COPPE UFRJ
Instituto Alberto Luiz Coimbra de Pós-Graduação e Pesquisa de Engenharia

THESIS TITLE

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 Braida do Carmo

Rio de Janeiro
Maio de 2024

THESIS TITLE

Ygor de Mello Canalli

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
Filipe Braida do Carmo

Aprovada por: Prof. Leandro Guimarães Marques Alvim
Prof. Nome do Segundo Examinador Sobrenome
Prof. Nome do Terceiro Examinador Sobrenome
Prof. Nome do Quarto Examinador Sobrenome
Prof. Nome do Quinto Examinador Sobrenome

RIO DE JANEIRO, RJ – BRASIL
MAIO DE 2024

de Mello Canalli, Ygor

Thesis Title/Ygor de Mello Canalli. – Rio de Janeiro: UFRJ/COPPE, 2024.

XI, 70 p.: il.; $29,7$cm.

Orientadores: Geraldo Zimbrão da Silva

Filipe Braida do Carmo

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

Referências Bibliográficas: p. 58 – 70.

1. Aprendizado de Máquina Justo. 2. Ruído em Aprendizado de Máquina. 3. Terceira palavra-chave. 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.)

LABEL NOISE TECHNIQUES TO FAIRNESS IN MACHINE LEARNING

Ygor de Mello Canalli

Maio/2024

Orientadores: Geraldo Zimbrão da Silva
Filipe Braida do Carmo

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.)

THESIS TITLE

Ygor de Mello Canalli

May/2024

Advisors: Geraldo Zimbrão da Silva
Filipe Braida do Carmo

Department: Systems Engineering and Computer Science

In this work, we present ...

Contents

List of Figures

List of Tables

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 hyperarameters, 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.

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. To the best of our knowledge, this work is the first of its kind to apply label noise techniques directly within the model to mitigate unfairness.

1.1 Contextualization

1.2 Objectives

1.3 Contributions

1.4 Results summary

1.5 Thesis structure

Chapter 2

Fair Machine Learning Review

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 array of societal domains, including criminal justice, healthcare, finance, and employment, the imperative 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.

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 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 deployers of ML systems to be answerable for the outcomes of these systems (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 enhance accountability in ML, like auditing, 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 stakeholders, including end-users, regulators, and the broader public, to understand how ML systems make decisions. Transparency is crucial for building trust, 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 isses as the trade-off between explainability and model performance, and discussing emerging approaches to enhance interpretability without sacrificing effectiveness.

The triad of Fairness, Accountability, and Transparency (FAT) alongwith data privacy forms the cornerstone of Trustworty 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 understayfoster environments where AI systems can be scrutinized, understood, and corrected, thereby aligning their functionality with societal norms and values.

In this context of Trustworty 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 crucial for building AI systems that are beneficial and do not cause 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 legitimised access to data.

Transparency the data, system and AI business models should be transparent. Traceability mechanisms can help achieving this. Moreover, AI systems and 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 vulnerable 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 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 enables the assessment of algorithms, data and design processes plays a key role therein, especially in critical applications. Moreover, adequate an 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 the aspect of fairness and negative social bias. Fairness is not only crucial for the development of just and equitable technological solutions but also imperative for maintaining the legitimacy and acceptability of AI systems in diverse societal contexts. In delving into the multifaceted dimensions of fairness, this text aims to unpack the theoretical underpinnings, practical challenges, and potential pathways to achieving fairer AI systems, thereby contributing to the broader discourse on ethical AI.

2.1 Sources and types of algorithmic unfairness

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 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 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;

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 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 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;

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 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 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 mapinupation;

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 Biasese:

- **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 delves into the different types of discrimination that arise from these biases. Understanding these types of discrimination is crucial 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. 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 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 refered 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;

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 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 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) and CATON e HAAS (2023), providing a comprehensive overview for understanding and navigating the multifaceted dimensions of fairness in ML systems. Initially, the discourse will delve into general considerations and intuitive aspects of fairness, setting the stage for a deeper understanding. This preliminary discussion is crucial as it lays the groundwork for grasping 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, it's crucial to understand 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 entails 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 underscores a crucial differentiation within fairness metrics: some are predicated solely on predicted values (such as Demographic Parity), while others derive from the comprehensive landscape of the confusion matrix (Table 2.1), incorporating true 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, offering 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

<table>
<tr><td colspan="2"></td><th colspan="2">Predicted</th></tr>
<tr><td colspan="2"></td><td>Positive</td><td>Negative</td></tr>
<tr><th rowspan="2">Actual</th><td>Positive</td><td>TP</td><td>FN</td></tr>
<tr><td>Negative</td><td>FP</td><td>TN</td></tr>
</table>

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 offer complementary insights into the performance of the classifier. These derived metrics, such as Positive Predictive Value (PPV), False Discovery Rate (FDR), and others, leverage the basic elements of the confusion matrix to quantify the reliability of the predictions in 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.*

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.*

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.*

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.*

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.*

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.*

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.*

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.*

As we transition from foundational metrics that directly stem from the confusion matrix, such as PPV and TPR, we now delve into 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.

Definition 9 (Accuracy (Acc.))**.** *Probably the most widely usaed performence 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 can be expressed as*

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

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.

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.

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.*

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 when this metric is lower.

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

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 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.

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

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.*

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

Predictive Equality according Equation 2.3.

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

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.

$$\frac{1}{2}\begin{bmatrix} |P(\hat{Y}|A=0,Y=1)-P(\hat{Y}|A=1,Y=1)|+ \\ |P(\hat{Y}|A=0,Y=0)-P(\hat{Y}|A=1,Y=0)| \end{bmatrix} \tag{2.4}$$

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, asessing 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, it becomes crucial to examine the potential shortcomings and complications that can arise when implementing fairness metrics across different demographic groups.

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,

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

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.

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.

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.

While the concept of counterfactual fairness is compelling, its implementation poses significant challenges KASIRZADEH e SMART (2021). Building accurate causal models that reflect the true causal relationships in the data is non-trivial and requires deep domain knowledge. Also, access to comprehensive data that sufficiently captures the causal dependencies is crucial, 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.

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 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;

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 Reweighing 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.

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

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 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 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 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 adhere to fairness constraints like Statistical Parity (Definition 13), Equal Opportunity (Definition 14), and Equalized 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 ilustrating a scenario where the refered fairness gerrymandering occurs.

> *Imagine a setting with two binary features, corresponding to race (say black and white) and gender (say male and female), both ofwhich 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 ifone 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 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 oftwo attributes.*

Their approach involves defining fairness not only for a small number of predefined groups but for exponentially or infinitely many subgroups defined by a structured class of functions over the protected attributes. 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 inherent in this approach are addressed by drawing connections to weak agnostic learning, thereby leveraging practical machine learning heuristics for effective auditing and learning in practice.

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 Reweighing HU *et al.* (2023) method introduces a systematic approach to enhance the 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 Reweighing method ensures that the classifier learns to make fairer decisions that generalize well to unseen data, addressing the limitations of traditional reweighing 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, enhancing fairness on both finetuned pre-trained models and newly traisned models. This technique can be classified as an hybrid approach, performing the traditional pre-processing instance reweighing 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 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 **??**.

2.4 Fairness and multi-objective optimization

Here 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 2.6, 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).

$$\begin{aligned} \max\ & (\rho(\lambda), \varphi(\lambda)) \\ \text{subject to}\ & \lambda \in \Lambda \end{aligned} \tag{2.6}$$

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 2.6 according Equation 2.7. 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.

$$\arg\max_{\lambda \in \Lambda} G(\lambda) = (\rho(\lambda), \varphi(\lambda)) \tag{2.7}$$

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 stochastic multi-gradient to Multi-Objective Optimization.

Another remarkable use of MOO in Fair Machine Learning is to perform a Fair Hyperparameter Optimization, which can offer a model agnostic approach with flexibility to apply in multiple machine learning pipelines. A time-efficient Baysian Optimization approach can be found in SCHMUCKER *et al.* (2020), combining 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 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 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 2.8, where 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.

$$G(\lambda) = \alpha \cdot \rho(\lambda) + (1 - \alpha) \cdot \varphi(\lambda) \tag{2.8}$$

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 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 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.

Chapter 3

Fair Transition Loss

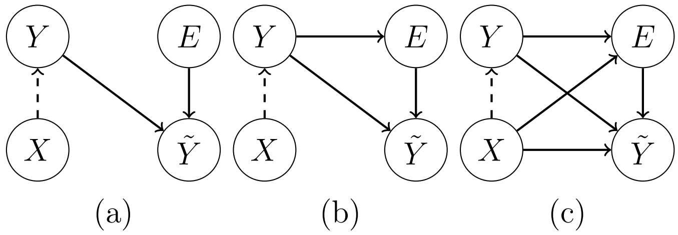
3.1 Preliminaries

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.

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 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.

Many label noise robustness methods can be found on literature, in this work we 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-

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.



rection techniques considers a NAR label noise, which is described by a transition matrix T such as

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

where $\mathcal{Y} = \{y_1, y_2, \ldots, y_c\}$ is the set of all possible class labels. Transition matrix 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.

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

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 linear combination of the loss values for each observed label, using coefficients that depends on the probability that each observed label reflects the true class. Intuitively, 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 back-propagation algorithm, making it flexible to include within different training techniques and data pipelines.

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

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 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 the need of matrix inversion.

Now we delve into 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.

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 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). 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 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.

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 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 deals with fairness under semi-supervised settings considering censored data, that is, for some individuals the class label is not available due 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 uses 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 privileged group ($A = 0$). Simultaneously, the negative class ($Y = 0$) could be more frequently flipped into positives observed labels ($\tilde{Y} = 1$) within privileged/unprotected group ($A = 0$). This scenario could lead to a undetected higher false negative rate to protected group and higher false positive rate to privileged group. In this case the Noisy Not at Random data would be a source of negative social bias.

As refered 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 (FRÉNAY 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.

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

3.2 Proposal

We propose a novel fair classification method inspired by techniques used for classification in the presence of label noise. By leveraging the features of certain label noise methods that redistribute probabilities for unbalanced noise across classes, our approach re-weights prediction probabilities to reduce disparities in favorable and unfavorable outcomes across social groups.

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 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 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, 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 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 not only one transition matrix, but a different matrix to each group of sensitive features. In this scenario, each group of sensitive feature have its own correction, with its own rates for each 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 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

$$T_{a,i,j} = P(\tilde{Y} = y_j | Y = y_i, A = a), \tag{3.4}$$

where label space $\mathcal{Y} = \{y_1, y_2, \ldots, y_c\}$, 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 inversible link function $\psi : \Delta^{c-1} \to \mathbb{R}^c$, 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 $\ell_\psi : \mathcal{Y} \times \mathbb{R}^c \to \mathbb{R}$ if it can be written as a decomposition of ψ^{-1}, that is,

$$\ell_\psi(Y, h(X)) = \ell(Y, \psi^{-1}(h(X)), \tag{3.5}$$

where $h : \mathcal{X} \to \mathbb{R}^c$ is a standard artificial neural network with multiple layers using activation functions, and $h(X)$ is the output o 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

$$\arg\min_h \mathbb{E}_{X,Y} \ell_\psi(Y, h(X|A = a)) = \psi(P(Y|X, A = a)). \tag{3.6}$$

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*

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

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:

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

Proof. First notice that:

$$\begin{aligned}\mathrm{FTL}_\psi(Y, h(X|A = a)) &= \ell(Y, T_a^\top \psi^{-1}(h(X|A = a))) \\ &= \ell_\phi(Y, h(X|A = a)),\end{aligned} \tag{3.7}$$

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 3.7 is proper composite with link ϕ. Finally, from Equation 3.6, the loss minimizer over the unfair distribution is

$$\begin{aligned}
\arg\min_h \mathbb{E}_{X,\tilde{Y}} \ell_\phi(Y, h(X|A=a)) &= \phi(P(\tilde{Y}|X, A=a)) &(3.8)\\
&= \psi((T_a^{-1})^\top)P(\tilde{Y}|X, A=a) &(3.9)\\
&= \psi(P(Y|X, A=a)), &(3.10)
\end{aligned}$$

that proves the Theorem by Equation 3.6 once again. □

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 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

$$\mathrm{FTL}(P(\tilde{Y}|X)) = (1 - A) \cdot \ell(T_0^\top P(\tilde{Y}|X)) + A \cdot \ell(T_1^\top P(\tilde{Y}|X)), \tag{3.11}$$

which in a standard batch learning, consists in alternating the transition matrix 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 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

$$T_0 = \begin{bmatrix} 1 - d_0 & d_0 \\ p_0 & 1 - p_0 \end{bmatrix}, \; T_1 = \begin{bmatrix} 1 - d_1 & d_1 \\ p_1 & 1 - p_1 \end{bmatrix}, \tag{3.12}$$

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 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 2.6. Here, the hyperparameter configuration is $\lambda = (d_0, p_0, d_1, p_1)$. 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 $\varphi(\lambda)$.

$$G(\lambda) = \rho(\lambda) - \varphi(\lambda). \tag{3.13}$$

Following some MOO approaches to fair machine learning, we will use a linear scalarization setup to define the optimization metric (PETROVIĆ *et al.*, 2021; 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 3.13 to maximize without the parameter α, i.e., giving same importance to fairness and performance. In Equation 3.13 we establish a simple objective to optimize, but one might need to consider a different 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 Debiasing (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 output, 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 Adversarial Debiasing classifier the hidden size is 100 units. Additionally, we compare the 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.

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 deliver better solutions to multi-objective hyperparameter optimization in the same number of trials than conventional approaches like Grid Search and Random Search (MORALES-HERNÁNDEZ *et al.*, 2023). At each trial 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 3.13. This same experimental procedure can be adapted to utilize other hyperparameter tuning algorithms such as FairRandom Search, Fair TPE, and Fairband (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 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 3.13 can be achieved with different performance and fairness metrics, we compare the proposed method with other relevant in-processing techniques from literature in different optimization scenarios. For predictive performance, we evaluate not only Accuracy (Acc.) but also the Mathews Correlation Coefficient (MCC), which has advantages over F1 score and Accuracy in binary classification evaluation (CHICCO e JURMAN, 2020), where 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 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.

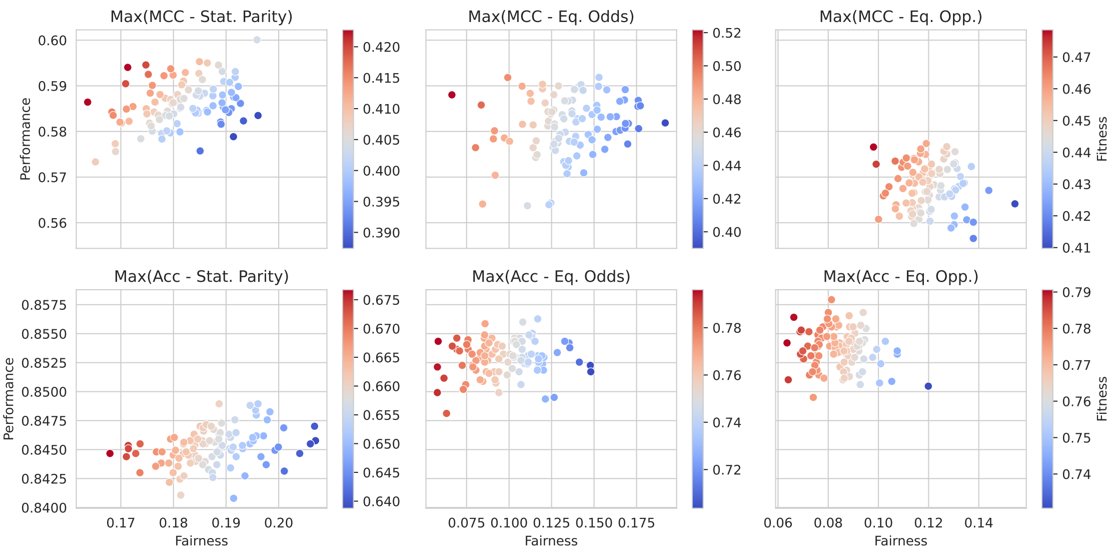
Table 3.1 presents the methods hyperparameters along with their corresponding search ranges or options. While each method may possess a varying number of hyperparameters and range sizes, all are optimized under the same conditions and number of configurations to guarantee a balanced comparison. In Figure 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

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 *et al.*, 2012)	η	[0.0, 50.0]
Adversarial Debasing (ZHANG *et al.*, 2018)	α	[0.0, 1.0]
Gerry Fair Classifier (KEARNS *et al.*, 2018)	C	[0.0, 20.0]
	γ	{0.1, 0.01, 0.001}
Adaptive Priority Reweighting (HU *et al.*, 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]

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 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 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.

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 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 suitable 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, 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 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 Recidivism* (JEFF LARSON e ANGWIN, 2016). We use the dataset readers available in the AI Fairness 360 toolkit (BELLAMY *et al.*, 2018) with its standard configurations. Instances with missing data are removed.

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

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 Recidivsm	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

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).

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

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 Appendix **??** 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.

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

Objective	d_0 Priv. Dem.	p_0 Priv. Prom.	d_1 Prot. Dem.	p_1 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

When optimizing for MCC and Statistical Parity, there's a notable high value for protected promotion. This value is compatible with the high corresponding fairness metric for this dataset, approximately 0.19 without correction. This enhances 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.

There is a remarkable difference between hyperparameters found through optimizing 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 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 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, 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 regularized logistic regression, it is a smaller model than FTL, which can explain the also smaller variance.

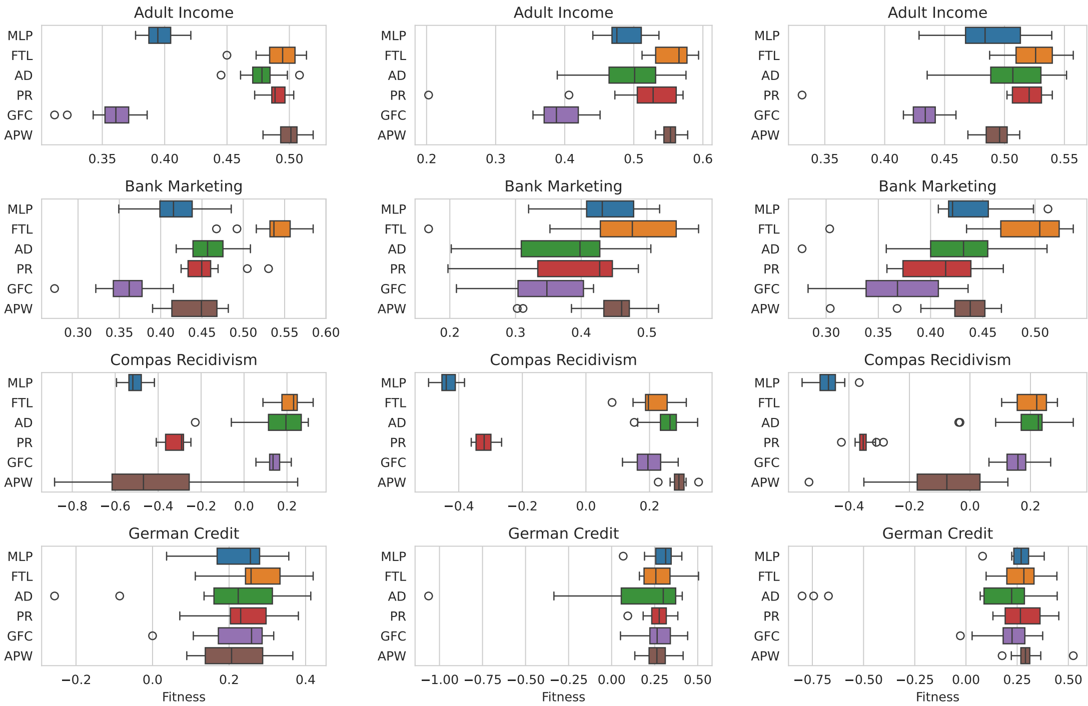
When comparing the optimization results for MCC and Equal Opportunity, FTL consistently outperforms its counterparts in most scenarios. Here we have only one discrepancy, as APW presents slightly advantage over FTL on COMPAS dataset. Also, on German Credit dataset all methods achieved similar results. 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 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 limited 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 persists across subsequent classification tasks involving this dataset. Interestingly, also with the exception of AD, we notice that the variance in results for most optimization 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.

When optimizing for MCC and Equalized Odds, we find that the results are

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

(a) Statistical Parity (b) Equal Opportunity (c) Equalized Odds



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 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 fairness metrics evaluated, it's evident that FTL consistently delivers superior results compared 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 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 performance could be due to an insufficiently large model to navigate such a highdimensional 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. 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 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 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 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.

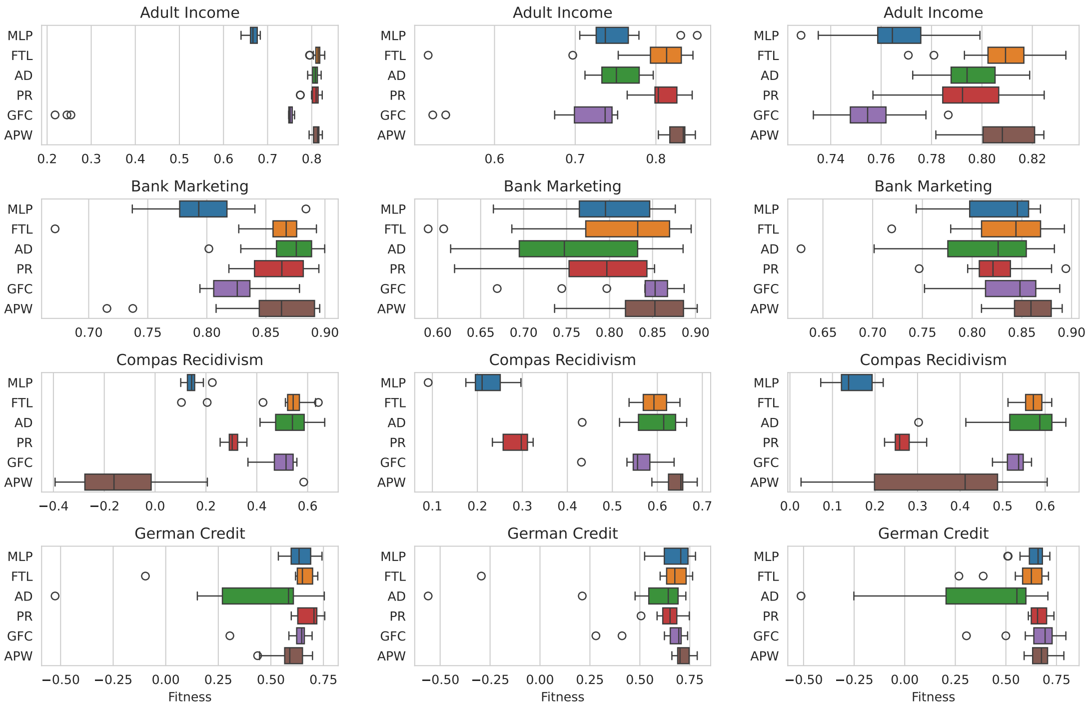
Fair Transition Loss consistently demonstrates effective bias mitigation, it does so by absorbing the nuances of the dataset and the fairness metric through its transition matrices, resulting in stochastically superior fitness values in a significant number of scenarios. Additionally, the method has 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 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 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 3.12) during the FTL training, keeping all other FTL hyperparameters

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

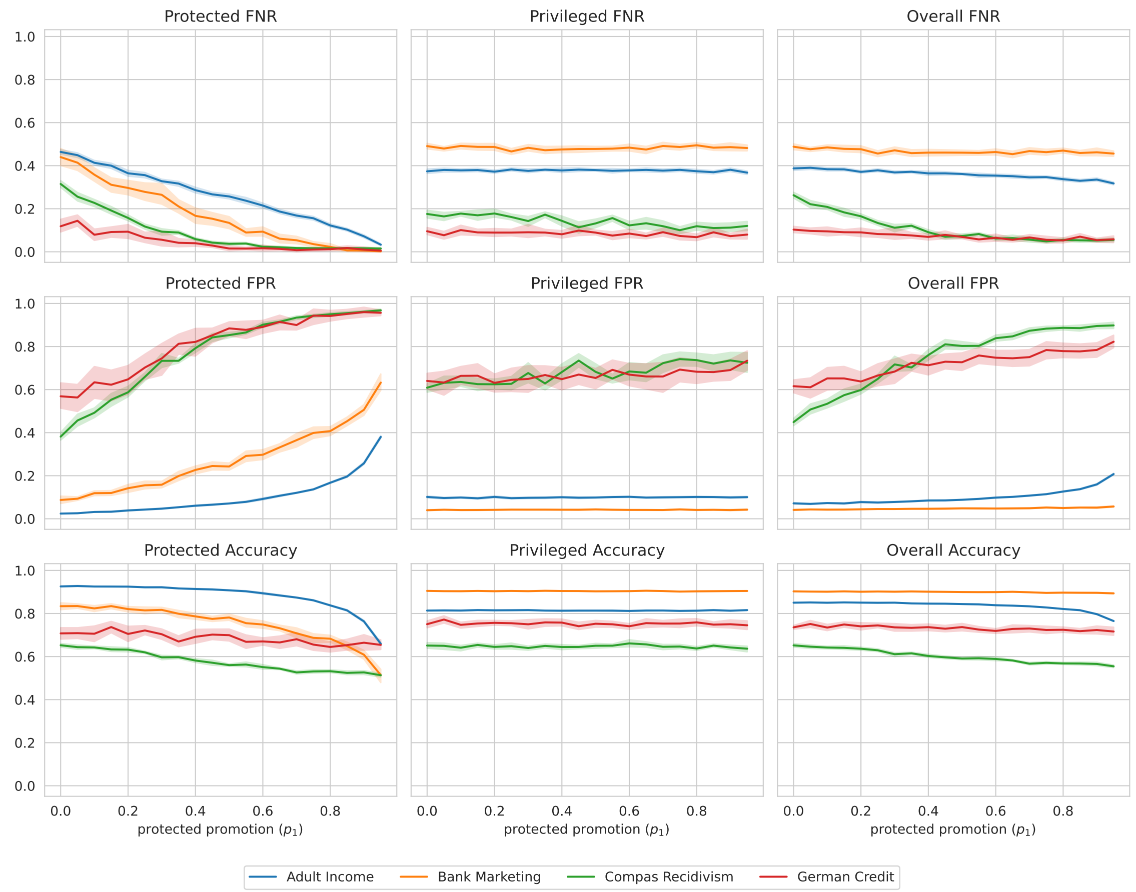
(a) Statistical Parity (b) Equal Opportunity (c) Equalized Odds



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 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

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



model.

In this appendix we present complete results with mean fitness (Equation 3.13), performance and fairness across multiple resampling, to a proper trade-off comparison. To each objective function and dataset methods results are ordered from higher to lower fitness mean, with corresponding standard deviation presented between parenthesis. Results corresponding each optimization scenario can be found on tables A.1, A.2, A.3, A.4, A.5, and A.6.

Table A.1: Complete results optimizing MCC and Statistical Parity.

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

Table A.2: Complete results optimizing MCC and Equal Opportunity.

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

Table A.3: Complete results optimizing MCC and Equalized Odds.

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

Table A.4: Complete results optimizing Accuracy and Statistical Parity.

Dataset	Method	Fitness	Accuracy	Stat. Parity
Adult Income	Fair Transition Loss	0.814(±0.01)	0.828(±0.01)	0.014(±0.01)
	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)
	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)
	Gerry Fair Classifier	0.651(±0.21)	0.721(±0.07)	0.070(±0.14)
Bank Marketing	Adversarial Debiasing	0.869(±0.03)	0.901(±0.00)	0.031(±0.02)
	Prejudice Remover	0.860(±0.02)	0.898(±0.00)	0.038(±0.02)
	Fair Transition Loss	0.854(±0.05)	0.889(±0.01)	0.035(±0.05)
	Adaptive Priority Reweighting	0.851(±0.06)	0.900(±0.00)	0.049(±0.06)
	Gerry Fair Classifier	0.824(±0.02)	0.895(±0.00)	0.071(±0.02)
	Standard MLP (baseline)	0.799(±0.04)	0.902(±0.00)	0.103(±0.03)
COMPAS Recidivism	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)
	Adaptive Priority Reweighting	−0.105(±0.26)	0.584(±0.03)	0.689(±0.23)
German Credit	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)
	Gerry Fair Classifier	0.621(±0.09)	0.712(±0.12)	0.090(±0.04)
	Fair Transition Loss	0.616(±0.20)	0.715(±0.06)	0.098(±0.17)
	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)

Table A.5: Complete results optimizing Accuracy and Equal Opportunity.

Dataset	Method	Fitness	Accuracy	Eq. Opp.
Adult Income	Adaptive Priority Reweighting	0.808(±0.01)	0.837(±0.00)	0.029(±0.01)
	Fair Transition Loss	0.808(±0.02)	0.842(±0.01)	0.034(±0.02)
	Adversarial Debiasing	0.796(±0.01)	0.849(±0.00)	0.052(±0.01)
	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)
	Gerry Fair Classifier	0.756(±0.01)	0.788(±0.03)	0.032(±0.04)
Bank Marketing	Adaptive Priority Reweighting	0.858(±0.02)	0.897(±0.00)	0.039(±0.03)
	Gerry Fair Classifier	0.837(±0.04)	0.895(±0.00)	0.058(±0.04)
	Fair Transition Loss	0.833(±0.05)	0.892(±0.01)	0.059(±0.05)
	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)
	Adversarial Debiasing	0.807(±0.07)	0.902(±0.00)	0.095(±0.07)
COMPAS Recidivism	Fair Transition Loss	0.572(±0.03)	0.631(±0.04)	0.059(±0.03)
	Adversarial Debiasing	0.553(±0.09)	0.669(±0.01)	0.116(±0.09)
	Gerry Fair Classifier	0.530(±0.03)	0.637(±0.04)	0.107(±0.05)
	Adaptive Priority Reweighting	0.356(±0.18)	0.643(±0.02)	0.287(±0.18)
	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)
	Prejudice Remover	0.664(±0.05)	0.748(±0.02)	0.084(±0.04)
	Gerry Fair Classifier	0.662(±0.12)	0.719(±0.12)	0.057(±0.07)
	Standard MLP (baseline)	0.638(±0.06)	0.738(±0.04)	0.101(±0.05)
	Fair Transition Loss	0.599(±0.12)	0.711(±0.05)	0.112(±0.11)
	Adversarial Debiasing	0.368(±0.38)	0.685(±0.10)	0.317(±0.30)

Table A.6: Complete results optimizing Accuracy and Equalized Odds.

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

Chapter 4

Chatterjee Redlining Penalty

4.1 Preliminaries

The Pearson correlation coefficient, denoted by $\rho_{X,Y}$, is a measure of the linear relationship between two variables X and Y. It is defined in Equation 4.1, where $\text{Cov}(X, Y)$ is the covariance of X and Y, while σ_X and σ_Y are the standard deviations of X and Y, respectively. 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.

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

Spearman's rank correlation coefficient, denoted by ρ_s, measures the strength and direction of the monotonic relationship between two ranked variables. It is defined as the Pearson correlation coefficient between the ranked variables, as described in Equation 4.2, where $\text{rank}(X)$ and $\text{rank}(Y)$ are the ranks of X and Y respectively. Using the notation in Pearson correlation coefficient, $\text{Cov}(\text{rank}(X), \text{rank}(Y))$ is the covariance of the rank variables, while $\sigma_{\text{rank}(X)}$ and $\sigma_{\text{rank}(Y)}$ are the standard deviations of the ranks of X and Y, respectively. As like Pearson correlation coefficient, Spearman's rank 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.

$$\rho_s = \rho_{\text{rank}(X),\text{rank}(Y)} = \frac{\text{Cov}(\text{rank}(X), \text{rank}(Y))}{\sigma_{\text{rank}(X)} \sigma_{\text{rank}(Y)}} \tag{4.2}$$

Chatterjee's correlation coefficient, denoted by ξ, is designed to measure the degree of dependence between two variables without assuming any specific type of relationship. Given a dataset (X, Y) with n pairs, the coefficient is defined as:

$$\xi_n(X, Y) := 1 - \frac{3\sum_{i=1}^{n-1} |r_{i+1} - r_i|}{n^2 - 1} \tag{4.3}$$

where r_i is the rank of Y_i in the ordered sequence of Y values corresponding to the sorted X values. This coefficient ranges from 0 to 1, where 0 indicates independence and 1 indicates a perfect functional relationship. For the general case with ties, a more complex formula involving additional terms to handle the ties is used.

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, 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:

$$R(\mathbf{W}^{(l)}) = \frac{1}{2} \sum_{i=1}^{d_l} \sum_{j=1}^{h_l} \left(W_{ij}^{(l)} \right)^2, \tag{4.4}$$

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.

The total regularization term for the entire network, considering all layers, is:

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

where L is the total number of layers in the network.

The total loss function L for the MLP, incorporating the L2 regularization term, is defined as:

$$L = L_0 + \alpha R(\mathbf{W}), \tag{4.6}$$

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

By adding this regularization term, the optimization process not only aims to minimize the primary loss L_0 but also to keep the weights small, thereby helping to reduce 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 Proposal

The Chatterjee Redlining Penalty is defined as a regularization term that penalizes the weights associated with features that are highly correlated with the sensitive attribute. This penalty term is incorporated into the loss function of the neural network in order to produce fairer predictions. Consider a dataset $X \in \mathbb{R}^{n\times d}$ where n represents the number of instances 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, $\lambda \in \mathbb{R}^d$ is a vector representing the regularization strengths for each feature, and λ is a scalar that controls the overall strength of the regularization.

The regularization term $R(\mathbf{W}^{(1)})$ applied to the weight matrix $\mathbf{W}^{(1)}$ of the first hidden layer is defined in Equation 4.8

$$R(\mathbf{W}^{(1)}) = \sum_{i=1}^{d} \xi_n(X_i,\, A) \sum_{j=1}^{h} (W_{ij}^{(1)})^2, \tag{4.7}$$

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

The total loss function L for the multilayer perceptron (MLP), incorporating the sensitive-feature-specific L_2 regularization, is given by Equation 4.9

$$L = L_0 + \lambda\, R(\mathbf{W}^{(1)}), \tag{4.8}$$

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

4.3 Experimental setup

4.4 Results and discussion

Chapter 5

Conclusions

In this study, we present Fair Transition Loss, a novel in-processing technique for addressing fair classification problems. It leverages concepts from label noise robustness to mitigate social bias against underprivileged groups. We delve into 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 maintaining acceptable predictive performance levels. We benchmark this approach and compare to proeminent 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 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 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.

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 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.CRUZ *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 relying on hyperparameter tuning techniques.

With this work, we hope to establish Fair Transition Loss as a valuable 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.

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`>.

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.

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.

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`>.

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.

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.

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`>.

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.

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`>.

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.

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.

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>.

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>.

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.

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/>.

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>.

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.

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.

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>.

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>.

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>.

ADEL, T., VALERA, I., GHAHRAMANI, 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>.

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.

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.

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.

HU, Z., XU, Y., TIAN, X. "Adaptive Priority Reweighing 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>.

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.

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>.

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>.

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>.

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>.

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/

J.KNOSYS.2023.110777. Disponível em: <https://doi.org/10.1016/j.knosys.2023.110777>.

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>.

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>.

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>.

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>.

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>.

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/

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>.

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.

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>.

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: 21622388. doi: 10.1109/TNNLS.2013.2292894.

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. Disponível em: <http://www.sciencedirect.com/science/article/pii/0004370294000948>.

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>.

PATRINI, G., ROZZA, A., MENON, A. K., et al. "Making deep neural networks robust to label noise: A loss correction approach", *Proceedings - 30th IEEE*

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.

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`>.

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`>.

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`>.

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`>.

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: <`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, 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>.

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>.

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.

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>.

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>.

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/>.

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>.

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>.

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.

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.>.

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>.

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>.

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.

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>.

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>.

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>.