2019). These principles are fundamental for developing and maintaining decision making systems that are beneficial and avoid unintended harm. The seven principles are as follows:

Human agency and oversight AI systems should empower human beings, allowing them to make informed decisions and fostering their fundamental rights. At the same time, proper oversight mechanisms need to be ensured, which can be achieved through human-in-the-loop, human-on-the-loop, and human-in-command approaches

Technical Robustness and safety AI systems need to be resilient and secure. They need to be safe, ensuring a fall back plan in case something goes wrong, as well as being accurate, reliable and reproducible. That is the only way to ensure that also unintentional harm can be minimized and prevented.

Privacy and data governance besides ensuring full respect for privacy and data protection, adequate data governance mechanisms must also be ensured, taking into account the quality and integrity of the data, and ensuring legitimized access to data.

Transparency the data, system and AI business models should be transparent. Traceability mechanisms can help achieving this. Moreover, AI systems and

265 their decisions should be explained in a manner adapted to the stakeholder
concerned. Humans need to be aware that they are interacting with an AI
system, and must be informed of the system’s capabilities and limitations.

Diversity, non-discrimination and fairness Unfair bias must be avoided, as it
could have multiple negative implications, from the marginalization of vulner-
270 able groups, to the exacerbation of prejudice and discrimination. Fostering
diversity, AI systems should be accessible to all, regardless of any disability,
and involve relevant stakeholders throughout their entire life circle.

Societal and environmental well-being AI systems should benefit all human
beings, including future generations. It must hence be ensured that they are
275 sustainable and environmentally friendly. Moreover, they should take into
account the environment, including other living beings, and their social and
societal impact should be carefully considered.

Accountability Mechanisms should be put in place to ensure responsibility and
accountability for AI systems and their outcomes. Auditability, which en-
280 ables the assessment of algorithms, data and design processes plays a key role
therein, especially in critical applications. Moreover, adequate and accessible
redress should be ensured.

Although there are many aspects to consider to an ethical automated decision
system with social impacts, the present text will concentrate predominantly on
285 the aspect of fairness and negative social bias. Fairness is not only desirable for
the development of just and equitable technological solutions but an obligation for
maintaining the trustworthiness and acceptability of AI systems in diverse societal
contexts.

2.1 Sources and types of algorithmic unfairness

290 The comprehensive survey conducted by MEHRABI *et al.* (2021) elucidates the
multitude of biases that can pervade artificial intelligence applications, potentially
leading to unfair outcomes. This analysis categorizes the various sources of bias,
illustrating the multifaceted ways in which such biases can infiltrate different stages
of machine learning processes, ranging from the initial data collection phase to the
295 final algorithmic processing. The following exposition provides a short delineation
of these sources of bias. To a rich discussion on this topic - including references,
examples and real cases where each source of bias can emerge - we recommend the
reading of the original work. The discussion here is with the purpose of proper

describing the complexity and multifaceted nature of unfairness in machine learning
300 models.

Historical Bias This is the existing societal bias that reflects past and present inequalities and prejudices. Historical bias is present in the data even before any machine learning model has interacted with it, due to inherent social and cultural inequalities;

305 **Representation Bias** Occurs when the data sample does not accurately represent the entire population or certain subgroups within it. This can lead to machine learning models that perform well on majority groups but poorly on underrepresented groups;

Measurement Bias Arises when the data collected does not accurately measure
310 the real-world constructs it purports to measure. This type of bias can occur due to flawed data collection instruments or processes that systematically misrepresent certain groups;

Evaluation Bias This type of bias occurs during the performance evaluation of machine learning models, where the evaluation criteria or methods may favor
315 one group over others, leading to biased assessments of model performance;

Aggregation Bias Happens when incorrect assumptions are made about the homogeneity of groups within the data. Aggregation bias can lead to misleading conclusions if the differences within and between groups or subgroups are not properly accounted for;

320 **Population Bias** Similar to representation bias, population bias occurs when statistics, demographics, representatives, and user characteristics are different in the user population represented in the dataset, leading to models that are not generalizable across different demographic groups;

Simpson's Paradox This is a statistical phenomenon where a trend appears in
325 several different groups of data but disappears or reverses when these groups are combined;

Longitudinal Data Fallacy Occurs when cross-sectional data is treated as longitudinal, leading to incorrect conclusions about data trends over time;

Sampling Bias Introduced by non-random sampling procedures, where certain
330 members of the intended population are less likely to be included in the sample than others, leading to skewed data that does not accurately represent the entire population;

Behavioral Bias Arises from variations in user behavior that differ across different platforms or contexts, affecting the data's representation of real-world phenomena;

Content Production Bias Results from differences in how content is generated by different groups, with structural, lexical, semantic, and syntactic differences, influencing the data available for machine learning models;

Linking Bias Occurs in networked data, where the connections between nodes can misrepresent the true attributes or behavior of the nodes;

Temporal Bias Reflects changes in data characteristics over time, due changes in representation or behaviors, which may not be accounted for in static machine learning models;

Popularity Bias Occurs when popular items are more likely to be recommended or rated highly, not necessarily because of their quality but simply because of their initial, higher visibility and these popularity metrics are subject of manipulation;

Algorithmic Bias Introduced by the algorithms themselves, when they add bias that was not present in the input data:

User Interaction Bias Results from the way system design influences user behavior in biased ways. This source of bias can be influenced by other types or subtypes, such as Presentation and Ranking Biases:

Presentation Bias This bias occurs when the way information is presented influences the outcomes. In machine learning, this can manifest through the design of user interfaces or the manner in which data is displayed, affecting user decisions and interactions;

Ranking Bias Arises when algorithms prioritize certain data points over others in ranked lists or search results, which can distort visibility and perpetuate certain preferences or discriminations;

Social Bias Social biases are the preconceived notions and stereotypes held by societies, where individual actions or contents are socially influenced. These biases often find their way into data through collective social behaviors and decisions, influencing the training data used for machine learning models;

Emergent Bias Emerges during the operation of a system, particularly as a result of changes in population, cultural values, or societal knowledge in the data over time. This type of bias is dynamic and can occur even if the initial model was unbiased, due to changes in the underlying data or context;

Self-Selection Bias Occurs when the individuals selected for a study or dataset have self-selected in some way, producing a sample that is not representative of the general population. This can skew results and make the data less generalizable;

Omitted Variable Bias Happens when a model overlooks certain relevant variables that are correlated with both the independent and dependent variables. Omitting these variables can lead to incorrect inferences about correlations and effects;

Cause-Effect Bias This bias is a misunderstanding in the determination of causation; it can occur when correlations are mistaken for causal relationships without proper justification through causal inference techniques;

Observer Bias Introduced by the expectations or preconceptions of those collecting or processing data, which can influence the outcomes subconsciously;

Funding Bias Refers to the influence that the source of funding can have on the conduct of research or development of algorithms. This type of bias can lead to results that favor the interests of the funding source, consciously or unconsciously.

These biases can pervade various stages of machine learning, from data collection to model evaluation and deployment, highlighting the importance of understanding and mitigating bias to achieve fairness in AI systems. Furthermore, the presence of biases can lead to feedback loops that exacerbate these inequalities over time. When biased data influence the decisions made by an AI system, these decisions can then be used to generate more data, which, if used to retrain the model, may reinforce and even amplify the existing biases. This cycle can create a self-perpetuating loop, making initial biases more entrenched and difficult to correct. Addressing feedback loops is critical, as they can progressively deteriorate the fairness of the system, leading to increasingly skewed outcomes that are harder to rectify. Effective strategies to break these loops include rigorous monitoring of model decisions, regular updates to training datasets to ensure diversity and representativeness, and the implementation of mechanisms that can detect and correct for emerging biases

Having outlined the various sources of unfairness in machine learning, MEHRABI *et al.* (2021) also explores different types of discrimination that arise from these biases. Understanding these types of discrimination is pivotal as they elucidate how biases, whether direct, indirect, systemic, statistical, explainable, or unexplainable, can culminate in unfair outcomes. Each type of discrimination demonstrates a distinct pathway through which biases embedded in data or algorithms manifest

in practices and decisions, thus potentially perpetuating unfairness in AI systems.
405 This comprehensive analysis helps in identifying targeted strategies to mitigate these discriminatory effects and underscores the importance of developing automated decision systems that are both just and equitable.

Direct Discrimination occurs when outcomes are directly affected by sensitive attributes such as race, gender, or age. This type of discrimination happens
410 explicitly and is frequently legally prohibited;

Indirect Discrimination manifests when proxy attributes indirectly linked to sensitive attributes influence outcomes. For example, using zip codes in decision-making processes might inadvertently reflect racial biases because residential areas often correlate with racial demographics. This phenomena is also referred
415 as redlining effect (PEDRESCHI *et al.*, 2008);

Systemic Discrimination involves policies or practices entrenched within an organization that perpetuate disadvantage for certain groups. This can stem from cultural biases embedded in the decision-making processes, often reflecting the preferences or biases of dominant groups;

420 **Statistical Discrimination** refers to the use of general statistics on a group to make inferences about individuals from that group. This type of discrimination might arise when decision-makers use visible characteristics as proxies for other traits, leading to biased assessments;

Explainable Discrimination is considered legally permissible if the differences
425 in treatment or outcomes can be justified through legitimate and relevant attributes. For instance, differences in pay might be justified by the number of hours worked if this factor significantly influences earnings;

Unexplainable Discrimination occurs when there is no justifiable reason for the
430 disparate treatment or outcomes, making it illegal and ethically unacceptable. This type of discrimination requires interventions to ensure fairness and equality in decision-making processes.

2.2 Fairness definitions and metrics

This section aims to present some widely used definitions and metrics of fairness, as described by VERMA e RUBIN (2018) and summarized by MEHRABI *et al.* (2021)
435 and CATON e HAAS (2023), providing a comprehensive overview for understanding and navigating the multifaceted dimensions of fairness in ML systems. Initially, we explore general considerations and intuitive aspects of fairness, setting the stage

for a deeper understanding. This preliminary discussion lays the groundwork for understanding the nuanced nature of fairness notions within the context of ML. Following this, we will transition into formal definitions, where we will dissect and explain those metrics and concepts.

Even before this discussion, we emphasize that no single fairness definition universally applies to all scenarios. The choice of a particular fairness definition and metric should be informed by ethical considerations grounded in the social context in which the model would be deployed (ALER TUBELLA *et al.*, 2022). Selecting a fairness definition is not a purely technical matter, as it inevitably requires ethical and social considerations that should not be neglected (ALVES *et al.*, 2023). Building fair machine learning models requires an interdisciplinary approach that engages all stakeholders, including specially those who are typically marginalized or underrepresented (WEINBERG, 2022).

A prevalent taxonomy within fairness literature differentiates fairness notions into group metrics and individual metrics. Group Fairness Metrics hinge on the principle that statistical measures — such as error rates, precision, and recall — ought to be equitably distributed across groups demarcated by sensitive attributes like race, gender, or age. The core premise of these metrics is that fairness is actualized when an algorithm exhibits consistent performance across diverse demographic segments.

Demographic Parity, for example, mandates uniformity in the rate of positive algorithmic outcomes across different groups, a standard that remains agnostic to the underlying base rates within each population segment. On the other hand, Equal Opportunity and Equalized Odds introduce a nuance to this conversation by tethering fairness to the true condition of outcomes. This refinement delineates a central differentiation within fairness metrics: some are predicated solely on predicted values (such as Demographic Parity), while others derive from the scope of the confusion matrix (Table 2.1), also incorporating true value conditions (as seen in Equal Opportunity and Equalized Odds).

Individual Fairness Metrics, in contrast, introduce a more granular perspective to fairness, advocating that similar individuals should be treated similarly by the ML system. This approach diverges from group-level considerations, focusing instead on ensuring that the algorithm’s treatment is consistent for individuals who are alike in relevant aspects, barring their membership in different demographic categories. Individual fairness seeks to ensure a personalized sense of justice, where the algorithmic outcomes are solely reflective of pertinent attributes rather than biased by irrelevant factors associated with sensitive attributes. This concept champions the notion that fairness extends beyond group identities to recognize and respect the uniqueness of individual experiences and qualifications.

To establish the foundation for discussing fairness definitions and metrics, we commence with an examination of the confusion matrix, which is an essential instrument in machine learning to assessing the performance of classification algorithms. It constitutes a tabular visualization that delineates the correspondence between the true labels and the predicted outcomes generated by a model. For binary classification tasks, the confusion matrix is structured into four principal components: True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN), as outlined in Table 2.1. By providing a clear breakdown of these outcomes, the confusion matrix allows to calculate many key performance metrics such as accuracy, precision, recall, and the F1 score, providing comprehensive insights into the strengths and weaknesses of the classification model. Also, the computation of those metrics forms the basis for evaluating fairness across distinct demographic groups.

Table 2.1: Confusion matrix of binary classification outcomes

		Predicted	
		Positive	Negative
Actual	Positive	TP	FN
	Negative	FP	TN

True Positives (TP) can be defined as the probability that the predictor correctly identifies a positive outcome when the true condition is positive. Using the conditional probability notation, it is expressed as $P(\hat{Y} = 1|Y = 1)$, indicating the probability that the predicted class \hat{Y} is positive given that the actual class Y is positive.

False Positives (FP) represent the probability that the predictor incorrectly identifies a positive outcome when the true class is negative. It is denoted as $P(\hat{Y} = 1|Y = 0)$, reflecting the probability that the predicted class \hat{Y} is positive when the actual class Y is negative.

False Negatives (FN) are defined as the probability that the predictor incorrectly identifies a negative outcome when the true class is positive. This is given by $P(\hat{Y} = 0|Y = 1)$, the probability that the predicted class \hat{Y} is negative given that the actual class Y is positive.

True Negatives (TN) correspond to the probability that the predictor correctly identifies a negative outcome when the true condition is negative. In conditional probability terms, it is $P(\hat{Y} = 0|Y = 0)$, indicating the probability that the predicted class \hat{Y} is negative given that the actual class Y is negative.

Now we proceed to more complex metrics that provides complementary insights into the performance of the classifier. These derived metrics, such as Positive Predictive Value (PPV), False Discovery Rate (FDR), and others, constitutes the basic elements of the confusion matrix to quantify the reliability of the predictions in

510 various ways. By expressing these metrics in terms of conditional probabilities and confusion matrix components, we facilitate a comprehensive analysis of the classifier's behavior, providing resources to a proper evaluation of its fairness across different demographic groups.

Definition 1 (Positive Predictive Value (PPV)). *PPV, or precision, measures the proportion of correctly identified positive outcomes among all predicted positives. It is defined as the probability that the true condition is positive given the predicted condition is positive, $P(Y = 1|\hat{Y} = 1)$. In terms of the confusion matrix, PPV is calculated as $\frac{TP}{TP+FP}$, the ratio of true positives to the sum of true positives and false positives.*

520 **Definition 2** (False Discovery Rate (FDR)). *FDR quantifies the rate of incorrect positive predictions. It is the probability that the true condition is negative when the predicted condition is positive, $P(Y = 0|\hat{Y} = 1)$. From the confusion matrix, FDR is computed as $\frac{FP}{TP+FP}$, indicating the proportion of false positives out of all predicted positives.*

525 **Definition 3** (Negative Predictive Value (NPV)). *NPV assesses the accuracy of negative predictions, representing the probability that the true condition is negative given the predicted condition is negative, $P(Y = 0|\hat{Y} = 0)$. NPV is derived from the confusion matrix as $\frac{TN}{TN+FN}$, the number of true negatives over the sum of true negatives and false negatives.*

530 **Definition 4** (False Omission Rate (FOR)). *FOR indicates the likelihood of a false negative prediction. It corresponds to the probability that the true condition is positive when the predicted condition is negative, $P(Y = 1|\hat{Y} = 0)$. In the confusion matrix context, FOR is $\frac{FN}{TN+FN}$, representing the number of false negatives relative to all predicted negatives.*

535 **Definition 5** (True Positive Rate (TPR)). *TPR, or recall, measures the proportion of actual positives that are correctly predicted. It is the probability that the predicted condition is positive given the true condition is positive, $P(\hat{Y} = 1|Y = 1)$. TPR is calculated as $\frac{TP}{TP+FN}$ in the confusion matrix, the ratio of true positives to the sum of true positives and false negatives.*

540 **Definition 6** (False Negative Rate (FNR)). *FNR quantifies the rate of missed positive predictions. It is defined as the probability that the predicted condition is negative when the true condition is positive, $P(\hat{Y} = 0|Y = 1)$. FNR is derived from the confusion matrix as $\frac{FN}{TP+FN}$, indicating the proportion of false negatives out of the actual positives.*

545 **Definition 7** (True Negative Rate (TNR)). *TNR, or specificity, indicates the accuracy of negative predictions, representing the probability that the predicted condition is negative given the true condition is negative, $P(\hat{Y} = 0|Y = 0)$. From the confusion matrix, TNR is computed as $\frac{TN}{TN+FP}$, the number of true negatives to the sum of true negatives and false positives.*

550 **Definition 8** (False Positive Rate (FPR)). *FPR assesses the likelihood of incorrect negative predictions, calculated as the probability that the predicted condition is positive when the true condition is negative, $P(\hat{Y} = 1|Y = 0)$. FPR is given by $\frac{FP}{TN+FP}$ in the confusion matrix, the ratio of false positives to the sum of true negatives and false positives.*

555 As we transition from foundational metrics that directly stem from the confusion matrix, such as PPV and TPR, we now discuss standard performance metrics that assess classification models in a more comprehensive manner. These metrics, such as Accuracy and F1 are distinguished by their reliance on both classes to provide a more holistic evaluation.

560 **Definition 9** (Accuracy (Acc.)). *Probably the most widely used performance metric to classification problems, Accuracy is the proportion of true results, both true positives and true negatives, among the total number of cases examined. In terms of conditional probabilities, accuracy reflects the probability that the predicted condition is correct, both as a positive and negative outcome, given the actual conditions, and*
 565 *can be expressed as*

$$P(\hat{Y} = Y) = P(\hat{Y} = 1|Y = 1) \cdot P(Y = 1) + P(\hat{Y} = 0|Y = 0) \cdot P(Y = 0).$$

Using the confusion matrix, accuracy is computed as

$$\frac{TP + TN}{TP + TN + FP + FN}.$$

Definition 10 (Balanced Accuracy (Bal. Acc.)). *Balanced accuracy is an average of the true positive rate (TPR) and the true negative rate (TNR), which compensates for class imbalance by treating both classes equally. Using conditional probabilities, it can be expressed as*

$$\frac{1}{2} \left[P(\hat{Y} = 1|Y = 1) + P(\hat{Y} = 0|Y = 0) \right],$$

where each term represents the conditional probability of correctly predicting the

respective class. In terms of the confusion matrix, balanced accuracy is calculated as

$$\frac{1}{2} \left[\frac{TP}{TP + FN} + \frac{TN}{TN + FP} \right].$$

Definition 11 (F1 Score). *The F1 score is the harmonic mean of precision and recall, providing a balance between the PPV and TPR. It is calculated as $2 \cdot \frac{PPV \cdot TPR}{PPV + TPR}$. Using conditional probabilities and confusion matrix terms, the F1 score can be expressed as*

$$2 \cdot \frac{P(Y = 1|\hat{Y} = 1) \cdot P(\hat{Y} = 1|Y = 1)}{P(Y = 1|\hat{Y} = 1) + P(\hat{Y} = 1|Y = 1)},$$

and calculated using terms from confusion matrix as

$$\frac{2 \cdot TP}{2 \cdot TP + FP + FN}.$$

Definition 12 (Matthews Correlation Coefficient (MCC)). *MCC is a measure of the quality of binary classifications, producing a value between -1 and 1 where 1 is a perfect prediction, 0 no better than random prediction, and -1 indicates total disagreement between prediction and observation. The MCC is defined as*

$$\frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN)}}.$$

In terms of conditional probabilities, MCC considers all four quadrants of the confusion matrix, correlating the true and predicted conditions. It can be seen as a correlation coefficient between the observed and predicted binary classifications, providing a more informative measure than simple accuracy in the presence of class imbalance.

570

Now we describe the most widely used group fairness definitions, including statistical parity, equal opportunity, predictive equality, and equalized odds. Demographic parity requires that the likelihood of a positive outcome is the same across different groups, irrespective of their sensitive attributes. Equal opportunity extends this concept to the true positive rate, ensuring that individuals from different groups have an equal chance of being correctly classified as positive. Predictive equality, on the other hand, focuses on the true negative rate, ensuring that individuals from different groups have an equal chance of being correctly classified as negative. Equalized odds combines the principles of equal opportunity and predictive equality, ensuring that both true positive and true negative rates are equal across different groups.

580

Definition 13 (Statistical Parity). *The likelihood of a positive, i.e., favorable, outcome should be the same in every group of the sensitive attribute (DWORK et al.,*

2012; KUSNER et al., 2017). A binary predictor \hat{Y} satisfies Statistical Parity (a.k.a. Demographic Parity) if $P(\hat{Y}|A = 0) = P(\hat{Y}|A = 1)$, where A is a protected attribute.

585 For example, the credit approval probability should be the same for the male and female groups. Demographic Parity does not depend on true class Y , only on prediction \hat{Y} . We can measure Demographic Parity (Definition 13) for a protected attribute A as the absolute difference between $P(\hat{Y}|A = 0)$ and $P(\hat{Y}|A = 1)$, as seen in Equation 2.1. According Demographic Parity, the predictor is considered fairer
590 when this metric is lower.

$$|P(\hat{Y}|A = 0) - P(\hat{Y}|A = 1)| \quad (2.1)$$

By analyzing the confusion matrix, we can determine the absolute difference between the rates of $(TP + FP)/(TP + FP + TN + FN)$ for both protected and unprotected groups.

Definition 14 (Equal Opportunity). *The probability of a person in a positive class being assigned to a positive, i.e., favorable, outcome should be the same in every group of the sensitive attribute (HARDT et al., 2016). A binary predictor \hat{Y} satisfies Equal Opportunity if $P(\hat{Y}|A = 0, Y = 1) = P(\hat{Y}|A = 1, Y = 1)$, where Y is true class and A is a protected attribute.*

Definition 14 claims that protected and unprotected, i.e., privileged, groups
600 should have equal true positive rates. Mathematically, a classifier with equal true positive rates will also have equal false negative rates, so we can analyze the confusion matrix checking whether a predictor has equal $(TP)/(TP + FN)$ or $(FN)/(TP + FN)$ in each group of the sensitive attribute. Like in Demographic Parity, we can measure Equal Opportunity as an absolute difference between protected and privileged groups, as defined in Equation 2.2.
605

$$|P(\hat{Y}|A = 0, Y = 1) - P(\hat{Y}|A = 1, Y = 1)| \quad (2.2)$$

Definition 15 (Predictive Equality). *The probability of a person in a negative class being assigned to a negative outcome should be the same in every group of the sensitive attribute. A binary predictor \hat{Y} satisfies Predictive Equality if $P(\hat{Y}|A = 0, Y = 0) = P(\hat{Y}|A = 1, Y = 0)$, where Y is true class and A is a protected attribute.*
610

Definition 15 establishes that that both the protected and privileged groups should have the same true negative rates, which consequently results in equal false positive rates. Using a confusion matrix definition, we check the absolute difference of $(TN)/(TN + FP)$ or $(FP)/(TN + FP)$ between the groups. So, we can measure

615 Predictive Equality according Equation 2.3.

$$|P(\hat{Y}|A = 0, Y = 0) - P(\hat{Y}|A = 1, Y = 0)| \quad (2.3)$$

Definition 16 (Equalized Odds). *Both probabilities of the person in a positive class being assigned to a positive outcome and of a person in a negative class being assigned to a negative outcome should be the same in every group of the sensitive attribute (HARDT et al., 2016). A binary predictor \hat{Y} satisfies Equalized Odds (a.k.a. Average Odds Difference) if $P(\hat{Y}|A = 0, Y) = P(\hat{Y}|A = 1, Y)$, where Y is true class and A is a protected attribute.*

Equalized Odds is a combination of the principles from Definition 14 and Definition 15, i.e., protected and unprotected groups should have equal true positive and true negative rates, therefore equal false positive and false negative rates. Using a confusion matrix definition, we check the absolute difference between $(TP)/(TP + FN)$ and $(TN)/(TN + FP)$ of predictor in protected and unprotected groups. Equation 2.4 describes how to measure Equalized Odds as the average between Equal Opportunity and Predictive Equality. According to Definition 16, the predictor is considered fairer when this metric is lower.

$$\frac{1}{2} \left[|P(\hat{Y}|A = 0, Y = 1) - P(\hat{Y}|A = 1, Y = 1)| + |P(\hat{Y}|A = 0, Y = 0) - P(\hat{Y}|A = 1, Y = 0)| \right] \quad (2.4)$$

630 Using the same logic, it is possible to define group fairness metrics based derived from any binary classification metric from confusion matrix. The procedure is the same, assessing the absolute difference from those metrics between protected and unprotected groups.

While individual fairness metrics strive to ensure equitable treatment of individuals based on their specific attributes and circumstances, they may not always capture broader systemic inequalities that affect entire groups. As we pivot our discussion towards group fairness, we examine the potential drawbacks and complications that can arise when pursuing fairness metrics across different demographic groups.

640 In this context a key challenge is the Simpson's Paradox (BLYTH, 1972), where trends apparent in separate groups disappear or reverse when these groups are combined. This can lead to misleading conclusions in aggregated data, potentially obscuring significant disparities within subgroups that are averaged out in the analysis. Furthermore, group fairness metrics may inadvertently mask discrimination within protected groups. For instance, a model could satisfy group fairness criteria overall while still discriminating against specific subgroups within a protected class due

to the heterogeneity within larger groups that isn't captured by broader fairness assessments.

Additionally, implementing group fairness often involves trade-offs GOH *et al.* (2016); KOMIYAMA *et al.* (2018); PETROVIĆ *et al.* (2021); F.CRUZ *et al.* (2021); LIU e VICENTE (2022) that can impact the overall performance of the predictive model. Balancing fairness with accuracy can lead to difficult choices, especially in high-stakes applications such as healthcare or criminal justice, where the cost of errors is significant. For example, efforts to reduce false positive rates in one group might inadvertently increase false negatives in another, adversely affecting the model's overall predictive utility. Another issue arises from the conflict between different fairness definitions, where improving fairness according to one metric might worsen it according to another. Achieving demographic parity, which calls for equal outcomes across groups, might conflict with ensuring equal opportunity, which demands equal true positive rates across groups. Such conflicts necessitate careful consideration to determine which fairness criteria are most appropriate for specific applications.

Lastly, standard group fairness metrics often overlook intersectionality—the complex, cumulative way in which multiple forms of discrimination, such as race, gender, and class, intersect and affect individuals (KEARNS *et al.*, 2017, 2019). Ignoring this aspect can result in policies and models that do not fully address the nuanced ways in which bias manifests. This oversight underscores the importance of individual fairness, a principle that seeks to ensure equitable treatment by focusing on the uniqueness of each individual rather than merely categorizing them into groups. Individual fairness advocates for algorithms to treat similar individuals similarly, regardless of their group membership MEHRABI *et al.* (2021), thus acknowledging and addressing the multifaceted nature of discrimination and ensuring that each person is considered on their own merits. By integrating individual fairness into our models, it is possible to better capture and mitigate the intersecting and often overlapping biases that group fairness metrics might miss, providing a more comprehensive approach to fairness in AI systems

Fairness Through Awareness (DWORK *et al.*, 2012) is a concept which focuses on treating similar individuals similarly. It emphasizes the importance of fairness at the individual level by defining a metric of similarity between individuals based on relevant characteristics, and ensuring that the algorithm's decisions are consistent for individuals deemed similar by this metric. This approach is rooted in the idea that fairness can be achieved by explicitly considering the sensitive attributes through a carefully defined similarity function, ensuring that decisions are justifiable and tailored to individual circumstances.

Definition 17 (Fairness Through Awareness). *A predictor \hat{Y} satisfies Fairness*

Through Awareness if for any two individuals $x, x' \in X$, where X is the domain of individuals, the distance metric $d(x, x')$ under which the individuals are considered similar enforces that $|\hat{Y}(x) - \hat{Y}(x')| \leq d(x, x')$. Here, d is a task-specific metric that measures similarity relevant to the decision-making process, incorporating sensitive attributes where necessary.

This definition implies that the algorithm must incorporate a nuanced understanding of what it means for two individuals to be similar, which goes beyond merely ignoring sensitive attributes. Instead, it considers these attributes in a way that respects individual differences and upholds fairness.

Fairness Through Unawareness CORBETT-DAVIES *et al.* (2018), on the other hand, is a more straightforward approach where an algorithm is considered fair if it does not explicitly use sensitive attributes (such as race, gender, etc.) in the decision-making process. This method assumes that the exclusion of sensitive attributes will prevent discriminatory practices. However, this approach can be naive as it fails to consider that biases can be encoded in other, non-sensitive attributes that are correlated with the sensitive ones MEHRABI *et al.* (2021); CATON e HAAS (2023); HORT *et al.* (2023).

Definition 18 (Fairness Through Unawareness). *A predictor \hat{Y} satisfies Fairness Through Unawareness if the decision function \hat{Y} does not explicitly include any sensitive attribute A as part of the input. In other words, \hat{Y} is constructed without direct knowledge of A .*

Another example of Individual Fairness Metric is the notion of Counterfactual Fairness (KUSNER *et al.*, 2017), which introduce a causal reasoning framework into the fairness discourse. These metrics are based on the concept that a decision is fair towards an individual if the same decision would have been made in a counterfactual world where the individual belonged to a different demographic group but all other characteristics remained constant. This approach hinges on causal models that specify how sensitive attributes affect other features and the outcome. Counterfactual fairness aims to address the individual-level biases that group fairness metrics might overlook, providing a nuanced approach that considers the hypothetical scenarios of individuals belonging to different demographic categories. By employing counterfactual analysis, one can assess whether the disparities in ML predictions stem from legitimate factors or unjust biases. Relevant works approaching this notion include WU *et al.* (2022), MA *et al.* (2023), and GRARI *et al.* (2023).

Definition 19 (Counterfactual Fairness). *A predictor \hat{Y} is counterfactually fair with respect to a protected attribute A if, under any context $X = x$ and $A = a$, the distribution of \hat{Y} is the same in the actual world and a counterfactual world where*

A is set to any permissible value. That is,

$$P(\hat{Y}_{A \leftarrow a}(U) = y \mid X = x, A = a) = P(\hat{Y}_{A \leftarrow a'}(U) = y \mid X = x, A = a),$$

720 *for all y and any value a' of A , where X are the features not causally dependent on A , and U denotes the background variables.*

This definition roots itself in the idea that fairness should be preserved across hypothetical alterations of the sensitive attribute, reflecting a robust stance against biases that might otherwise emerge due to such attributes.

725 Implementing counterfactual fairness involves constructing a causal model that maps how inputs (features including sensitive attributes) influence the outputs (predictions). One must identify which attributes are causally independent of the sensitive attribute and ensure that the predictions are invariant when the sensitive attribute’s values are modified hypothetically.

730 This approach is particularly pertinent when decisions have substantial impacts on individuals, such as in hiring, loan approval, or healthcare settings. By ensuring that predictions remain consistent regardless of changes to sensitive attributes, models can be designed to mitigate unfair discriminatory practices that could otherwise affect outcomes based on irrelevant attributes.

735 While the concept of counterfactual fairness is intuitive and persuasive, its implementation poses significant challenges KASIRZADEH e SMART (2021). Building models that reflect the true causal relationships in the data is non-trivial and requires deep domain knowledge. Also, access to good quality data that sufficiently captures the causal dependencies is necessary, which can be a limiting factor in many practical scenarios. Finally, the complexity of calculating counterfactuals, especially in large datasets with many attributes, can be computationally demanding. Despite these challenges, counterfactual fairness pushes the boundaries of fairness in machine learning by providing a framework that directly tackles the underlying causal mechanisms leading to biased decisions.

745 In the context of fairness definitions and metrics there is a relevant problem to be considered, the Impossibility Theorem. As elucidated simultaneously by KLEINBERG *et al.* (2017) and CHOULDECHOVA (2017) with further contributions by SARAVANAKUMAR (2020), BELL *et al.* (2023) and BEIGANG (2023), articulates a fundamental challenge in the domain of algorithmic fairness: the concurrent satisfaction of distinct fairness metrics is inherently unfeasible under certain conditions. This theorem, also referred to as the Incompatibility of Fairness Criteria, delineates the intrinsic conflicts arising amongst prevalent fairness constructs.

The Impossibility Theorem in the context of algorithmic fairness articulates a fundamental challenge: it is not feasible to simultaneously satisfy multiple fairness

755 criteria in certain realistic settings. This theorem highlights the inherent conflict that arises when attempting to meet several well-intentioned fairness metrics such as counterfactual fairness, equalized odds, and predictive parity at the same time.

According to the theorem, if a predictive model is designed to achieve counterfactual fairness, it will likely conflict with the criteria of equalized odds or predictive parity. Counterfactual fairness demands that the model’s prediction for an individual would remain unchanged in hypothetical scenarios where the individual’s protected characteristics (such as race or gender) are altered but all other variables are held constant. In contrast, equalized odds require that error rates across different groups are similar, while predictive parity necessitates comparable predictive values across these groups. When protected characteristics are causally relevant to the predicted outcomes, aligning the model with one fairness metric may inadvertently breach another.

This theorem thus underscores the practical dilemmas in fair machine learning models. Achieving comprehensive fairness in ML systems often requires navigating complex trade-offs, necessitating a thoughtful prioritization of fairness criteria tailored to the specific context and ethical considerations of each use case. The Impossibility Theorem serves as a critical reminder of the limitations and careful considerations required in the pursuit of fair decision making systems, highlighting the importance of making informed, contextually sensitive decisions when implementing fairness metrics.

2.3 Fair classification

In this section, we review pertinent literature on fair machine learning, placing a particular emphasis on in-processing techniques. Fairness intervention methods can be classified into three categories based on the stage at which they occur, as proposed by MEHRABI *et al.* (2021) and ALER TUBELLA *et al.* (2022):

Pre-processing intervene before learning, modifying the data to reduce existing biases;

In-processing intervene during learning by modifying the objective functions or imposing constraints to the model in order to mitigate discriminatory effects;

785 **Post-processing** affects predictions produced by the model after learning to change possibly unfair outcomes.

One notable pre-processing method is the reweighting approach proposed by KAMIRAN e CALDERS (2012), which adjusts the weights of different samples in

the training data to ensure that underrepresented groups are fairly represented during training. Another significant pre-processing technique is the Fair Representation Learning by ZEMEL *et al.* (2013), which learns a latent representation of the data that obfuscates sensitive attributes while retaining the information necessary for accurate predictions. An example of a post-processing method is the Reject Option Classification by KAMIRAN *et al.* (2012), which changes the decisions of the classifier for individuals near the decision boundary. Another example is Equalized Odds and Equal Opportunity post-processing technique by HARDT *et al.* (2016), which adjusts the classifier’s predictions to equalize the true positive and false positive rates across different demographic groups.

In this work, we incorporate information about disparities among social groups in the dataset into our model by modifying the loss function through the use of a transition matrix. This fairness intervention is thus classified as an in-processing technique. Other relevant in-processing strategies for fair classification include Naive Bayes approaches for discrimination-free classification (CALDERS e VERWER, 2010), Fairness Through Awareness Framework (DWORK *et al.*, 2012), Fairness-Aware Classifier with Prejudice Remover Regularizer KAMISHIMA *et al.* (2012), α -discriminatory empirical risk minimizer (WOODWORTH *et al.*, 2017), Disparate Impact and Disparate Mistreatment frameworks for margin-based classifiers (ZAFAR *et al.*, 2017a,b), Weak Agnostic Learning to Auditing Subgroup Fairness (KEARNS *et al.*, 2019, 2018), One-Network Adversarial Fairness (ADEL *et al.*, 2019), FairGan⁺ (XU *et al.*, 2019), Monte Carlo policy gradient method (PETROVIĆ *et al.*, 2021), Fairness-accuracy Pareto (WEI e NIETHAMMER, 2022), and Pareto front stochastic multi-gradient (LIU e VICENTE, 2022) based in original stochastic multi-gradient (MERCIER *et al.*, 2018) to Multi-Objective Optimization and the hybrid Adaptive Priority Reweighting approach HU *et al.* (2023).

In KAMISHIMA *et al.* (2012) the authors proposes the Prejudice Remover (PR) which is a regularizer to logistic regression models. It introduces an additional term in the loss function to penalize the model for making decisions based on sensitive features. The objective function to be minimized is available on Equation 2.5, where Θ is the model parameters, $L(D; \Theta)$ the log-likelihood, $R(D, \Theta)$ the prejudice remover regularizer, η a regularization parameter controlling the trade-off between fairness and accuracy, and λ a parameter for the L2 regularizer.

$$-L(D; \Theta) + \eta R(D, \Theta) + \frac{\lambda}{2} \|\Theta\|^2 \quad (2.5)$$

The regularizer $R(D, \Theta)$ aims to minimize the mutual information between the predicted outcomes and the sensitive features, thereby reducing the model’s reliance on sensitive information. The mutual information is approximated using sample

825 means to make the computation feasible for large datasets. The authors compared the proposed method with Calders-Verwer 2-naïve-Bayes method (CALDERS e VERWER, 2010), showing that the PR effectively reduced bias, though sometimes at the cost of decreased accuracy.

As an alternative to mitigating unwanted bias, ZHANG *et al.* (2018) proposes an 830 adversarial method for reducing bias in machine learning models, namely Adversarial Debiasing (AD). This technique involves training a neural network predictor to forecast an outcome variable from inputs while an adversary network simultaneously attempts to predict a sensitive attribute, which should not influence the outcome.

The method utilizes an adversarial network architecture where the main predictor’s task is complemented by an adversarial model that tries to learn the sensitive 835 attribute. By integrating the adversarial model’s feedback into the training process, the predictor learns to make decisions that are increasingly independent of the sensitive attribute. This setup allows the model to attain to fairness constraints like Statistical Parity (Definition 13), Equal Opportunity (Definition 14), and Equalized 840 Odds (Definition 16).

In a similar vein, KEARNS *et al.* (2018) propose a framework for ensuring subgroup fairness, addressing the issue of fairness gerrymandering. Below, a toy example given by the authors illustrating a scenario where the referred fairness gerrymandering occurs.

845 *Imagine a setting with two binary features, corresponding to race (say black and white) and gender (say male and female), both of which are distributed independently and uniformly at random in a population. Consider a classifier that labels an example positive if and only if it corresponds to a black man, or a white woman. Then the classifier will appear to be equitable when one considers either protected attribute alone, in the sense that it labels both men and women as positive 50% of the time, and labels both black and white individuals as positive 50% of the time. But if one looks at any conjunction of the two attributes (such as black women), then it is apparent that the classifier maximally violates the statistical parity fairness constraint. Similarly, if examples 855 have a binary label that is also distributed uniformly at random, and independently from the features, the classifier will satisfy equal opportunity fairness with respect to either protected attribute alone, even though it maximally violates it with respect to conjunctions of two attributes.*

860 Their approach involves defining fairness for exponentially or infinitely many subgroups defined by a structured class of functions over the protected attributes, not only for a small number of pre-defined groups as considered in hegemonic fair

classification approaches. This framework is formalized as a two-player zero-sum game between a Learner (the primal player) and an Auditor (the dual player), where the Learner aims to minimize classification error while the Auditor seeks to identify and penalize fairness violations. The computational challenges of this approach are mitigated by connecting it to weak agnostic learning, which allows the use of practical machine learning heuristics for effective auditing and learning in real-world applications.

The algorithms derived from this framework provably converge to the best fair distribution over classifiers, given access to oracles capable of optimally solving the agnostic learning problem. These algorithms include a variant based on the no-regret Follow the Perturbed Leader algorithm and another using Fictitious Play, both of which have been implemented and evaluated on real datasets, demonstrating their efficacy in achieving subgroup fairness.

The Adaptive Priority Reweighting HU *et al.* (2023) method introduces a systematic approach to increase fairness and generalizability of classifiers by dynamically adjusting sample weights based on their proximity to the decision boundary. Initially, the training samples are divided into subgroups according to their sensitive attributes and classifier predictions. Each sample’s distance to the decision boundary is then computed and updated iteratively. During each iteration, subgroup weights are recalibrated by comparing the observed probability of the positive prediction rate within each subgroup to the expected probability under statistical independence. This comparison helps to assign higher weights to samples closer to the decision boundary, thereby prioritizing them in the training process. The weighted loss function is optimized using a stochastic gradient descent algorithm, which iteratively adjusts the classifier to reduce bias while maintaining accuracy. By continuously updating the weights and distances, the Adaptive Priority Reweighting method ensures that the classifier learns to make fairer decisions that generalize well to unseen data, addressing the limitations of traditional reweighting methods that often fail to generalize beyond the training set.

The method is evaluated on benchmark datasets, outperforming many state-of-art pre-processing, in-processing and post-processing fair classification techniques, promoting fairness on both finetuned pre-trained models and newly trained models. This technique can be classified as an hybrid approach, performing the traditional pre-processing instance reweighting through an adaptive training algorithm.

Recently, special attention has been given in fair machine learning research topics like addressing multiple sensitive attributes or multiple classes D’ALOISIO *et al.* (2023); LIU *et al.* (2023), loss balancing techniques KIM *et al.* (2023); KHALILI *et al.* (2023), where the objective is to balance the loss across different groups instead of predictive metrics, adversarial approaches MA *et al.* (2023); GRARI *et al.*

(2023); LIANG *et al.* (2023); ZHANG *et al.* (2023a); MOUSAVI *et al.* (2023); WEI *et al.* (2023) and the privacy concerns involving fairness under federated learning settings CHEN *et al.* (2024); VUCINICH e ZHU (2023). Another relevant research
905 topic in fair machine learning is learning under censored data ZHANG e WEISS (2022); ZHANG *et al.* (2023b); ZHANG e WEISS (2023); ZHANG *et al.* (2023c), which we will discuss in section ??.

Chapter 3

Fair loss function from noise robustness

910

In this chapter we propose a novel loss function inspired by label noise robustness methods to fair classification problems. The core of this proposal consists in using transition matrices as parameters to a proper loss correction PATRINI *et al.* (2017), with matrices values as factors to redistribute unbalanced bias across different groups within the sensitive feature.

915

3.1 Preliminaries

920

The presence of noise in data can substantially decrease model performance in classification problems. Noise can be defined as non systematic errors that obscures the relationship between features of an instance and its class (FRÉNAY e VERLEYSEN, 2014; HICKEY, 1996; QUINLAN, 1986). Two types of noise are found in literature, in features (or attributes) and in labels (or classes). Feature noise affects observed values, e.g. by adding a small Gaussian noise on each feature during measurement. Likewise, label noise change the observed label assigned to an instance, e.g. by randomly inverting labels in a binary classification problem. Although feature noise could affect model performance, label noise is potentially more harmful, since we frequently have many features and only one label. Note that in label noise only the observed label of an instance is affected, its true class remains the same.

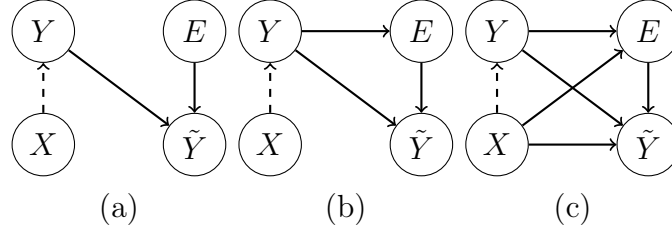
925

930

The label noise taxonomy considers three types of noise: Noisy Completely at Random, Noisy at Random, and Noisy Not at Random (FRÉNAY e VERLEYSEN, 2014). Figure 3.1 presents the statistical dependency between features X , class Y , observed label \tilde{Y} and the occurrence of error E , i.e. $E = 1$ when $Y \neq \tilde{Y}$. The simplest type is Noisy Completely at Random, where the occurrence of error E not depend on X and Y , e.g. randomly flipping labels on a binary classification problem.

In Noisy at Random, the occurrence of error E depends only on Y , e.g. randomly
 935 flipping labels on binary classification with different rates for positives and negatives
 classes. Noisy Not at Random considers the occurrence of error E depending on both
 Y and X , e.g. flipping labels on binary classification with different rates for each
 group of instances of a certain feature.

Figure 3.1: Noise taxonomy from a statistical perspective. (a) completely random noise (NCAR), (b) random noise (NAR) and (c) non-random noise (NNAR). The arrows correspond to the statistical dependencies. For clarity, the dependency between X and Y was placed as a dashed arrow.



Many label noise robustness methods can be found on literature, in this work we
 940 highlight the *backward* and *forward* loss corrections, proposed by PATRINI *et al.*
 (2017) using concepts of loss factorization (PATRINI *et al.*, 2016). Those loss cor-
 rection techniques considers a NAR label noise, which is described by a transition
 matrix T such as

$$T_{i,j} = P(\tilde{Y} = y_j | Y = y_i), \quad (3.1)$$

where $\mathcal{Y} = \{y_1, y_2, \dots, y_c\}$ is the set of all possible class labels. Transition matrix
 945 includes corruption probabilities for every possible label combination, each value
 represents the probability of one label be corrupted onto another. This matrix is
 row-stochastic and not necessarily symmetric across the classes.

$$\ell^{\leftarrow}(P(\tilde{Y}|X)) = T^{-1}\ell(P(\tilde{Y}|X)) \quad (3.2)$$

The backward loss correction is defined by Equation 3.2 to an arbitrary loss
 function ℓ and a transition matrix T . The backward loss correction involves a lin-
 950 ear combination of the loss values for each observed label, using coefficients that
 depends on the probability that each observed label reflects the true class. Intu-
 itively, we are reweighting the loss according to the noise probabilities of each label
 using the inverse of T and thus somehow going one step back, reverting the noise
 effects. This corrected loss is unbiased and can be minimized with any conventional
 955 back-propagation algorithm, making it flexible to include within different training
 techniques and data pipelines.

$$\ell^{\rightarrow}(P(\tilde{Y}|X)) = \ell(T^{\top}P(\tilde{Y}|X)) \quad (3.3)$$

However, backward correction requires matrix inversion, which may not exist or may lead to numerical instabilities if the transition matrix T is ill-conditioned. Although there is possible solutions to a bad condition number of T , one should
 960 consider using the forward correction, a backward variation proposed by PATRINI *et al.* (2017) to avoid this issue, as defined in Equation 3.3. While backward acts on the loss itself, forward corrects model predictions. Forward correction does not have the same theoretical guarantees as backward, but offers a label noise robustness, ensuring that the learned model is the minimize over the clean distribution without
 965 the need of matrix inversion.

Now we discuss classification methodologies that operate in the presence of label noise. While our research does not directly tackle fairness problems in the presence of label noise, we highlight relevant works that, akin to ours, bridge the domains of fairness and noise in machine learning research.

970 Some recent works deal with fairness problems in the presence of noise. For example, the sensitive attribute available could be noisy, which could distort the effects of fairness intervention. In this context, LAMY *et al.* (2019) uses noise-rate estimators from the label noise literature to change a fairness model. Also, FOGLIATO *et al.* (2020) proposes a framework for assessing how assumptions on
 975 the noise across groups affect the predictive bias properties in risk assessment models. Furthermore, WANG *et al.* (2020) considers the consequences of naively relying on noisy protected group labels while proposing two new optimization approaches with sensitive attribute noise robustness. A denoised version of the selection problem to deal with noisy sensitive attributes is proposed in MEHROTRA e CELIS (2021).
 980 Lastly, CELIS *et al.* (2021) proposes an optimization framework for classification in the presence of noisy protected attributes.

There is also the perspective of dealing with the proxy features divergence or covariance. A theoretical approach to this issue identifying potential sources of errors can be found in PROST *et al.* (2021). The problem of measuring group fairness in
 985 ranking based on divergence with proxy features is investigated by GHAZIMATIN *et al.* (2022). A framework of fair semi-supervised learning in the pre-processing phase can be found in ZHANG *et al.* (2022), which includes predicting labels for unlabeled data, a resampling method, and ensemble learning to improve accuracy and decrease discrimination.

990 Another research direction is considering how fair models perform in the presence of NNAR label noise, where error rates of corruption depend both on the label class and the membership of a protected subgroup. In this scenario WANG *et al.* (2021) addresses the problem of fair classification and WU *et al.* (2022) provides a general framework for rewriting the classification risk and the fairness metric in terms of
 995 noisy data and thereby building robust classifiers. In GHOSH *et al.* (2023) a study

about the presence of noise in the protected attribute can be found.

Furthermore, many recent works deal with fairness under semi-supervised settings considering censored data, that is, for some individuals the class label is not available due to censorship ZHANG e WEISS (2022); ZHANG *et al.* (2023b); ZHANG e WEISS (2023); ZHANG *et al.* (2023c). In this scenario, the main approach is to use some technique to estimate the missing data instead of removing the instance from training data. This is closely related to the previous problems of fair learning under noisy data. In censored fairness problems noise can be interpreted as a kind of censorship, as the original data affected by noise is not available.

Bias and noise are two related phenomena, both corrupt data affecting models trained with this data. For example, if noise disproportionately affects different groups this potentially produces unfairness in models that use this data in training (WANG *et al.*, 2021). For example, we could have positive true class ($Y = 1$) flipped into negative labels ($\tilde{Y} = 0$) more frequently in the protected group ($A = 1$) than in the privileged group ($A = 0$). Simultaneously, the negative class ($Y = 0$) could be more frequently flipped into positive observed labels ($\tilde{Y} = 1$) within the privileged/unprotected group ($A = 0$). This scenario could lead to a higher undetected false negative rate for the protected group and a higher false positive rate for the privileged group. In this case the Noisy Not at Random data would be a source of negative social bias.

As referred to before, in MEHRABI *et al.* (2021) a non-exhaustive list of bias types was presented. In the scenario described above, the incorrect measurement of the true class resulted in a different observed label ($Y \neq \tilde{Y}$), which could be classified as a *Measurement Bias*. Similarly, a Noisy Not at Random data could lead to a *Population Bias*, where the characteristics of the population represented in the data differ from those of the original target population.

It can be challenging to distinguish between label noise and bias in certain scenarios, specially when noise disproportionately affects different social groups. Although there is some overlapping, they are distinct phenomena. Label noise is a stochastic process that is considered independent and unintentional (FRÉNEY e VERLEYSEN, 2014), whereas bias is rooted in historical and social issues and could be intentional. Furthermore, even noise-free data, correctly represented by observed features and labels, may be unfair since the social phenomena that produce this data could be biased against some groups.

Previous studies on fair machine learning have largely concentrated on understanding how noisy or censored data affects fair learning and on mitigating these effects. Thus, the objective of this work is not to theoretically deal with fair machine learning as a label noise problem or incorporate noisy classes or attributes in fairness problems. In contrast, our approach is inspired by label noise techniques, but with

1035 a distinct goal: not merely to analyze or mitigate the impact of noise or censorship,
but to directly address and reduce unfairness itself.

To achieve a competitive trade-off between performance and fairness in the proposed method, we discuss multi-objective optimization within the context of fair machine learning. A model that substantially decreases model performance to reduce unfairness may not be a viable option, as low performance could harm all groups affected by the model’s decisions, including protected groups. Similarly, a model projected to be a fair alternative that keeps performance almost intact, but with little or even no gain in fairness, is not practically relevant. It is possible that fine-tuning this trade-off could result in a fairer solution that achieves better performance than traditional methods, but this is not the case for most practical problems. Achieving this balance is one of the most challenging tasks in fair machine learning.

In this context, an interesting approach is to deal with fair machine learning as a Multi-Objective Optimization (MOO) problem, where predictive performance and fairness metric are the objectives, which could be defined according Equation 3.4, where λ is a parameter configuration in the space Λ , $\rho : \Lambda \mapsto [0, 1]$ is a model performance metric and $\varphi : \Lambda \mapsto [0, 1]$ a fairness metric. The set of all optimal solutions is called Pareto front, where one objective cannot be improved without sacrificing another. In this setting there is no single λ^* optimal solution, but a set of solutions forming a Pareto front (PARETO, 1906).

$$\begin{aligned} & \max (\rho(\lambda), \varphi(\lambda)) \\ & \text{subject to } \lambda \in \Lambda \end{aligned} \tag{3.4}$$

1055 One of the most frequent approach to deal with MOO problems like these is to combine the multiple function outputs to a single scalar, which is called scalarization. Therefore, we could describe a general scalarization setup to Equation 3.4 according Equation 3.5. The effectiveness of this approach is that is also possible to use single objective optimization techniques to tackle the MOO optimization problem. In this scenario a relevant issue is to select a scalarization setup capable of promote a proper trade-off of all the objectives thorough the optimization process given the optimization method.

$$\arg \max_{\lambda \in \Lambda} G(\lambda) = (\rho(\lambda), \varphi(\lambda)) \tag{3.5}$$

The fairness-accuracy Pareto front is formally described in WEI e NIETHAMMER (2022), which demonstrate that many existing fairness methods are performing a linear scalarization scheme and argues that it has several limitations in recovering Pareto optimal solutions. Instead, authors proposes a Chebyshev scalarization

scheme, that is theoretically superior than linear scheme. A characterization of the accuracy-fairness trade-off as a Pareto front can be found in LIU e VICENTE (2022). Also, MERCIER *et al.* (2018) proposes a stochastic multi-gradient based in original
1070 stochastic multi-gradient to Multi-Objective Optimization.

Another remarkable use of MOO in Fair Machine Learning is to perform a Fair Hyperparameter Optimization, which provides a model agnostic approach with flexibility to apply in multiple machine learning pipelines. A time-efficient Bayesian Optimization approach can be found in SCHMUCKER *et al.* (2020), combining
1075 scalarization techniques with the bandit-inspired Hyperband (LI *et al.*, 2017) algorithm to Hyperparameter Optimization in context of fairness.

A general objective function to be used with some popular off-the-shelf hyperparameters optimization techniques combining model performance and fairness in a flexible setting can be found in F.CRUZ *et al.* (2021). The authors argues that
1080 in fairness context the Pareto front is most often convex, thus proposes a simple scalarizing function that could be applied to reduce G to a single scalar with weighed l_p -norm. Also, they argue that GIAGKIOZIS e FLEMING (2015) demonstrate the use of l_p -norms with a high p value leads to slower convergence. Thus, the optimization metric $g(\lambda) = \|G(\lambda)\|_1$ is optimized according Equation 3.6, where
1085 α is the relative importance of predictive performance and fairness and λ is a parameter configuration in the space Λ . In experiments, α is fixed at 0.5, giving same importance to both objectives.

$$G(\lambda) = \alpha \cdot \rho(\lambda) + (1 - \alpha) \cdot \varphi(\lambda) \quad (3.6)$$

A Multi-objective SVMOptimizer with Dataset Constraints is proposed by GOH *et al.* (2016), where the objective is to minimize multiple objectives on real-world
1090 datasets, such as misclassification error and positive prediction at specific rate to some population. A custom reinforcement learning algorithm directly modeling performance and fairness as objectives is proposed by PETROVIĆ *et al.* (2021). Authors proposes using as reward function the difference between model performance (Area Under the ROC Curve) and three different fairness metrics (Statistical Parity, Equal Opportunity and Equalized Odds), each one with its respective importance
1095 coefficient. In experimental setups only one of those coefficient are different from zero. Thus, the optimized metric could be written as $G(\lambda) = \rho(\lambda) - \alpha \cdot \varphi(\lambda)$, where α is the relative importance of fairness.

3.2 Fair Transition Loss

1100 We propose a novel fair classification method inspired by techniques used for classification in the presence of label noise. By using some features of label noise methods that redistribute probabilities for unbalanced noise across classes, our approach reweights prediction probabilities to reduce disparities in favorable and unfavorable outcomes across social groups. This proposal was originally inspired by ?, that
1105 applies label noise techniques in order to achieve label noise robust collaborative filtering,

Whereas forward loss correction (PATRINI *et al.*, 2017) uses a transition matrix with corruption probabilities for every label combination in the case of NAR, fair classification problems are more related to NNAR. While forward loss correction
1110 uses a transition matrix with corruption probabilities for each label combination, as in the case of NAR, fair classification problems align more with NNAR scenarios. In NNAR, the probability of corruption depends not only on the true class but also on features, analogous to how bias in fairness problems is directed against certain groups. Here our correction does not revert a random label corruption from the
1115 true class, but a potentially unfair prediction. While noise label techniques, like forward (PATRINI *et al.*, 2017), aims to correct the prediction targeting a unknown true class using the available noisy label, analogously the proposed technique focus on correcting predictions chasing the unknown fair class using the available unfair label. Despite those are distinct phenomena, the corrections works the same way,
1120 adjusting the probabilities of predictions produced by a machine learning model during the training.

Thus, our proposal is a prediction probability loss reweighting technique that accounts different rates to each group of the sensitive feature, instead of using the same correction to every individual. A correction method that incorporates different
1125 probabilities for protected and unprotected groups could be more effective in mitigating bias during the learning phase. Specifically, we want a forward-based correction that takes into account a different matrix to each group of sensitive features, not only one transition matrix as used in label noise techniques. In this scenario, each group of sensitive feature have its own correction, with its own rates for each
1130 class combination. Ideally, if we can find an appropriate transition matrix that describes the bias to each group in a specific problem, we can apply a correction that attenuates those negative effects by reweighting model’s predictions in the learning process.

Next, we formally present Fair Transition Loss. For purpose of clarity we follow
1135 the same structure available at (PATRINI *et al.*, 2017), with the pertinent changes to our scope. The Fairness Transition Matrix T_a is defined with some abuse of

notation to the group $A = a$ of the sensitive feature as

$$T_{a,i,j} = P(\tilde{Y} = y_j | Y = y_i, A = a), \quad (3.7)$$

where label space $\mathcal{Y} = \{y_1, y_2, \dots, y_c\}$, c the number of classes, $Y = y_i$ is the unknown fair class and $\tilde{Y} = y_j$ is the available and possibly unfair label. Here, $T_{a,i,j}$ is the probability of the fair class $Y = y_i$ being unfairly labeled as $\tilde{Y} = y_j$ to an individual of the group $A = a$ due negative social bias. Therefore, suppose that there is an invertible link function $\psi : \Delta^{c-1} \rightarrow \mathbb{R}^c$, where $\Delta^{c-1} \subset [0, 1]^c$ is the c -simplex, the simplex in a c -dimensional space. Thus, a composite loss function, denoted by $\ell_\psi : \mathcal{Y} \times \mathbb{R}^c \rightarrow \mathbb{R}$ if it can be written as a decomposition of ψ^{-1} , that is,

$$\ell_\psi(Y, h(X)) = \ell(Y, \psi^{-1}(h(X))), \quad (3.8)$$

where $h : \mathcal{X} \rightarrow \mathbb{R}^c$ is a standard artificial neural network with multiple layers using activation functions, and $h(X)$ is the output of this neural network to a given input X . For example, to cross entropy loss function the softmax is the inverse link function. Proper loss functions are those that can be directly used to estimate class probabilities. The minimizer of a proper composite loss has the particular form of the link function applied to the conditional class probabilities $P(Y|X)$. Adding a new conditioning to this formulation, to an individual from group $A = a$ we have

$$\arg \min_h \mathbb{E}_{X,Y} \ell_\psi(Y, h(X|A = a)) = \psi(P(Y|X, A = a)). \quad (3.9)$$

Fair Transition Loss consists in correcting model's predictions with the same technique as forward, but taking into account the sensitive attribute value when choosing the transition matrix. In Theorem 1 the Fair Transition Loss is formally defined, with a guarantee about its minimizers.

Theorem 1. *Suppose that the Fairness Transition Matrix T_a for a given sensitive attribute $A = a$ is non-singular. Given a proper composite loss ℓ_ψ , define the Fair Transition Loss as*

$$\text{FTL}_\psi(h(X|A = a)) = \ell(T_a^\top \psi^{-1}(h(X|A = a))).$$

Then, the minimizer of the corrected loss under the unfair distribution is the same as the minimizer of the original loss under the fair distribution:

$$\arg \min_h \mathbb{E}_{X,\tilde{Y}} \text{FTL}_\psi(Y, h(X|A = a)) = \arg \min_h \mathbb{E}_{X,Y} \ell_\psi(Y, h(X|A = a)).$$

Proof. First notice that:

$$\begin{aligned}\text{FTL}_\psi(Y, h(X|A=a)) &= \ell(Y, T_a^\top \psi^{-1}(h(X|A=a))) \\ &= \ell_\phi(Y, h(X|A=a)),\end{aligned}\tag{3.10}$$

where we denote $\phi^{-1} = \psi^{-1} \circ T_a^\top$. Equivalently, $\phi = (T_a^{-1})^\top \circ \psi$ is invertible by composition of invertible functions, its domain is Δ^{c-1} as of ψ and its codomain is \mathbb{R}^c . The last loss in Equation 3.10 is proper composite with link ϕ . Finally, from
1165 Equation 3.9, the loss minimizer over the unfair distribution is

$$\arg \min_h \mathbb{E}_{X, \tilde{Y}} \ell_\phi(Y, h(X|A=a)) = \phi(P(\tilde{Y}|X, A=a))\tag{3.11}$$

$$= \psi((T_a^{-1})^\top) P(\tilde{Y}|X, A=a)\tag{3.12}$$

$$= \psi(P(Y|X, A=a)),\tag{3.13}$$

that proves the Theorem by Equation 3.9 once again. \square

Considering a common scenario with only two groups in sensitive attributes (protected and privileged), we can correct the model's predictions using two different fair transition matrices. One with rates applied while learning instances from the
1170 protected group, and the other with rates applied while learning instances from the privileged group. Formally, to the sensitive feature $A \in \{0, 1\}$, let T_0 the transition matrix associated with privileged/unprotected group ($A = 0$) and T_1 with the protected group ($A = 1$), FTL can be computed as

$$\text{FTL}(P(\tilde{Y}|X)) = (1 - A) \cdot \ell(T_0^\top P(\tilde{Y}|X)) + A \cdot \ell(T_1^\top P(\tilde{Y}|X)),\tag{3.14}$$

which in a standard batch learning, consists in alternating the transition matrix
1175 applied according instance's sensitive attribute.

Furthermore, to a common binary classification problem, where there is a positive (favorable) class and a negative (unfavorable) class, and two groups from sensitive feature (protected and privileged), we have two 2×2 transition matrices. Intuitively we are choosing rates to increase or decrease the probability of each group to be
1180 classified with the positive or negative prediction. We name those rates associated with increasing the probability to achieve the positive outcome as *promotion* rate, and those associated with increasing the probability to receive the negative outcome as *demotion* rate. As the transition matrix is row-stochastic, we can describe T_0 and T_1 as

$$T_0 = \begin{bmatrix} 1 - d_0 & d_0 \\ p_0 & 1 - p_0 \end{bmatrix}, \quad T_1 = \begin{bmatrix} 1 - d_1 & d_1 \\ p_1 & 1 - p_1 \end{bmatrix},\tag{3.15}$$

1185 where d_0 is the privileged demotion rate, p_0 the privileged promotion rate, the d_1 protected demotion rate, and p_1 the protected promotion rate. With an appropriate combination of d_0 , p_0 , d_1 , p_1 we can define a transition matrix pair that should be able to reweight model’s predictions with *FTL* to achieve fairer results with a reasonable model performance. The central problem in our methodology thus
 1190 relies in choosing these rates, which can be seen as an hyperparameter optimization problem.

Our hyperparameter optimization problem consists in finding an optimal trade-off between fairness and performance, which can be described as a MOO problem, as defined in Equation 3.4. Here, the hyperparameter configuration is $\lambda =$
 1195 (d_0, p_0, d_1, p_1) . Since the transition matrix is row stochastic these parameters are sufficient to define T_0 and T_1 . We want to maximize model performance $\rho(\lambda)$ and minimize fairness metric $\varphi(\lambda)$.

$$G(\lambda) = \rho(\lambda) - \varphi(\lambda). \quad (3.16)$$

Following some MOO approaches to fair machine learning, we will use a linear scalarization setup to define the optimization metric (PETROVIĆ *et al.*, 2021;
 1200 SCHMUCKER *et al.*, 2020). As we yet have four hyperparameter to fine-tune, and in F.CRUZ *et al.* (2021) the relative importance α is fixed at 0.5, we choose a simple and intuitive objective function in Equation 3.16 to maximize without the parameter α , i.e., giving same importance to fairness and performance. In Equation 3.16 we establish a simple objective to optimize, but one might need to consider a different
 1205 formulation depending on the specific problem at hand.

3.3 Experimental setup

In this section, we detail the experimental setup employed to benchmark our model against relevant in-processing fair classification models found in standard fairness toolkits, namely, Prejudice Remover (KAMISHIMA *et al.*, 2012), Adversarial De-
 1210 biasing (ZHANG *et al.*, 2018), and Gerry Fair Classifier (KEARNS *et al.*, 2018). We use the implementation of these methods from AI Fairness 360 toolkit (BELLAMY *et al.*, 2018). The baseline is a Standard MLP using two hidden layers with 100 hidden units each, *ReLU* activation function, batch size of 64, 50 epochs early stopped at 3 epochs without improvement (LI *et al.*, 2020) and softmax in out-
 1215 put, trained with ADAM optimizer (KINGMA e BA, 2015) with learning rate at $3e-4$. The only difference between baseline MLP and Fair Transition Loss MLP is that baseline uses standard Binary Cross Entropy Loss. The Gerry Fair Classifier implementation uses the False Negative Rate as its fairness definition and in Adver-

serial Debiasing classifier the hidden size is 100 units. Additionally, we compare the
1220 Fair Transition Loss within the Adaptive Priority Reweighting HU *et al.* (2023), a
promising fairness promoting technique focused on improving generalization, which
outperformed many recent methods such as JIANG e NACHUM (2020), MROUEH
et al. (2021), and ROH *et al.* (2021).

Our methodology consists of two phases: hyperparameter tuning and testing.
1225 In the hyperparameter tuning phase we perform a Bandit-Based pruning approach
using HyperBand (LI *et al.*, 2018) with Tree-structured Parzen Estimator Sampler
(TPE) (BERGSTRA *et al.*, 2011) over 100 trials. Those techniques achieves better
solutions to multi-objective hyperparameter optimization in the same number of tri-
als than conventional approaches like Grid Search and Random Search (MORALES-
1230 HERNÁNDEZ *et al.*, 2023). At each trial fitness function is evaluated by performing
a complete training and validation, where both model performance and fairness met-
rics are assessed. The fitness function is computed based on the objective defined in
Equation 3.16. This same experimental procedure can be adapted to utilize other
hyperparameter tuning algorithms such as FairRandom Search, Fair TPE, and Fair-
1235 band (F.CRUZ *et al.*, 2021).

Once the best hyperparameters are selected, we proceed to the testing phase,
where a new training is conducted using those optimal hyperparameters. After this
training, we evaluate the model’s performance on a separate test set that was not
used during the hyperparameter tuning phase, which are reported. This complete
1240 tuning-training-testing described is repeated 15 times with dataset re-sampling then
we proceed to comparison. Here the re-sampling consists in shuffling the whole
dataset before splitting, which is better described further in this section.

As the objective defined in Equation 3.16 can be achieved with different perfor-
mance and fairness metrics, we compare the proposed method with other relevant in-
1245 processing techniques from literature in different optimization scenarios. In addition
to Accuracy (Acc.) as performance metric, we also evaluate the Mathews Correla-
tion Coefficient (MCC), which has advantages over F1 score and Accuracy in binary
classification evaluation (CHICCO e JURMAN, 2020). To this performance metric,
1 means a perfect prediction according true class, -1 a complete inversion and 0 an
average random outcome. As fairness metric we consider Statistical Parity (Stat.
Parity, Definition 13), Equal Opportunity (Eq. Opp., Definition 14) and Equalized
Odds (Eq. Odds, Definition 16). Thus we have the following optimization scenar-
ios: MCC and Statistical Parity; MCC and Equal Opportunity; MCC and Equalized
Odds; Accuracy and Statistical Parity; Accuracy and Equal Opportunity; Accuracy
1250 and Equalized Odds.

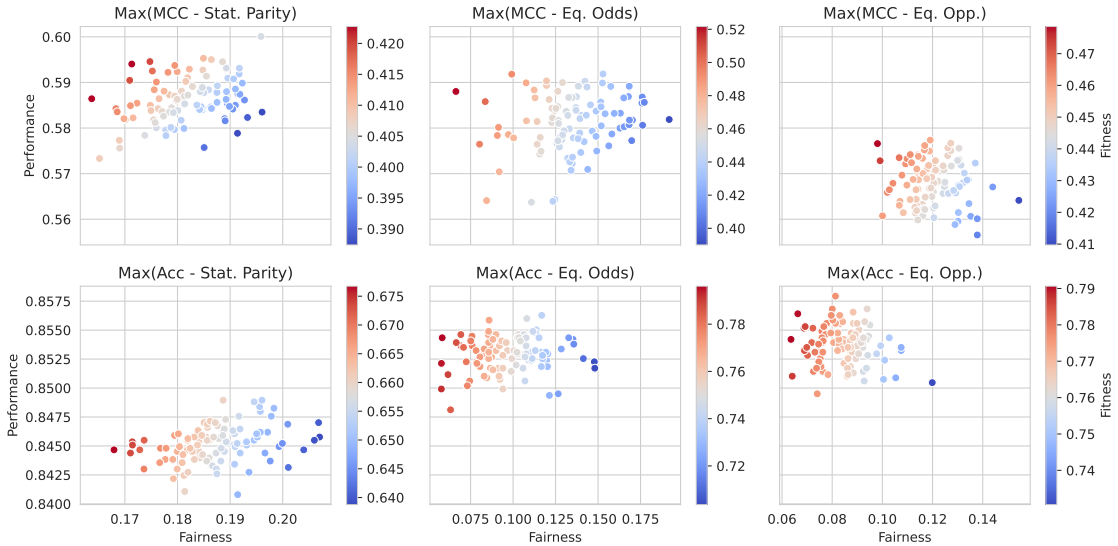
Table 3.1 presents the methods hyperparameters along with their corresponding
search ranges or options. While each method may possess a varying number of

Table 3.1: Hyperparameters search ranges or options of each method.

Method	Parameter	Range/options
Standard MLP (baseline)	dropout	[0.0, 0.2]
Prejudice Remover (KAMISHIMA <i>et al.</i> , 2012)	η	[0.0, 50.0]
Adversarial Debasing (ZHANG <i>et al.</i> , 2018)	α	[0.0, 1.0]
Gerry Fair Classifier (KEARNS <i>et al.</i> , 2018)	C	[0.0, 20.0]
	γ	{0.1, 0.01, 0.001}
Adaptive Priority Reweighting (HU <i>et al.</i> , 2023)	α	[0.0, 10000.0]
	η	[0.5, 3.0]
Fair Transition Loss	d_0, p_0, d_1, p_1	[0.0, 1.0]
	dropout	[0.0, 0.2]

hyperparameters and range sizes, all are optimized under the same conditions and number of configurations to guarantee a balanced comparison. In Figure 3.2 we present a brief sensibility analysis on the fitness functions with different fairness and performance values over those six optimization scenarios listed before. Here we perform a complete hyperparameter tuning with HyperBand and TPE over 100 trials using the baseline model (Standard MLP) with the Adult Income dataset optimizing the hyperparamers reported on Table 3.1. This sensibility analysis aims better present the linear objective function behavior within performance and fairness metrics evaluated in this study.

Figure 3.2: Sensibility analysis on optimized fitness functions within different performance and fairness metrics. Results from complete hyperparameter tuning through 100 trials with baseline model over the Adult Income dataset.



Each plot in Figure 3.2 illustrates the fitness value corresponding to specific performance and fairness metrics. The color scheme in the plots represents the fitness values, with higher values in red and lower values in blue. On the x -axis, we

1270 have the fairness metric, where a lower value is preferable. The y -axis represents
the performance metric, with higher values being preferred. The color gradient in
these plots demonstrates the linear relationship between fitness and variations in
the corresponding performance and fairness metrics. It is important to note that
the scale of the fairness metric is significantly smaller than that of the performance
1275 metric. However, it is sufficiently to act as a penalization. The general fitness
function in the scenarios described is capable of producing results with reduced bias
while maintaining similar performance levels. Although each metric combination
has different scales, each hyperparameter tuning experiment uses only one metric
combination at a time, ensuring consistency in the optimization process.

1280 On these plots we used the fitness and performance levels obtained through the
TPE sampler with HyperBand pruning during the hyperparameter tuning phase
using the baseline model, as previously described. In this setting, the solution (i.e.,
the combination of hyperparameters) that yields the best fitness value is selected
for a new complete training phase. This involves assessing metrics on test data not
1285 used in the previous phase. To ensure robust evaluation, the dataset is reshuffled,
re-split into train, validation and test segments, and this entire process is repeated
over 15 iterations.

To properly compare this set of 15 results of each method, we conduct an Almost
Stochastic Order (ASO) test (DROR *et al.*, 2019), which is a significance test suit-
1290 able for comparing complex machine learning models with various hyperparameters.
The ASO test involves evaluating a set of metrics through multiple samplings of a
Collection of Statistics (in this case assessed in test phase using random resampling)
to compare one method against another. The $ASO(A, B)$ function yields a value
in the range $[0, 1]$, given two methods A and B . If $ASO(A, B)$ is lesser than 0.5,
1295 we can reject the null hypothesis and conclude that method A outperforms method
 B in the given task. That is, method A produces stochastically larger values than
method B for a given metric. The lower the $ASO(A, B)$ value, the stronger the
evidence that A is superior to B in that particular task, which can be interpreted
as a confidence interval. Therefore, we perform pairwise comparisons between all
1300 methods for each optimization scenario outlined previously and for each dataset.

Our experiments uses common datasets used in Fair Classification problems,
namely *Adult Income* (BECKER e KOHAVI, 1996), *German Credit* (HOFMANN,
1994), *Bank Marketing* (S. MORO e CORTEZ, 2012), and *COMPAS Recidi-*
vism (JEFF LARSON e ANGWIN, 2016). We use the dataset readers available
1305 in the AI Fairness 360 toolkit (BELLAMY *et al.*, 2018) with its standard configu-
rations. Instances with missing data are removed.

Table 3.2 present dataset details used in this work, including the number of fea-
tures before pre-processing, the count of valid instances, the proportion of positive

Table 3.2: Dataset details used in this work, including performance and fairness metrics assessed to a standard classifier without tuning, and the maximum correlation between sensitive feature and the other features.

Dataset	Adult Income	Bank Marketing	COMPAS Recidivism	German Credit
# Features	102	57	401	58
# Instances	45222	30488	6167	1000
Sensitive Attribute	sex	age	race	sex
Positives	24.78%	12.66%	54.45%	70.00%
Negatives	75.22%	87.34%	45.55%	30.00%
Privileged	67.50%	97.17%	34.05%	69.00%
Unprivileged	32.50%	2.83%	65.95%	31.00%
Accuracy	0.846	0.906	0.358	0.685
MCC	0.572	0.553	-0.275	0.000
Stat. Parity.	0.192	0.106	0.172	0.074
Equal Opportunity	0.094	0.145	0.120	0.043
Equalized Odds	0.092	0.094	0.163	0.122
Maximum Correlation	0.527	0.364	0.826	0.593

and negative labels, the sensitive feature considered in experiments, the proportion
of privileged and unprivileged groups within the corresponding sensitive feature,
reference performance and fairness metrics using a standard Random Forest Clas-
sifier with 1000 classifiers without tuning, and the maximum correlation coefficient
between the sensitive feature and the other features. The maximum correlation is
useful to assess whether it is possible to use another feature as proxy to the sensitive
feature, which is commonly referred as redlining effect (PEDRESCHI *et al.*, 2008).

The *Adult Income* dataset presents a classification task to predict whether an
individual earns more than 50,000 per year. This dataset consists of 48,842 instances
sourced from the U.S. 1994 Census database. The sensitive attribute used in this
task is sex, with the male group considered privileged and the female group protected
(unprivileged). In the *German Credit* dataset, the task consists of classifying 1,000
individuals described by a set of attributes as good or bad credit risks. Similar to
the *Adult Income* dataset, here we use sex as the sensitive attribute, with the male
group considered privileged and the female group protected. The *Bank Marketing*
classification task aims to predict whether 45,211 clients will subscribe to a term
deposit after direct marketing campaigns (phone calls) by a Portuguese banking
institution. In this case, the sensitive feature is age, where individuals under the
age of 25 are considered unprivileged, while those aged 25 and older are considered
privileged. The *COMPAS* dataset presents around 80,000 criminal records from
the Broward County Clerk’s Office. The task here is to predict whether a defendant
will recidivate in the next two years. The sensitive feature in this case is race, with
Caucasians as the privileged group and non-Caucasians (Black and Hispanic) as
unprivileged.

For all datasets, the data preparation process is the same, one-hot encoding for categorical features and standardize the numerical features. We perform a random split, with 80% allocated for the hyperparameter tuning phase and the remaining 20% reserved for evaluating metrics in the test phase. Within the hyperparameter tuning phase, this corresponding fraction of data is further randomly split, with 80% assigned to training and 20% to validation. The validation set allows us to assess metrics and compute the fitness function for each hyperparameter configuration. In datasets where there is originally some kind of split (e.g., train set and test set in separate files), all available data is merged and then shuffled to produce new splits at each run.

3.4 Results and discussion

This section summarizes our results, comparing Fair Transition Loss (FTL) with the baseline Standard MLP (MLP) and some relevant fair in-processing methods: Adversarial Debiasing (AD), Prejudice Remover (PR), Gerry Fair Classifier (GFC) and Adaptive Priority Reweighting (APW).

As we have multiple optimization scenarios with different objective functions and datasets, and to each of them multiple runs, we present in Table 3.3 the results of the ASO test described before, which allow us to properly compare each method to FTL. Values under 0.5 (in bold) mean that we can reject the null hypothesis, i.e., FTL produces stochastically larger fitness than method in respective column for a objective and dataset. Lower values indicate stronger evidence. The complete results with mean and standard deviation of fitness, performance and fairness can be found in tables 3.5 to 3.10 to a fairness-performance trade-off analysis.

In 69 of 120 comparison scenarios from Almost Stochastic Order test (Table 3.3), it is possible to claim that FTL outperforms its competitor, i.e., FTL produces stochastically higher fitness values. Despite these positive results, one can argue that the proposed technique only adds extra hyperparameters that increase models flexibility to achieve higher fitness values. In other words, are we effectively describing bias in datasets by transition matrices as claimed before? To address this, we showcase various FTL hyperparameter combinations selected during the tuning phase described in Section 3.3, comparing with corresponding dataset information available at Table 3.2. We perform this analysis using *Adult Income* dataset. The corresponding hyperparameters can be found in Table 3.4. Here, high values mean that FTL alters the corresponding probabilities, while values close to zero indicate minimal interference by the method.

When optimizing for MCC and Statistical Parity, there’s a notable high value for protected promotion. This value is compatible with the high corresponding fair-

Table 3.3: Almost Stochastic Order test comparing Fair Transition Loss fitness. Values under 0.5 (in bold) mean that FTL outperforms corresponding method in such optimization scenario.

Fairness/Performance Metric	Dataset	MLP	AD	PR	GFC	APW
Statistical Parity MCC	Adult Income	0.00	0.15	1.00	0.00	1.00
	Bank Marketing	0.01	0.00	0.00	0.00	0.00
	Compas Recidivism	0.01	0.25	0.00	0.02	0.00
	German Credit	0.28	0.30	0.39	0.21	0.28
Equal Opportunity MCC	Adult Income	0.01	0.00	0.05	0.00	0.93
	Bank Marketing	0.81	0.18	0.24	0.09	0.77
	Compas Recidivism	0.00	1.00	0.00	0.66	1.00
	German Credit	1.00	0.23	0.84	0.78	0.76
Equalized Odds MCC	Adult Income	0.03	0.28	0.42	0.00	0.00
	Bank Marketing	0.46	0.18	0.12	0.02	0.18
	Compas Recidivism	0.01	0.58	0.00	0.07	0.00
	German Credit	1.00	0.07	1.00	0.31	1.00
Statistical Parity Accuracy	Adult Income	0.01	0.26	0.32	0.00	0.53
	Bank Marketing	0.25	1.00	1.00	0.76	0.82
	Compas Recidivism	0.00	1.00	0.10	1.00	0.00
	German Credit	1.00	0.26	1.00	1.00	1.00
Equal Opportunity Accuracy	Adult Income	0.89	0.97	1.00	0.23	0.98
	Bank Marketing	1.00	0.39	0.81	1.00	1.00
	Compas Recidivism	0.01	0.78	0.00	0.10	1.00
	German Credit	1.00	0.64	1.00	1.00	1.00
Equalized Odds Accuracy	Adult Income	0.01	0.21	0.19	0.00	1.00
	Bank Marketing	0.76	0.40	0.82	1.00	1.00
	Compas Recidivism	0.01	0.45	0.00	0.01	0.00
	German Credit	1.00	0.12	1.00	1.00	1.00

ness metric for this dataset, approximately 0.19 without correction. This increases the likelihood that an unprivileged instance receives a favorable outcome. Since statistical parity only compare the probability of a positive outcome across groups (ignoring true class) this is enough. The other hyperparameters presents low values. In contrast to the previous case, optimizing Equal Opportunity requires compatible false negative rates. Optimizing this fairness metric within MCC produces the effect of promoting both privileged and protected, although protected with higher values. This produces the effect of reducing false negatives at all, since the method enhances the probability of a positive outcome. This effect is counterbalanced with intermediate demotion rates to both groups through a finetuning to keep MCC. Note that Equal Opportunity values without correction to this dataset is not as high as Statistical Parity. To optimize Equalized Odds within MCC it is necessary to keep both false negatives and false positives comparable across groups, which lead to a less intense intervention when compared to Equal Opportunity. Here remains the high values to protected promotion to achieve fairness.

Table 3.4: Fair Transition Loss hyperparameters chosen by optimizing different metrics in *Adult Income* dataset.

Objective	d_0	p_0	d_1	p_1
	Priv. Dem.	Priv. Prom.	Prot. Dem.	Prot. Prom.
MCC and Stat. Parity	0.056	0.076	0.043	0.878
MCC and Eq. Opp.	0.292	0.455	0.329	0.575
MCC and Eq. Odds	0.037	0.165	0.005	0.432
Acc. and Stat. Parity	0.470	0.110	0.023	0.446
Acc. and Eq. Opp.	0.389	0.326	0.311	0.530
Acc. and Eq. Odds	0.497	0.286	0.228	0.094

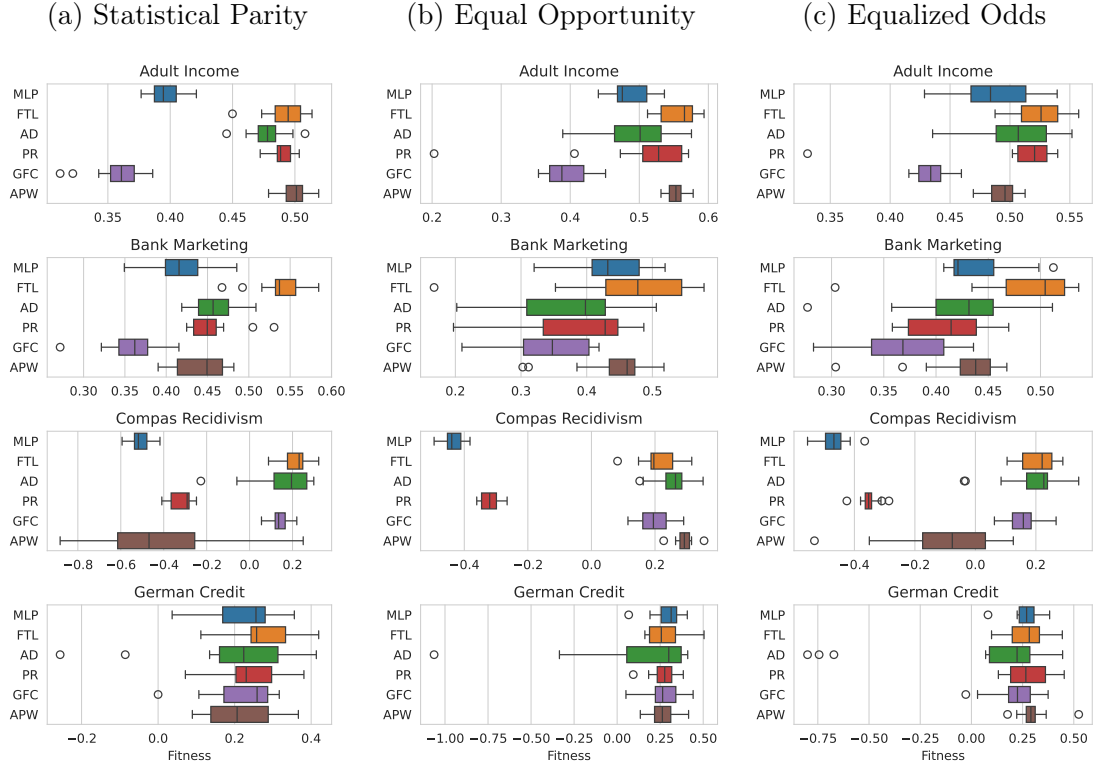
1385 There is a remarkable difference between hyperparameters found through op-
 timizing Accuracy and MCC. While MCC handles unbalanced classes effectively,
 Accuracy only measures the probability of correctly predicting an instance. If the
 dataset is unbalanced it is possible to achieve high Accuracy only by predicting the
 label of the more frequent class. In this dataset, only about a quarter of the instances
 1390 are positives, which can lead to more frequent negative outcomes to achieve higher
 Accuracy. Results optimizing for Accuracy show significantly higher demotion rates
 compared to those from MCC optimization, both to privileged and protected groups.
 From this analysis, it’s evident that the proposed methodology effectively describes
 and mitigates bias in a dataset according to a given fairness definition and while
 1395 keeping targeted performance metric at a reasonable level.

Now we discuss the results to each objective, starting with MCC and Statistical
 Parity. To this objective Fair Transition Loss consistently outperforms all methods
 at all classification tasks, except Prejudice Remover (PR) and Adaptive Priority
 Reweighting (APW) in *Adult Income*. Figure 3.3a presents a box-plot comparison,
 1400 where we can see that FTL, PR and APW are effectively drawn. While FTL and
 APW has little bit higher values, PR presents smaller variance. AS PR is a regular-
 ized logistic regression, it is a smaller model than FTL, which can explain the also
 smaller variance.

When comparing the optimization results for MCC and Equal Opportunity, FTL
 1405 consistently outperforms its counterparts in most scenarios. Here we have only one
 discrepancy, as APW presents slightly advantage over FTL on COMPAS dataset.
 Also, on German Credit dataset all methods achieved similar results. Given the
 small size of this dataset, we theorize that all methods, barring AD, have reached
 the Pareto front — meaning, any further improvements in fairness would necessitate
 1410 a proportional sacrifice in performance. This equilibrium between baseline (MLP),
 FTL, PR, GFC and APW is evident in Figure 3.3b.

The sub-optimal results of AD can likely be attributed to the dataset’s lim-
 ited size; with merely 1000 instances, this dataset might be too small for the an
 adversarial model like AD train effectively. This pattern of AD underperforming

Figure 3.3: Fitness values optimizing MCC and multiple fairness metrics.



1415 persists across subsequent classification tasks involving this dataset. Interestingly,
also with the exception of AD, we notice that the variance in results for most op-
timization scenarios is smaller than in other classification tasks. This observation
further underscores our Pareto front hypothesis and suggests that the classification
task’s simplicity may contribute to the reduced variability in outcomes.

1420 When optimizing for MCC and Equalized Odds, we find that the results are
consistent, with FTL outperforming its counterparts in most scenarios. Notably,
within the *German Credit* dataset, FTL surpasses not just AD but also GFC. Since
GFC primarily relies on the False Negative Rate for its fairness definition, it has a
natural advantage when optimizing for Equal Opportunity compared to Equalized
1425 Odds, which requires maintaining equitable False Positive and False Negative Rates.
As observed in the previous comparisons, Figure 3.3c underscores that the baseline,
FTL, PR and APW seem to hit the Pareto front for this dataset.

Upon examining FTL’s results when optimizing for MCC across all the fair-
ness metrics evaluated, it’s evident that FTL consistently superior results compared
1430 to its counterparts. Specifically, FTL achieves stochastically higher fitness values
in 44 out of the 60 scenarios evaluated. Given the inherent challenges associated
with optimizing MCC compared to Accuracy, we attribute FTL’s dominance in the
MCC optimization to its capability of effectively capturing the bias idiosyncrasies
of the dataset and the specified performance and fairness metrics trough transition

1435 matrices.

Furthermore, it’s noteworthy that both the baseline and PR models substantially underperform across all optimization scenarios in the *COMPAS Recidivism* classification task. This dataset, characterized by its complexity with 401 features, might be at the heart of these subpar results. We theorize that the this lack of
1440 performance could be due to an insufficiently large model to navigate such a high-dimensional space, especially when we observe, as indicated in Table 3.2, that the standard performance on this dataset is relatively low.

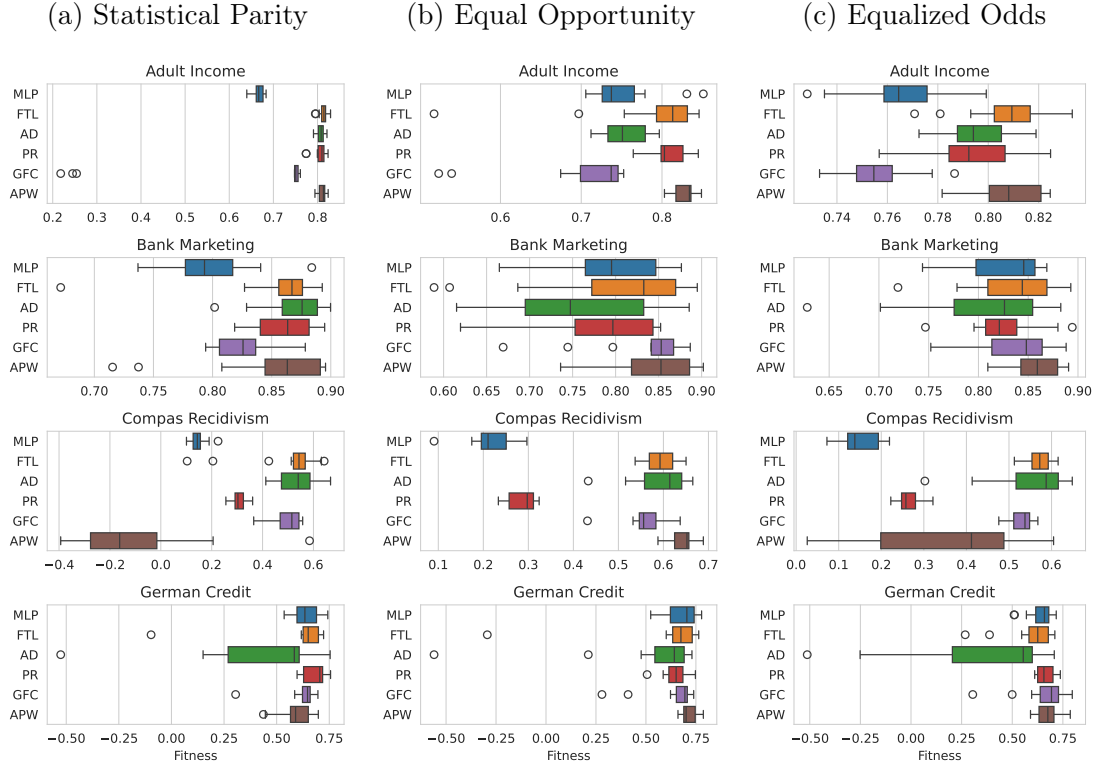
When turning our attention to results obtained by optimizing Accuracy, we must first reiterate its inherent simplicity as a performance metric compared to MCC.
1445 Given its nature, it allows models to attain high values simply by predicting the label of the predominant class. In such circumstances, it is comparatively easier to reach the Pareto front. Even under these conditions, FTL displays commendable competitiveness. Although it achieves stochastically higher fitness values in 25 of the 60 scenarios, this rate is notably less than what we observed when optimizing
1450 for MCC. By juxtaposing Figures 3.4a, 3.4b, and 3.4c with Table 3.3, we discern that in scenarios where FTL does not have the upper hand, it still competes closely with its counterparts. This very close results are primarily attributed to multiple methods simultaneously approaching the Pareto front. Likely when optimizing for MCC using Equal Opportunity as fairness metric, APW presents slightly advantage
1455 over FTL on COMPAS dataset.

A particularly straightforward fair classification task emerges when optimizing for Accuracy and Statistical Parity within the *Adult Income* dataset. With this pronounced class imbalance, models can lean into over-predicting the majority class, thereby aligning the probabilities of positive predictions across the groups. This
1460 strategy results in a exceptionally low variance across all methodologies. A similar, albeit reduced, effect can be observed within the *German Credit* dataset, as previously highlighted.

Fair Transition Loss consistently demonstrates effective bias mitigation, it does so by absorbing the nuances of the dataset and the fairness metric through its
1465 transition matrices, resulting in stochastically superior fitness values in a significant number of scenarios. Additionally, the method has the capacity to effectively handle datasets with unbalanced classes when optimizing for metrics like MCC. However, it’s important to recognize that Fair Transition Loss requires fine-tuning multiple hyperparameters. We thus consider that this technique is especially beneficial in
1470 setups where hyperparameter optimization is an inherent part of the prediction pipeline.

A key concern is about the potential for fairness-promoting techniques to inadvertently shift the burden onto the very group they aim to protect. This arises from

Figure 3.4: Fitness values optimizing Accuracy and multiple fairness metrics.

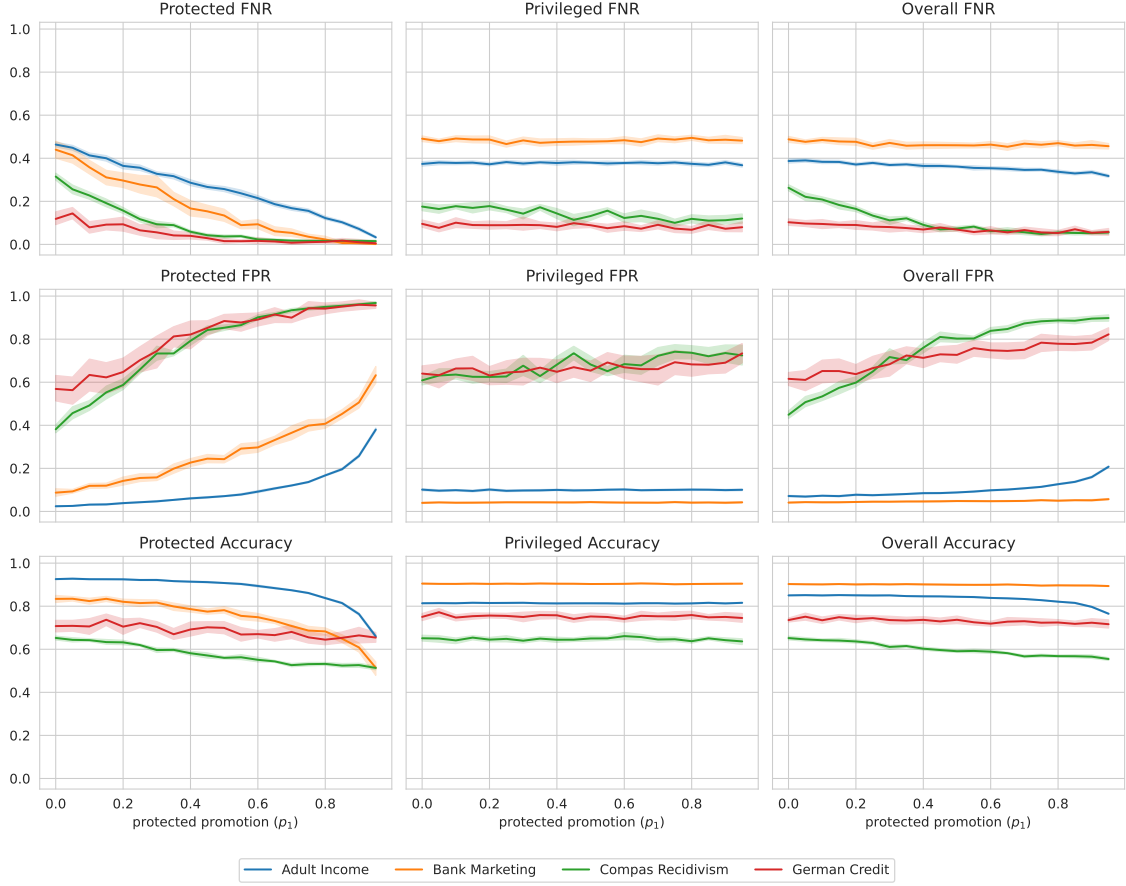


the possibility that by imposing additional constraints, the method might unintentionally learn alternative ways to reproduce and even reinforce the negative social biases present in the data, thus harming the individuals it intends to safeguard.

To evaluate the capability of FTL to address this risk, we present another experiment, where we adjusted only the protected promotion hyperparameter (p_1 , Equation 3.15) during the FTL training, keeping all other FTL hyperparameters at zero and dropout at 0.2. Here we follow the same re-sampling procedure, each experiment is performed over 15 repetitions, shuffling the dataset before splitting. This experiment was conducted using the four datasets previously analyzed, and we reported the following metrics assessed over the training set: protected false negative rate, protected false positive rate, protected accuracy, privileged false negative rate, privileged false positive rate, privileged accuracy, overall false negative rate, overall false positive rate and overall accuracy.

The results, presented in Figure 3.5, show that increasing the protected promotion hyperparameter value leads to a decreased false negative rate for the protected group. Meanwhile, most other monitored metrics tended towards stability to reasonable hyperparameter levels (p_1 under 0.95). An exception is the false positive rate, specially to protected group, which increased as the false negative rate decreased to keep accuracy. This pattern was consistent across all evaluated datasets. The primary aim of this experiment was to demonstrate that our proposed technique

Figure 3.5: Results of false negatives and false positives within groups on protected promotion (p_1) parameter at increasing levels.



1495 does not inadvertently penalize the protected group. Rather, the overall impact of increasing the aforementioned parameter is to effectively promote fairness for the protected group without detriment to either the privileged or protected groups. The behavior of the remaining FTL parameters is analogous, necessitating proper fine-tuning to achieve a balanced outcome. This underscores the efficacy of our method in achieving its intended purpose of reducing bias and promoting fairness in the model.

1500 Additionally, we present performance and fairness results along with the fitness, in order to provide material to a trade-off analysis. Also, those metric values enable the reader to compare this results with different experimental setup and fitness objective from literature. Results corresponding each optimization scenario can be found on tables 3.5 to 3.10, presenting metric means and standard deviation values across multiple resample run. Best result of each metric within evaluation scenario are in bold, and standard deviation values are presented between parenthesis. Up arrow (\uparrow) indicates that the referred metric should be maximized while down arrow (\downarrow) that the metric should be minimized. To provide a visual resource to this comparison, the distribution of those metrics across multiple resample runs comparing

the propose method (FTL) with baseline (MLP) and current state-of-art (APW) are presented in joint plot format in figures 3.6 to 3.11. Here each joint plot is present within the corresponding table to a better comprehension.

Table 3.5: Mean and standard deviation metric values optimizing MCC and Statistical Parity in comparison with Fair Transition Loss across multiple resample runs.

Dataset	Method	↑ Fitness	↑ MCC	↓ Stat. Parity
Adult Income	Adaptive Priority Reweighting	0.499 (± 0.01)	0.510 (± 0.01)	0.011 (± 0.01)
	Adversarial Debiasing	0.478 (± 0.01)	0.501 (± 0.02)	0.024 (± 0.02)
	Fair Transition Loss	0.492 (± 0.02)	0.512 (± 0.01)	0.020 (± 0.01)
	Gerry Fair Classifier	0.357 (± 0.02)	0.512 (± 0.02)	0.154 (± 0.03)
	Prejudice Remover	0.491 (± 0.01)	0.500 (± 0.01)	0.009 (± 0.01)
	Standard MLP (baseline)	0.395 (± 0.01)	0.581 (± 0.01)	0.185 (± 0.01)
Bank Marketing	Adaptive Priority Reweighting	0.441 (± 0.03)	0.482 (± 0.02)	0.041 (± 0.04)
	Adversarial Debiasing	0.459 (± 0.03)	0.505 (± 0.02)	0.046 (± 0.02)
	Fair Transition Loss	0.539 (± 0.03)	0.579 (± 0.01)	0.040 (± 0.03)
	Gerry Fair Classifier	0.358 (± 0.04)	0.428 (± 0.02)	0.070 (± 0.03)
	Prejudice Remover	0.454 (± 0.03)	0.487 (± 0.02)	0.033 (± 0.02)
	Standard MLP (baseline)	0.419 (± 0.04)	0.522 (± 0.02)	0.102 (± 0.03)
COMPAS Recidivism	Adaptive Priority Reweighting	-0.412 (± 0.35)	0.194 (± 0.07)	0.606 (± 0.29)
	Adversarial Debiasing	0.157 (± 0.14)	0.322 (± 0.02)	0.165 (± 0.14)
	Fair Transition Loss	0.220 (± 0.06)	0.276 (± 0.03)	0.057 (± 0.05)
	Gerry Fair Classifier	0.141 (± 0.04)	0.289 (± 0.06)	0.148 (± 0.06)
	Prejudice Remover	-0.318 (± 0.05)	-0.276 (± 0.03)	0.042 (± 0.03)
	Standard MLP (baseline)	-0.511 (± 0.05)	-0.299 (± 0.03)	0.212 (± 0.04)
German Credit	Adaptive Priority Reweighting	0.217 (± 0.09)	0.321 (± 0.05)	0.105 (± 0.06)
	Adversarial Debiasing	0.200 (± 0.17)	0.368 (± 0.06)	0.168 (± 0.15)
	Fair Transition Loss	0.272 (± 0.08)	0.354 (± 0.07)	0.083 (± 0.04)
	Gerry Fair Classifier	0.221 (± 0.09)	0.291 (± 0.11)	0.071 (± 0.06)
	Prejudice Remover	0.234 (± 0.09)	0.329 (± 0.05)	0.095 (± 0.06)
	Standard MLP (baseline)	0.223 (± 0.10)	0.330 (± 0.07)	0.107 (± 0.07)

Figure 3.6: Metric distribution optimizing MCC and Statistical Parity in comparison with Fair Transition Loss across multiple resample runs. Corresponding values available at Table 3.5.

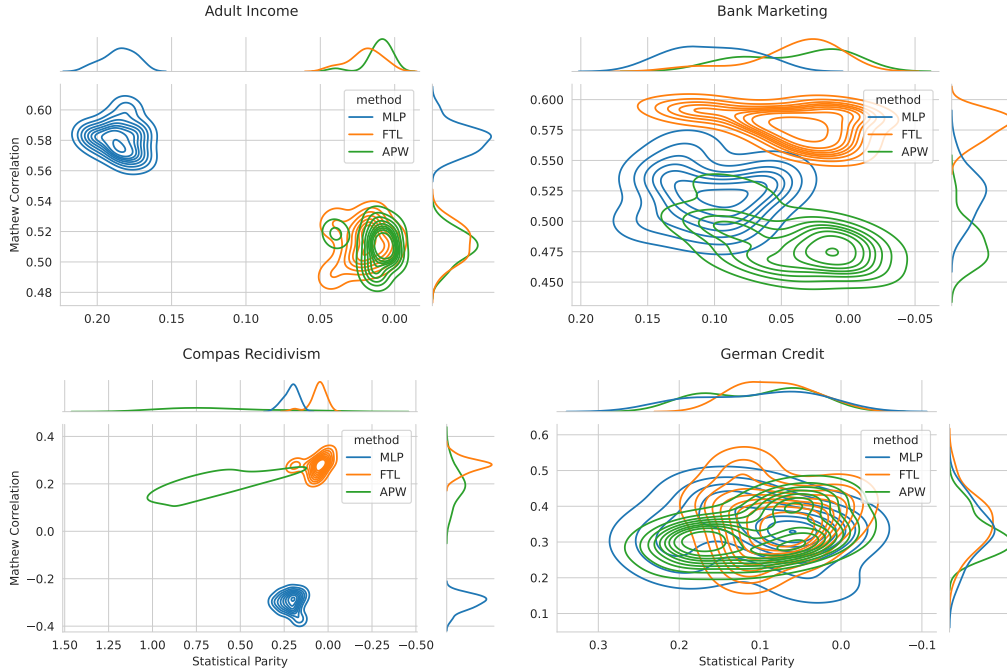


Table 3.6: Mean and standard deviation metric values optimizing MCC and Equal Opportunity in comparison with Fair Transition Loss across multiple resample runs.

Dataset	Method	\uparrow Fitness	\uparrow MCC	\downarrow q. Opp.
Adult Income	Adaptive Priority Reweighting	0.493 (± 0.01)	0.523 (± 0.01)	0.030 (± 0.01)
	Adversarial Debiasing	0.509 (± 0.03)	0.565 (± 0.02)	0.056 (± 0.02)
	Fair Transition Loss	0.523 (± 0.02)	0.576 (± 0.02)	0.052 (± 0.02)
	Gerry Fair Classifier	0.434 (± 0.01)	0.523 (± 0.01)	0.089 (± 0.01)
	Prejudice Remover	0.509 (± 0.05)	0.558 (± 0.02)	0.049 (± 0.03)
	Standard MLP (baseline)	0.489 (± 0.03)	0.576 (± 0.01)	0.087 (± 0.03)
Bank Marketing	Adaptive Priority Reweighting	0.424 (± 0.04)	0.474 (± 0.02)	0.050 (± 0.04)
	Adversarial Debiasing	0.426 (± 0.06)	0.512 (± 0.02)	0.086 (± 0.05)
	Fair Transition Loss	0.485 (± 0.06)	0.569 (± 0.01)	0.084 (± 0.06)
	Gerry Fair Classifier	0.371 (± 0.04)	0.423 (± 0.02)	0.052 (± 0.03)
	Prejudice Remover	0.413 (± 0.04)	0.485 (± 0.02)	0.072 (± 0.04)
	Standard MLP (baseline)	0.439 (± 0.03)	0.514 (± 0.02)	0.075 (± 0.03)
COMPAS Recidivism	Adaptive Priority Reweighting	-0.111 (± 0.18)	0.260 (± 0.04)	0.371 (± 0.17)
	Adversarial Debiasing	0.191 (± 0.11)	0.324 (± 0.03)	0.133 (± 0.10)
	Fair Transition Loss	0.208 (± 0.06)	0.283 (± 0.02)	0.074 (± 0.05)
	Gerry Fair Classifier	0.155 (± 0.05)	0.274 (± 0.06)	0.120 (± 0.04)
	Prejudice Remover	-0.352 (± 0.03)	-0.278 (± 0.02)	0.073 (± 0.03)
	Standard MLP (baseline)	-0.471 (± 0.05)	-0.294 (± 0.02)	0.176 (± 0.04)
German Credit	Adaptive Priority Reweighting	0.299 (± 0.08)	0.373 (± 0.06)	0.075 (± 0.05)
	Adversarial Debiasing	0.040 (± 0.41)	0.301 (± 0.13)	0.261 (± 0.30)
	Fair Transition Loss	0.273 (± 0.10)	0.386 (± 0.08)	0.113 (± 0.08)
	Gerry Fair Classifier	0.218 (± 0.11)	0.321 (± 0.10)	0.103 (± 0.05)
	Prejudice Remover	0.283 (± 0.10)	0.391 (± 0.07)	0.107 (± 0.06)
	Standard MLP (baseline)	0.270 (± 0.07)	0.352 (± 0.05)	0.082 (± 0.04)

Figure 3.7: Metric distribution optimizing MCC and Equal Opportunity in comparison with Fair Transition Loss across multiple resample runs. Corresponding values available at Table 3.6.

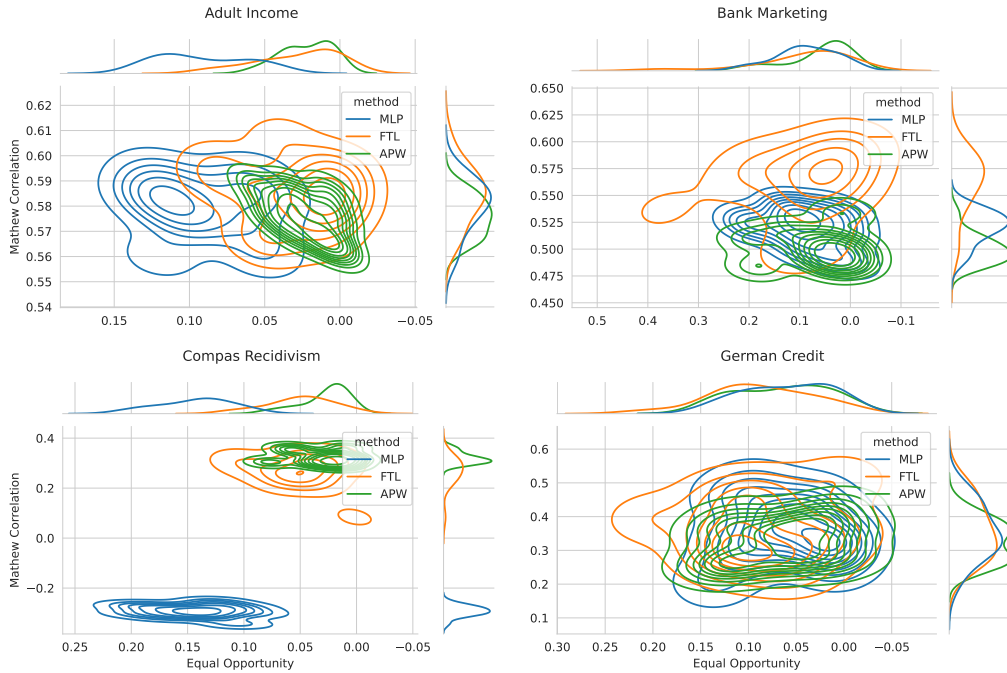


Table 3.7: Mean and standard deviation metric values optimizing MCC and Equalized Odds in comparison with Fair Transition Loss across multiple resample runs.

Dataset	Method	↑ Fitness	↑ MCC	↓ Eq. Odds
Adult Income	Adaptive Priority Reweighting	0.553 (± 0.01)	0.576 (± 0.01)	0.022 (± 0.02)
	Adversarial Debiasing	0.493 (± 0.05)	0.573 (± 0.01)	0.080 (± 0.05)
	Fair Transition Loss	0.556 (± 0.03)	0.584 (± 0.01)	0.029 (± 0.03)
	Gerry Fair Classifier	0.394 (± 0.03)	0.515 (± 0.02)	0.121 (± 0.02)
	Prejudice Remover	0.505 (± 0.09)	0.560 (± 0.02)	0.055 (± 0.08)
	Standard MLP (baseline)	0.489 (± 0.03)	0.580 (± 0.01)	0.091 (± 0.03)
Bank Marketing	Adaptive Priority Reweighting	0.441 (± 0.06)	0.500 (± 0.01)	0.059 (± 0.06)
	Adversarial Debiasing	0.373 (± 0.09)	0.508 (± 0.02)	0.136 (± 0.09)
	Fair Transition Loss	0.467 (± 0.11)	0.560 (± 0.03)	0.093 (± 0.10)
	Gerry Fair Classifier	0.344 (± 0.07)	0.422 (± 0.02)	0.078 (± 0.06)
	Prejudice Remover	0.392 (± 0.09)	0.490 (± 0.02)	0.098 (± 0.08)
	Standard MLP (baseline)	0.432 (± 0.06)	0.520 (± 0.02)	0.087 (± 0.06)
COMPAS Recidivism	Adaptive Priority Reweighting	0.292 (± 0.03)	0.319 (± 0.02)	0.027 (± 0.02)
	Adversarial Debiasing	0.258 (± 0.05)	0.329 (± 0.03)	0.070 (± 0.05)
	Fair Transition Loss	0.213 (± 0.06)	0.264 (± 0.06)	0.050 (± 0.03)
	Gerry Fair Classifier	0.201 (± 0.05)	0.290 (± 0.04)	0.089 (± 0.05)
	Prejudice Remover	-0.319 (± 0.03)	-0.289 (± 0.03)	0.030 (± 0.02)
	Standard MLP (baseline)	-0.435 (± 0.03)	-0.292 (± 0.02)	0.143 (± 0.03)
German Credit	Adaptive Priority Reweighting	0.261 (± 0.08)	0.326 (± 0.06)	0.065 (± 0.05)
	Adversarial Debiasing	0.116 (± 0.40)	0.311 (± 0.14)	0.195 (± 0.28)
	Fair Transition Loss	0.274 (± 0.10)	0.361 (± 0.08)	0.087 (± 0.05)
	Gerry Fair Classifier	0.273 (± 0.10)	0.361 (± 0.06)	0.087 (± 0.06)
	Prejudice Remover	0.271 (± 0.07)	0.324 (± 0.06)	0.054 (± 0.04)
	Standard MLP (baseline)	0.295 (± 0.09)	0.354 (± 0.08)	0.060 (± 0.04)

Figure 3.8: Metric distribution optimizing MCC and Equalized Odds in comparison with Fair Transition Loss across multiple resample runs. Corresponding values available at Table 3.7.

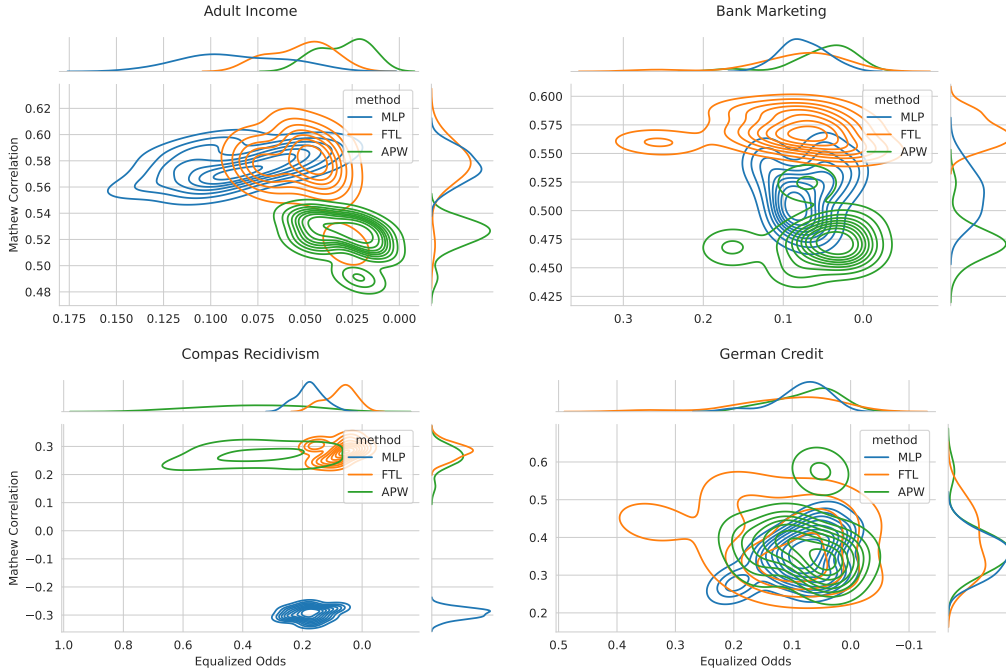


Table 3.8: Mean and standard deviation metric values optimizing Accuracy and Statistical Parity in comparison with Fair Transition Loss across multiple resample runs.

Dataset	Method	↑ Fitness	↑ Accuracy	↓ Stat. Parity
Adult Income	Adaptive Priority Reweighting	0.811 (± 0.01)	0.822 (± 0.01)	0.011 (± 0.01)
	Adversarial Debiasing	0.808 (± 0.01)	0.830 (± 0.01)	0.022 (± 0.01)
	Fair Transition Loss	0.814 (± 0.01)	0.828 (± 0.01)	0.014 (± 0.01)
	Gerry Fair Classifier	0.651 (± 0.21)	0.721 (± 0.07)	0.070 (± 0.14)
	Prejudice Remover	0.807 (± 0.01)	0.825 (± 0.00)	0.018 (± 0.01)
	Standard MLP (baseline)	0.666 (± 0.01)	0.851 (± 0.00)	0.184 (± 0.01)
Bank Marketing	Adaptive Priority Reweighting	0.851 (± 0.06)	0.900 (± 0.00)	0.049 (± 0.06)
	Adversarial Debiasing	0.869 (± 0.03)	0.901 (± 0.00)	0.031 (± 0.02)
	Fair Transition Loss	0.854 (± 0.05)	0.889 (± 0.01)	0.035 (± 0.05)
	Gerry Fair Classifier	0.824 (± 0.02)	0.895 (± 0.00)	0.071 (± 0.02)
	Prejudice Remover	0.860 (± 0.02)	0.898 (± 0.00)	0.038 (± 0.02)
	Standard MLP (baseline)	0.799 (± 0.04)	0.902 (± 0.00)	0.103 (± 0.03)
COMPAS Recidivism	Adaptive Priority Reweighting	-0.105 (± 0.26)	0.584 (± 0.03)	0.689 (± 0.23)
	Adversarial Debiasing	0.538 (± 0.07)	0.670 (± 0.02)	0.132 (± 0.08)
	Fair Transition Loss	0.501 (± 0.15)	0.600 (± 0.05)	0.099 (± 0.14)
	Gerry Fair Classifier	0.501 (± 0.05)	0.614 (± 0.05)	0.113 (± 0.07)
	Prejudice Remover	0.308 (± 0.03)	0.359 (± 0.01)	0.052 (± 0.02)
	Standard MLP (baseline)	0.146 (± 0.03)	0.354 (± 0.02)	0.208 (± 0.02)
German Credit	Adaptive Priority Reweighting	0.589 (± 0.08)	0.682 (± 0.03)	0.093 (± 0.08)
	Adversarial Debiasing	0.430 (± 0.33)	0.713 (± 0.09)	0.283 (± 0.26)
	Fair Transition Loss	0.616 (± 0.20)	0.715 (± 0.06)	0.098 (± 0.17)
	Gerry Fair Classifier	0.621 (± 0.09)	0.712 (± 0.12)	0.090 (± 0.04)
	Prejudice Remover	0.684 (± 0.05)	0.757 (± 0.02)	0.073 (± 0.06)
	Standard MLP (baseline)	0.639 (± 0.06)	0.752 (± 0.02)	0.113 (± 0.06)

Figure 3.9: Metric distribution optimizing Accuracy and Statistical Parity in comparison with Fair Transition Loss across multiple resample runs. Corresponding values available at Table 3.8.

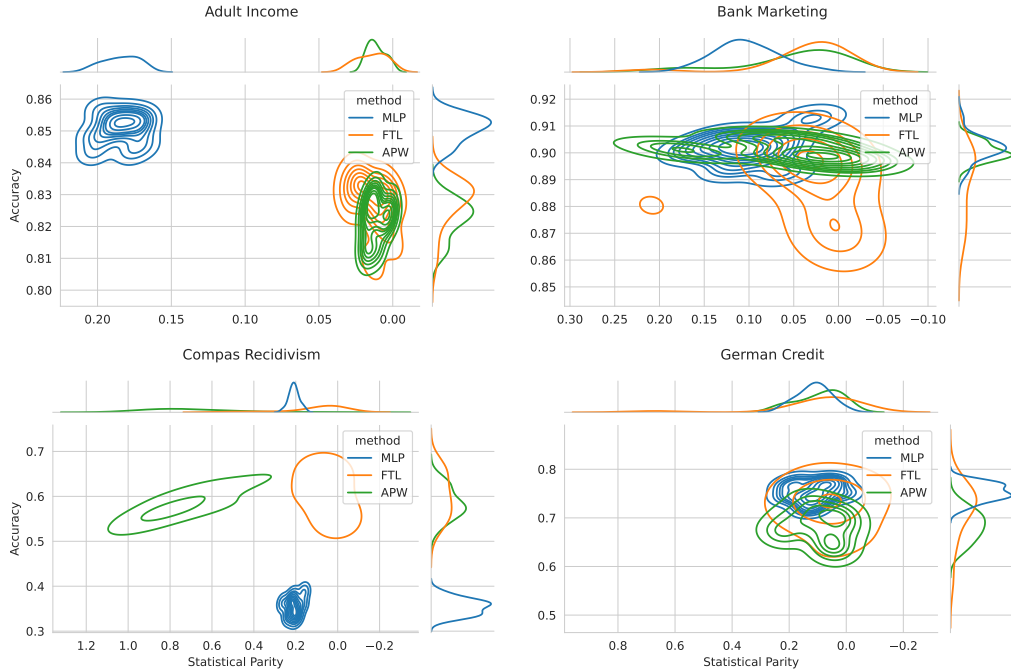


Table 3.9: Mean and standard deviation metric values optimizing Accuracy and Equal Opportunity in comparison with Fair Transition Loss.

Dataset	Method	↑ Fitness	↑ Accuracy	↓ Eq. Opp.
Adult Income	Adaptive Priority Reweighting	0.808 (± 0.01)	0.837 (± 0.00)	0.029 (± 0.01)
	Adversarial Debiasing	0.796 (± 0.01)	0.849 (± 0.00)	0.052 (± 0.01)
	Fair Transition Loss	0.808 (± 0.02)	0.842 (± 0.01)	0.034 (± 0.02)
	Gerry Fair Classifier	0.756 (± 0.01)	0.788 (± 0.03)	0.032 (± 0.04)
	Prejudice Remover	0.794 (± 0.02)	0.845 (± 0.01)	0.051 (± 0.01)
	Standard MLP (baseline)	0.765 (± 0.02)	0.850 (± 0.00)	0.084 (± 0.02)
Bank Marketing	Adaptive Priority Reweighting	0.858 (± 0.02)	0.897 (± 0.00)	0.039 (± 0.03)
	Adversarial Debiasing	0.807 (± 0.07)	0.902 (± 0.00)	0.095 (± 0.07)
	Fair Transition Loss	0.833 (± 0.05)	0.892 (± 0.01)	0.059 (± 0.05)
	Gerry Fair Classifier	0.837 (± 0.04)	0.895 (± 0.00)	0.058 (± 0.04)
	Prejudice Remover	0.827 (± 0.04)	0.898 (± 0.00)	0.071 (± 0.04)
	Standard MLP (baseline)	0.826 (± 0.04)	0.901 (± 0.00)	0.075 (± 0.04)
COMPAS Recidivism	Adaptive Priority Reweighting	0.356 (± 0.18)	0.643 (± 0.02)	0.287 (± 0.18)
	Adversarial Debiasing	0.553 (± 0.09)	0.669 (± 0.01)	0.116 (± 0.09)
	Fair Transition Loss	0.572 (± 0.03)	0.631 (± 0.04)	0.059 (± 0.03)
	Gerry Fair Classifier	0.530 (± 0.03)	0.637 (± 0.04)	0.107 (± 0.05)
	Prejudice Remover	0.264 (± 0.03)	0.357 (± 0.01)	0.093 (± 0.02)
	Standard MLP (baseline)	0.155 (± 0.04)	0.350 (± 0.02)	0.195 (± 0.04)
German Credit	Adaptive Priority Reweighting	0.674 (± 0.06)	0.750 (± 0.03)	0.076 (± 0.04)
	Adversarial Debiasing	0.368 (± 0.38)	0.685 (± 0.10)	0.317 (± 0.30)
	Fair Transition Loss	0.599 (± 0.12)	0.711 (± 0.05)	0.112 (± 0.11)
	Gerry Fair Classifier	0.662 (± 0.12)	0.719 (± 0.12)	0.057 (± 0.07)
	Prejudice Remover	0.664 (± 0.05)	0.748 (± 0.02)	0.084 (± 0.04)
	Standard MLP (baseline)	0.638 (± 0.06)	0.738 (± 0.04)	0.101 (± 0.05)

Figure 3.10: Metric distribution optimizing Accuracy and Equal Opportunity in comparison with Fair Transition Loss across multiple resample runs. Corresponding values available at Table 3.9.

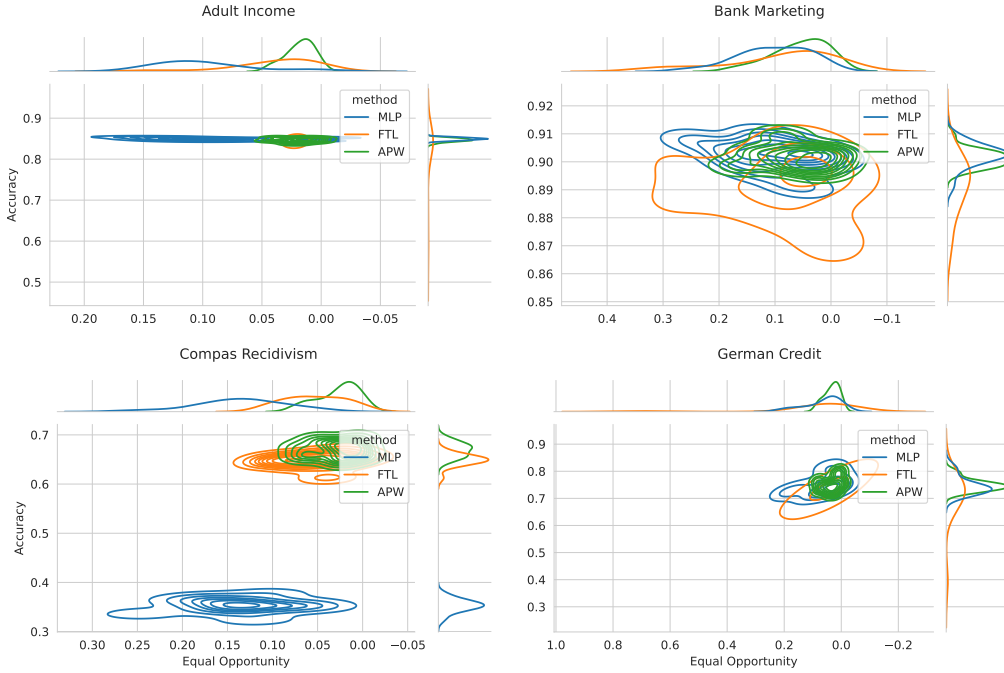
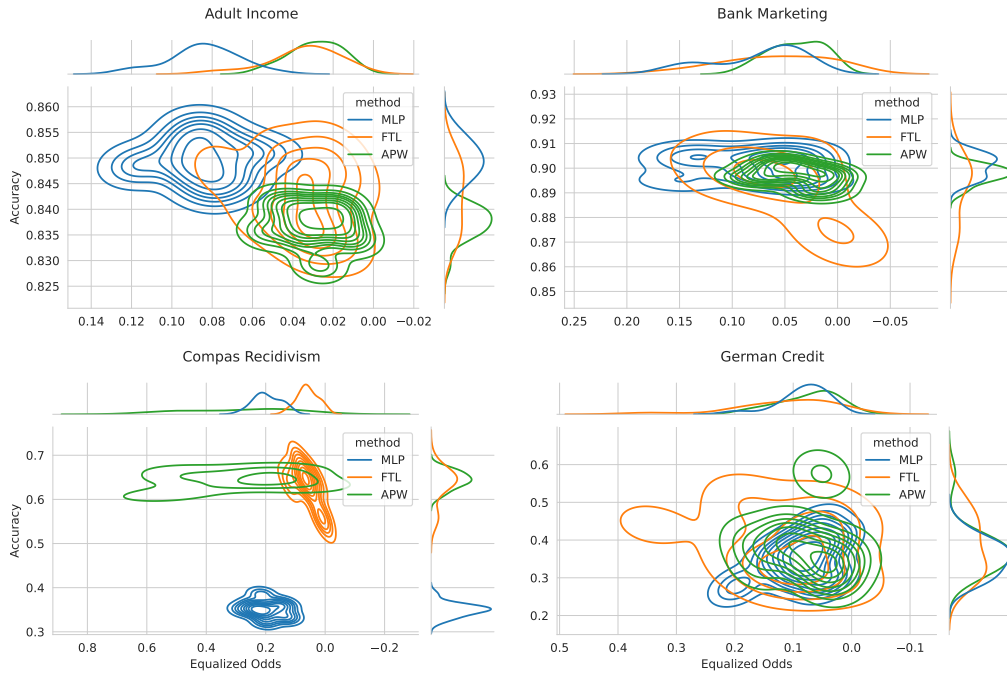


Table 3.10: Mean and standard deviation metric values optimizing Accuracy and Equalized Odds in comparison with Fair Transition Loss across multiple resample runs.

Dataset	Method	\uparrow Fitness	\uparrow Accuracy	\downarrow Eq. Odds
Adult Income	Adaptive Priority Reweighting	0.829 (± 0.01)	0.847 (± 0.00)	0.018 (± 0.01)
	Adversarial Debiasing	0.756 (± 0.03)	0.848 (± 0.00)	0.092 (± 0.03)
	Fair Transition Loss	0.787 (± 0.08)	0.826 (± 0.07)	0.039 (± 0.04)
	Gerry Fair Classifier	0.705 (± 0.07)	0.751 (± 0.09)	0.046 (± 0.05)
	Prejudice Remover	0.810 (± 0.02)	0.846 (± 0.00)	0.036 (± 0.02)
	Standard MLP (baseline)	0.752 (± 0.04)	0.849 (± 0.00)	0.097 (± 0.04)
Bank Marketing	Adaptive Priority Reweighting	0.846 (± 0.05)	0.901 (± 0.00)	0.055 (± 0.05)
	Adversarial Debiasing	0.750 (± 0.09)	0.900 (± 0.00)	0.150 (± 0.09)
	Fair Transition Loss	0.799 (± 0.10)	0.891 (± 0.01)	0.092 (± 0.10)
	Gerry Fair Classifier	0.837 (± 0.06)	0.893 (± 0.00)	0.057 (± 0.06)
	Prejudice Remover	0.781 (± 0.07)	0.899 (± 0.00)	0.118 (± 0.07)
	Standard MLP (baseline)	0.800 (± 0.06)	0.902 (± 0.00)	0.102 (± 0.06)
COMPAS Recidivism	Adaptive Priority Reweighting	0.642 (± 0.03)	0.669 (± 0.01)	0.027 (± 0.02)
	Adversarial Debiasing	0.594 (± 0.07)	0.672 (± 0.02)	0.078 (± 0.06)
	Fair Transition Loss	0.594 (± 0.04)	0.648 (± 0.01)	0.054 (± 0.03)
	Gerry Fair Classifier	0.558 (± 0.05)	0.647 (± 0.02)	0.088 (± 0.04)
	Prejudice Remover	0.287 (± 0.03)	0.342 (± 0.01)	0.055 (± 0.03)
	Standard MLP (baseline)	0.218 (± 0.05)	0.353 (± 0.01)	0.135 (± 0.05)
German Credit	Adaptive Priority Reweighting	0.716 (± 0.04)	0.750 (± 0.02)	0.034 (± 0.03)
	Adversarial Debiasing	0.530 (± 0.33)	0.713 (± 0.10)	0.183 (± 0.24)
	Fair Transition Loss	0.622 (± 0.26)	0.705 (± 0.10)	0.083 (± 0.17)
	Gerry Fair Classifier	0.643 (± 0.13)	0.707 (± 0.13)	0.063 (± 0.04)
	Prejudice Remover	0.648 (± 0.06)	0.743 (± 0.03)	0.095 (± 0.06)
	Standard MLP (baseline)	0.681 (± 0.08)	0.747 (± 0.03)	0.066 (± 0.06)

Figure 3.11: Metric distribution optimizing Accuracy and Equalized Odds in comparison with Fair Transition Loss across multiple resample runs. Corresponding values available at Table 3.10.



Chapter 4

Correlation based penalty function

In this chapter, we propose a novel regularization factor to penalize the use of features highly correlated with a protected attribute by a machine learning model, aiming to avoid the redlining effect (PEDRESCHI *et al.*, 2008). The proposed penalization factor is applied directly to the input weights of a Multi-Layer Perceptron, using a strategy that proportionately penalizes features use based on their correlation with the sensitive feature.

4.1 Preliminaries

The phenomena known as redlining effect (PEDRESCHI *et al.*, 2008) consists in the unintended use of proxy variables to the sensitive feature by the model, which can lead model to produce indirect discrimination in their outcomes. Thus, the insight here is to penalize the use of those proxy features by the model proportionally according its correlation with sensitive feature in order to avoid redlining effed. The proposed regularization approach uses the recently described Chatterjee’s xi correlation coefficient ξ , which robustly asses whether a random variable can be described as a function of another one, as a measure of the potential of given feature to be used by the model as a proxy to the sensitive feature.

Before discussing this approach, we start defining some common correlation coefficients and providing a proper comparison within Chatterjee’s correlation. For purpose of simplicity we describe only the most commonly used form, more complex formulations involving additional terms depending on available data should be considered to most of coefficients. Also, we present some related approaches, specially those ones that, like ours, use regularization and penalty factors to avoid indirect discrimination.

The Pearson correlation coefficient, denoted by ρ and defined in Definition 20, is a measure of the linear relationship between two random variables. The Pearson correlation coefficient ranges from -1 to 1 , where 1 indicates a perfect positive linear

relationship, -1 indicates a perfect negative linear relationship, and 0 indicates no linear relationship.

Definition 20 (Pearson correlation coefficient). *Let X and Y random variables. The Pearson correlation coefficient ρ between X and Y can be defined as*

$$\rho = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}, \quad (4.1)$$

where $\text{Cov}(X, Y)$ is the covariance between X and Y , while σ_X and σ_Y the standard deviations of X and Y , respectively.

Spearman's rank correlation coefficient, denoted by ρ_s and defined in Definition 21, measures the strength and direction of the monotonic relationship between two ranked (ordered) variables. As like Pearson correlation coefficient, Spearman's rank correlation coefficient ranges from -1 to 1 , where 1 indicates a perfect positive relationship, -1 indicates a perfect negative relationship, and 0 indicates no relationship.

Definition 21 (Spearman's correlation coefficient). *Let X and Y random variables, n the sample size and X_i, Y_i the i -th observations to $i = 1 \dots n$. Let R_{X_i}, R_{Y_i} the rank of observations X_i, Y_i , i.e. the position of X_i, Y_i by ordering the samples, respectively. The Spearman's rank correlation coefficient ρ_s between X and Y can be defined to non-repeated observation values as*

$$\rho_s = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)}, \quad (4.2)$$

where $d_i = R_{X_i} - R_{Y_i}$.

Kendall's rank correlation coefficient, denoted by τ and defined in Definition 22, indicates the strength and direction of association between two variables. It is based on the relative ordering of pairs of observations rather than their actual values. Here we interpret the correlation values as like in Spearman's and Pearson's correlation coefficients, ranging from -1 to 1 , where 1 indicates a perfect positive relationship, -1 a perfect negative relationship, and 0 no linear relationship.

Definition 22 (Kendall's correlation coefficient). *Let two variables X and Y sampled with n pairs of observations (X_i, Y_i) for $i = 1, 2, \dots, n$. A pair of observations (X_i, Y_i) and (X_j, Y_j) is **concordant** if the ranks (order) of both elements agree, i.e., to $i < j$ either $X_i > X_j$ and $Y_i > Y_j$ or $X_i < X_j$ and $Y_i < Y_j$. Otherwise they are considered **discordant**, i.e., either $X_i > X_j$ and $Y_i < Y_j$ or $X_i < X_j$ and $Y_i > Y_j$.*

The Kendall's rank correlation coefficient τ between X and Y can be defined as

$$\tau = \frac{2(N_c - N_d)}{n(n-1)}, \quad (4.3)$$

where N_c is the number of concordant observation and N_d the number of discordant observations.

Chatterjee's rank correlation coefficient, denoted by ξ and defined in Definition 23, is designed to robustly measure the degree of dependence between two variables without assuming any specific type of relationship and capturing noise nuances. The previously described correlation coefficients are not effective on detecting associations that are not monotonic, even in complete absence of noise.

This correlation coefficient assesses whether a random variable can be described as a function of another one. Differently from correlation coefficients described before, this coefficient ranges from 0 to 1, where 0 indicates independence and 1 indicates a perfect functional relationship. Also this correlation is not symmetric, i.e., the correlation between X and Y may differ from between Y and X .

Definition 23 (Chatterjee's correlation coefficient). *Let two variables X and Y , where Y is not a constant, sampled with n pairs of observations (X_i, Y_i) for $i = 1, 2, \dots, n$. The Chatterjee's rank correlation coefficient τ between X and Y when there are no ties among Y can be defined as*

$$\xi = 1 - \frac{3 \sum_{i=1}^{n-1} |R_{i+1} - R_i|}{n^2 - 1}, \quad (4.4)$$

where R_i is the rank of Y_i in the ordered sequence of Y values corresponding to the sorted X values.

To illustrate some characteristics of Chatterjee's rank correlation coefficient we compare those results on Anscombe's quartet (?) along with Pearson's, Spearman's and Kendall's correlation coefficients. The Anscombe's quartet (Table 4.1) is a set of four different datasets that have nearly identical simple descriptive statistics, yet appear very different when graphed.

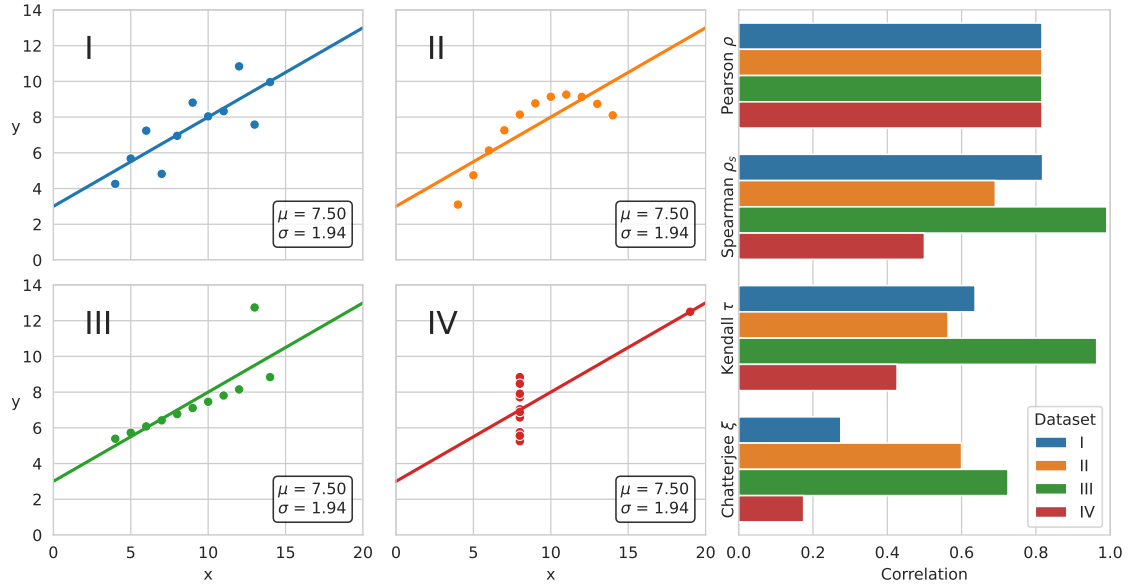
L2 regularization, also known as weight decay, is a common technique used to prevent overfitting in machine learning models, including Multi-Layer Perceptrons (MLPs). In the context of an MLP, L2 regularization adds a penalty term to the loss function that is proportional to the sum of the squares of the model parameters (weights). This encourages the model to keep the weights small, which can help improve generalization.

Let $\mathbf{W}^{(l)}$ represent the weight matrix for the l -th layer of the MLP, and let $\mathbf{b}^{(l)}$ denote the corresponding bias vector. The primary loss function of the network, L_0 ,

Table 4.1: Anscombe's quartet

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Figure 4.1: Multiple correlation coefficients on Anscombe's quartet.



could be any suitable loss function such as the mean squared error for regression or the cross-entropy loss for classification. The L2 regularization term for a single layer is given by

$$R(\mathbf{W}^{(l)}) = \frac{1}{2} \sum_{i=1}^{d_l} \sum_{j=1}^{h_l} \left(W_{ij}^{(l)} \right)^2, \quad (4.5)$$

where d_l and h_l are the dimensions of the weight matrix $\mathbf{W}^{(l)}$, and $W_{ij}^{(l)}$ is the weight connecting the i -th input neuron to the j -th neuron in the l -th layer. Thus, the total

regularization term for the entire network, considering all layers, is

$$R(\mathbf{W}) = \frac{1}{2} \sum_{l=1}^L \sum_{i=1}^{d_l} \sum_{j=1}^{h_l} \left(W_{ij}^{(l)} \right)^2, \quad (4.6)$$

where L is the total number of layers in the network. Furthermore, the total loss
 1610 function L for the MLP, incorporating the L2 regularization term, is defined as

$$L = L_0 + \lambda R(\mathbf{W}), \quad (4.7)$$

where λ is a scalar hyperparameter that controls the overall strength of the regularization.

By adding this regularization term, the optimization process aims to minimize the primary loss L_0 along with keeping the weights small, thereby helping to reduce
 1615 the model complexity and prevent overfitting. The gradient descent updates for the weights will be adjusted to account for the regularization term, effectively shrinking the weights during the training process.

4.2 Chatterjee Redlining Penalty

The Chatterjee Redlining Penalty (CRP) is a regularization term that penalizes the
 1620 weights of features highly correlated with the sensitive attribute, as measured by Chatterjee’s Xi Rank Correlation Coefficient. By incorporating this penalty into the loss function of the neural network, the model is encouraged to reduce its reliance on sensitive attributes, thus promoting fairer predictions.

As referred before, the Chatterjee Xi Correlation Coefficient distinguish from
 1625 many other by providing a measure of how much one random variable can be expressed as a function of another, with a range from 0 to 1. This characteristics is relevant to the proposed use, to capture redlining effect, as of this phenomena happens exactly when the sensitive feature can be inferred by another one, producing the same harmful effects whether the correlation is positive or negative. Naturally
 1630 it would be possible to use the absolute value of a correlation coefficient such as Pearson or Spearman, but this strategy could lead to some kind of information loss. Another relevant characteristics of this correlation coefficient it is the ability to capture sophisticated nonlinearities and the effects of noise on variable’s distribution, conditions frequently present in proxy feature relationships.

Thus, let $X \in \mathbb{R}^{n \times d}$ be a dataset where n represents the number of instances
 1635 and d represents the number of features. Let $X_i \in \mathbb{R}^n$ denote the i -th feature of the dataset, and let $A = X_i \in \mathbb{R}^n$ be a sensitive (protected) feature for some i . In this neural network, $\mathbf{W}^{(1)} \in \mathbb{R}^{d \times h}$ is the weight matrix for the first hidden layer, with h

being the number of neurons in this layer. Additionally, λ is a scalar that controls
1640 the overall strength of the regularization.

Thus, the proposed regularization term $R(\mathbf{W}^{(1)})$ applied to the weight matrix $\mathbf{W}^{(1)}$ of the first hidden layer is defined as

$$R(\mathbf{W}^{(1)}) = \sum_{i=1}^d \xi_n(X_i, A) \sum_{j=1}^h (W_{ij}^{(1)})^2, \quad (4.8)$$

where $W_{ij}^{(1)}$ are the weights connecting the i -th input feature to the j -th neuron in
the first hidden layer. Here, the Chatterjee's Xi Correlation Coefficient $\xi_n(X_i, A)$
1645 between the i -th input feature X_i and the sensitive feature A acts as the regulariza-
tion strength for the i -th input feature. The greater i -th input feature dependence
on sensitive feature the greater the penalization factor enforcing lower values to
those weights.

Thus, the total loss function L to a MLP, incorporating the sensitive-feature-
1650 specific L_2 regularization, is defined as

$$L = L_0 + \lambda R(\mathbf{W}^{(1)}), \quad (4.9)$$

where L_0 is the primary loss function of the network. This formulation ensures
that the model's learning process penalizes weights with high values associated with
features highly correlated to the sensitive attribute, thereby reducing the potential
for biased decisions influenced by redlining effect.

This formulation differs from similar approaches to regularization in order to
1655 achieve fairness in machine learning by two fundamental points. The first one is
that here the regularization acts as a focused penalty factor through the use of the
correlation of each feature to the sensitive one, reducing it's effect on less correlated
features and boosting on the highly correlated ones. The second point is that this
1660 proposed penalty does not acts as a common regularizer by imposing small values
on every model's parameter. Rather, it penalizes the use of each feature according
it's correlation to the sensitive one by exclusively acting on the input layer weights,
that is, the usage level of each feature by hidden units.

With the combination of characteristics tailored by this formulation and the
1665 chosen correlation coefficient we do expects effectively avoiding the use of indirect
sensitive feature predictors within the model, thus mitigating redlining effect.

4.3 Experimental setup

In this section, we detail the experimental setup employed to benchmark the proposed regularization within both Standard MLP with Cross Entropy Loss and a MLP using Fair Transition Loss. Both MLP model uses two hidden layers with 100 hidden units each, *ReLU* activation function, batch size of 64, 50 epochs early stopped at 3 epochs without improvement (LI *et al.*, 2020) and softmax in output, trained with ADAM optimizer (KINGMA e BA, 2015) and learning rate at $3e-4$.

The experimental methodology follows same principles of the one used on Fair Transition Loss Evaluation. There are two phases: hyperparameter tuning and testing. In the hyperparameter tuning phase we perform a Bandit-Based pruning approach using HyperBand (LI *et al.*, 2018) with Tree-structured Parzen Estimator Sampler (TPE) (BERGSTRA *et al.*, 2011) over 100 trials. At each trial a fitness function is evaluated by performing a complete training and validation, where both model performance and fairness metrics are assessed. The fitness function is computed based on the objective defined in Equation 3.16.

Once the best hyperparameters are selected, we proceed to the testing phase, where a new training is conducted using those optimal hyperparameters. After this training, we evaluate the model’s performance on a separate test set that was not used during the hyperparameter tuning phase, which are reported. This complete tuning-training-testing described is repeated 15 times with dataset re-sampling then we proceed to comparison. Here the re-sampling consists in shuffling the whole dataset before splitting, which is better described further in this section.

As the objective defined in Equation 3.16 can be achieved with different performance and fairness metrics, we compare the proposed regularization schema within Standard MLP and MLP with FTL in different optimization scenarios, considering as performance metrics Accuracy (Acc.) and Mathews Correlation Coefficient (MCC), while the fairness metrics considered are Statistical Parity (Stat. Parity), Equal Opportunity (Eq. Opp.) and Equalized Odds (Eq. Odds). Those pursued metrics lead us to the following optimization scenarios: MCC and Statistical Parity; MCC and Equal Opportunity; MCC and Equalized Odds; Accuracy and Statistical Parity; Accuracy and Equal Opportunity; Accuracy and Equalized Odds.

Table 4.2 presents the methods hyperparameters along with their corresponding search ranges or options. While each method may possess a varying number of hyperparameters and range sizes, all are optimized under the same conditions and number of configurations to guarantee a balanced comparison.

To properly compare this set of 15 results of each method, we conduct an Almost Stochastic Order (ASO) test (DROR *et al.*, 2019), a significance test suitable for comparing complex machine learning models with various hyperparameters. The

Table 4.2: Hyperparameters search ranges or options of each method.

Method	Parameter	Range/options
Standard MLP without regularization (baseline)	dropout	[0.0, 0.2]
Standard MLP with CRP	dropout	[0.0, 0.2]
	λ	$\{1e^{-2}, 1e^{-3}, 1e^{-4}\}$
Fair Transition Loss without regularization	d_0, p_0, d_1, p_1	[0.0, 1.0]
	dropout	[0.0, 0.2]
Fair Transition Loss with CRP	d_0, p_0, d_1, p_1	[0.0, 1.0]
	dropout	[0.0, 0.2]
	λ	$\{1e^{-2}, 1e^{-3}, 1e^{-4}\}$

1705 ASO test involves evaluating a set of metrics through multiple samplings of a Col-
 lection of Statistics (in this case assessed in test phase using random resampling)
 to compare one method against another. The $ASO(A, B)$ function yields a value
 in the range $[0, 1]$, given two methods A and B . If $ASO(A, B)$ is lesser than 0.5,
 we can reject the null hypothesis and conclude that method A outperforms method
 1710 B in the given task. That is, method A produces stochastically larger values than
 method B for a given metric. The lower the $ASO(A, B)$ value, the stronger the
 evidence that A is superior to B in that particular task, which can be interpreted
 as a confidence interval. Therefore, we perform pairwise comparisons between all
 methods for each optimization scenario outlined previously and for each dataset.

1715 Our experiments uses common datasets used in Fair Classification problems,
 namely *Adult Income* (BECKER e KOHAVI, 1996), *German Credit* (HOFMANN,
 1994), *Bank Marketing* (S. MORO e CORTEZ, 2012), and *COMPAS Recidi-*
vism (JEFF LARSON e ANGWIN, 2016). We use the dataset readers available
 in the AI Fairness 360 toolkit (BELLAMY *et al.*, 2018) with its standard configu-
 1720 rations. Instances with missing data are removed. To a proper description of those
 datasets please check Section 3.3

For all datasets, the data preparation process is the same, one-hot encoding for
 categorical features and standardize the numerical features. We perform a random
 split, with 80% allocated for the hyperparameter tuning phase and the remaining
 1725 20% reserved for evaluating metrics in the test phase. Within the hyperparameter
 tuning phase, this corresponding fraction of data is further randomly split, with 80%
 assigned to training and 20% to validation. The validation set allows us to assess
 metrics and compute the fitness function for each hyperparameter configuration. In
 datasets where there is originally some kind of split (e.g., train set and test set in
 1730 separate files), all available data is merged and then shuffled to produce new splits
 at each run.

4.4 Results and discussion

Table 4.3: Almost Stochastic Order test comparing the use of Chatterjee Redlining Penalty on Standard MLP and MLP with Fair Transition Loss fitness. Values under 0.5 (in bold) mean that the model with CRP outperforms the same model without regularization on each optimization scenario.

Fitness Rule	MLP				FTL			
	Adult	Bank	Compas	German	Adult	Bank	Compas	German
MCC - S. Parity	0.79	0.99	0.03	1.00	0.34	1.00	1.00	0.31
MCC - Eq. Opp.	0.26	1.00	0.01	1.00	0.35	1.00	0.19	1.00
MCC - Eq. Odds	0.08	1.00	0.20	1.00	0.30	1.00	0.56	1.00
Acc. - S. Parity	0.75	1.00	0.39	0.87	0.44	1.00	0.02	1.00
Acc. - Eq. Opp.	1.00	0.18	0.07	0.24	0.58	0.67	1.00	1.00
Acc. - Eq. Odds	0.33	1.00	0.01	0.62	0.40	1.00	0.20	1.00

Figure 4.2: Fitness values of CRP optimizing MCC and multiple fairness metrics.

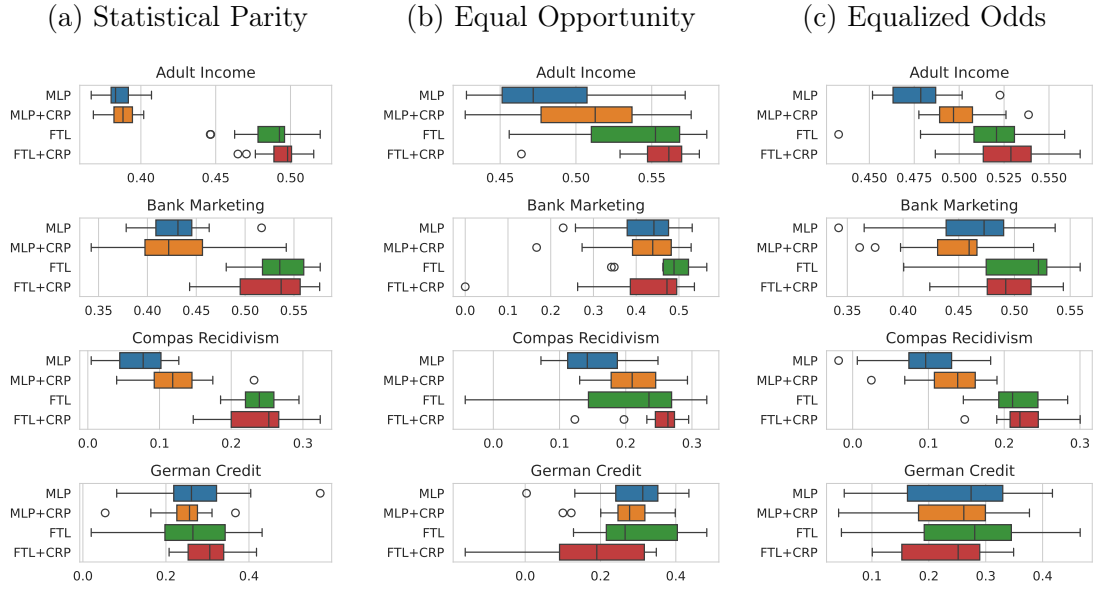


Table 4.4: Mean and standard deviation metric values optimizing MCC and Statistical Parity in comparison with Chatterjee Redlining Penalty.

Dataset	Method	Fitness	MCC	Stat. Parity
Adult Income	FTL+CRP	0.494(± 0.01)	0.517(± 0.02)	0.023(± 0.02)
	FTL	0.487(± 0.02)	0.509(± 0.02)	0.022(± 0.02)
	MLP+CRP	0.388(± 0.01)	0.578(± 0.01)	0.191(± 0.01)
	MLP	0.386(± 0.01)	0.576(± 0.01)	0.191(± 0.01)
Bank Marketing	FTL	0.534(± 0.03)	0.569(± 0.01)	0.035(± 0.03)
	FTL+CRP	0.523(± 0.04)	0.569(± 0.01)	0.046(± 0.04)
	MLP	0.429(± 0.03)	0.521(± 0.02)	0.092(± 0.02)
	MLP+CRP	0.426(± 0.05)	0.528(± 0.02)	0.102(± 0.04)
Compas Recidivism	FTL	0.239(± 0.03)	0.276(± 0.03)	0.036(± 0.03)
	FTL+CRP	0.236(± 0.05)	0.294(± 0.03)	0.058(± 0.04)
	MLP+CRP	0.121(± 0.05)	0.329(± 0.03)	0.208(± 0.03)
	MLP	0.074(± 0.03)	0.283(± 0.02)	0.209(± 0.04)
German Credit	FTL+CRP	0.302(± 0.06)	0.371(± 0.05)	0.069(± 0.05)
	MLP	0.266(± 0.10)	0.329(± 0.09)	0.064(± 0.05)
	FTL	0.256(± 0.12)	0.355(± 0.08)	0.099(± 0.06)
	MLP+CRP	0.246(± 0.07)	0.329(± 0.05)	0.084(± 0.06)

Table 4.5: Mean and standard deviation metric values optimizing MCC and Equal Opportunity in comparison with Chatterjee Redlining Penalty.

Dataset	Method	Fitness	MCC	Eq. Opp.
Adult Income	FTL+CRP	0.554(± 0.03)	0.581(± 0.02)	0.027(± 0.02)
	FTL	0.540(± 0.04)	0.576(± 0.01)	0.036(± 0.03)
	MLP+CRP	0.506(± 0.05)	0.577(± 0.01)	0.071(± 0.05)
	MLP	0.480(± 0.04)	0.582(± 0.01)	0.103(± 0.04)
Bank Marketing	FTL	0.483(± 0.07)	0.567(± 0.02)	0.084(± 0.06)
	MLP	0.420(± 0.08)	0.524(± 0.02)	0.104(± 0.08)
	MLP+CRP	0.420(± 0.10)	0.529(± 0.02)	0.109(± 0.09)
	FTL+CRP	0.416(± 0.14)	0.519(± 0.15)	0.104(± 0.08)
Compas Recidivism	FTL+CRP	0.251(± 0.04)	0.309(± 0.03)	0.058(± 0.04)
	MLP+CRP	0.211(± 0.05)	0.325(± 0.02)	0.114(± 0.04)
	FTL	0.195(± 0.11)	0.281(± 0.03)	0.086(± 0.09)
	MLP	0.150(± 0.05)	0.282(± 0.03)	0.132(± 0.05)
German Credit	FTL	0.293(± 0.12)	0.368(± 0.10)	0.074(± 0.05)
	MLP	0.290(± 0.09)	0.355(± 0.07)	0.065(± 0.06)
	MLP+CRP	0.275(± 0.09)	0.341(± 0.06)	0.066(± 0.05)
	FTL+CRP	0.166(± 0.16)	0.284(± 0.14)	0.117(± 0.07)

Table 4.6: Mean and standard deviation metric values optimizing MCC and Equalized Odds in comparison with Chatterjee Redlining Penalty.

Dataset	Method	Fitness	MCC	Eq. Odds
Adult Income	FTL+CRP	0.526(± 0.02)	0.575(± 0.01)	0.048(± 0.02)
	FTL	0.515(± 0.03)	0.572(± 0.02)	0.057(± 0.02)
	MLP+CRP	0.500(± 0.02)	0.578(± 0.01)	0.078(± 0.02)
	MLP	0.478(± 0.02)	0.575(± 0.01)	0.097(± 0.02)
Bank Marketing	FTL	0.495(± 0.05)	0.574(± 0.01)	0.078(± 0.05)
	FTL+CRP	0.490(± 0.03)	0.571(± 0.01)	0.081(± 0.03)
	MLP	0.461(± 0.05)	0.522(± 0.02)	0.061(± 0.04)
	MLP+CRP	0.446(± 0.04)	0.539(± 0.02)	0.093(± 0.04)
Compas Recidivism	FTL+CRP	0.225(± 0.04)	0.282(± 0.03)	0.056(± 0.03)
	FTL	0.216(± 0.04)	0.288(± 0.02)	0.072(± 0.04)
	MLP+CRP	0.129(± 0.05)	0.303(± 0.02)	0.174(± 0.03)
	MLP	0.098(± 0.05)	0.283(± 0.03)	0.185(± 0.03)
German Credit	FTL	0.278(± 0.12)	0.383(± 0.08)	0.105(± 0.06)
	MLP	0.249(± 0.10)	0.352(± 0.08)	0.102(± 0.06)
	MLP+CRP	0.231(± 0.10)	0.332(± 0.07)	0.102(± 0.06)
	FTL+CRP	0.228(± 0.08)	0.337(± 0.07)	0.109(± 0.06)

Table 4.7: Mean and standard deviation metric values optimizing Accuracy and Statistical Parity in comparison with Chatterjee Redlining Penalty.

Dataset	Method	Fitness	Accuracy	Stat. Parity
Adult Income	FTL+CRP	0.811(± 0.01)	0.825(± 0.01)	0.015(± 0.01)
	FTL	0.806(± 0.02)	0.827(± 0.01)	0.022(± 0.02)
	MLP+CRP	0.666(± 0.01)	0.850(± 0.00)	0.183(± 0.02)
	MLP	0.664(± 0.01)	0.850(± 0.00)	0.185(± 0.01)
Bank Marketing	FTL	0.828(± 0.14)	0.887(± 0.01)	0.059(± 0.14)
	FTL+CRP	0.800(± 0.25)	0.897(± 0.01)	0.096(± 0.24)
	MLP	0.799(± 0.03)	0.902(± 0.00)	0.103(± 0.03)
	MLP+CRP	0.777(± 0.03)	0.905(± 0.00)	0.127(± 0.04)
Compas Recidivism	FTL+CRP	0.606(± 0.03)	0.651(± 0.01)	0.045(± 0.03)
	FTL	0.520(± 0.12)	0.619(± 0.04)	0.100(± 0.11)
	MLP+CRP	0.462(± 0.03)	0.659(± 0.01)	0.197(± 0.03)
	MLP	0.439(± 0.04)	0.646(± 0.01)	0.207(± 0.04)
German Credit	FTL	0.684(± 0.07)	0.723(± 0.03)	0.040(± 0.05)
	FTL+CRP	0.655(± 0.07)	0.725(± 0.03)	0.070(± 0.06)
	MLP+CRP	0.631(± 0.07)	0.733(± 0.02)	0.102(± 0.06)
	MLP	0.628(± 0.06)	0.742(± 0.03)	0.114(± 0.06)

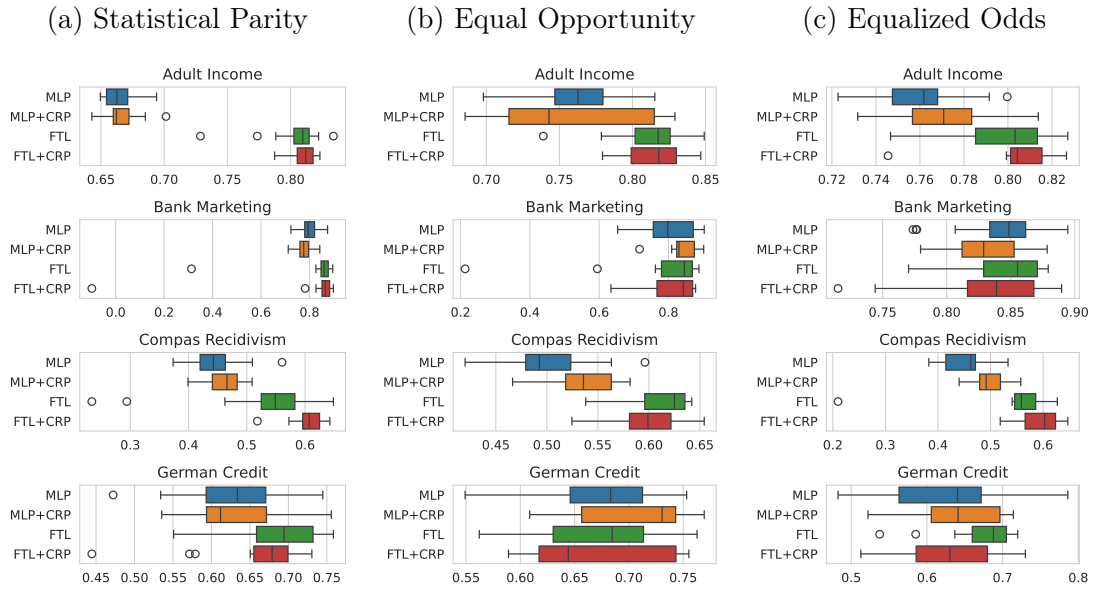
Table 4.8: Mean and standard deviation metric values optimizing Accuracy and Equal Opportunity in comparison with Chatterjee Redlining Penalty.

Dataset	Method	Fitness	Accuracy	Eq. Opp.
Adult Income	FTL+CRP	0.815(± 0.02)	0.847(± 0.00)	0.031(± 0.02)
	FTL	0.812(± 0.03)	0.845(± 0.01)	0.034(± 0.02)
	MLP	0.758(± 0.04)	0.848(± 0.00)	0.090(± 0.04)
	MLP+CRP	0.757(± 0.05)	0.849(± 0.00)	0.091(± 0.05)
Bank Marketing	MLP+CRP	0.839(± 0.05)	0.903(± 0.00)	0.063(± 0.04)
	MLP	0.803(± 0.07)	0.902(± 0.00)	0.099(± 0.07)
	FTL+CRP	0.802(± 0.08)	0.892(± 0.01)	0.090(± 0.09)
	FTL	0.781(± 0.17)	0.883(± 0.02)	0.102(± 0.17)
Compas Recidivism	FTL	0.614(± 0.03)	0.645(± 0.03)	0.031(± 0.02)
	FTL+CRP	0.598(± 0.03)	0.639(± 0.02)	0.041(± 0.03)
	MLP+CRP	0.536(± 0.03)	0.663(± 0.01)	0.127(± 0.03)
	MLP	0.498(± 0.04)	0.646(± 0.01)	0.148(± 0.04)
German Credit	MLP+CRP	0.702(± 0.05)	0.750(± 0.02)	0.048(± 0.04)
	FTL	0.676(± 0.06)	0.751(± 0.02)	0.075(± 0.06)
	MLP	0.673(± 0.06)	0.739(± 0.03)	0.066(± 0.05)
	FTL+CRP	0.665(± 0.06)	0.736(± 0.03)	0.071(± 0.05)

Table 4.9: Mean and standard deviation metric values optimizing Accuracy and Equalized Odds in comparison with Chatterjee Redlining Penalty.

Dataset	Method	Fitness	Accuracy	Eq. Odds
Adult Income	FTL+CRP	0.805(± 0.02)	0.843(± 0.01)	0.038(± 0.02)
	FTL	0.798(± 0.02)	0.841(± 0.01)	0.043(± 0.02)
	MLP+CRP	0.772(± 0.02)	0.849(± 0.00)	0.077(± 0.02)
	MLP	0.760(± 0.02)	0.849(± 0.00)	0.089(± 0.02)
Bank Marketing	FTL	0.846(± 0.03)	0.890(± 0.01)	0.044(± 0.04)
	MLP	0.845(± 0.03)	0.901(± 0.00)	0.057(± 0.03)
	FTL+CRP	0.831(± 0.05)	0.892(± 0.01)	0.061(± 0.05)
	MLP+CRP	0.830(± 0.03)	0.902(± 0.00)	0.072(± 0.03)
Compas Recidivism	FTL+CRP	0.594(± 0.04)	0.647(± 0.02)	0.053(± 0.04)
	FTL	0.545(± 0.10)	0.631(± 0.05)	0.086(± 0.08)
	MLP+CRP	0.497(± 0.04)	0.667(± 0.01)	0.170(± 0.03)
	MLP	0.449(± 0.04)	0.649(± 0.01)	0.200(± 0.03)
German Credit	FTL	0.669(± 0.05)	0.712(± 0.02)	0.043(± 0.07)
	MLP+CRP	0.640(± 0.06)	0.748(± 0.02)	0.107(± 0.06)
	FTL+CRP	0.631(± 0.06)	0.721(± 0.03)	0.090(± 0.08)
	MLP	0.619(± 0.07)	0.740(± 0.03)	0.121(± 0.07)

Figure 4.3: Fitness values of CRP optimizing Accuracy and multiple fairness metrics.



Chapter 5

Conclusions

1735 In this study, we present Fair Transition Loss, a novel in-processing technique for addressing fair classification problems. It hoisting concepts from label noise robustness to mitigate social bias against underprivileged groups. We explore the intersection of these two research areas, highlighting both their similarities and differences. Our approach tackles the fairness-performance trade-off as a multi-objective optimization problem, employing a linear relaxed objective function to reduce bias while
1740 maintaining acceptable predictive performance levels. We benchmark this approach and compare to prominent in-processing techniques in common fair classification tasks, using the Almost Stochastic Order test to evaluate results through multiple resampling iterations. This ensures that all methods operate under the same
1745 conditions, maximizing their potential within the scope of hyperparameter tuning.

This is the first technique that models fair classification problems by drawing insights from classification in the presence of label noise. Our experiments indicate that Fair Transition Loss consistently outperforms its competitors in most optimization scenarios. Even in those cases that the proposed method isn't the outright
1750 leader, it performs at least as well as evaluated alternatives. Therefore, this novel approach can significantly mitigate bias while keeping model performance, particularly in scenarios optimizing balanced performance metrics like MCC. The proposed technique particularly stands out in setups where hyperparameter tuning is an integral component of the prediction pipeline.

1755 While our proposed method seems competitive in problems involving hyperparameter optimization for binary fair classification tasks using a simple Multi-Layer Perceptron, we can outline some potential research directions: evaluate Fair Transition Loss within different neural network architectures, such as Deep Neural Networks; investigate whether the proposed method can effectively address multi-class
1760 fair classification problems and handle multiple sensitive attributes, as theoretically possible; evaluate FTL within different multi objective optimization schemes, such as the Fair Hyperparameter Tuning techniques proposed by F.CRUIZ *et al.* (2021) or

the non-linear Chebyshev scalarization scheme proposed by WEI e NIETHAMMER (2022); explore approaches to estimating or initializing transition matrices without
1765 relying on hyperparameter tuning techniques.

With this work, we hope to establish Fair Transition Loss as a relevant tool in fair classification tasks and pave the way for novel approaches that draw insights from label noise for various fair machine learning problems, including regression, recommender systems, ranking and language models.

1770 5.1 Considerations on the proposal

5.2 Contributions

5.3 Results summary

colocar uma tabela pequena para FTL e outra para regularização aqui

5.4 Research directions

References

- MEHRABI, N., MORSTATTER, F., SAXENA, N., et al. “A Survey on Bias and Fairness in Machine Learning”, *ACM Comput. Surv.*, v. 54, n. 6, jul 2021. ISSN: 0360-0300. doi: 10.1145/3457607. Disponível em: <<https://doi.org/10.1145/3457607>>.
- 1780 HUTCHINSON, B., MITCHELL, M. “50 Years of Test (Un)fairness: Lessons for Machine Learning”, *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2018. Disponível em: <<https://api.semanticscholar.org/CorpusID:53782832>>.
- CATON, S., HAAS, C. “Fairness in Machine Learning: A Survey”, *ACM Comput. Surv.*, aug 2023. ISSN: 0360-0300. doi: 10.1145/3616865. Disponível em: <<https://doi.org/10.1145/3616865>>. Just Accepted.
- 1785 CATON, S., HAAS, C. “Fairness in Machine Learning: A Survey”, *ACM Comput. Surv.*, aug 2023. ISSN: 0360-0300. doi: 10.1145/3616865. Disponível em: <<https://doi.org/10.1145/3616865>>. Just Accepted.
- ZAFAR, M. B., VALERA, I., ROGRIGUEZ, M. G., et al. “Fairness Constraints: Mechanisms for Fair Classification”. In: Singh, A., Zhu, J. (Eds.), *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, v. 54, *Proceedings of Machine Learning Research*, pp. 962–970. PMLR, 20–22 Apr 2017a. Disponível em: <<https://proceedings.mlr.press/v54/zafar17a.html>>.
- 1790 ZAFAR, M. B., VALERA, I., ROGRIGUEZ, M. G., et al. “Fairness Constraints: Mechanisms for Fair Classification”. In: Singh, A., Zhu, J. (Eds.), *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, v. 54, *Proceedings of Machine Learning Research*, pp. 962–970. PMLR, 20–22 Apr 2017a. Disponível em: <<https://proceedings.mlr.press/v54/zafar17a.html>>.
- MEMARIAN, B., DOLECK, T. “Fairness, Accountability, Transparency, and Ethics (FATE) in Artificial Intelligence (AI) and higher education: A systematic review”. 1 2023. ISSN: 2666920X.
- 1795 MEMARIAN, B., DOLECK, T. “Fairness, Accountability, Transparency, and Ethics (FATE) in Artificial Intelligence (AI) and higher education: A systematic review”. 1 2023. ISSN: 2666920X.
- HUTCHINSON, B., SMART, A., HANNA, A., et al. “Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure”. In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT ’21, p. 560–575, New York, NY, USA, 2021. Association for Computing Machinery. ISBN: 9781450383097. doi: 10.1145/3442188.3445918. Disponível em: <<https://doi.org/10.1145/3442188.3445918>>.
- 1800 HUTCHINSON, B., SMART, A., HANNA, A., et al. “Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure”. In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT ’21, p. 560–575, New York, NY, USA, 2021. Association for Computing Machinery. ISBN: 9781450383097. doi: 10.1145/3442188.3445918. Disponível em: <<https://doi.org/10.1145/3442188.3445918>>.

- BURKART, N., HUBER, M. F. “A Survey on the Explainability of Supervised Machine Learning”, *J. Artif. Int. Res.*, v. 70, pp. 245–317, may 2021. ISSN: 1076-9757. doi: 10.1613/jair.1.12228. Disponível em: <<https://doi.org/10.1613/jair.1.12228>>.
- HLEG, A. “Ethics guidelines for trustworthy AI”. abr. 2019. Disponível em: <<https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>>.
- PEDRESCHI, D., RUGGIERI, S., TURINI, F. “Discrimination-aware data mining”, *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD 08*, p. 560, 2008. ISSN: 0309-0167 (Print). doi: 10.1145/1401890.1401959. Disponível em: <<http://dl.acm.org/citation.cfm?doid=1401890.1401959>>.
- VERMA, S., RUBIN, J. “Fairness Definitions Explained”, *IEEE/ACM International Workshop on Software Fairness*, v. 18, 2018. doi: 10.1145/3194770.3194776. Disponível em: <<https://doi.org/10.1145/3194770.3194776>>.
- ALER TUBELLA, A., BARSOTTI, F., KOÇER, R. G., et al. “Ethical implications of fairness interventions: what might be hidden behind engineering choices?” *Ethics and Information Technology*, v. 24, n. 1, pp. 1–11, mar 2022. ISSN: 15728439. doi: 10.1007/S10676-022-09636-Z/TABLES/4. Disponível em: <<https://link.springer.com/article/10.1007/s10676-022-09636-z>>.
- ALVES, G., BERNIER, F., COUCEIRO, M., et al. “Survey on fairness notions and related tensions”, *EURO Journal on Decision Processes*, v. 11, pp. 100033, 2023. ISSN: 2193-9438. doi: <https://doi.org/10.1016/j.ejdp.2023.100033>. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S2193943823000067>>.
- WEINBERG, L. “Rethinking Fairness: An Interdisciplinary Survey of Critiques of Hegemonic ML Fairness Approaches”, *Journal of Artificial Intelligence Research*, v. 74, pp. 75–109, 2022. doi: 10.1613/jair.1.13196.
- DWORK, C., HARDT, M., PITASSI, T., et al. “Fairness through Awareness”. In: *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, p. 214–226, New York, NY, USA, 2012. Association for Computing Machinery. ISBN: 9781450311151. doi: 10.1145/2090236.2090255. Disponível em: <<https://doi.org/10.1145/2090236.2090255>>.

- 1840 KUSNER, M., LOFTUS, J., RUSSELL, C., et al. “Counterfactual Fairness”. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS’17, p. 4069–4079, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN: 9781510860964.
- 1845 HARDT, M., PRICE, E., SREBRO, N. “Equality of Opportunity in Supervised Learning”. In: *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS’16, p. 3323–3331, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN: 9781510838819.
- 1850 BLYTH, C. R. “On Simpson’s Paradox and the Sure-Thing Principle”, *Journal of the American Statistical Association*, v. 67, n. 338, pp. 364–366, 1972. ISSN: 01621459. Disponível em: <<http://www.jstor.org/stable/2284382>>.
- 1855 GOH, G., COTTER, A., GUPTA, M., et al. “Satisfying Real-world Goals with Dataset Constraints”. In: Lee, D., Sugiyama, M., Luxburg, U., et al. (Eds.), *Advances in Neural Information Processing Systems*, v. 29. Curran Associates, Inc., 2016. Disponível em: <https://proceedings.neurips.cc/paper_files/paper/2016/file/dc4c44f624d600aa568390f1f1104aa0-Paper.pdf>.
- KOMIYAMA, J., TAKEDA, A., HONDA, J., et al. “Nonconvex optimization for regression with fairness constraints”, *35th International Conference on Machine Learning, ICML 2018*, v. 6, pp. 4280–4294, 2018.
- 1860 PETROVIĆ, A., NIKOLIĆ, M., JOVANOVIĆ, M., et al. “Fair classification via Monte Carlo policy gradient method”, *Engineering Applications of Artificial Intelligence*, v. 104, n. February, pp. 104398, 2021. ISSN: 09521976. doi: 10.1016/j.engappai.2021.104398. Disponível em: <<https://doi.org/10.1016/j.engappai.2021.104398>>.
- 1865 F.CRUZ, A., SALEIRO, P., BELÉM, C., et al. “Promoting Fairness through Hyperparameter Optimization”. In: *2021 IEEE International Conference on Data Mining (ICDM)*, pp. 1036–1041, 2021. doi: 10.1109/ICDM51629.2021.00119.
- 1870 LIU, S., VICENTE, L. N. “Accuracy and fairness trade-offs in machine learning: a stochastic multi-objective approach”, *Computational Management Science*, v. 19, pp. 513–537, 7 2022. ISSN: 16196988. doi: 10.1007/s10287-022-00425-z.

- 1875 KEARNS, M., NEEL, S., ROTH, A., et al. “Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness”, 11 2017. Disponível em: <<http://arxiv.org/abs/1711.05144>>.
- 1880 KEARNS, M., NEEL, S., ROTH, A., et al. “An Empirical Study of Rich Subgroup Fairness for Machine Learning”. In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, p. 100–109, New York, NY, USA, 2019. Association for Computing Machinery. ISBN: 9781450361255. doi: 10.1145/3287560.3287592. Disponível em: <<https://doi.org/10.1145/3287560.3287592>>.
- 1885 CORBETT-DAVIES, S., GOEL, S., CHOHLAS-WOOD, A., et al. *The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning* *. Relatório técnico, 2018. Disponível em: <<https://arxiv.org/pdf/1808.00023.pdf>>.
- HORT, M., CHEN, Z., ZHANG, J. M., et al. “Bias Mitigation for Machine Learning Classifiers: A Comprehensive Survey”, *ACM Journal on Responsible Computing*, pp. 1–52, 2023. doi: 10.1145/3631326.
- 1890 WU, S., HAN, B., LIU, Y., et al. “Fair Classification with Instance-dependent Label Noise”, *Proceedings of Machine Learning Research*, v. 140, pp. 1–17, 2022. Disponível em: <<https://www.mturk.com/>>.
- 1895 MA, Y., FRAUEN, D., MELNYCHUK, V., et al. “Counterfactual Fairness for Predictions using Generative Adversarial Networks”, *CoRR*, v. abs/2310.17687, 2023. doi: 10.48550/ARXIV.2310.17687. Disponível em: <<https://doi.org/10.48550/arXiv.2310.17687>>.
- GRARI, V., LAMPRIER, S., DETYNIECKI, M. “Adversarial learning for counterfactual fairness”, *Mach. Learn.*, v. 112, n. 3, pp. 741–763, 2023. doi: 10.1007/S10994-022-06206-8. Disponível em: <<https://doi.org/10.1007/s10994-022-06206-8>>.
- 1900 KASIRZADEH, A., SMART, A. “The use and misuse of counterfactuals in ethical machine learning”, *FACCT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 228–236, 2021. doi: 10.1145/3442188.3445886.
- 1905 KLEINBERG, J., MULLAINATHAN, S., RAGHAVAN, M. “Inherent trade-offs in the fair determination of risk scores”, *Leibniz International Proceedings in Informatics, LIPIcs*, v. 67, pp. 1–23, 2017. ISSN: 18688969. doi: 10.4230/LIPIcs.ITCS.2017.43.

- CHOULDECHOVA, A. “Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments”, *Big Data*, v. 5, n. 2, pp. 153–163, 2017. doi: 10.1089/big.2016.0047. Disponível em: <<https://doi.org/10.1089/big.2016.0047>>. PMID: 28632438.
- SARAVANAKUMAR, K. K. “The Impossibility Theorem of Machine Fairness – A Causal Perspective”, 2020. Disponível em: <<http://arxiv.org/abs/2007.06024>>.
- BELL, A., BYNUM, L., DRUSHCHAK, N., et al. “The Possibility of Fairness: Revisiting the Impossibility Theorem in Practice”, *ACM International Conference Proceeding Series*, v. 1, n. 1, pp. 400–422, 2023. doi: 10.1145/3593013.3594007.
- BEIGANG, F. “Yet Another Impossibility Theorem in Algorithmic Fairness”, *Minds and Machines*, v. 33, n. 4, pp. 715–735, 2023. ISSN: 15728641. doi: 10.1007/s11023-023-09645-x. Disponível em: <<https://doi.org/10.1007/s11023-023-09645-x>>.
- KAMIRAN, F., CALDERS, T. “Data preprocessing techniques for classification without discrimination”, *Knowledge and Information Systems*, v. 33, pp. 1–33, 2012. ISSN: 02193116. doi: 10.1007/s10115-011-0463-8.
- ZEMEL, R., WU, Y., SWERSKY, K., et al. “Learning Fair Representations”. In: Dasgupta, S., McAllester, D. (Eds.), *Proceedings of the 30th International Conference on Machine Learning*, v. 28, *Proceedings of Machine Learning Research*, pp. 325–333, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. Disponível em: <<https://proceedings.mlr.press/v28/zemel13.html>>.
- KAMIRAN, F., KARIM, A., ZHANG, X. “Decision Theory for Discrimination-Aware Classification”. In: *2012 IEEE 12th International Conference on Data Mining*, pp. 924–929, 2012. doi: 10.1109/ICDM.2012.45.
- CALDERS, T., VERWER, S. “Three naive Bayes approaches for discrimination-free classification”, *Data Mining and Knowledge Discovery*, v. 21, pp. 277–292, 2010. ISSN: 13845810. doi: 10.1007/s10618-010-0190-x.
- KAMISHIMA, T., AKAHO, S., ASOH, H., et al. “Fairness-aware classifier with prejudice remover regularizer”, *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, v. 7524 LNAI, n. PART 2, pp. 35–50, 2012. ISSN: 03029743. doi: 10.1007/978-3-642-33486-3.

- 1945 WOODWORTH, B., GUNASEKAR, S., OHANNESSIAN, M. I., et al. “Learning Non-Discriminatory Predictors”. In: Kale, S., Shamir, O. (Eds.), *Proceedings of the 2017 Conference on Learning Theory*, v. 65, *Proceedings of Machine Learning Research*, pp. 1920–1953. PMLR, 07–10 Jul 2017. Disponível em: <<https://proceedings.mlr.press/v65/woodworth17a.html>>.
- 1950 ZAFAR, M. B., VALERA, I., GOMEZ RODRIGUEZ, M., et al. “Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment”. In: *Proceedings of the 26th International Conference on World Wide Web*, WWW ’17, p. 1171–1180, Republic and Canton of Geneva, CHE, 2017b. International World Wide Web Conferences Steering Committee. ISBN: 9781450349130. doi: 10.1145/3038912.3052660. Disponível em: <<https://doi.org/10.1145/3038912.3052660>>.
- 1955 KEARNS, M., NEEL, S., ROTH, A., et al. “Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness”. In: Dy, J., Krause, A. (Eds.), *Proceedings of the 35th International Conference on Machine Learning*, v. 80, *Proceedings of Machine Learning Research*, pp. 2564–2572. PMLR, 10–15 Jul 2018. Disponível em: <<https://proceedings.mlr.press/v80/kearns18a.html>>.
- 1960 ADEL, T., VALERA, I., GHAMRAMANI, Z., et al. “One-Network Adversarial Fairness”, *Proceedings of the AAAI Conference on Artificial Intelligence*, v. 33, n. 01, pp. 2412–2420, Jul. 2019. doi: 10.1609/aaai.v33i01.33012412. Disponível em: <<https://ojs.aaai.org/index.php/AAAI/article/view/4085>>.
- 1965 XU, D., YUAN, S., ZHANG, L., et al. “FairGAN+: Achieving Fair Data Generation and Classification through Generative Adversarial Nets”, *Proceedings - 2019 IEEE International Conference on Big Data, Big Data 2019*, pp. 1401–1406, 2019. doi: 10.1109/BigData47090.2019.9006322.
- 1970 WEI, S., NIETHAMMER, M. “The fairness-accuracy Pareto front”, *Statistical Analysis and Data Mining*, v. 15, pp. 287–302, 6 2022. ISSN: 19321872. doi: 10.1002/SAM.11560.
- 1975 MERCIER, Q., POIRION, F., DÉSIDÉRI, J. A. “A stochastic multiple gradient descent algorithm”, *European Journal of Operational Research*, v. 271, pp. 808–817, 12 2018. ISSN: 03772217. doi: 10.1016/j.ejor.2018.05.064.

- 1980 HU, Z., XU, Y., TIAN, X. “Adaptive Priority Reweighting for Generalizing Fairness Improvement”. In: *International Joint Conference on Neural Networks, IJCNN 2023, Gold Coast, Australia, June 18-23, 2023*, pp. 1–8. IEEE, 2023. doi: 10.1109/IJCNN54540.2023.10191757. Disponível em: <<https://doi.org/10.1109/IJCNN54540.2023.10191757>>.
- 1985 ZHANG, B. H., LEMOINE, B., MITCHELL, M. “Mitigating Unwanted Biases with Adversarial Learning”. pp. 335–340. Association for Computing Machinery, Inc, 12 2018. ISBN: 9781450360128. doi: 10.1145/3278721.3278779.
- 1990 D’ALOISIO, G., D’ANGELO, A., DI MARCO, A., et al. “Debiaser for Multiple Variables to enhance fairness in classification tasks”, *Information Processing and Management*, v. 60, n. 2, pp. 103226, 2023. ISSN: 03064573. doi: 10.1016/j.ipm.2022.103226. Disponível em: <<https://doi.org/10.1016/j.ipm.2022.103226>>.
- 1995 LIU, T., WANG, H., WANG, Y., et al. “SimFair: A Unified Framework for Fairness-Aware Multi-Label Classification”, *Proceedings of the AAAI Conference on Artificial Intelligence*, v. 37, n. 12, pp. 14338–14346, 2023. doi: 10.1609/aaai.v37i12.26677. Disponível em: <<https://ojs.aaai.org/index.php/AAAI/article/view/26677>>.
- 2000 KIM, D., PARK, S., HWANG, S., et al. “Fair classification by loss balancing via fairness-aware batch sampling”, *Neurocomputing*, v. 518, pp. 231–241, 2023. ISSN: 0925-2312. doi: <https://doi.org/10.1016/j.neucom.2022.11.018>. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S0925231222013984>>.
- 2005 KHALILI, M. M., ZHANG, X., ABROSHAN, M. “Loss Balancing for Fair Supervised Learning”. In: Krause, A., Brunskill, E., Cho, K., et al. (Eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, v. 202, *Proceedings of Machine Learning Research*, pp. 16271–16290. PMLR, 2023. Disponível em: <<https://proceedings.mlr.press/v202/khalili23a.html>>.
- 2010 LIANG, Y., CHEN, C., TIAN, T., et al. “Fair classification via domain adaptation: A dual adversarial learning approach”, *Frontiers in Big Data*, v. 5, 2023. ISSN: 2624-909X. doi: 10.3389/fdata.2022.1049565. Disponível em: <<https://www.frontiersin.org/articles/10.3389/fdata.2022.1049565>>.
- ZHANG, T., ZHU, T., LI, J., et al. “Revisiting model fairness via adversarial examples”, *Knowl. Based Syst.*, v. 277, pp. 110777, 2023a. doi: 10.1016/

- 2015 J.KNOSYS.2023.110777. Disponível em: <<https://doi.org/10.1016/j.knosys.2023.110777>>.
- 2020 MOUSAVI, S. A., MOUSAVI, H., DANESHTALAB, M. “FARMUR: Fair Adversarial Retraining to Mitigate Unfairness in Robustness”. In: Abelló, A., Vassiliadis, P., Romero, O., et al. (Eds.), *Advances in Databases and Information Systems - 27th European Conference, ADBIS 2023, Barcelona, Spain, September 4-7, 2023, Proceedings*, v. 13985, *Lecture Notes in Computer Science*, pp. 133–145. Springer, 2023. doi: 10.1007/978-3-031-42914-9_10. Disponível em: <https://doi.org/10.1007/978-3-031-42914-9_10>.
- 2025 WEI, Z., WANG, Y., GUO, Y., et al. “CFA: Class-wise Calibrated Fair Adversarial Training”, *CoRR*, v. abs/2303.14460, 2023. doi: 10.48550/ARXIV.2303.14460. Disponível em: <<https://doi.org/10.48550/arXiv.2303.14460>>.
- 2030 CHEN, H., ZHU, T., ZHANG, T., et al. “Privacy and Fairness in Federated Learning: On the Perspective of Tradeoff”, *ACM Comput. Surv.*, v. 56, n. 2, pp. 39:1–39:37, 2024. doi: 10.1145/3606017. Disponível em: <<https://doi.org/10.1145/3606017>>.
- 2035 VUCINICH, S., ZHU, Q. “The Current State and Challenges of Fairness in Federated Learning”, *IEEE Access*, v. 11, pp. 80903–80914, 2023. doi: 10.1109/ACCESS.2023.3295412. Disponível em: <<https://doi.org/10.1109/ACCESS.2023.3295412>>.
- 2040 ZHANG, W., WEISS, J. C. “Longitudinal Fairness with Censorship”. In: *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pp. 12235–12243. AAAI Press, 2022. doi: 10.1609/AAAI.V36I11.21484. Disponível em: <<https://doi.org/10.1609/aaai.v36i11.21484>>.
- 2045 ZHANG, W., HERNANDEZ-BOUSSARD, T., WEISS, J. “Censored Fairness through Awareness”, *Proceedings of the AAAI Conference on Artificial Intelligence*, v. 37, n. 12, pp. 14611–14619, Jun. 2023b. doi: 10.1609/aaai.v37i12.26708. Disponível em: <<https://ojs.aaai.org/index.php/AAAI/article/view/26708>>.
- ZHANG, W., WEISS, J. C. “Fairness with censorship and group constraints”, *Knowl. Inf. Syst.*, v. 65, n. 6, pp. 2571–2594, 2023. doi: 10.1007/

- 2050 S10115-023-01842-5. Disponível em: <<https://doi.org/10.1007/s10115-023-01842-5>>.
- ZHANG, W., KIM, J., WANG, Z., et al. “Individual Fairness Guarantee in Learning with Censorship”, *CoRR*, v. abs/2302.08015, 2023c. doi: 10.48550/ARXIV.2302.08015. Disponível em: <<https://doi.org/10.48550/arXiv.2302.08015>>.
- 2055 PARETO, V. “Manuale di economica politica, societa editrice libraria”, *Manual of political economy*, v. 1971, 1906.
- SCHMUCKER, R., DONINI, M., PERRONE, V., et al. “Multi-objective multi-fidelity hyperparameter optimization with application to fairness”. In: *NeurIPS 2020 Workshop on Meta-learning*, 2020.
- 2060 LI, L., JAMIESON, K., DESALVO, G., et al. “Hyperband: A Novel Bandit-Based Approach to Hyperparameter Optimization”, *J. Mach. Learn. Res.*, v. 18, n. 1, pp. 6765–6816, jan 2017. ISSN: 1532-4435.
- GIAGKIOZIS, I., FLEMING, P. J. “Methods for multi-objective optimization: An analysis”, *Information Sciences*, v. 293, pp. 338–350, 2015. ISSN: 0020-0255. doi: <https://doi.org/10.1016/j.ins.2014.08.071>. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S0020025514009074>>.
- 2065 FRÉNAY, B., VERLEYSEN, M. “Classification in the presence of label noise: A survey”, *IEEE Transactions on Neural Networks and Learning Systems*, v. 25, n. 5, pp. 845–869, 2014. ISSN: 2162-2388. doi: 10.1109/TNNLS.2013.2292894.
- 2070 HICKEY, R. J. “Noise modelling and evaluating learning from examples”, *Artificial Intelligence*, v. 82, n. 1–2, pp. 157–179, apr 1996. ISSN: 0004-3702. doi: [http://dx.doi.org/10.1016/0004-3702\(94\)00094-8](http://dx.doi.org/10.1016/0004-3702(94)00094-8). Disponível em: <<http://www.sciencedirect.com/science/article/pii/0004370294000948>>.
- 2075 QUINLAN, J. R. “Induction of decision trees”, *Machine Learning 1986 1:1*, v. 1, n. 1, pp. 81–106, mar 1986. ISSN: 1573-0565. doi: 10.1007/BF00116251. Disponível em: <<https://link.springer.com/article/10.1007/BF00116251>>.
- 2080 PATRINI, G., ROZZA, A., MENON, A. K., et al. “Making deep neural networks robust to label noise: A loss correction approach”, *Proceedings - 30th IEEE*

- 2085 *Conference on Computer Vision and Pattern Recognition, CVPR 2017*,
v. 2017-Janua, pp. 2233–2241, 2017. doi: 10.1109/CVPR.2017.240.
- PATRINI, G., NIELSEN, F., NOCK, R., et al. “Loss Factorization, Weakly Supervised Learning and Label Noise Robustness”. In: *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML’16, p. 708–717. JMLR.org, 2016.
- 2090 LAMY, A., ZHONG, Z., VERMA, N., et al. “Noise-tolerant fair classification”,
Advances in Neural Information Processing Systems, v. 32, 1 2019. ISSN:
10495258. doi: 10.48550/arxiv.1901.10837. Disponível em: <<https://arxiv.org/abs/1901.10837v4>>.
- 2095 FOGLIATO, R., CHOULDECHOVA, A., G’SELL, M. “Fairness Evaluation in
Presence of Biased Noisy Labels”. In: Chiappa, S., Calandra, R. (Eds.),
Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics, v. 108, *Proceedings of Machine Learning Research*, pp. 2325–2336. PMLR, 26–28 Aug 2020. Disponível em:
<<https://proceedings.mlr.press/v108/fogliato20a.html>>.
- 2100 WANG, S., GUO, W., NARASIMHAN, H., et al. “Robust Optimization for
Fairness with Noisy Protected Groups”, *Advances in Neural Information Processing Systems*, v. 2020-December, feb 2020. ISSN: 10495258. doi:
10.48550/arxiv.2002.09343. Disponível em: <<https://arxiv.org/abs/2002.09343v3>>.
- 2105 MEHROTRA, A., CELIS, L. E. “Mitigating Bias in Set Selection with Noisy
Protected Attributes”, *FAccT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 237–248, nov
2021. doi: 10.48550/arxiv.2011.04219. Disponível em: <<https://arxiv.org/abs/2011.04219v2>>.
- 2110 CELIS, L. E., HUANG, L., KESWANI, V., et al. “Fair Classification with
Noisy Protected Attributes: A Framework with Provable Guarantees”. In: Meila, M., Zhang, T. (Eds.), *Proceedings of the 38th International Conference on Machine Learning*, v. 139, *Proceedings of Machine Learning Research*, pp. 1349–1361. PMLR, 18–24 Jul 2021. Disponível em:
2115 <<https://proceedings.mlr.press/v139/celis21a.html>>.
- PROST, F., AWASTHI, P., BLUMM, N., et al. “Measuring Model Fairness under Noisy Covariates: A Theoretical Perspective”. In: *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES ’21, p. 873–883,

- 2120 New York, NY, USA, 2021. Association for Computing Machinery. ISBN: 9781450384735. doi: 10.1145/3461702.3462603. Disponível em: <<https://doi.org/10.1145/3461702.3462603>>.
- 2125 GHAZIMATIN, A., KLEINDESSNER, M., RUSSELL, C., et al. “Measuring Fairness of Rankings under Noisy Sensitive Information”. In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT ’22, p. 2263–2279, New York, NY, USA, 2022. Association for Computing Machinery. ISBN: 9781450393522. doi: 10.1145/3531146.3534641. Disponível em: <<https://doi.org/10.1145/3531146.3534641>>.
- 2130 ZHANG, T., ZHU, T., LI, J., et al. “Fairness in Semi-Supervised Learning: Unlabeled Data Help to Reduce Discrimination”, *IEEE Transactions on Knowledge and Data Engineering*, v. 34, n. 4, pp. 1763–1774, apr 2022. ISSN: 15582191. doi: 10.1109/TKDE.2020.3002567.
- 2135 WANG, J., CRUZ, U. S. C. S., LIU, U. Y., et al. “Fair classification with group-dependent label noise”, *FAccT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 526–536, 3 2021. doi: 10.1145/3442188.3445915. Disponível em: <<https://doi.org/10.1145/3442188.3445915>>.
- 2140 GHOSH, A., KVITCA, P., WILSON, C. “When Fair Classification Meets Noisy Protected Attributes”. In: *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES ’23, p. 679–690, New York, NY, USA, 2023. Association for Computing Machinery. ISBN: 9798400702310. doi: 10.1145/3600211.3604707. Disponível em: <<https://doi.org/10.1145/3600211.3604707>>.
- 2145 BELLAMY, R. K. E., DEY, K., HIND, M., et al. “AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias”. out. 2018. Disponível em: <<https://www.ibm.com/opensource/open/projects/ai-fairness-360/>>.
- 2150 LI, M., SOLTANOLKOTABI, M., OYMAK, S. “Gradient Descent with Early Stopping is Provably Robust to Label Noise for Overparameterized Neural Networks”. In: Chiappa, S., Calandra, R. (Eds.), *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, v. 108, *Proceedings of Machine Learning Research*, pp. 4313–4324. PMLR, 26–28 Aug 2020. Disponível em: <<https://proceedings.mlr.press/v108/li20j.html>>.

- 2155 KINGMA, D. P., BA, J. “Adam: A Method for Stochastic Optimization”. In: Bengio, Y., LeCun, Y. (Eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. Disponível em: <<http://arxiv.org/abs/1412.6980>>.
- 2160 JIANG, H., NACHUM, O. “Identifying and correcting label bias in machine learning”. In: *International Conference on Artificial Intelligence and Statistics*, pp. 702–712. PMLR, 2020.
- MROUEH, Y., OTHERS. “Fair Mixup: Fairness via Interpolation”. In: *International Conference on Learning Representations*, 2021.
- 2165 ROH, Y., LEE, K., WHANG, S. E., et al. “Fairbatch: Batch selection for model fairness”, 2021.
- LI, L., JAMIESON, K., ROSTAMIZADEH, A., et al. “Hyperband: A Novel Bandit-Based Approach to Hyperparameter Optimization”. 2018. Disponível em: <<http://jmlr.org/papers/v18/16-558.html>>.
- 2170 BERGSTRA, J., BARDENET, R., BENGIO, Y., et al. “Algorithms for Hyper-Parameter Optimization”. In: Shawe-Taylor, J., Zemel, R., Bartlett, P., et al. (Eds.), *Advances in Neural Information Processing Systems*, v. 24. Curran Associates, Inc., 2011. Disponível em: <https://proceedings.neurips.cc/paper_files/paper/2011/file/86e8f7ab32cfd12577bc2619bc635690-Paper.pdf>.
- 2175 MORALES-HERNÁNDEZ, A., VAN NIEUWENHUYSE, I., ROJAS GONZALEZ, S. *A survey on multi-objective hyperparameter optimization algorithms for machine learning*, v. 56. New York, NY, USA, Springer Netherlands, 2023. ISBN: 0123456789. doi: 10.1007/s10462-022-10359-2. Disponível em: <<https://doi.org/10.1007/s10462-022-10359-2>>.
- 2180 CHICCO, D., JURMAN, G. “The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation”, *BMC genomics*, v. 21, n. 1, pp. 1–13, 2020.
- 2185 DROR, R., SHLOMOV, S., REICHART, R. “Deep Dominance - How to Properly Compare Deep Neural Models”. In: Korhonen, A., Traum, D. R., Màrquez, L. (Eds.), *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28-August 2, 2019, Volume 1: Long Papers*, pp. 2773–2785. Association for

Computational Linguistics, 2019. doi: 10.18653/v1/p19-1266. Disponível em: <<https://doi.org/10.18653/v1/p19-1266>>.

2190 BECKER, B., KOHAVI, R. “Adult”. UCI Machine Learning Repository, 1996. Disponível em: <<https://archive.ics.uci.edu/dataset/2/adult>>.

HOFMANN, H. “Statlog (German Credit Data)”. UCI Machine Learning Repository, 1994. Disponível em: <<https://archive.ics.uci.edu/dataset/144/statlog+german+credit+data>>.

2195 S. MORO, P. R., CORTEZ, P. “Bank Marketing”. UCI Machine Learning Repository, 2012. Disponível em: <<https://archive.ics.uci.edu/dataset/222/bank+marketing>>.

JEFF LARSON, SURYA MATTU, L. K., ANGWIN, J. “COMPAS Dataset”. ProPublica, 2016. Disponível em: <<https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm>>.

2200