Using a Fairness-Utility Trade-off Metric to Systematically Benchmark Non-Generative Fair Adversarial Learning Strategies

Luiz Fernando Fonsêca Pinheiro de Lima

CENTRO DE INFORMÁTICA
UNIVERSIDADE FEDERAL DA PARAÍBA

João Pessoa - PB

2022

Luiz Fernando Fonsêca Pinheiro de Lima

Using a Fairness-Utility Trade-off Metric to Systematically Benchmark Non-Generative Fair Adversarial Learning Strategies

Dissertation presented to the Graduate Program in Informatics - Master's level, from the Informatics Center of the Federal University of Paraíba.

Advisors:
Clauirton de Albuquerque Siebra
Danielle Rousy Dias Ricarte

João Pessoa - PB

2022

Catalogação na publicação
Seção de Catalogação e Classificação

L732u Lima, Luiz Fernando Fonsêca Pinheiro de.
Using a Fairness-Utility Trade-off Metric to Systematically Benchmark Non-Generative Fair Adversarial Learning Strategies / Luiz Fernando Fonsêca Pinheiro de Lima. - João Pessoa, 2022.
108 f. : il.

Orientação: Clauirton de Albuquerque Siebra.
Coorientação: Danielle Rousy Dias Ricarte.
Dissertação (Mestrado) - UFPB/CI.

1. Aprendizado adversário. 2. Aprendizado de máquina. 3. Benchmark. 4. Trade-off. I. Siebra, Clauirton de Albuquerque. II. Ricarte, Danielle Rousy Dias. III. Título.

UFPB/BC CDU 004.85(043)

Elaborado por RUSTON SAMMEVILLE ALEXANDRE MARQUES DA SILVA - CRB-15/386

UNIVERSIDADE FEDERAL DA PARAÍBA
CENTRO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

Ata da Sessão Pública de Defesa de Dissertação de Mestrado de Luiz Fernando Fonsêca Pinheiro de Lima, candidato ao título de Mestre em Informática na Área de Sistemas de Computação, realizada em 26 de agosto de 2022.

Aos vinte e seis dias do mês de agosto, do ano de dois mil e vinte e dois, às quatorze horas, no Centro de Informática da Universidade Federal da Paraíba, em Mangabeira, reuniram-se os membros da Banca Examinadora constituída para julgar o Trabalho Final do Sr. Luiz Fernando Fonsêca Pinheiro de Lima, vinculado a esta Universidade sob a matrícula nº 20201003407, candidato ao grau de Mestre em Informática, na área de "Sistemas de Computação", na linha de pesquisa "Processamento de Sinais e Sistemas Gráficos", do Programa de Pós-Graduação em Informática, da Universidade Federal da Paraíba. A comissão examinadora foi composta pelos professores: Clauirton de Albuquerque Siebra (PPGI-UFPB), Orientador e Presidente da Banca, Thais Gaudencio do Rego (PPGI-UFPB), Examinadora Interna, Danielle Rousy Dias Ricarte (UFPB), Examinadora Externa ao Programa, Geber Lisboa Ramalho (UFPE), Examinador Externo à Instituição. Dando início aos trabalhos, o Presidente da Banca cumprimentou os presentes, comunicou a finalidade da reunião e passou a palavra ao candidato para que ele fizesse a exposição oral do trabalho de dissertação intitulado "Usando uma Métrica de Compensação de Utilidade e Justiça para Comparar Sistematicamente Estratégias de Aprendizado Adversárias Justas e Não Generativas". Concluída a exposição, o candidato foi arguido pela Banca Examinadora que emitiu o seguinte parecer: "**aprovado**". Do ocorrido, eu, Fernando Menezes Matos, Coordenador do Programa de Pós-Graduação em Informática, lavrei a presente ata que vai assinada por mim e pelos membros da banca examinadora. João Pessoa, 26 de agosto de 2022.

Prof. Fernando Menezes Matos

Prof. Clauirton de Albuquerque Siebra
Orientador (PPGI-UFPB)

Prof. Thais Gaudencio do Rego
Examinadora interna (PPGI-UFPB)

Prof. Danielle Rousy Dias Ricarte
Examinadora Externa ao Programa (UFPB)

Prof. Geber Lisboa Ramalho
Examinador Externo à Instituição (UFPE)

DEDICATORY

I dedicate this work to every educator that made it possible for me to be here.

ACKNOWLEDGEMENTS

First, I would appreciate it if you excuse me from writing these acknowledgments in Portuguese so all mentioned people can receive the message. E isto se refere especialmente sobre minha mãe, Dona Eliane, que apesar de não ter tido oportunidade de se dedicar aos estudos, sempre fez de tudo para eu ter a melhor educação e poder me dedicar exclusivamente a minha graduação e parte desse curso de mestrado. Outra grande mulher que deve estar no topo dos meus agradecimentos é minha avó, Dona Mariinha. Ela, juntamente com meu avô, ajudaram a construir boa parte da educação que tenho.

Também agradeço à toda minha família que sempre esteve preocupada comigo e sendo altamente compreensível com a falta de tempo para alguns eventos, principalmente nos meses finais.

Não posso deixar de agradecer minha namorada. Ingrid foi pessoa que mais esteve ao meu lado nesse processo desde os planos para ingressar no mestrado, passando pelos aperreios do meio do caminho, até chegar nessa etapa de conclusão. Ela foi a pessoa que mais escutou minhas ideias, meus raciocínios e meus lamentos e sempre me motivou e colocou meu trabalho numa perspectiva acima do que eu enxergava.

Tenho à agradecer a muitos amigos. Todos eles, em diversos momentos, foram meu ponto de escape para esquecer algumas preocupações durante esses 2 anos e meio. E eu sou muito grato por ser rodeado de muitos bons amigos, os amigos do colégio, os amigos da graduação, os amigos do mestrado, os primos de Ingrid, os amigos do LUMO, e meus mais novos amigos do CESAR (com quem eu compartilhei os últimos 6 meses de correria). Mas, em especial, gostaria de agradecer à Jéssica, Iury e Eudis, por terem me acolhido tão cedo na graduação e nos projetos do LUMO e terem me ensinado o que significa fazer ciência da computação. Talvez, sem vocês eu não estaria aqui hoje.

Agradeço à todos os professores, dos ensinos fundamental, médio e superior, que contribuíram de alguma forma com meu aprendizado, graças ao que vocês me ensinaram eu pude chegar aqui. Agradeço especialmente os meus orientadores, Danielle Rousy e Clauirton Siebra. Eu não tenho nem palavras para descrever o quão sou grato por ter trabalhado com vocês nesses 2 anos e meio. Sou muito grato por vocês terem sido tão abertos para trabalharmos com um tema que ainda não tinham contato, por terem sido tão gentis com seus *feedbacks*, por terem sido sempre muito abertos as minhas ideias e por sempre terem acreditado tanto no nosso trabalho e olhado com grande potencial (as vezes até mais que eu mesmo). Vocês foram essenciais para eu estar terminando esse mestrado minimamente são.

Por fim, agradeço a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) por ter apoiado financeiramente o desenvolvimento deste trabalho.

ABSTRACT

Artificial intelligence systems for decision-making have become increasingly popular in several areas. However, it is possible to identify biased decisions in many applications, which have become a concern for the computer science, artificial intelligence, and law communities. Therefore, researches are proposing solutions to mitigate bias and discrimination in decision-makers. Some explored strategies are based on generative adversarial networks to generate fair data. Others are based on adversarial learning to achieve fairness in machine learning by encoding fairness constraints through an adversarial model. Moreover, it is usual for each proposal to assess its model with a specific metric, making the comparison of current approaches a complex task. Therefore, this work proposes a benchmark procedure with a systematical method to assess the fair machine learning models. In this sense, we define the *FU-score* metric to evaluate the utility-fairness trade-off, the utility and fairness metrics to compose this assessment, the used dataset and applied data preparation, and the statistical test. We also performed this benchmark evaluation for the non-generative adversarial models, analyzing the literature models from the same metric perspective. This assessment could not indicate a single model which better performs for all datasets. However, we built an understanding of how each model performs on each dataset which statistical confidence.

Key-words: Adversarial Learning, Benchmark, Machine Learning, Fairness, Trade-off.

RESUMO

Os sistemas de inteligência artificial para tomada de decisão têm se tornado cada vez mais populares em diversas áreas. Entretanto, é possível identificar decisões enviesadas em muitas aplicações, que se tornaram uma preocupação para as comunidades de ciência da computação, inteligência artificial e direito. Portanto, as pesquisas vêm propondo soluções para mitigar o viés e a discriminação presente nos tomadores de decisão. Algumas estratégias exploradas são baseadas em redes adversários generativas para gerar dados justos. Outros são baseados no aprendizado adversário para alcançar a justiça no aprendizado de máquina codificando restrições de justiça por meio de um componente adversário. Além disso, é comum que cada proposta avalie seu modelo com uma métrica específica, tornando a comparação das abordagens atuais uma tarefa complexa. Portanto, este trabalho propõe um procedimento de *benchmark* com um método sistemático para avaliar os modelos de aprendizado de máquina justo. Nesse sentido, definimos a métrica *FU-score* para avaliar o *trade-off* de utilidade e justiça, as métricas de utilidade e justiça para compor essa avaliação, o conjunto de dados utilizado e a preparação aplicada e o teste estatístico. Também realizamos esta avaliação de *benchmark* para os modelos adversários não generativos, analisando os modelos da literatura sob a mesma métrica. Essa avaliação não pôde apontar um único modelo com melhor desempenho para todos os conjuntos de dados. No entanto, construímos um entendimento de como cada modelo funciona em cada conjunto de dados com confiança estatística.

Palavras-chave: Aprendizado Adversário, Aprendizado de Máquina, *Benchmark*, Justiça, *Trade-off*.

LIST OF FIGURES

1 Artificial intelligence subareas, and learning strategies in ML 18
2 10-fold cross-validation example 20
3 Illustration of a simple GAN models 28
4 LAFTR model from MADRAS et al. (2018) (adapted) 29
5 ZHANG et al. (2018) general architecture (adapted) 30
6 Model from BEUTEL et al. (2017) (adapted) 32
7 FairGAN model from XU et al. (2018) (adapted) 33
8 FairGAN$^+$ model from XU et al. (2019) (adapted) 35
9 Cabin attribute filling proportion 45
10 Distributions for protected and label attributes of Titanic dataset 46
11 Saving accounts and checking accounts attributes filling proportion 47
12 Distributions for protected and label attributes of German dataset 48
13 Distributions for protected (sex) and label attributes of Adult dataset . . . 50
14 Distributions for protected (race) and label attributes of Adult dataset . . 50

LIST OF TABELS

1 Contingency table example 21
2 Summary of works main characteristics 28
3 Features in the Titanic dataset 44
4 Features in the German dataset 47
5 Features in the UCI Adult dataset, adapted from ZHANG et al. (2018) . . 49
6 Models' results for Titanic dataset 55
7 Models' results for German dataset 60
8 Models' results for Adult (sex) dataset 64
9 Models' results for Adult (race) dataset 69
10 Accuracy t-test results for Titanic dataset 83
11 DemDisp t-test results for Titanic dataset 84
12 DispEqOdds t-test results for Titanic dataset 85
13 DispEqOpp t-test results for Titanic dataset 86
14 FU-score (DemDisp) t-test results for Titanic dataset 87
15 FU-score (DispEqOdds) for Titanic dataset 88
16 FU-score (DispEqOpp) for Titanic dataset 89
17 Accuracy t-test results for German dataset 90
18 DemDisp t-test results for German dataset 91
19 DispEqOdds t-test results for German dataset 92
20 DispEqOpp t-test results for German dataset 93
21 FU-score (DemDisp) t-test results for German dataset 94
22 FU-score (DispEqOdds) for German dataset 95
23 FU-score (DispEqOpp) for German dataset 96
24 Accuracy t-test results for Adult dataset 97
25 DemDisp t-test results for Adult dataset 98
26 DispEqOdds t-test results for Adult dataset 99
27 DispEqOpp t-test results for Adult dataset 100
28 FU-score (DemDisp) t-test results for Adult dataset 101
29 FU-score (DispEqOdds) for Adult dataset 102
30 FU-score (DispEqOpp) for Adult dataset 103
31 Accuracy t-test results for Adult (race) dataset 104
32 DemDisp t-test results for Titanic dataset 105
33 DispEqOdds t-test results for Adult (race) dataset 106
34 DispEqOpp t-test results for Adult (race) dataset 107
35 FU-score (DemDisp) t-test results for Adult (race) dataset 108
36 FU-score (DispEqOdds) for Adult (race) dataset 109
37 FU-score (DispEqOpp) for Adult (race) dataset 110

LIST OF ABBREVIATIONS

AI – Artificial Intelligence

DemDisp – Demographic Disparity

DispEqOdds – Disparity in Equal Odds

DispEqOpp – Disparity in Equal Opportunity

FN – False Negatives

FNR – False Negative Rate

FP – False Positives

FPR – False Positive Rate

GAN – Generative Adversarial Network

LAFTR – Learning Adversarially Fair and Transferable Representations

ML – Machine Learning

stdev – Standard Deviation

Summary

1 Introduction **14**
1.1 Objectives . . . 16
1.2 Contributions . . . 16
1.3 Dissertation Structure . . . 17

2 Theoretical Fundamentals **18**
2.1 Machine Learning . . . 18
2.2 Statistical Tests for Comparing Machine Learning Models . . . 19
2.3 Fairness in Machine Learning . . . 21
2.3.1 Bias . . . 22
2.3.2 Discrimination . . . 23
2.3.3 Fairness: Types of Approaches, and Definitions . . . 24
2.4 Endings . . . 25

3 Related Works **27**
3.1 Fair Adversarial Strategies . . . 27
3.2 The Learning Adversarially Fair and Transferable Representations Model . . . 29
3.3 Zhang's Adversarial Debiasing Architecture . . . 30
3.4 Beutel's Fair Representations . . . 31
3.5 FairGAN and FairGAN$^+$ Models . . . 32
3.6 Evaluation Metrics . . . 36
3.7 Discussion . . . 39

4 Research Method **41**
4.1 Proposed Approach . . . 41
4.2 Metrics . . . 41
4.2.1 Fairness-Utility Trade-off Metric . . . 41
4.2.2 Fairness and Performance Metrics . . . 42
4.3 Statistical Test for Model Comparison . . . 43
4.4 Datasets . . . 43
4.4.1 Titatic Dataset . . . 44
4.4.2 German Dataset . . . 46
4.4.3 Adult Dataset . . . 48

5 Implementation Details **51**
5.1 Baseline . . . 51
5.2 Implementations based on ZHANG et al. (2018) . . . 51
5.3 Implementations based on MADRAS et al. (2018) . . . 52
5.4 Implementations based on BEUTEL et al. (2017) . . . 52

5.5 Other parameters and resources . . . 53

6 Results and Discussion **55**
6.1 Results for Titanic Dataset . . . 55
6.1.1 Utility . . . 55
6.1.2 Fairness . . . 56
6.1.3 *FU-score* . . . 58
6.1.4 Discussion . . . 59
6.2 Results for German Dataset . . . 60
6.2.1 Utility . . . 60
6.2.2 Fairness . . . 61
6.2.3 *FU-score* . . . 63
6.2.4 Discussion . . . 64
6.3 Results for Adult (sex) Dataset . . . 64
6.3.1 Utility . . . 64
6.3.2 Fairness . . . 65
6.3.3 *FU-score* . . . 67
6.3.4 Discussion . . . 68
6.4 Results for Adult (race) Dataset . . . 68
6.4.1 Utility . . . 69
6.4.2 Fairness . . . 69
6.4.3 *FU-score* . . . 71
6.4.4 Discussion . . . 72

7 Conclusions **74**
7.1 Final Remarks . . . 74
7.2 Limitations and Future Work . . . 76

REFERENCES **77**

A HYPOTHESIS TESTS' RESULTS **82**

1 Introduction

The increase of available data enabled better results in machine learning (ML) algorithms (RUSSEL and NORVIG, 2021), making ML models a common approach for building decision-making software in the most diverse areas such as health, finances, security, and education. However, this increasing importance of these models as decision-making resources, mainly in critical areas, brought about some problems embedded in such algorithms. Bringing the DAL'EVEDOVE and FUJITA (2009) idea to the artificial intelligence (AI) and ML era, we must understand that, despite the widespread use of these algorithms being irreversible, we must debate the social impacts of AI and how we can reduce the negative ones. These concerns raised a new research area focused on socio-algorithmic problems in AI solutions, such as fairness, transparency, accountability, explainability, and privacy (KEARNS and ROTH, 2019).

Building models that mitigate bias and discrimination problems in algorithms is the central concern of the fairness area (MEHRABI et al., 2019). We consider a model fair when it can avoid discrimination in its results (i.e., it is not biased). Discrimination can be understood, in general, as the fact of having a prejudice against an individual or a group in decision-making based on some characteristic, e.g., gender, sexual orientation, ZIP code, and race.

We observe discrimination problems in the most diverse applications. For example, ANGWIN et al. (2016) showed how a decision system about crime recidivism used in the United States of America had its decisions biased with racial prejudice. In addition, GARCIA (2016) demonstrated how applications to determine online advertisement delivery had a sexist bias. While these applications delivered job ads to men, they also delivered clothing and accessories ads to women, even though both men and women have the same characteristics. BOLUKBASI et al. (2016) also demonstrated sexism in the computational task of generating analogies in natural language processing.

Recently, we could observe a case of algorithmic discrimination while the United Kingdom universities incorporated a system for students admission due to the coronavirus pandemic (HAO, 2020). In this case, the system affected 40% of students, giving them lower grades than expected. It was also observed that most of these students were from the working class or disadvantaged groups. On the other hand, some students from private schools had an advantage by increasing their grades.

Therefore, researchers have tried to define bias and fairness to build fair machine learning solutions. For example, the study of LEAVY (2018) aimed to describe a process for reducing sexist bias in natural language processing. Similarly, the work of BOLUKBASI et al. (2016) defined a framework for treating sexist bias in word embeddings. Moreover, the study of LUM and JOHNDROW (2016) used a statistical strategy to reduce

racial discrimination in predictions about criminal recidivism.

The literature presents different formal definitions for fairness, such as demographic parity and equalized odds. Therefore, we might implement these fairness definitions as constraints in our ML models. In this sense, the model will learn to maximize its performance (e.g., accuracy). However, it will limit its learning process to ensure it will not violate the implemented constraint.

Some works have demonstrated how we can implement these constraints through an adversarial model. These works are based on the use of the adversarial learning strategy for representation learning tasks and generative adversarial networks (GANs).

Adversarial learning has been used in representation learning tasks and shown to help increase models' predictive performances for different tasks (BOUSMALIS et al., 2016; GANIN et al., 2016). We refer to adversarial learning as the learning process that uses a second predictor, the adversary, that plays a minimax game with the main predictor (i.e., the one which aims to learn how to predict Y given the attributes X). This minimax game occurs because the adversary aims to maximize its performance while the main predictor aims to minimize it. Moreover, the main predictor wants to maximize its performance.

We can encode a chosen fairness constraint in the adversary component using the adversarial learning process. The works of BEUTEL et al. (2017), ZHANG et al. (2018) and MADRAS et al. (2018) are examples of that. Both BEUTEL et al. (2017) and MADRAS et al. (2018) worked in fair models focusing on learning fair representation. On the other hand, ZHANG et al. (2018) worked on structuring a model-agnostic adversarial debiasing architecture. In general, they use an adversary model and a classifier model, where the adversary aims to correctly predict the protected attribute (i.e., an attribute containing information about groups or individuals that can be used as a discrimination resource) from a fair representation of the classifier's outcomes. These are also non-generative adversarial approaches to fair encoding fairness.

Other works consider treating biased data before the model's learning process. This is motivated by biased models usually being built due to their training from biased data. The recent approaches use fair models based on GANs (GOODFELLOW et al., 2014) to mitigate these biases problems in data. As mentioned, GANs can also be trained following some definition(s) of fairness (i.e., using a fairness definition as a constraint) (XU et al., 2018, 2019). Thus, the generated data follows the real data distribution but does not reproduce the bias presented, helping to promote fairness for the models trained with this generated data.

New fair ML approaches are rapidly emerging in literature. However, each work assesses its proposal using a different methodology, dataset, and metrics. This lack of a

standard procedure is a gap in fairness research, specifically in works of fair adversarial learning approaches. Moreover, it makes comparing the literature models to themselves challenging. It is also complex to compare new approaches to the literature works. In order, JONES et al. (2020) presented a benchmark model for evaluating fair ML algorithms, however, this work does not include adversarial strategies and, principally, presents some weaknesses discussed in Chapter 3.

As known, benchmarks are necessary for the maturity of research in any area, but especially in those new ones (WAZLAWICK, 2020), such as machine learning fairness. Thus, developing a benchmark that includes the adversarial learning approaches to evaluate these proposals systematically, proposals with other strategies, and new proposals that emerge is essential.

1.1 Objectives

This study mainly aims to develop a benchmark to assess fair machine learning strategies, more specifically the non-generative fair adversarial strategies, using a performance-fairness trade-off metric, helping in the fairness area maturity. In order to achieve that, the following specific objectives were considered:

- Define a trade-off metric to evaluate the fair strategies systemically;
- Define the benchmark procedure;
- Access the non-generative adversarial strategies through the proposed benchmark.

1.2 Contributions

This master dissertation presented as the main contribution to the research community a benchmark of the non-generative adversarial models, providing an assessment ruler for the new approaches that may emerge. The presented procedure also can be used to assess other fairness strategies beyond the adversaries. We can point out the work's primary contributions as:

- Presenting an overview on the use of adversarial approaches to encoding fairness in ML models;
- Definition of the *FU-score* metric to compute trade-off between models' utility and fairness;
- Definition of the systematic benchmark procedure, specifying all necessary steps (datasets, data preparation, statistical tests, used models, and implementation details).

We can also point out other technical contributions of the work as:

- Implementation of models and making the code available in an open repository. Thereby, other researchers can reuse this code in their works;
- Implementation of ZHANG et al. (2018) architecture for the demographic parity and equal opportunity fairness constraints. The original work presented only the model implementation for the equal odds definition;
- Expanding and assessing the adversarial strategies for a non-binary protected attribute.

We presented part of these contributions in peer-reviewed conferences. Previous results from this benchmark and the proposition of the *FU-score* metric were presented in the ENIAC 2021 paper "*Assessing Fair Machine Learning Strategies Through a Fairness-Utility Trade-off Metric*" (LIMA et al., 2021). Another paper related to this work was presented in the SBSI 2022 and contained part of the overview presented in this dissertation (LIMA et al., 2022). In this late paper, we also present a further discussion on future works.

1.3 Dissertation Structure

This work is presented in 7 chapters: this Introduction, Theoretical Fundamentals, Related Works, Methodology, Implementation Details, Results and Discussion, and, finally, Conclusions. Chapter 2 presents the baselines to understand the fairness area (e.g., machine learning, bias, discrimination, and fairness concepts). Chapter 3 presents fair adversarial works, summarizes their main characteristics, and presents a discussion on building a benchmark to evaluate them. Chapter 4 presents the methods and details behind the benchmark implementation. Chapter 5 presents the implementation details for the assessed models. Chapter 6 presents the results and discusses the benchmark results for the chosen approaches. Finally, Chapter 7 ends the work by presenting its conclusions, discussions, and future works.

2 Theoretical Fundamentals

This chapter presents the main concepts related to this work. Section 2.1 summarizes machine learning and its tasks. Section 2.2 presents the basis on statistical tests for comparing machine learning models. Finally, Section 2.3 presents the main aspects of the fairness research in machine learning, its formal definitions, and related concepts such as bias, discrimination, and types of approaches.

2.1 Machine Learning

This work aims to build a benchmark of fair adversarial machine learning models for classification tasks. For better understanding, we explain the concepts of machine learning, supervised learning, and classification tasks as follows.

Machine learning is a subarea of artificial intelligence where the machine, based on data, builds a model that is a hypothesis about the represented world in data; this model is also a software that can solve problems for which it was trained (RUSSEL and NORVIG, 2021).

According to RUSSEL and NORVIG (2021), the three main learning strategies are supervised learning, unsupervised learning, and reinforcement learning. The type of feedback characterizes each approach. Figure 1 summarizes the relation between AI and ML and the main learning strategies in ML.

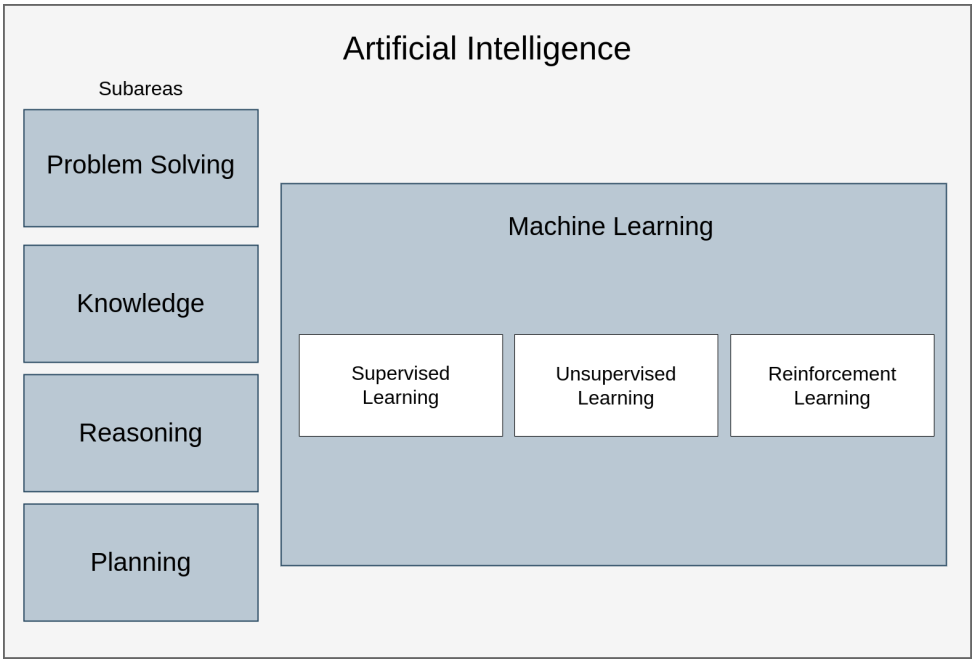


Figure 1: Artificial intelligence subareas, and learning strategies in ML

In supervised learning, there is a dataset where each data instance has a set of X attributes and its associated label Y. For example, we can have a set of attributes X

indicating an individual gender, marital status, education status, age, occupation, and hours per week worked, and a binary label Y that says if a person will present an income higher than 50 thousand dollars in the year or not. The term "supervised" comes from simulating the presence of an "external supervisor" who knows the true value of the label (FACELI et al., 2021).

Supervised learning aims to build a model that maximizes some objective, e.g., predicting a person's income, predicting a house's price, or predicting a student's performance in an exam. When the task is to predict a real value, i.e., the label in data is a real value, we call this task, regression. Otherwise, the label is a class in a set of possible values. We call this task, classification. For example, while predicting tomorrow's temperature is a regression task, predicting tomorrow's weather is a classification task.

Unsupervised learning, however, is applied when we cannot supervise the model training, i.e., our dataset has no associated labels to the attributes. In other words, the machine finds and learns patterns without any feedback. Tasks of unsupervised learning are clustering, summarizing, and association.

Finally, we can build a model through rewards and punishments. Given a set of possible actions, the scenario, and the environment, the machine will decide which action it will choose. If this world interaction is good, the machine receives a reward, i.e., it will actively learn a model through its interactions with the world. We call this approach reinforcement learning.

2.2 Statistical Tests for Comparing Machine Learning Models

For any machine learning task, it is common to train some models that we want to compare and choose the one that presents the better performance. Using only metrics such as accuracy to evaluate our models cannot guarantee that the model with better performance is the best model for any scenario.

Our work intends to build a benchmark to assess and compare the fair adversarial learning approaches. Therefore, we want statistical confidence that a trained model performs better than others or that they perform equally.

We can apply statistical tests to verify, with some confidence, these comparative scenarios. BROWNLEE (2019) presents an introduction to the methods we can use for comparing and selecting a ML model. We will revisit the standard methods (Student's t-test and McNemar's test), pointing out their strengths and weaknesses.

The Student's t-test is a parametric statistical test, i.e., it makes assumptions on the data distribution. The Student's t-test is commonly used in research to compare ML models with the k-fold cross-validation. For example, we evaluate our models using

10-fold cross-validation, take the mean of the accuracy over this 10-fold distribution and apply the Student's t-test.

However, when we use cross-validation to assess our models, we split the data into folds, for example, 10 folds (Figure 2). Then, we train and test our models at the same number of folds. In each train/test iteration, we use 9 folds to train our models and the 1 hold-out fold to test them.

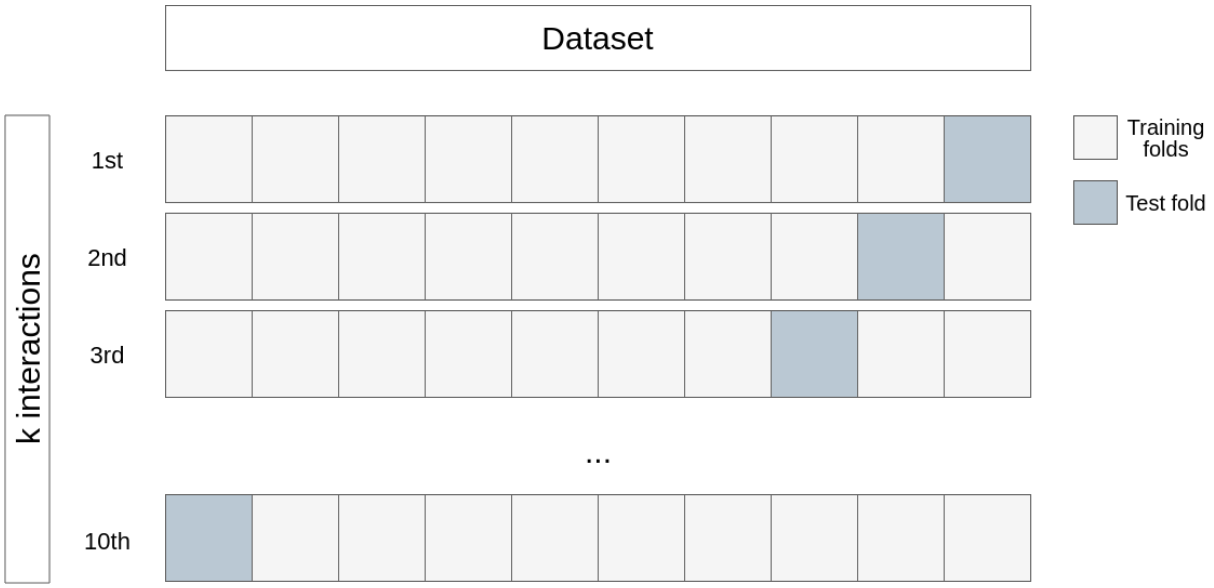


Figure 2: 10-fold cross-validation example

Thus, we understand that we do not have independent data for each fold iteration. In the k-fold cross-validation procedure, each fold will be part of training data $k-1$ times. Therefore, we do not have a guarantee of independent data, violating the Student's t-test assumption that the observations in each sample are independent. Consequently, "the estimated skill scores are dependent, not independent, and in turn that the calculation of the t-statistic in the test will be misleadingly wrong along with any interpretations of the statistic and p-value" BROWNLEE (2019).

On the other hand, with this approach, we can present good repeatability relative to the other methods. Then, we can trade-off the 10-fold cross-validation with Student's t-test strengths with the independence data violation and still choose this approach, knowing that the method has its limitations.

However, the literature presents modifications to this method and other methods to mitigate this problem. For example, the 5x2-fold cross-validation where we train the ML models 5 times, making a resample on data splitting it into 2 folds, and applying the Student's t-test on the results. The fold number is chosen to ensure that samples are observed only in the train and test sets. Moreover, this approach is recommended when we have computer power, or the algorithm is efficient to run the 5 times we need

(BROWNLEE, 2019).

Another recommendation to avoid using Student's t-test with the data independence violation is using McNemar's test instead, especially when we can run the algorithms only once due to expensive costs, e.g., deep learning models (BROWNLEE, 2019).

McNemar's test is similar to a Chi-Squared test and does not say only which model is better. Instead, McNemar's test assesses whether the models' errors are statistically similar. When applying McNemar's test, we split data into train and test samples, run the algorithms and build the contingency table used as input to compute the p-value.

Therefore, when comparing 2 models with a binary label, the contingency table is a 2x2 table that counts the correct and incorrect predictions of the models. The primary diagonal of the table is the intersection where the models are both correct and incorrect. Table 1 shows an example of contingency table. In this example, models A and B make both correct predictions for the same data example 10 times and make both incorrect predictions for the same data example 5 times. For 3 data examples, model A makes correct predictions while model B mistakes the correct value (b in equation 1). Finally, for 2 data points, model A mistakes predictions, and model B gets them right (c in equation 1).

	Model B correct	Model B incorrect
Model A correct	10	3
Model A incorrect	2	5

Table 1: Contingency table example

The McNemar's test statistics are then calculated using Equation 1, where b and c are values in the contingency table. Using the example in Table 1, the equation would be written as in Equation 2.

$$statistic = \frac{(b-c)^2}{b+c} \tag{1}$$

$$statistic = \frac{(3-5)^2}{3+5} \tag{2}$$

2.3 Fairness in Machine Learning

Our work aims to build a benchmark of fair machine learning approaches, specifically, the non-generative fair adversarial algorithms. These are recent proposals to mitigate bias and discrimination in machine learning. However, the study of fairness is not recent.

For example, HUTCHINSON and MITCHELL (2019) summarized in their work 50 years of study about what is fair and unfair from the perspective of testing in education and hiring communities. The authors show how the earlier definitions are similar or identical to the current definitions in the recent years of machine learning fairness area and point "the way towards future research, and measurement of (un)fairness that builds from our modern understanding of fairness while incorporating insights from the past".

Bias and discrimination are two fundamental concepts of fairness in machine learning. Using biased data is a common way to build unfair models, i.e., models that cannot avoid discrimination in their results. MEHRABI et al. (2019) present in their survey the definition of bias, and its possible sources, definitions for discrimination, and fairness. We will reinforce these definitions in the following subsections, presenting the taxonomies pointed out by MEHRABI et al. (2019).

2.3.1 Bias

A heterogeneous dataset, extracted from different group contexts, temporal or spatial, will possibly present biased data. Models trained with this dataset will also reproduce these biases. We can find different bias types in the literature. Following the MEHRABI et al. (2019) survey, here we list some of them:

- **Historical bias** is the bias that reflects the social, cultural, and technical issues existing in the world.
- **Representation bias** is caused by how to define a population sample.
- **Measurement bias** happens from how a particular attribute is defined, used, and measured.
- **Evaluation bias** occurred when we chose biased benchmarks to evaluate trained models.
- **Aggregation bias** occurs when we draw false conclusions for a subgroup based on other subgroups' observations.
- **Population bias** arises when statistics, demographic data, and user characteristics are different in the population of users represented in the dataset and the original target population.
- **Sample bias** occurs through non-random sampling of subgroups, so the estimated trends for a population may not generalize to the collected data from a new population.

- **Algorithmic bias** is defined when the bias is not present in the training data but in the algorithm.

2.3.2 Discrimination

Discrimination can be understood, in general, as having prejudice or harm against an individual or a group in decision-making. A dataset can contain some attributes with specific information about individuals or groups. A model trained with this data could use these attributes as a discrimination source. Then, we should consider these attributes as protected or sensitive attributes in the learning process.

Considering a classification problem, there is a dataset where each instance has a set of X unprotected attributes, a set of protected attributes A, and its associated Y label. Examples of protected attributes are gender ($A \in \{Male, Female\}$), race ($A \in \{African\ American, Caucasian, Hispanic\}$), sexual orientation, and ZIP code. As surveyed by MEHRABI et al. (2019), many literature works use these constructs to formulate definitions for discrimination as:

- **Direct Discrimination** happens when the protected attributes of individuals explicitly result in outcomes that are not favorable to them.
- **Indirect Discrimination** occurs when the model apparently treats individuals based on neutral and unprotected attributes. However, protected groups continue to have the wrong treatment due to implicit associations from their protected attributes.
- **Systemic Discrimination** refers to discrimination against certain social groups perpetuated in organizations' culture and structure through policies, customs, and behavior.
- **Statistical Discrimination** is the phenomenon in which decision-makers use statistics of a group means to judge unfairly an individual who belongs to that group.
- **Explainable Discrimination** is considered when differences in treatment and outcomes can be justified and explained by some attributes in the dataset; according to some regulations, it is a kind of discrimination considered legal.
- **Unexplainable Discrimination** has the opposite definition of explainable discrimination. This kind of discrimination is considered illegal because there is no justification for discrimination against a group, and it is considered illegal.

2.3.3 Fairness: Types of Approaches, and Definitions

To mitigate bias and discrimination and ensure fairness for machine learning models, some techniques have been defined. To categorize these techniques, one can look at their time of application and separate them into three categories (MEHRABI et al., 2019):

- **Pre-processing** techniques try to transform the data to remove discrimination before the learning process.
- **In-processing** techniques seek to modify the state of art learning algorithms to remove discrimination during the learning process.
- **Post-processing** techniques run after the training process with a not yet seen dataset, seeking to evaluate, and debias a trained model.

However, before defining techniques to mitigate discrimination, it is necessary to define the concept of fairness. Formal definitions are how we translate the human understatement of fairness to the machine. Regarding ML fairness, different definitions have been formulated and presented in the literature, so there is no universal definition. Some commonly used fairness definitions are:

- **Fairness Through Unawareness** is a naive concept of fairness in which a fair algorithm is defined when it does not use any protected attribute in the training process. We consider this definition naive because hiding the protected attribute does not guarantee fairness. The model may learn discrimination by using other attributes with a high correlation to protected attributes (LUM and JOHNDROW, 2016; CALDERS and VERWER, 2010).
- **Fairness Through Awareness** uses the idea of similarity between individuals, measured by some distance metric, and defines a fair algorithm when it presents similar predictions for similar individuals (DWORK et al., 2012).
- **Demographic Parity** (or **Statistical Parity**) defines a fair model by $P(\hat{Y} = 1|A = 0) = P(\hat{Y} = 1|A = 1)$, that is, the probability of the predictions must be equal for both groups of the protected attribute, being the decision independent of the protected attribute (CALDERS et al., 2009; DWORK et al., 2012).
- **Equalized Odds** (or **Equal Odds**) defines that the rates of true positives and false positives must be equal for the two groups of the protected attribute (HARDT et al., 2016). Mathematically it is defined as $P(\hat{Y} = 1|A = 0, Y = y) = P(\hat{Y} = 1|A = 1, Y = y)$. Thus, Equalized Odds impose equal bias and accuracy for all groups, punishing models that perform well only for most individuals.

- **Equal Opportunity**, also defined by HARDT et al. (2016), is a more specific case of equal odds when working on "advantage" problems. For example, we understand the advantage when $Y = 1$ in problems such as university admission, promotion receipt, and credit release. In this case, the true positive rates must be equal for the two groups of the protected attribute. Mathematically, $P(\hat{Y} = 1|A = 0, Y = 1) = P(\hat{Y} = 1|A = 1, Y = 1)$.

- **Treatment Equality** is satisfied when the ratio of false negatives (FN), and false positives (FP) is the same for all groups of the protected attribute, mathematically, in an example where A presents two groups, $\frac{FN}{FP}\ for\ the\ first\ group = \frac{FN}{FP}\ for\ the\ second\ group$ (BERK et al., 2018).

MEHRABI et al. (2019) summarized other definitions of bias and fairness. The study of VERMA and RUBIN (2018) also summarizes and presents different fairness definitions, in addition, to evaluating a logistic regression classifier for the UCI German Credit dataset[1] with respect to those fairness definitions.

In another line of thought, Floridi et al. (2018) consider mitigating the real-world unfair discrimination by AI and ML models as a justice ethical principle to be followed. Although both works talk about fair AI/ML, this reasoning differs from that presented by MEHRABI et al. (2019). Floridi et al. (2018) says about using AI to achieve social justice, while in their survey, MEHRABI et al. (2019) defines a fair ML model as a model that mitigates discrimination in decision-making.

Jobin et al. (2019) also pointed out this divergence in a broader view. The authors could observe how works express justice in terms of fairness, mitigation of unwanted biases, respect for diversity, inclusion, and equity, and how some works focus on preserving and promoting (social) justice.

2.4 Endings

Our work aims to build a benchmark of fair adversarial machine learning models. We understand a fair model as described by MEHRABI et al. (2019), i.e., a model that follows a fairness constraint to mitigate discrimination in its decision-making process despite does not aim to promote justice.

In this sense, all fair models assessed in this work were built to perform classification tasks. Thus, we comprehend all used datasets as described in Section 2.3.2. Each data example presents a set of X unprotected attributes, a set of protected attributes A, and its associated Y label. We described the datasets, their attributes, and pre-processing applied in Chapter 4.

[1]Link: https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data).

The assessed fair adversarial models are described on the basis of their type of approach and selected fairness definitions to encode. These aspects follow the categories of strategy and some of the fairness definitions presented in Section 2.3.3. We better explore and present these models' aspects in Chapter 3.

3 Related Works

This chapter presents an overview on the use of fair adversarial strategies used as the baseline in our work. To select this scope and works, we conducted our research as follows: we started our research with the work of MEHRABI et al. (2019), their survey briefly describes the fair adversarial strategies proposed by XU et al. (2018) and ZHANG et al. (2018); then, we searched for papers that included the terms "adversarial learning" and "fairness" in any title, abstract or body; finally, we selected papers that present the fair adversarial learning approaches and reviewed its references to identify additional papers. We previously presented part of this overview in our paper LIMA et al. (2022).

Section 3.1 presents a view of how the fair adversarial approaches work. Sections 3.2 to 3.5 presents the selected works and their adversarial approaches. Section 3.6 focuses on the works' evaluation metrics (for performance and fairness). We conclude the chapter with a discussion of the works and a literature benchmark pointing out gaps we identified to attack in our work.

3.1 Fair Adversarial Strategies

As pointed out in chapter 1, some proposals use adversarial learning to build fairer models. The main idea of this approach is to encode a fairness definition through an adversary component. These works are mainly based on the use of adversarial in representation learning tasks (BOUSMALIS et al., 2016; GANIN et al., 2016) and the generative adversarial networks (GOODFELLOW et al., 2014).

The works of XU et al. (2018, 2019) are examples of generative fair adversarial works. The main idea of this fair GAN approaches is to use GANs ability to generate data with a distribution close to the distribution of the real data and are composed of two models, a generator (G) and a discriminator (D) (Figure 3). While G aims to generate data from random noise, D aims to correctly classify whether an example of data is real or generated. Thus, a GAN runs a minimax game since G wants to minimize the accuracy of D, trying to fool D with the generated data, and D wants to continue maximizing its accuracy, correctly classifying the real and generated examples.

Therefore, XU et al. (2018, 2019) attempt to modify the basis of a GAN structure to add adversarial/discriminator models to encode a chosen fairness constraints. Thus, the generated data follows the real data distribution but tends not to reproduce the bias presented, helping to promote fairness for the models trained with this generated data.

Moreover, the works of BEUTEL et al. (2017), ZHANG et al. (2018) and MADRAS et al. (2018) are examples of non-generative fair adversarial works that include an adversary into the ML model to encode a fairness constraint as an in-processing approach. In

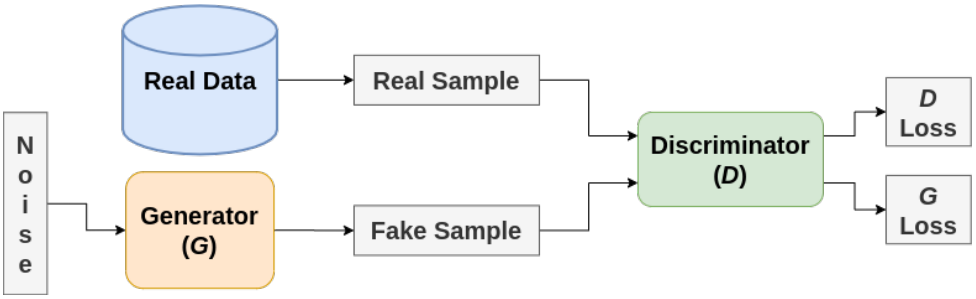


Figure 3: Illustration of a simple GAN models

this approach, a ML model comprises a predictor and an adversary. While the predictor aims to learn how to predict Y given X, the adversary aims to correctly predict the protected attribute A given $\hat{Y}$. Thus, like the other GANs strategies, These model structures play a minimax game, i.e., the adversary aims to maximize its performance while the main predictor aims to minimize it, which characterizes the adversarial learning process.

Each of these fair adversarial works describes their proposals on the basis of its type of approach, selected fairness definitions to encode, and the datasets and metrics used for assessment. We analyzed these aspects for each work and describe them in following sections. Table 2 also presents the summary of the works' main characteristics. In general, these works used the UCI Adult Income dataset[2] to evaluate their proposals. To assess the model's transfer learning ability, the study of MADRAS et al. (2018) also used the Heritage Health dataset[3]. For the word embedding task evaluated by ZHANG et al. (2018), they used embeddings trained from Wikipedia to generate input data from the Google Analogy dataset[4].

Table 2: Summary of works main characteristics

Work	Approach	Fairness Constraint	Fairness Metrics	Utility Metrics	Dataset
Xu et al. [30]	Pre-processing	Demographic Parity	Risk Difference and ϵ-fairness	External Classifier's Accuracy	UCI Adult Income
Zhang et al. [32]	In-processing	Demographic Parity, Equalized Odds and Equal Opportunity	False Positive and False Negative Rates	Classifier's Accuracy	UCI Adult Income and Google Analogy
Madras et al. [23]	In-processing	Demographic Parity, Equalized Odds and Equal Opportunity	Fair Statistical Distances	Classifier's Accuracy	UCI Adult Income and Heritage Health
Beutel et al. [3]	In-processing	Demographic Parity and Equal Opportunity	Parity Gap and Equality Gap	Classifier's Accuracy	UCI Adult Income
Xu et al. [31]	Pre and In-processing	Demographic Parity, Equalized Odds and Equal Opportunity	Risk Difference, Differences in True Positive and in False Positive Rates	Built-in Classifier's Accuracy	UCI Adult Income

[2]UCI Adult Income dataset present 48,842 records from the 1994 American Census database. The attribute sex is commonly used as the protected attribute. Link: http://archive.ics.uci.edu/ml/datasets/Adult.

[3]The Heritage Health dataset contains records related to health and hospitalization of over 60,000 patients, binarized age was used as a sensitive attribute. Link: https://kaggle.com/c/hhp.

[4]Link: https://code.google.com/archive/p/word2vec/source/default/source.

3.2 The Learning Adversarially Fair and Transferable Representations Model

In MADRAS et al. (2018), the authors present the Learning Adversarially Fair and Transferable Representations (LAFTR) model. LAFTR (Figure 4) uses an encoder ($f(X)$) to learn fair representations Z from the input attributes X. It also uses a Decoder ($k(Z, A)$) that can reconstruct X from Z and the sensitive attribute A. To predict A, an adversary ($h(Z)$) is trained, as well as a classifier ($g(Z)$) to predict Y.

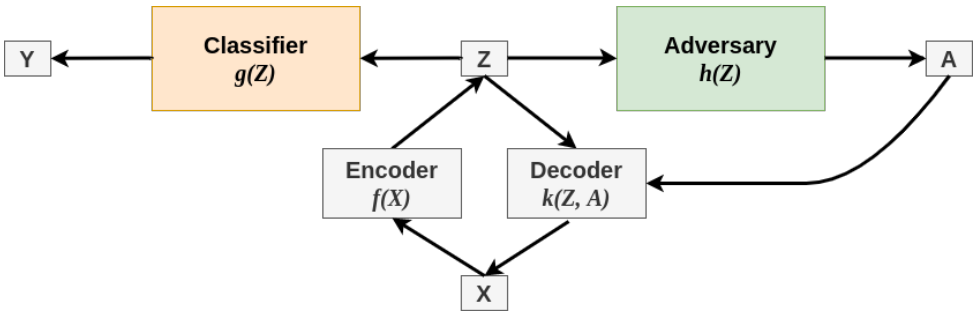


Figure 4: LAFTR model from MADRAS et al. (2018) (adapted)

In the LAFTR model, the adversary aims to maximize its objective. In contrast, the encoder, decoder, and classifier jointly aim to minimize the classification loss and reconstruction error and also, to minimize the adversary's objective.

All LAFTR model elements are neural networks that alternate gradient decent and ascent steps to optimize their parameters according to Equation 3, where L_C is the classifier loss, L_{Dec} denotes the reconstruction loss and L_{Adv} is the adversary loss. Firstly f, g and k take a gradient step to minimize L while the adversary h is fixed. Then h takes a step to maximize L with fixed f, g and k. The hyperparameters α, β, γ in Eq. 3 respectively specify a desired balance between utility, reconstruction of the inputs, and fairness.

$$\begin{aligned} L(f, g, h, k) = \alpha L_C(g(f(X, A)), Y) + \\ \beta L_{Dec}(k(f(X, A), A, X) - \\ \gamma L_{Adv}(h(f(X, A)), A) \end{aligned} \tag{3}$$

Demographic parity, equalized odds, and equal opportunity are the fairness definitions encoded into LAFTR's learning process. The choice of which fairness constraint is encoded is defined by the suitable adversarial objective that varies its functional form depending on the desired fairness criteria.

For demographic parity, the adversarial objective is the average absolute difference between each protected group D_0 and D_1 (Eq. 4). When we desire to follow equalized odds, L_{Adv} is defined by Eq. 5. This formulation considers the average absolute difference

on each protected group-label combination $D_0^0, D_1^0, D_0^1, D_1^1$, where $D_i^j = \{(x, y, a) \in D | a = i, y = j\}$. Finally, to encode equal opportunity, we consider the same formulation for equal odds, however only summing terms corresponding to the positive outcome $Y = 1$.

$$L_{Adv}^{DP} = 1 - \sum_{i \in \{0,1\}} \frac{1}{|D_i|} \sum_{(x,a) \in D_i} |h(f(x, a)) - a)| \tag{4}$$

$$L_{Adv}^{EqOdds} = 2 - \sum_{(i,j) \in \{0,1\}^2} \frac{1}{|D_i^j|} \sum_{(x,a) \in D_i^j} |h(f(x, a)) - a)| \tag{5}$$

The fair representation learned in the LAFTR model was able to train the model's classifier with good results for the trade-offs between an accuracy and fairness. All fair models trained could achieve accuracy $\approx 84\%$ and fair metrics between 0 and 0.2 (the target was 0). Moreover, LAFTR achieved its second goal, to be a model for fair transfer learning. That means it can produce representations that transfer utility to new tasks and yield fairness improvements.

3.3 Zhang's Adversarial Debiasing Architecture

The study of ZHANG et al. (2018) presents a general architecture for achieving fairness through the adversarial process. The model (Figure 5) consists of training a predictor, with the objective to predict Y from X, and an adversary, with the objective to predict A from $\hat{Y}$. Different input data is used for the adversary to achieve each fairness definition.

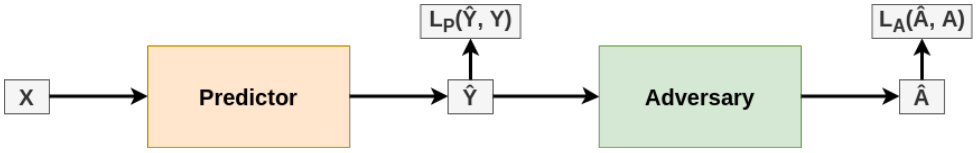


Figure 5: ZHANG et al. (2018) general architecture (adapted)

The predictor is associated with its weights W and the adversary with its weights U. The model is trained by attempting to modify weights W to minimize the predictor loss $L_P(\hat{Y}, Y)$, using a gradient-based method such as stochastic gradient descent. The prediction $\hat{Y}$ is then used as the input to the adversary, which attempts to predict A. In addition to the weights U, the adversary has the loss term $L_A(\hat{A}, A)$.

To achieve demographic parity, the adversary uses only the predicted $\hat{Y}$ labels. In addition to $\hat{Y}$ for equalized odds, the adversary also uses the real labels Y as input. For equal opportunity, for a given class y, the adversary's training is restricted to training data where $Y = y$. For example, when treating advantage problems, we restrict the training data to the examples where $Y = 1$.

ZHANG et al. (2018) define the weights update formulation for U and W. In each training step, U is updated to minimize L_A according to the gradient $\nabla_U L_A$. W is updated according to Equation 6. The term $proj_{\nabla_U L_A} \nabla_W L_P$ prevents the predictor from moving in a direction that helps the adversary decrease its loss. Furthermore, the last term, $\alpha \nabla_W L_A$, attempts to increase the adversary's loss, α is a tunable hyperparameter to balance this attempt.

$$W = W - \nabla_W L_P - proj_{\nabla_U L_A} \nabla_W L_P - \alpha \nabla_W L_A \tag{6}$$

Thus, the model presented by ZHANG et al. (2018) has three main characteristics. First, generality, since different fairness definitions can be achieved depending on the adversary's input data. Second, it is a model-agnostic approach since this strategy can be applied to any classifier model, as long the model is trained using a gradient-based method. Finally, the model is optimality since, if the predictor converges, it converges to a model that satisfies the desired fairness definition.

ZHANG et al. (2018) evaluated the fairness and utility of this model for two scenarios, debiasing word embeddings to perform analogies and a supervised learning task. The authors could demonstrate the model's ability to reduce bias and perform well for the tasks in both scenarios.

In their proposal, when looking at the classification task, ZHANG et al. (2018) evaluated the model by looking at the overall accuracy. They also assessed the false positive rate (FPR) and false negative rate (FNR) for each protected attribute group and used the UCI Adult Income dataset. They observed an accuracy decrease for the debiased model (86% to 84.5%) and achieved approximately values for FPR and FNR across sex subgroups, respectively, $0.4458 \approx 0.4349$, and $0.0647 \approx 0.0701$.

3.4 Beutel's Fair Representations

BEUTEL et al. (2017) consider scenarios in which the protected attribute's values cannot be accessed for all data examples, such as a recommendation system that cannot observe some user attributes. Thus the authors presented a strategy based on the use of adversarial training to create a latent representation that does not contain information about the protected attribute.

The model defined by BEUTEL et al. (2017) is presented in Figure 6 and is composed of three main elements: the encoder of the latent representations ($g(X) = H$), the predictor of the class label from latent representations ($f(H) = \hat{Y}$), and the predictor of the sensitive attribute from the latent representations ($a(H) = \hat{A}$). The goal of the learning process is to make $f(H)$ and $a(H)$ correctly predict, respectively, Y and A. However

we also want that $g(X)$ makes this task hard for $a(H)$.

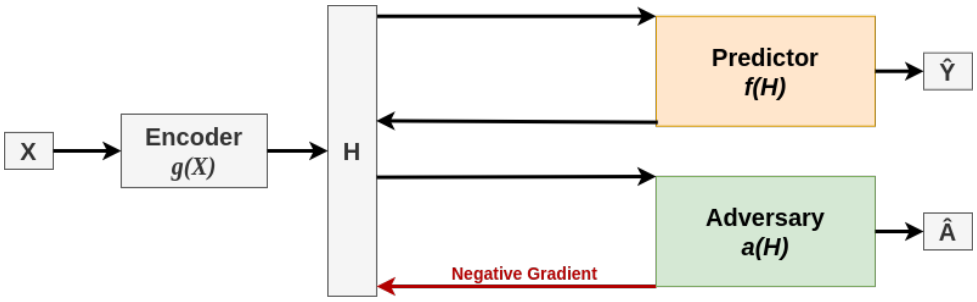


Figure 6: Model from BEUTEL et al. (2017) (adapted)

Therefore, for a classification task, the authors consider a cross-entropy loss for the classifier with form $L_Y(f(g(X)), Y)$ and a cross-entropy loss $L_A(f(a(X)), A)$ for the adversary. To guarantee that minimizing $L_Y + L_A$ will discourage $g(X)$ to produce a representation that makes it easier to predict A, the loss term for the adversary was changed to include the J_λ term, an identity function with a negative gradient. Thus, the loss term have form of $L_A(a(J_\lambda(g(X))), A)$ that means $J(g(X)) = g(X)$ and $\frac{\partial J}{\partial X} = -\lambda\frac{\partial g(X)}{\partial X}$.

For this reason, while $a(H)$ is trained to minimize the classification error, $g(X)$ is trained to maximize the classification error for the adversary. Therefore, $g(X)$ is trained from L_Y to predict Y and from L_A to not encode any information allowing the model to predict A. λ is a hyperparameter that determines the trade-off between accuracy and the model capability of removing information about the protected attribute, which we can consider as a trade-off between the predictive and fairness performances.

This model was evaluated under different distributions of the protected attribute and the class label, in addition to the necessary amount of data to learn a fair latent representation. Tests with balanced data concerning the protected attribute showed that this characteristic positively affects the adversarial training and improves the fairness results of the model, despite decreasing the predictor's accuracy. In addition, the authors demonstrated that the model could achieve fairness even using few training samples.

3.5 FairGAN and FairGAN$^+$ Models

The work of XU et al. (2018) presents the FairGAN model. FairGAN aims to generate a dataset that respects the demographic parity constraint for the protected attribute and ensures a fair classifier as long it is trained from the fair generated dataset.

The FairGAN model (Figure 7) consists of a generator (G) and two discriminators (D_1 and D_2). G generates a fake pair $(\hat{x}, \hat{y})$ following the conditional distribution $P_G(x, y|a)$ from a noise variable z and the protected attribute used as input to the gen-

erator. To ensure that the generated dataset achieves fairness, a rule, that aims to keep $P_G(x, y|a = 1) = P_G(x, y|a = 0)$, is applied.

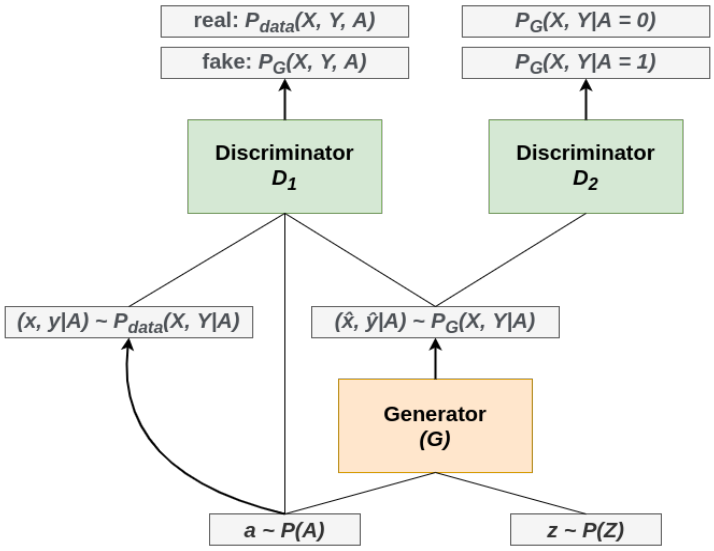


Figure 7: FairGAN model from XU et al. (2018) (adapted)

While D_1 is trained as a classic discriminator to classify the data between real and generated, D_2 is trained to discriminate the protected attribute of the generated data, $P_G(x, y|a = 1)$ and $P_G(x, y|a = 0)$. Thus, the value function of the minimax game is described by Equation 7.

$$\min_G \max_{D_1, D_2} V(G, D_1, D_2) = V_1(G, D_1) + \lambda V_2(G, D_2) \tag{7}$$

V_1 represents the generator objective of learning the joint distribution $P_G(x, y, a)$ over real data $P_{data}(x, y, a)$ by first drawing $\hat{a}$ from $P_G(a)$ and then drawing $\hat{x}, \hat{y}$ from $P_G(x, y|a)$ given a noise variable. On the other hand, the V_2 function value encodes the objective of avoiding the generated samples encoding some information that supports the value prediction of the protected attribute a. λ is a hyperparameter that specifies the trade-off between predictive utility and fairness of data generation. V_1 and V_2 are formally defined, respectively, by Equations 8 and 9.

$$\begin{aligned} &V_1(G, D_1) = \\ &\mathbb{E}_{a \sim P_{data}(A), (x,y) \sim P_{data}(X,Y|A)}[log D_1(x, y, a)] + \\ &\mathbb{E}_{\hat{a} \sim P_G(A), (\hat{x},\hat{y}) \sim P_G(X,Y|A)}[log(1 - D_1(\hat{x}, \hat{y}, \hat{a}))] \end{aligned} \tag{8}$$

$$\begin{aligned} &V_2(G, D_2) = \\ &\mathbb{E}_{(\hat{x},\hat{y}) \sim P_G(X,Y|A=1)}[log D_2(\hat{x}, \hat{y})] + \\ &\mathbb{E}_{(\hat{x},\hat{y}) \sim P_G(X,Y|A=0)}[log(1 - D_2(\hat{x}, \hat{y}))] \end{aligned} \tag{9}$$

In addition to FairGAN, XU et al. (2018) also present the NaiveFairGAN variation. This naive variation achieves only fair data generation but not fair classification, so the NaiveFairGAN is a regular GAN without an additional fairness constraint. In this approach, the protected attribute is removed from the real dataset, the GAN generates the data, and the values for the protected attribute are randomly allocated, preserving only the ratio between the protected group and the unprotected group from real data. We can understand this approach as an attempt to build a fair model by the fairness through unawareness definition.

Data generated by FairGAN, in addition to presenting a good approximation of real data's distribution, also present good results for fairness and utility. The experimental results showed the generated data's utility (euclidean distance ≈ 0.0233) and fairness (≈ 0.0411). The SVM classifier trained with the generated data and assessed with real data also presented good utility (accuracy $\approx 82.17\%$) and fairness (≈ 0.0461) results.

In their second work, XU et al. (2019) presented an improved version of FairGAN. The FairGAN$^+$ model is based on an extended version of GANs called Auxiliary Classifier Generative Adversarial Network (ODENA et al., 2017). A classifier is trained when building an ACGAN in addition to the generator. Thus, FairGAN$^+$ aims to generate fair data and train a fair classifier simultaneously. Another improvement that FairGAN$^+$ brings over FairGAN is the addition of other fairness definitions to the model, specifically, equalized odds and equal opportunity beyond demographic parity.

FairGAN$^+$ model (Figure 8) consists of a generator (G), a classifier ($\eta(X)$) and three discriminators (D_1, D_2 and D_3). G generates samples $\hat{x}$ from random noise z following the distribution $P_G(X|Y,S)$. Each $\hat{x}$ generated has an associated pair $a \sim P_{data}(A)$ and $y \sim P_{data}(Y)$. The classifier $\eta(X)$ is trained for both, accurately predicting the label Y and being fair. G plays an adversarial game with D_1, which is trained to distinguish between real and generated data. To satisfy the fairness notion in generated data, G also plays an adversarial game with D_2, which is trained to distinguish values for the protected attribute of each sample ($P_G(X,Y|A=0)$ and $P_G(X,Y|A=1)$). Finally, D_3 plays an adversarial game with the classifier, where D_3 is trained to distinguish protected attribute values from the prediction made by $\eta(X)$ ($P(\eta(X)=1|A=1)$ and $P(\eta(X)=1|A=0)$).

Therefore, Equation 10 describes the objective function of FairGAN$^+$, J, where V is the described function of the minimax game, and L is the classifier objective function.

$$J = V + L \tag{10}$$

Similar to the FairGAN minimax game, Equation 11 describes the minimax game

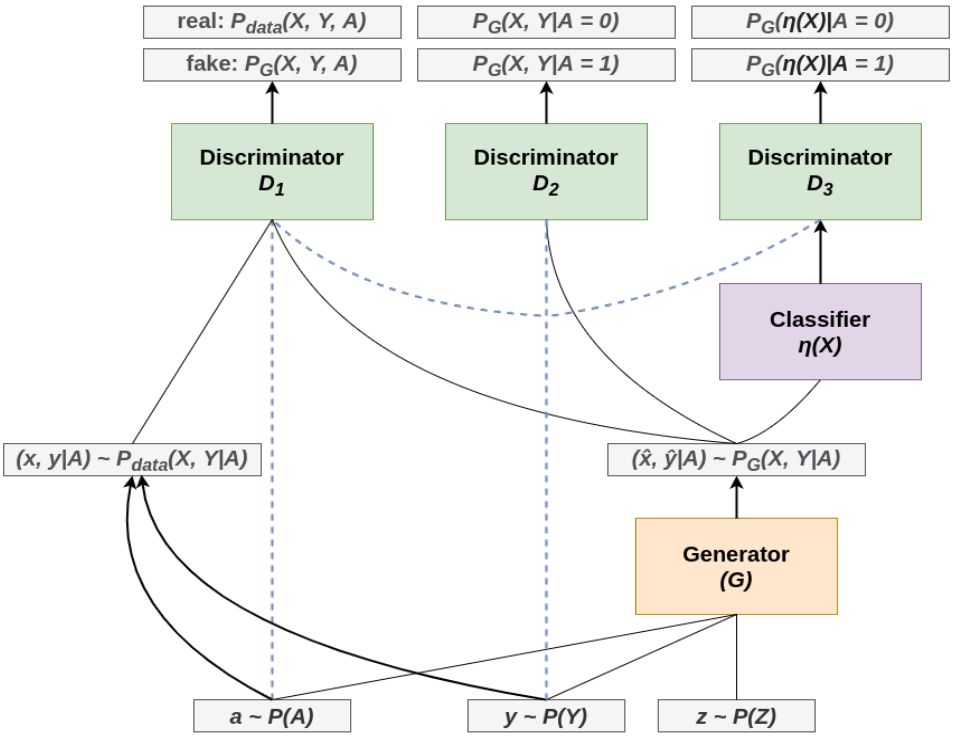


Figure 8: FairGAN$^+$ model from XU et al. (2019) (adapted)

for the FairGAN$^+$ model. Here, λ is also the hyperparameter that specifies the trade-off between utility and fairness of data generation. μ is a hyperparameter that specifies a trade-off between the classifier's accuracy and fairness performances.

$$\min_{G,\eta} \max_{D_1,D_2,D_3} V(G, \eta, D_1, D_2, D_3) = V_1(G, D_1) + \lambda V_2(G, D_2) + \mu V_3(\eta, D_3) \tag{11}$$

V_1, like in their first work, defines the generator objective of learning a distribution that matches the real data distribution. In this case, G needs to learn the joint distribution $P_G(x, y, a)$ over real data $P_{data}(x, y, a)$ by drawing $\hat{x}$ from $P_G(x|y, a)$ given a noise variable. Secondly, the V_2 function value also encodes the objective of avoiding the generated samples encoding some information that supports the value prediction of the protected attribute a.

The model's novelty is in the V_3 function. V_3 defines the objective of making the predictions from $\eta(x)$ in such a way that it does not encode any information that supports predicting the value of the protected attribute a. In that sense, D_3 is trained to correctly predict a given a sample, while the classifier η aims to fool that discriminator. Therefore, once the prediction of η cannot be used to predict the protected attribute a, the correlation between $\eta(x)$ and a is removed, and the desired fairness notion is achieved.

Equations 12, 13, and 14, respectively, define V_1, V_2, and V_3 for the FairGAN$^+$

model. Equation 15 describes the classifier objective function L. The classifier objective is to maximize the log-likelihood of the correct class labels during the training step.

$$
\begin{aligned}
&V_1(G, D_1) = \\
&\mathbb{E}_{a\sim P(A), y\sim P(Y), x\sim P_{data}(X|Y,A)}[logD_1(x, y, a)] + \\
&\mathbb{E}_{a\sim P(A), y\sim P(Y), \hat{x}\sim P_G(X|Y,A)}[log(1 - D_1(\hat{x}, y, a))]
\end{aligned} \tag{12}
$$

$$
\begin{aligned}
&V_2(G, D_2) = \\
&\mathbb{E}_{y\sim P(Y), \hat{x}\sim P_G(X|Y,A=1)}[logD_2(\hat{x}, y)] + \\
&\mathbb{E}_{y\sim P(Y), \hat{x}\sim P_G(X|Y,A=0)}[log(1 - D_2(\hat{x}, y))]
\end{aligned} \tag{13}
$$

$$
\begin{aligned}
&V_3(\eta, D_3) = \\
&\mathbb{E}_{x\sim P(X|Y,A=1)}[logD_3(\eta(x))] + \\
&\mathbb{E}_{x\sim P(X|Y,A=0)}[log(1 - D_3(\eta(x)))]
\end{aligned} \tag{14}
$$

$$
\begin{aligned}
&L(G, \eta) = \\
&\mathbb{E}_{y\sim P(Y), x\sim P_{data}(X|Y,A)}[y\ log\ \eta(x)] + \\
&\mathbb{E}_{y\sim P(Y), \hat{x}\sim P_G(X|Y,A)}[y\ log\ \eta(\hat{x})]
\end{aligned} \tag{15}
$$

FairGAN$^+$ can respect demographic parity, equalized odds, or equal opportunity. It is necessary to adapt the function of D_3 to determine which definition will be respected. This function is changed in mathematical terms according to the desired definition. For example, Equation 14 describes the model's objective that encodes demographic parity.

The experimental results point out that the FairGAN$^+$ model generates data with good utility/approximation of real data's distribution (euclidean distance ≈ 0.0208) and fairness (fair metrics $\approx$ 0.0106 and $\approx$ 0.3867). The FairGAN$^+$ built-in classifier was also evaluated, which presented satisfactory results both in terms of fairness (fair metrics ≈ 0.0141, ≈ 0.0312, and ≈ 0.0245) and accuracy ($\approx 81.78\%$).

The authors also re-evaluated the original FairGAN model to compare it with the FairGAN$^+$'s results. The authors found divergent results from the first report in this later experiment. While using the fair data generated by the FairGAN model, they could not guarantee that a fair classifier was trained due to the classifier results for the fair metric used in the assessment.

3.6 Evaluation Metrics

In the fair machine learning area, the proposals evaluate their models from two perspectives, utility and fairness, i.e., the model's predictive and fairness performances.

When measuring the model utility, the works use standard metrics, such as overall accuracy, false positive and false negative rates, and area under the ROC curve. Differently, each of these works presents specific metrics to measure the fairness in its proposed models.

The remaining of the section discusses the fairness metrics used in the adversarial approaches to achieve fairness. Both fairness and utility metrics used in the presented works are also summarized in Table 2.

In their works, XU et al. (2018, 2019) evaluate the FairGAN and FairGAN$^+$ models utility and fairness both for the generated data and the classifier. For the FairGAN model, they assessed the external classifier, and for the FairGAN$^+$ model, they assessed the built-in classifier. The utility of the generated data is measured by the closeness between these and the real data by calculating the Euclidean distance of joint and conditional probabilities ($P(x, y)$, $P(x, y, a)$ and $P(x, y|a)$).

The authors used the risk difference (Eq. 16) and ϵ-fairness (Eq. 17a) metrics to assess the fairness of generated data, where the balanced error rate (BER) is defined by Eq. 17b. Risk difference is the difference between the conditional probabilities of a positive outcome given the protected attribute for each group, i.e., the disparity when we look at the demographic parity definition.

$$RD(D) = P(y = 1|a = 1) - P(y = 1|a = 0) \tag{16}$$

A classifier is said to be ϵ-fair if it respects Eq. 17a, considering the ϵ-fairness. To evaluate the BER value, they compute the classifier average class-conditioned error on distribution D over the pair (X, A).

$$BER(f(X), A) > \epsilon \tag{17a}$$

$$BER(f(X), A) = \frac{P(f(X) = 0|A = 1) + P(f(X) = 1|A = 0)}{2} \tag{17b}$$

The fairness of the classifier trained with the data generated by FairGAN is measured by the risk difference, considering the classifier (Eq. 18). The fairness in the FairGAN$^+$'s built-in classifier is also measured by the risk difference when considering demographic parity. When considering the equalized odds definition, it is evaluated by the difference in true positive rates (Eq. 19a) and the difference in false positive rates (Eq. 19b). This approach is similar to looking at the disparity in equal odds, but separately.

$$RD(\eta) = P(\eta(x) = 1|a = 1) - P(\eta(x) = 1|a = 0) \tag{18}$$

$$DTPR = P(\eta(X) = 1|Y = 1, S = 1) - P(\eta(X) = 1|Y = 1, S = 0) \tag{19a}$$

$$DFPR = P(\eta(X) = 1|Y = 0, S = 1) - P(\eta(X) = 1|Y = 0, S = 0) \tag{19b}$$

For the LAFTR model, MADRAS et al. (2018) define their metrics based on statistical distance defined by COVER and THOMAS (2012) and incorporate them in the model training process. Then, the model is evaluated by the trade-off between its accuracy and its fairness metrics for demographic parity (Eq. 20a), equalized odds (Eq. 20b), and equal opportunity (Eq. 20c).

$$\Delta_{DP}(g) \triangleq d_g(\mathcal{Z}_0, \mathcal{Z}_1) = |\mathbb{E}_{\mathcal{Z}_0}[g] - \mathbb{E}_{\mathcal{Z}_1}[g]| \tag{20a}$$

$$\Delta_{EO}(g) \triangleq |\mathbb{E}_{\mathcal{Z}_0^0}[g] - \mathbb{E}_{\mathcal{Z}_1^0}[g]| + |\mathbb{E}_{\mathcal{Z}_0^1}[1-g] - \mathbb{E}_{\mathcal{Z}_1^1}[1-g]| \tag{20b}$$

$$\Delta_{EOpp}(g) \triangleq |\mathbb{E}_{\mathcal{Z}_0^0}[g] - \mathbb{E}_{\mathcal{Z}_1^0}[g]| \tag{20c}$$

Based on the metrics defined by Equations 21 and 22, to evaluate, respectively, demographic parity and equal opportunity, BEUTEL et al. (2017) defined two metrics to evaluate their proposal. Parity Gap (Eq. 23a) for demographic parity and Equality Gap (Eq. 23b) for equal opportunity. Moreover, they used accuracy to assess the model's utility.

$$ProbTrue_a = P(\hat{Y} = 1|A = a) = \frac{TP_a + FP_a}{N_a} \tag{21}$$

$$ProbCorrect_{1,a} = P(\hat{Y} = 1|A = a, Y = 1) = \frac{TP_a}{TP_a + FN_a} \tag{22a}$$

$$ProbCorrect_{0,a} = P(\hat{Y} = 0|A = a, Y = 0) = \frac{TN_a}{TN_a + FP_a} \tag{22b}$$

$$Parity\ Gap = |ProbTrue_1 - ProbTrue_0| \tag{23a}$$

$$Equality\ Gap_y = |ProbCorrect_{y,1} - ProbCorrect_{y,0}| \tag{23b}$$

Finally, in their proposal, ZHANG et al. (2018) evaluate the model's utility by looking at the overall accuracy. On the other hand, the model's fairness was assessed by looking at the false positive and false negative rates for each group of the protected

attribute but not computing the difference. Like in the FairGAN$^+$ work, this is similar to evaluating the value for equal odds, but separately.

3.7 Discussion

Section 3.6 presents different approaches to evaluate 5 different fair machine learning proposals. It is common for works in the fair ML area to evaluate their models in a specific way. That was the motivation for the work of JONES et al. (2020).

To bring to the fairness community a benchmark of fair models, JONES et al. (2020) evaluated 27 baseline and fairness algorithms considering 4 real datasets (Titanic, German, Adult, and Adult with race as the protected attribute) and 3 generated datasets. In their work, all considered datasets have only one binary protected attribute, and the target label is also binary. They also explicitly take into account a decision-threshold policy, i.e., the predicted value is compared to a threshold τ and the predicted label is given by $\bar{Y} = I(\hat{Y} > \tau)$, where I is the indicator function. Lastly, they consider models that present a fairness parameter λ, indicating the model trade-off between fairness and classification performance.

JONES et al. (2020) assess the algorithms through 3 different policies: Argmax policy, which fixes the decision threshold at 0.5; The PPR (positive predictive rate) policy, in which the threshold is determined to the positive predictive rate, matches a predetermined value of 20% within a fixed tolerance; and, finally, the Policy Free evaluation, that considers all possible values in a range for the threshold. For this latter aspect, they define and apply the fair efficiency metric (Equation 24).

$$\Theta_{p,f} = 2\frac{K_p K_f}{K_p + K_f} \tag{24}$$

The fair efficiency metric evaluates jointly the model classification performance p (e.g., accuracy, area under the ROC curve, positive and negative rates) and fairness f (e.g., demographic parity, equal odds, and equal opportunity) by computing the harmonic mean between K_p and K_f. K_m (Equation 25) is a additional integral that considers all possible values for m, i.e., the full range for all combinations of τ and λ. The fair efficiency metric penalizes models that score highly for fairness but are not highly useful, and vice versa. If the model is maximally unfair or non-useful, then $\Theta = 0$. Whereas if the model is maximally fair and useful, then $\Theta = 1$ and the model is optimal.

$$K_m = \int_0^1 \int_0^1 m(\lambda, \tau)\, d\tau\, d\lambda \tag{25}$$

The weakness of their work we intend to address is that any evaluated model is

an adversarial strategy. Moreover, the evaluation is limiting because JONES et al. (2020) consider that all fair model proposals present a λ to indicate the model trade-off between predictive performance and fairness, which is not valid. For example, the LAFTR model, presented in Section 3.2, does not have a unique parameter to address this trade-off. Instead, LAFTR considers 3 different parameters to take this trade-off into account.

Furthermore, the trade-off coefficients are considered tunable hyperparameters in most works this chapter presents. Thus, any comparative proposal needs to enable the assessment between different models or algorithms and between the same model or algorithm with this trade-off hyperparameter changed. Therefore, one could evaluate this hyperparameter's best value, which will assist in learning a better fair model.

In our work, Chapter 4, we define a benchmark procedure to address these weaknesses and provide a comparative ruler for the fair adversarial works. A new contribution is the *FU-score* metric defined in Section 4.2.

4 Research Method

This chapter presents the methodological aspects proposed for this work. Section 4.1 presents the proposed approach to achieve the dissertation's main goal. Thereon the chapter focus on how the evaluation occurs. Section 4.2 presents the fairness-utility trade-off metric and the used predictive performance and fairness metrics. Section 4.3 presents the chosen statistical test to compare the model's results. Finally, Section 4.4 presents the used datasets and the applied pre-processing for each dataset.

4.1 Proposed Approach

A concern in fair machine learning research, especially in fair adversarial learning works, is the nonexistence of a systematical assessment methodology. There is variability in chosen metrics and datasets, for example. Without this defined methodology, comparing the literature algorithms and emerging proposals is challenging. Moreover, we understand that benchmarks are necessary to increase the maturity of research (WAZLAWICK, 2020).

Thus we aim to define a systematic benchmark to assess fair machine learning proposals and use this methodology to assess the non-generative fair adversarial algorithms. In the following sections, we define the metrics, statistical tests, datasets, and the applied pre-processing that compose the proposed benchmark procedure.

4.2 Metrics

This section presents the metrics used for our benchmark procedure. The *FU-score* is a new trade-off metric we propose to evaluate the models for both fairness and utility. The utility and fairness metrics were selected from the literature and based on the fairness definitions we consider. This method was presented in our previous work (LIMA et al., 2021).

4.2.1 Fairness-Utility Trade-off Metric

To assess the literature models and the approach proposed by this work with a fairness-utility metric, we present the *FU-score* (Equation 26.). *FU-score* is a fairness-utility trade-off metric inspired by the F1-score[5], but is also a simplification of the fair efficiency metric proposed by JONES et al. (2020).

[5] *F1-score* is a utility metric commonly used to access machine learning models. *F1-score* takes the harmonic mean from two other performance metrics, *Precision* and *Recall*

$$FU\text{-}score = 2\frac{pf}{p+f} \tag{26}$$

Similar to the fair efficiency metric, *FU-score* jointly evaluates the model fairness f and predictive performance p by the harmonic mean of the chosen utility and fairness metrics. In this sense, also like the fair efficiency, *FU-score* penalizes models that score highly for fairness but do not present a good utility and vice versa. In addition, *FU-score* takes into account the fairness and utility metrics we want to maximize, i.e., achieve results near 1. Then, $FU\text{-}score = 0$ means that the model is maximally unfair or non-predictive. When the model is optimal, i.e., the model is maximally fair and useful, $FU\text{-}score = 1$.

FU-score does not consider the helper integral K_m proposed by JONES et al. (2020). Thus, we can use this metric to compare the same model, varying its tunable, fair hyperparameter. Being more general like this, *FU-score* can assist in the model's tune process where one could compare the same model to find a better value for the fair parameter. It also turns possible to assess models that use fair hyperparameters that are different from that considered in JONES et al. (2020) work like the LAFTR model proposed by MADRAS et al. (2018).

4.2.2 Fairness and Performance Metrics

We used the overall accuracy defined by Equation 27 to assess the models' predictive performance. The accuracy measures the overall model utility by looking at the prediction's hit rate over the total number of classifications.

$$Acc = \frac{TN + TP}{TN + FP + FN + TP} \tag{27}$$

In order to evaluate the model's fairness, we considered the disparities for the three commonly used fairness definitions. Thus, we can measure this by the demographic disparity (Eq. 28), disparity in equal odds (Eq. 29) and disparity in equal opportunity (Eq. 30).

$$DemDisp = |P(\hat{Y} = 1|A = 0) - P(\hat{Y} = 1|A = 1)| \tag{28}$$

$$DispEqOdds = |P(\hat{Y} = 1|A = 0, Y = y) - P(\hat{Y} = 1|A = 1, Y = y)| \tag{29}$$

$$DispEqOpp = |P(\hat{Y} = 1|A = 0, Y = 1) - P(\hat{Y} = 1|A = 1, Y = 1)| \tag{30}$$

But there is a problem with these disparity definitions (Equations 28, 29, and 30). The *FU-score* treats both fairness and utility metrics that we want to maximize, i.e., achieve values next to 1. However, our disparity metrics are defined as we want them

smallest as possible, i.e., next to 0. This can be easily solved by adding a difference of 1 in those metrics. Thus, we rewrite the fairness metrics as in Equations 31, 32, and 33. We can apply this modification to any fair definition or metric when necessary.

$$DemDisp = 1 - |P(\hat{Y} = 1|A = 0) - P(\hat{Y} = 1|A = 1)| \tag{31}$$

$$DispEqOdds = 1 - |P(\hat{Y} = 1|A = 0, Y = y) - P(\hat{Y} = 1|A = 1, Y = y)| \tag{32}$$

$$DispEqOpp = 1 - |P(\hat{Y} = 1|A = 0, Y = 1) - P(\hat{Y} = 1|A = 1, Y = 1)| \tag{33}$$

4.3 Statistical Test for Model Comparison

In Section 2.2 we presented some standard used statistical tests for comparing machine learning models, pointing out their weaknesses and strengths.

In our work, we used the 5x2 cross-validation approach within a paired Student's t-test. This approach tries to mitigate the data independence Student's t-test assumption violation. It is also better than McNemar's test because it compares whether the models' results are statistically similar. On the other hand, the late test compares whether the models' errors are statistically similar. Furthermore, McNemar's test is more suitable for models trained for binary classification tasks, and we aimed to define a broader benchmark method.

Then we made paired comparisons over each dataset, i.e., for each dataset, we trained the models and compared their results for accuracy, fair metrics, and *FU-score* metric in pairs to understand if any model is statistically better than the others.

For this statistical our null hypothesis says that the results of the paired models are equal, and the alternative hypothesis says it has a significant difference in the models' results. For these tests we used a significance value of 0.05, which means that our interpretations has a confidence level of 95%.

4.4 Datasets

In their work, JONES et al. (2020) evaluated the selected algorithms considering 4 real datasets (Titanic, German, Adult, and Adult with race as the protected attribute) and 3 generated datasets. In this work, we follow JONES et al. (2020) datasets choosing.

To perform our benchmark experiments, we used the Titanic, German and Adult datasets. The data examples in these datasets represent individual information and for all of them we consider the Sex attribute as the protected attribute. We also assess the models training considering the race as the protected attribute for the Adult dataset. All of these datasets has a binary label attribute which is suitable for a classification problem.

The following subsections presents details on the datasets, distributions over the label and the protected attribute, and the pre-processing step applied to each dataset.

4.4.1 Titatic Dataset

The data on Titanic dataset[6] has information about the Titanic passengers. The label attribute indicates if the passenger survived or not to the Titanic shipwreck. The linked dataset presents a split into train (with 891 examples) and test (with 418 examples). Table 3 provides details on the dataset features.

Table 3: Features in the Titanic dataset

Feature	Type	Description
PassengerId	Discrete	Passenger unique identifier
Pclass	Categorical	Ticket class (1 - 1st, 2 - 2nd, 3 - 3rd)
Name	Text	Passenger name
Sex	Categorical	Passenger sex (male, female)
Age	Continuous	Passenger age. It is fractional if less than 1. If the age is estimated, is it in the form of xx.5
SibSp	Discrete	# of siblings and/or spouses aboard the Titanic
Parch	Discrete	# of parents and/or children aboard the Titanic
Ticket	Text	Ticket number/identifier
Fare	Continuous	Passenger fare
Cabin	Text	Cabin number
Embarked	Categorical	Port of embarkation (C - Cherbourg, Q - Queenstown, S - Southampton)
Survived	Categorical	Label attribute (0 - did not survived, 1 - survived)

We aggregated both train and test files to start the data preparation. We removed the text attributes (name, ticket, and cabin), which are less suitable for the algorithms we assess in this work. Notably, the cabin attribute is tough because it is only filled for 22.54% of all data (Figure 9).

Then we treated the other attributes with missing values. The label attribute presented 418 not filled data points. We dropped these data points because we need these values to work with a classification problem. The embarked attribute presented 2 data examples with missing values, and we also dropped these data points. After these data removing, we kept 889 registers.

The age attribute presented 263 examples with any age filled. In this case, however, we did not discard these samples. Instead, we filled these with the value of -1 to indicate we have no accurate information for this attribute in these data examples.

For the age attribute, we also applied a floor operation to continuous values to discrete, and then we bucketed the age attribute at boundaries [-1, 2, 12, 18, 25, 35, 45, 55, 65, 75, 80].

[6]Link: https://www.kaggle.com/c/titanic

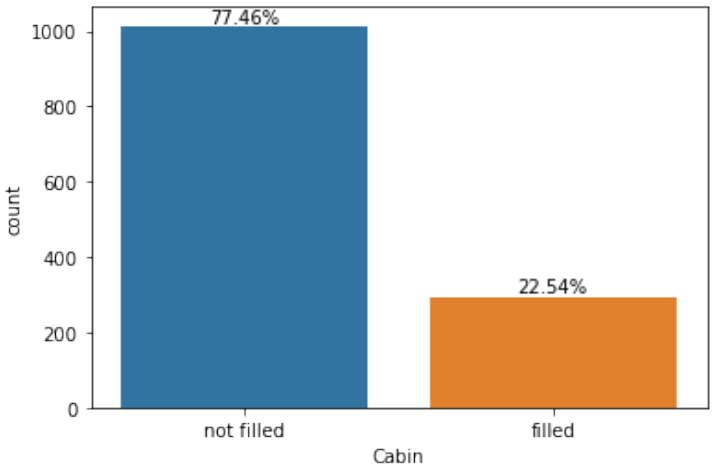
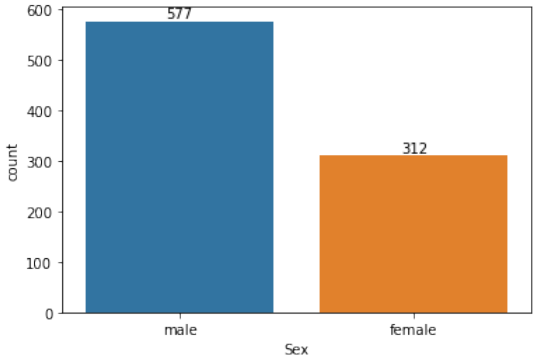


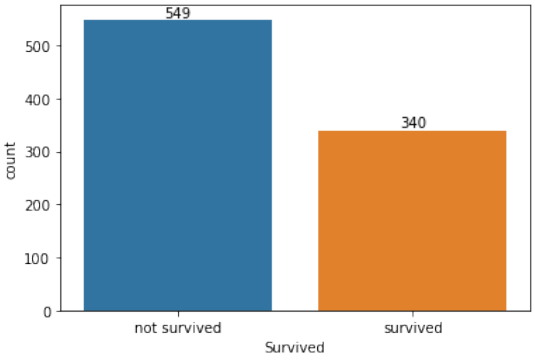
Figure 9: Cabin attribute filling proportion

The embarked attribute is a categorical attribute that indicates the port of passenger embarkation with the values C, Q, and S. We one-hot encoded this attribute for these presented values. The sex and survived attributes are categorical attributes filled, respectively, with the values male/female and 0/1. The label attribute is ready for use. Therefore, we binarized the sex attribute, mapping males as 0 and females as 1.

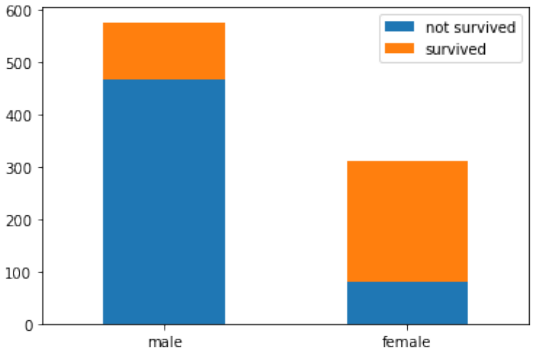
After this preparation, the dataset present 889 data examples. The sex attribute is skewed, presenting more examples for males than females (Figure 10a). The survived attribute is also skewed, presenting more examples of not surviving passengers than survived (Figure 10b). Figure 10c presents the distribution of survived passengers over the sex.



(a) Sex distribution of Titanic dataset



(b) Label distribution of Titanic dataset



(c) Titanic distribution of survived passengers over the sex

Figure 10: Distributions for protected and label attributes of Titanic dataset

4.4.2 German Dataset

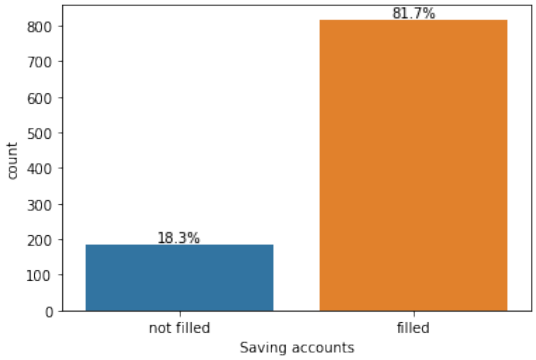
The data on German dataset[7] has information about individuals who take credit from a bank. The label attribute indicates if the person has a good or bad credit risk. The linked dataset presents 1000 data examples. Table 4 provides details on the dataset features.

[7]We used a simplified version of the German dataset. Link: `https://www.kaggle.com/datasets/uciml/german-credit`

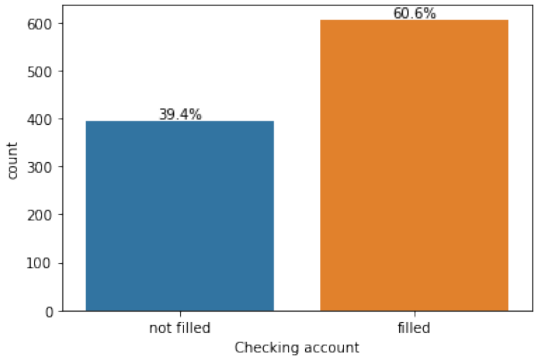
Table 4: Features in the German dataset

Feature	Type	Description
Age	Discrete	Person age
Sex	Categorical	Person sex (male, female)
Job	Categorical	Categorizes the person job into 0 - unskilled and non-resident, 1 - unskilled and resident, 2 - skilled, 3 - highly skilled
Housing	Categorical	Categorizes the person house int own, rent, or free
Saving accounts	Categorical	Categorizes the person savings into little, moderate, quite rich, rich
Checking accounts	Categorical	Categorizes the person savings into little, moderate, quite rich, rich
Credit amount	Continuous	Credit amount in Deutsche Mark
Duration	Discrete	Credit duration in months
Purpose	Categorical	Categorizes the credit purpose into car, furniture/equipment, radio/TV, domestic appliances, repairs, education, business, vacation/others
Risk	Categorical	Label attribute (good, bad)

For this dataset, we use all existing features. The saving accounts and checking accounts attributes present numerous missing values. Figures 11a and 11b present the filling proportion of both attributes. In this case, if we drop the data points with not filled values in any of these attributes, we would have only 522 registers. This data dropping would reduce the dataset by almost half.



(a) Saving accounts attribute filling proportion



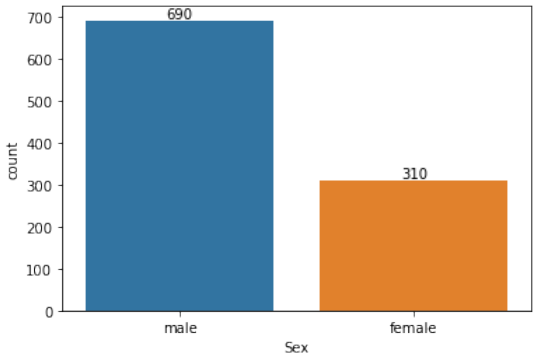
(b) Checking accounts attribute filling proportion

Figure 11: Saving accounts and checking accounts attributes filling proportion

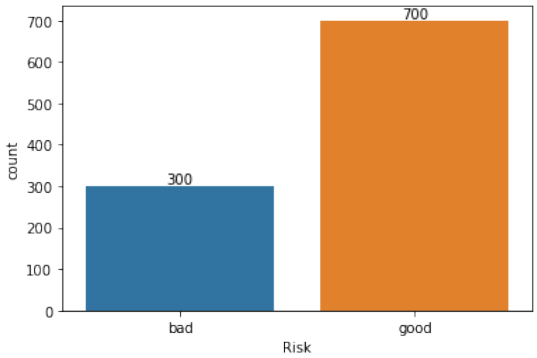
In this case, we filled the missing registers with the value "none" to indicate we have no accurate information for these attributes in these data examples. Then, we one-hot encoded all those categorical attributes (job, housing, saving accounts, checking accounts, and purpose).

We normalized the credit amount and duration attributes. We bucketed the age attribute at boundaries [25, 35, 60, 75]. Therefore, the sex and risk attributes are categorical attributes filled, respectively, with the values male/female and 0/1. The label attribute is ready for use. Therefore, we binarized the sex attribute, mapping males as 0 and females as 1.

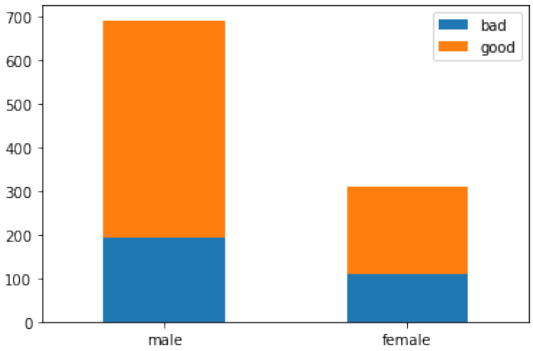
After this preparation, the dataset present 1000 data examples. The sex attribute is skewed, presenting more examples for males than females (Figure 12a). The risk attribute is also skewed, presenting more examples of people with good credit risk (Figure 12b). Figure 12c presents the distribution of person's credit risk over sex.



(a) Sex distribution of German dataset



(b) Label distribution of German dataset



(c) German distribution of person risk over the sex

Figure 12: Distributions for protected and label attributes of German dataset

4.4.3 Adult Dataset

The Adult Income dataset[8] has data on a person's income. The label attribute indicates if the person has an income less or greater than 50K dollars. The original Adult dataset is separated into two sets, a train set with 32561 examples and a test set with 16281 examples, which sums to 48842. Table 5 provides details on the dataset features.

[8]Link: http://archive.ics.uci.edu/ml/datasets/Adult.

Table 5: Features in the UCI Adult dataset, adapted from ZHANG et al. (2018)

Feature	Type	Description
Age	Discrete	Age of the individual
Capital gain	Continuous	Capital gains recorded
Capital-loss	Continuous	Capital losses recorded
Fnlwgt	Continuous	# of people census takers believe that observation represents
Education	Categorical	Highest level of education achieved
Education num	Categorical	Highest education level (numerical form)
Sex	Categorical	Female, Male
Relationship	Categorical	Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried
Marital status	Categorical	Marital status
Occupation	Categorical	Occupation
Hours per week	Continuous	Hours worked per week
Work-class	Categorical	Employer type
Race	Categorical	White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black
Native country	Categorical	Country of origin
Income	Categorical	Whether individual makes >$50K annually

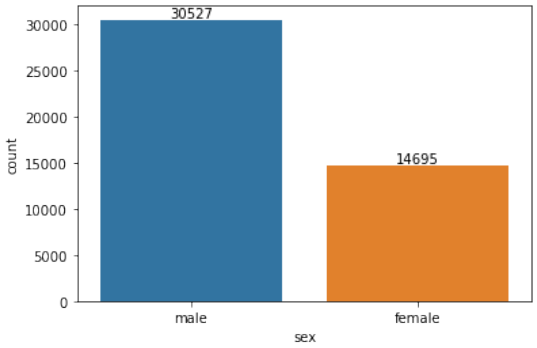
For this dataset, we first normalized the continuous features (capital gain, capital-loss, and hours per week). We removed the attributes "fnlwt" and education num, the last because it represents the same information as the education feature. Moreover, we bucketed at boundaries [18, 25, 30, 35, 40, 45, 50, 55, 60, 65].

Then we treated the missing values present in the dataset. In this case, we dropped all data points with missing values because of the amount of data. After these data removing, we kept 45222 registers. Therefore, the income attribute is categorical attribute filled, with the values $\leq$50K/>50K. Then, we binarized the target attribute, mapping $\leq$50K as 0 and >50K as 1.

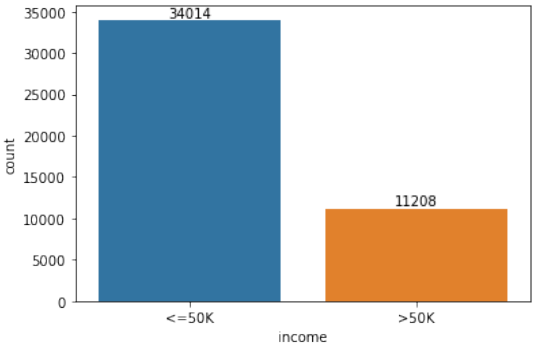
For this dataset, we consider the sex and the race features as protected attributes. The sex attribute presents the values male/female that we binarized, mapping males as 0 and females as 1. On the other hand, race is a non-binary categorical attribute. In this case, we one-hot encoded this feature. We also applied the same preparation to the other non-binary categorical features (work-class, education, marital status, occupation, relationship, and native country).

After this preparation, the dataset present 45222 data examples. The sex attribute is skewed, presenting more examples for males than females (Figure 13a). The income attribute is also skewed, presenting more examples of lower-income people (Figure 13b). Figure 13c presents the distribution of person income over sex.

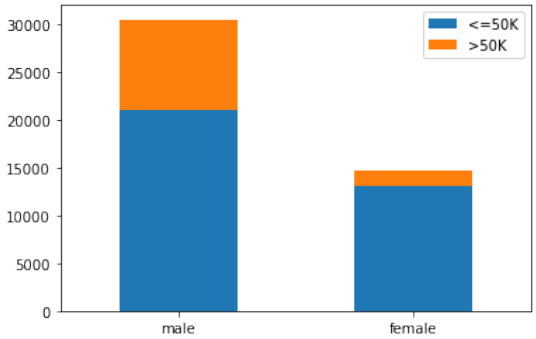
When we look at the race attribute, we understand that this dataset has mostly data about white people. There are 38903 data examples of white people, which represents $\approx$ 86% of all data (Figure 14a). Figure 14b presents the distribution of personal income over race.



(a) Sex distribution of Adult dataset

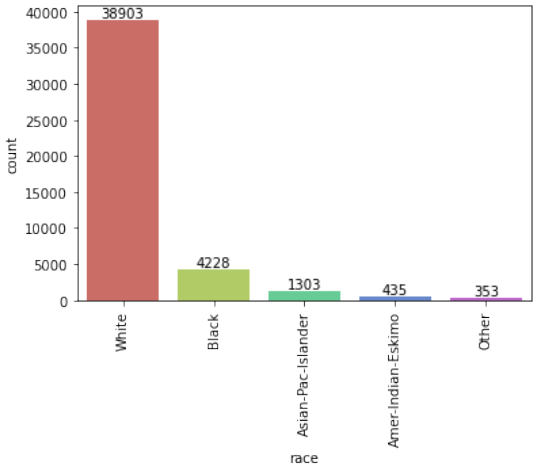


(b) Label distribution of Adult dataset

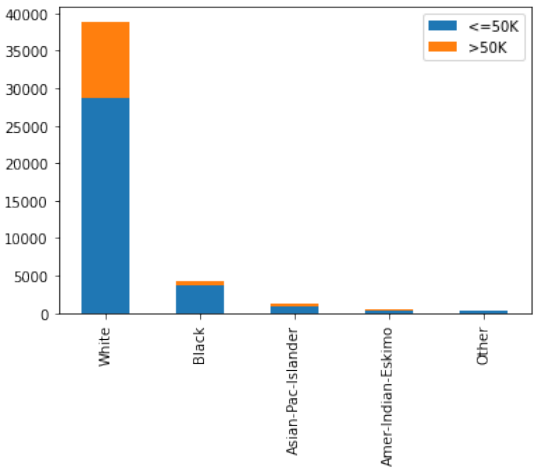


(c) Adult distribution of person income over the sex

Figure 13: Distributions for protected (sex) and label attributes of Adult dataset



(a) Race distribution of Adult dataset



(b) Adult distribution of person income over the Race

Figure 14: Distributions for protected (race) and label attributes of Adult dataset

5 Implementation Details

This chapter presents the implementation details for each model assessed in this work. We implemented and benchmarked 2 baseline models without any fairness constraint and the non-generative adversarial approaches (models proposed by MADRAS et al. (2018), ZHANG et al. (2018), and BEUTEL et al. (2017)). The following sections describe the details of which model.

5.1 Baseline

To compare the fair models to a baseline without any fair constraint, we implemented 2 logistic regression models that compute the predictions through Equation 34. The learning rate decay is the main difference between these two models. While the first uses the default rule for the decay, we apply a learning rate (LR) decrease rule for the second model as follows: for each epoch setting it to $LR = 0.001/t$, where t is the step/epoch counter (for some experiments the baseline model seemed to take advantage in the use of the approach adopted by ZHANG et al. (2018) to avoid local minimum problems). For all datasets we set the optimizer = Adam, batch size = 64, epochs = 100, and initial LR = 0.001.

$$\hat{y} = \sigma(wx + b) \tag{34}$$

5.2 Implementations based on ZHANG et al. (2018)

We provided three implementations for the approach proposed by ZHANG et al. (2018). In their work, they present the implementation of a model that enforces equal odds, which we reproduce here. In addition to the original model, we also implemented the models that enforce equal opportunity and demographic parity, following the theoretical statements presented in their work.

Firstly, we reproduced the model presented by ZHANG et al. (2018). This model has a predictor model like in Equation 34 and an adversarial model to predict the protected attribute defined by Equations 35a and 35b.

$$s = \sigma[(1 + |c|)\sigma^{-1}(\hat{y})] \tag{35a}$$

$$\hat{a} = u[s, sy, s(1 - y)] + b \tag{35b}$$

The model that enforces equal opportunity is similar to the last but differs by using only data examples where y = 1. Finally, the implementation to incorporate the

demographic parity constraint has an adversarial model that is a simplified model from the original. We define this adversary by Equations 36a and 36b.

$$s = \sigma[(1 + |c|)\sigma^{-1}(\hat{y})] \tag{36a}$$

$$\hat{a} = us + b \tag{36b}$$

In these Equations, σ is the sigmoid function, and σ^{-1} is its inverse function, known as the logit function. c is a learnable parameter that weighs the use of the prediction $\hat{y}$ and 1 is added to c to make sure the adversary does not try to ignore $\hat{y}$ by setting $c = 0$.

For all datasets we set the optimizer = Adam, batch size = 64, epochs = 100, and initial LR = 0.001. For all models based on ZHANG et al. (2018) work, we used the fairness parameter as $\alpha = 1/t$, where t is the step counter. This approach worked better than the $\alpha = \sqrt{t}$ used in the original work and kept the guarantee that $\alpha LR \to 0$.

5.3 Implementations based on MADRAS et al. (2018)

For the LAFTR model, we followed the implementation provided in its paper. We also have three neural network models, one for each fair definition. The network structure is similar to all implementations. A single hidden layer is used for each of our encoder, classifier, and adversary, with 8 hidden units and a latent space with dimension = 8. As an activation function for all layers, we applied the Leaky ReLU function (MAAS et al., 2013).

For the equal odds constraint, our adversary uses as input the latent representation and the real label y. Our adversaries use only the latent representation as input for demographic parity and equal opportunity constraints. However, to compute the loss function for the equal opportunity model, it considers only the examples with a positive outcome, i.e., $y = 1$.

For each LAFTR model, we kept the reconstruction coefficient $\beta = 0$ and the classifier coefficient $\alpha = 1$. Also trained and evaluated the model with different values for the fair/adversarial coefficient γ, these values were $\gamma = [0.2, 0.5, 0.7, 1]$.

For all datasets we set the optimizer = Adam, batch size = 64, epochs = 100, and initial LR = 0.001.

5.4 Implementations based on BEUTEL et al. (2017)

We followed as much as possible the implementation details provided by BEUTEL et al. (2017). Due to time limitations, we could provide only the model that encodes the

demographic parity constraint.

The network structure comprises the encoder with 128 hidden units and an output dimension equal to the input. A shared hidden layer connects the encoder's output to a hidden layer with 128 hidden units and has the output dimension equal to 1. The classifier and adversary take the logit from the shared hidden layer output and apply the proper activation function to compute the predictions of $\hat{Y}$ and $\hat{A}$. As an activation function for all intermediate layers, we applied the ReLU function.

This implementation is not suitable for non-binary features. The shared hidden layer output with dimension = 1 limits the adversary and classifier to perform only for binary attributes. Therefore, we did not access this model for the adult dataset, considering race as the protected attribute.

For all accessed datasets we set the optimizer = Adam, batch size = 32, epochs = 100, and initial LR = 0.01. We set the fairness parameter $\lambda = 1$.

5.5 Other parameters and resources

We applied the same weight initialization rule for all models. The weights u and w in Equations 34, 35b, and 36b and the weights for the layers in the neural networks were initialized with zeros. On the other hand, the b's in Equations 34, 35b, and 36b and c's in Equations 35a, 36a and the bias parameters for the neural networks were initialized with ones.

To make the predictions of $\hat{Y}$ and $\hat{A}$, we suited the last activation function for the length of the features. We applied the sigmoid function for binary features. For non-binary attributes, we used the softmax function. We applied the same idea to compute the losses. In this case, we used the binary cross-entropy for binary attributes and the categorical cross-entropy for non-binary attributes.

As presented in Chapter 4, we carried out the benchmark experiments following the 5x2 cross-validation approach and applying the paired t-test. To split the data, we used the scikit-learn `train_test_split` method keeping the proportions of 70% and 30% for the training and test sets. To guarantee reproducibility in the splitting process, we used the `random_state` parameter with the values = [13, 29, 42, 55, 73].

We used the paired t-test implementation from the stats module of the scipy package. We used the `ttest_rel` function for comparing the models' results (accuracies, fair metrics, and trade-offs with the *FU-score* metric) in pairs.

As technological resources for this implementation, we used the programming language Python (version 3.8.13) and the packages TensorFlow (version 2.4.1), NumPy (version 1.22.3), Scikit-learn (version 0.22.2), and Scipy (version 1.4.1). The parameters

related to these packages and not specified here were used as default. For reproducibility, we provide our code at this work GitHub repository[9].

[9]Project repository available in `https://github.com/limafernando/falsb`

6 Results and Discussion

This chapter presents and discusses the results of the benchmark assessment proposed in Chapter 4. The following sections present the understanding of the models' behaviors for each dataset, looking at the means and standard deviation (stdev) of the utility, fairness, and trade-off metrics. The paired t-test results are important for analyze and understand the statistical significance of the findings and point out which model has the best performance for each metric, on the other hand, to have greater fluidity in the text, these results are presented in the Appendix A.

6.1 Results for Titanic Dataset

In this section, we present and discuss the models' results for the titanic dataset considering sex as the protected attribute. This assessment brings the first view of how these models perform on this dataset. Table 6 presents the models' accuracies, fairness and $fu\text{-}score$ results for the titanic dataset. In the following subsections, we discuss each of these metrics.

Table 6: Models' results for Titanic dataset

Model	Accuracy	DemDisp	DispEqOdds	DispEqOpp	*FU-score* (DemDisp)	*FU-score* (DispEqOdds)	*FU-score* (DispEqOpp)
UnfairLR	0.7633 ± 0.0347	0.5274 ± 0.0402	0.6902 ± 0.0318	0.5697 ± 0.0333	0.6222 ± 0.0199	0.7241 ± 0.0196	0.6519 ± 0.0273
UnfairLR-decay	0.3711 ± 0.0381	0.9577 ± 0.0486	0.9765 ± 0.0126	0.9671 ± 0.0224	0.5344 ± 0.0448	0.5369 ± 0.0402	0.5353 ± 0.0394
Zhang4DP	0.7672 ± 0.0326	0.4881 ± 0.06	0.6345 ± 0.0466	0.5114 ± 0.0578	0.5937 ± 0.0392	0.693 ± 0.0227	0.6111 ± 0.0356
Zhang4EqOdds	0.7656 ± 0.0333	0.487 ± 0.0591	0.6302 ± 0.0399	0.5114 ± 0.0578	0.5924 ± 0.0378	0.69 ± 0.0177	0.6106 ± 0.035
Zhang4EqOpp	0.7664 ± 0.0293	0.4665 ± 0.0644	0.6081 ± 0.0599	0.4954 ± 0.0446	0.5767 ± 0.0429	0.6757 ± 0.0266	0.6002 ± 0.0306
LAFTR4DP-0.2	0.7562 ± 0.0256	0.5483 ± 0.0331	0.7243 ± 0.0404	0.7002 ± 0.0608	0.6349 ± 0.0185	0.7395 ± 0.0274	0.7256 ± 0.0287
LAFTR4DP-0.5	0.7305 ± 0.0236	0.6744 ± 0.039	0.8572 ± 0.0431	0.8205 ± 0.0729	0.7004 ± 0.0149	0.7882 ± 0.0212	0.7712 ± 0.0283
LAFTR4DP-0.7	0.7172 ± 0.0176	0.7218 ± 0.0266	0.8956 ± 0.0331	0.868 ± 0.0585	0.7191 ± 0.0141	0.7963 ± 0.0197	0.7848 ± 0.0287
LAFTR4DP-1.0	0.7203 ± 0.0239	0.7068 ± 0.0445	0.8924 ± 0.0471	0.8497 ± 0.0634	0.7126 ± 0.0217	0.7966 ± 0.0256	0.7793 ± 0.0379
LAFTR4EqOdds-0.2	0.7547 ± 0.0354	0.5455 ± 0.1079	0.7311 ± 0.0942	0.6834 ± 0.1098	0.6261 ± 0.0614	0.7384 ± 0.0406	0.7116 ± 0.0595
LAFTR4EqOdds-0.5	0.7547 ± 0.0354	0.5455 ± 0.1079	0.7311 ± 0.0942	0.6834 ± 0.1098	0.6261 ± 0.0614	0.7384 ± 0.0406	0.7116 ± 0.0595
LAFTR4EqOdds-0.7	0.7547 ± 0.0354	0.5455 ± 0.1079	0.7311 ± 0.0942	0.6834 ± 0.1098	0.6261 ± 0.0614	0.7384 ± 0.0406	0.7116 ± 0.0595
LAFTR4EqOdds-1.0	0.7547 ± 0.0354	0.5455 ± 0.1079	0.7311 ± 0.0942	0.6834 ± 0.1098	0.6261 ± 0.0614	0.7384 ± 0.0406	0.7116 ± 0.0595
LAFTR4EqOpp-0.2	0.707 ± 0.0311	0.7173 ± 0.0792	0.8706 ± 0.0706	0.8378 ± 0.0844	0.7091 ± 0.0321	0.7788 ± 0.03	0.7651 ± 0.0399
LAFTR4EqOpp-0.5	0.707 ± 0.0311	0.7173 ± 0.0792	0.8706 ± 0.0706	0.8378 ± 0.0844	0.7091 ± 0.0321	0.7788 ± 0.03	0.7651 ± 0.0399
LAFTR4EqOpp-0.7	0.707 ± 0.0311	0.7173 ± 0.0792	0.8706 ± 0.0706	0.8378 ± 0.0844	0.7091 ± 0.0321	0.7788 ± 0.03	0.7651 ± 0.0399
LAFTR4EqOpp-1.0	0.707 ± 0.0311	0.7173 ± 0.0792	0.8706 ± 0.0706	0.8378 ± 0.0844	0.7091 ± 0.0321	0.7788 ± 0.03	0.7651 ± 0.0399
BEUTEL4DP	0.4719 ± 0.0395	0.8212 ± 0.1911	0.7662 ± 0.2339	0.7923 ± 0.1766	0.5961 ± 0.0814	0.5775 ± 0.1008	0.5886 ± 0.0779

6.1.1 Utility

For all assessments, we expect the baseline model presents a higher accuracy than the fair models. However, in this evaluation, we observe a low increase in the accuracy of the results of ZHANG et al. (2018) based implementations. The UnfairLR-decay model did not perform well for this dataset, presenting only 37.1% of utility, which is worst than a random choice. The other fair models presented a decrease in accuracy compared to the UnfairLR baseline.

The statistical comparison between the UnfairLR model and the ZHANG et al. (2018) based implementations shows that all values are higher than our statistical significance level, which indicates the test failed to reject the null hypotheses. Therefore, we

can not assume if any ZHANG et al. (2018) based implementation performed better or worse than the UnfariLR baseline.

On the other hand, when observing the t-test result from the comparison between the UnfairLR and the fair models LAFTR4DP-0.5, LAFTR4DP-0.7, LAFTR4DP-1.0, LAFTR4EqOpp-0.2, LAFTR4EqOpp-0.5, LAFTR4EqOpp-0.7, LAFTR4EqOpp-1.0, BEUTEL4DP, we understand that the t-test rejected our null hypothesis. This result means that the unfair model outperforms these fair models.

The ZHANG et al. (2018) base implementations present a better accuracy, with statistical confidence, when compared with the other fair models, but LAFTR4DP-0.2, LAFTR4EqOdds (regardless of the value for the fairness coefficient).

6.1.2 Fairness

As opposed to the accuracies results, the models that better performed for utility presented lower demographic disparities (UnfairLR and ZHANG et al. (2018) based implementations). The models with the worst accuracies (UnfairLR-decay and BEUTEL4DP) outperformed all other models. This result does not necessarily mean these models learned to respect the demographic parity constraint. With the presented accuracies, this result could mean only that these models miss the correct prediction for most data points equally.

Excluding the UnfairLR-decay and BEUTEL4DP models, the LAFTR4DP-0.7 model presented the best result for this fair metric. This result is followed by the LAFTR4EqOpp (regardless of the value for the fairness coefficient). These fair models outperformed the UnfairLR implementation by ≈ 0.20 and 0.19, respectively.

When we look at the t-test results from comparing the UnfairLR model and the other implementations, we see that almost all tests reject the null hypothesis, which means that the models are worst or better than the UnfairLR model for this metric. The exceptions are the Zhang4EqOdds, LAFTR4DP-0.2, and LAFTR4EqOdds (regardless of the fair coefficient) models.

The paired t-test results also show that the LAFTR4DP-0.7 presents a worse result for the demographic disparity when compared with the UnfairLR-decay and BEUTEL4DP (with the presented concerns) and outperforms the other models for this metric. However, the t-test failed to reject the null hypothesis when comparing the LAFTR4DP-0.7 with the LAFTR4DP-1.0 and LAFTR4EqOpp (regardless of the value for the fairness coefficient) models.

We can apply a similar interpretation for the LAFTR4EqOpp (regardless of the value for the fairness coefficient). This model presents a worse result for the demographic

disparity compared with the UnfairLR-decay and BEUTEL4DP (with the presented concerns) and outperforms the other models for this metric. However, the t-test result failed to reject the null hypothesis when comparing the LAFTR4EqOpp with the LAFTR4DP-0.7 and LAFTR4DP-1.0 models.

For the disparity in equal odds, UnfairLR-decay almost reached the optimal value. However, we keep the previous understanding that this could mean that the model misses the correct prediction for most data points equally. Moreover, the BEUTEL4DP model reached a mean result DispEqOdds of ≈ 0.76, but with a high standard deviation of ≈ 0.23 (all other models reached a standard deviation between 0.01 and 0.1).

ZHANG et al. (2018) based implementations reached the lower results for DispEqOdds ($\approx$ 0.63, 0.63, and 0.60). This result is worse than the UnfairLR (≈ 0.69). Excluding the UnfairLR-decay, the LAFTR4DP-0.7 model presented the best result for this fair metric, followed by the LAFTR4DP-1.0. These both fair models outperformed the UnfairLR implementation by $\approx$ 0.20 and 0.18.

Comparing the UnfairLR model and the other implementation statistically, we see that almost all tests reject the null hypothesis, which means that the models are worst or better than the UnfairLR model for this metric. The exceptions are the LAFTR4DP-0.2, LAFTR4EqOdds (regardless of the fair coefficient), and BEUTEL4DP models.

The paired t-test results also show that the LAFTR4DP-0.7 presents a worse result for the disparity in equal odds when compared with the UnfairLR-decay (with the presented concerns) and outperforms the other models for this metric. However, the t-test result failed to reject the null hypothesis when comparing the LAFTR4DP-0.7 with the LAFTR4DP-1.0, LAFTR4EqOpp (regardless of the value for the fairness coefficient), and BEUTEL4DP models.

Finally, when we look at the models' results for the disparity in equal opportunity, we also see that the models that better performed for utility presented the lower results for this metric (UnfairLR and ZHANG et al. (2018) based implementations). The models with the worst accuracies (UnfairLR-decay and BEUTEL4DP) outperformed all other models. However, we keep the previous understanding that this could mean that the models miss the correct prediction for most data points equally. Moreover, the BEUTEL4DP model reached a mean result DispEqOdds of ≈ 0.79, but with a high standard deviation of ≈ 0.17 (all other models reached a standard deviation < 0.11).

Excluding the UnfairLR-decay and BEUTEL4DP models, the LAFTR4DP-0.7 model presented the better result for this fair metric, followed by LAFTR4DP-1.0. These fair models outperformed the UnfairLR implementation by $\approx$ 0.30 and 0.28, respectively.

When we look at the statistical experiments' results from the comparison between the UnfairLR model and the other implementations, we see that almost all tests reject

the null hypothesis, which means that the models are worst or better than the UnfairLR model for DispEqOpp. The exceptions are the Zhang4EqOdds, Zhang4DP, LAFTR4DP-0.2, LAFTR4EqOdds (regardless of the fair coefficient), and BEUTEL4DP models.

The paired t-test results also show that the LAFTR4DP-0.7 presents a worse result for the disparity in equal opportunity compared with the UnfairLR-decay (with the presented concerns) and outperforms the other models for this metric. However, the t-test result failed to reject the null hypothesis when comparing the LAFTR4DP-0.7 with the LAFTR4DP-1.0, LAFTR4EqOpp (regardless of the value for the fairness coefficient), and BEUTEL4DP models.

We can apply a similar interpretation for the LAFTR4DP-1.0. This model presents a worse result for the disparity in equal opportunity compared with the UnfairLR-decay (with the presented concerns) and outperforms the other models for this metric. However, the t-test result failed to reject the null hypothesis when comparing the LAFTR4DP-1.0 with the LAFTR4DP-0.5, LAFTR4DP-0.7, LAFTR4EqOpp (regardless of the value for the fairness coefficient), and BEUTEL4DP models.

6.1.3 *FU-score*

The trade-off results between accuracy and demographic disparity demonstrated how the *FU-score* penalizes models with low accuracies and/or fairness. The UnfairLR and ZHANG et al. (2018) based models presented the highest accuracies for the titanic dataset. However, these models did not perform well for the demographic disparity metric. Therefore, the *FU-score* penalizes these models, and their trade-off performances were ≈ 0.62, 0.59, 0.59, and 0.57. On the other hand, the UnfairLR-decay and BEUTEL4DP models presented the lowest accuracies but the highest fairness results. The *FU-score* also penalizes these models, and their trade-off performances were ≈ 0.53 and ≈ 0.59.

The model which achieved the higher trade-off performance was the LAFTR4DP-0.7 (≈ 0.719), followed by the LAFTR4DP-1.0 (≈ 0.712). When we look at the statistical comparisons, we see that this model outperforms almost all models for this trade-off assessment with a statistical significance. The exceptions are the LAFTR4DP-1.0, and LAFTR4EqOpp (regardless of the fair coefficient) models, in which the t-test failed to reject the null hypothesis.

The LAFTR4DP-0.7 and LAFTR4DP-1.0 models also demonstrated the best results for the trade-off between accuracy and disparity in equal odds. Both models achieved a trade-off result ≈ 0.796. The UnfairLR-decay and BEUTEL4DP models kept presenting the worst trade-off performances when considering the disparity in equal odds (≈ 0.53 and ≈ 0.57, respectively). In this case, the UnfairLR and ZHANG et al. (2018) based models presented a better trade-off performance, and their results were ≈ 0.72, 0.69, 0.68,

and 0.67.

The statistical comparisons showed that the LAFTR4DP-0.7 model outperforms almost all models for this trade-off assessment. The exceptions are the LAFTR4DP-1.0, and LAFTR4EqOpp (regardless of the fair coefficient) models, in which the t-test failed to reject the null hypothesis. The same occurs for the LAFTR4DP-1.0 model, which outperforms all models, but LAFTR4DP-0.7 and LAFTR4EqOpp (regardless of the fair coefficient) models.

We have a similar understanding in the trade-off results between accuracy and disparity in equal opportunity. The better performances were also demonstrated by the LAFTR4DP-0.7 (≈ 0.784) followed by LAFTR4DP-1.0 (also ≈ 0.779) and LAFTR4DP-0.5 (also ≈ 0.771) models. The UnfairLR-decay and BEUTEL4DP models kept presenting the worst trade-off performances when considering the disparity in equal odds (≈ 0.53 and ≈ 0.58, respectively). Moreover, in this case, the UnfairLR and ZHANG et al. (2018) based models returned to present lower trade-off performance, and their results were ≈ 0.65, 0.61, 0.61, and 0.60.

The LAFTR4DP-0.7 model outperforms, with statistical confidence, almost all models for this trade-off assessment. The exceptions are the LAFTR4DP-1.0, and LAFTR4EqOpp (regardless of the fair coefficient) models, in which the t-test failed to reject the null hypothesis. The same occurs for the LAFTR4DP-1.0 model, which outperforms all models, but LAFTR4DP-0.7 and LAFTR4EqOpp (regardless of the fair coefficient) models.

6.1.4 Discussion

The results of the models assessments for the titanic dataset showed that the UnfairLR model outperformed all other models in utility for this task, but the ZHANG et al. (2018) based implementations. When looking for the trade-off results, we observed how the *FU-score* penalizes models with low accuracy or fairness.

The overall understanding of the trade-off results shows that the LAFTR4DP-0.7 model outperforms most other models. For all trade-off results, the t-test results from comparing the LAFTR4DP-0.7 model with LAFTR4DP-1.0, and LAFTR4EqOpp (regardless of the fair coefficient) models, failed to reject the null hypothesis.

One could look at the utility and fair metrics individually to break the tie and choose which model to use. In this case, the LAFTR4DP-0.7 model does not outperform both models in accuracy and fairness (for any metric) with statistical significance.

6.2 Results for German Dataset

In this section, we present and discuss the models' results for the german dataset considering sex as the protected attribute. This assessment brings the first view of how these models perform on this dataset. Table 7 presents the models' accuracies, fairness and $fu\text{-}score$ results for the german dataset. In the following subsections, we discuss each of these metrics.

Table 7: Models' results for German dataset

Model	Accuracy	DemDisp	DispEqOdds	DispEqOpp	*FU-score* (DemDisp)	*FU-score* (DispEqOdds)	*FU-score* (DispEqOpp)
UnfairLR	0.7164 ± 0.0180	0.8982 ± 0.0503	0.8869 ± 0.0912	0.9454 ± 0.0152	0.7967 ± 0.0276	0.7911 ± 0.043	0.8151 ± 0.0164
UnfairLR-decay	0.7023 ± 0.0274	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	0.8249 ± 0.0191	0.8249 ± 0.0191	0.8249 ± 0.0191
Zhang4DP	0.7219 ± 0.0136	0.8712 ± 0.0614	0.8578 ± 0.0958	0.9313 ± 0.0267	0.789 ± 0.0316	0.7823 ± 0.046	0.8132 ± 0.0171
Zhang4EqOdds	0.7234 ± 0.0134	0.8720 ± 0.0662	0.8567 ± 0.0987	0.9354 ± 0.0262	0.79 ± 0.0324	0.7825 ± 0.0474	0.8157 ± 0.0156
Zhang4EqOpp	0.7219 ± 0.0211	0.8883 ± 0.0673	0.8679 ± 0.0938	0.9413 ± 0.034	0.796 ± 0.0386	0.7869 ± 0.0501	0.817 ± 0.0236
LAFTR4DP-0.2	0.7102 ± 0.0286	0.9107 ± 0.0651	0.8969 ± 0.0567	0.9374 ± 0.0376	0.7975 ± 0.0383	0.7921 ± 0.0329	0.8078 ± 0.0272
LAFTR4DP-0.5	0.7070 ± 0.0226	0.9257 ± 0.0310	0.9108 ± 0.0262	0.9583 ± 0.0278	0.8016 ± 0.0236	0.7958 ± 0.0169	0.8135 ± 0.0195
LAFTR4DP-0.7	0.7047 ± 0.0239	0.9340 ± 0.0427	0.9170 ± 0.0474	0.9672 ± 0.0279	0.803 ± 0.0257	0.7963 ± 0.0227	0.815 ± 0.0182
LAFTR4DP-1.0	0.7063 ± 0.0263	0.9219 ± 0.0322	0.9119 ± 0.0364	0.9479 ± 0.0243	0.7995 ± 0.0237	0.7954 ± 0.0175	0.8092 ± 0.022
LAFTR4EqOdds-0.2	0.7086 ± 0.0248	0.9257 ± 0.0314	0.9140 ± 0.0293	0.9538 ± 0.0159	0.8025 ± 0.0227	0.7979 ± 0.0174	0.8128 ± 0.0157
LAFTR4EqOdds-0.5	0.7086 ± 0.0248	0.9257 ± 0.0314	0.9140 ± 0.0293	0.9538 ± 0.0159	0.8025 ± 0.0227	0.7979 ± 0.0174	0.8128 ± 0.0157
LAFTR4EqOdds-0.7	0.7086 ± 0.0248	0.9257 ± 0.0314	0.9140 ± 0.0293	0.9538 ± 0.0159	0.8025 ± 0.0227	0.7979 ± 0.0174	0.8128 ± 0.0157
LAFTR4EqOdds-1.0	0.7086 ± 0.0248	0.9257 ± 0.0314	0.9140 ± 0.0293	0.9538 ± 0.0159	0.8025 ± 0.0227	0.7979 ± 0.0174	0.8128 ± 0.0157
LAFTR4EqOpp-0.2	0.7047 ± 0.0282	0.9191 ± 0.0280	0.9004 ± 0.0443	0.945 ± 0.0276	0.7976 ± 0.0261	0.79 ± 0.025	0.8069 ± 0.0203
LAFTR4EqOpp-0.5	0.7047 ± 0.0282	0.9191 ± 0.0280	0.9004 ± 0.0443	0.945 ± 0.0276	0.7976 ± 0.0261	0.79 ± 0.025	0.8069 ± 0.0203
LAFTR4EqOpp-0.7	0.7047 ± 0.0282	0.9191 ± 0.0280	0.9004 ± 0.0443	0.945 ± 0.0276	0.7976 ± 0.0261	0.79 ± 0.025	0.8069 ± 0.0203
LAFTR4EqOpp-1.0	0.7047 ± 0.0282	0.9191 ± 0.0280	0.9004 ± 0.0443	0.945 ± 0.0276	0.7976 ± 0.0261	0.79 ± 0.025	0.8069 ± 0.0203
BEUTEL4DP	0.6986 ± 0.0150	0.9998 ± 0.0004	0.9982 ± 0.0040	0.9993 ± 0.0017	0.8224 ± 0.0104	0.8219 ± 0.0099	0.8222 ± 0.0102

6.2.1 Utility

For this task, the baseline models achieved $\approx 71.6\%$ and 70.2% of accuracy. We can observe a low accuracy increase in the results of ZHANG et al. (2018) based implementations ($\approx 72\%$). The other fair models presented a decrease in accuracy compared to the UnfairLR baseline performing between 69% and 71% of accuracy.

Statistically comparing the baseline models and the ZHANG et al. (2018) based implementations, we see that all values are higher than our statistical significance test, which indicates the test failed to reject the null hypotheses. Therefore, we can not make assumptions on which model presents the better performance.

Moreover, when observing the t-test result from the comparison between the unfair models and the other fair approaches, we also understand that the t-test failed to reject the null hypothesis, indicating no statistical significance difference between the models' utility performance.

Looking at the paired t-test comparing ZHANG et al. (2018) based implementations, we understand that the ZHANG4DP model only outperforms the BEUTEL4DP model. The ZHANG4EqOdds model outperforms LAFTR4DP-0.5, LAFTR4DP-0.7, and BEUTEL4DP models. Finally, the ZHANG4EqOpp model outperforms the LAFTR4EqOpp (regardless of the fair coefficient) model.

6.2.2 Fairness

For the demographic disparity, all models presented results near the optimal. The models with the worst accuracies (UnfairLR-decay and BEUTEL4DP) outperformed all other models reaching, respectively, DemDisp = 1 and DemDisp $\approx$ 0.99. This result does not necessarily mean these models learned to respect the demographic parity constraint. This result could mean that these models only miss the correct prediction for most data points.

Excluding the UnfairLR-decay and BEUTEL4DP models, the LAFTR4DP-0.7 model presented the better result for the demographic disparity, followed by the LAFTR4DP-0.5. These fair models outperformed the UnfairLR implementation by $\approx$ 0.04 and 0.03, respectively.

When statistically comparing the UnfairLR model and the other implementation, we see that almost all tests failed to reject the null hypothesis. This result means we have no significant difference between the models' fairness. The exceptions are the UnfairLR-decay and BEUTEL4DP models, which outperformed the baseline model.

The paired t-test results also show that the LAFTR4DP-0.7 presents a worse result for the demographic disparity when compared with the UnfairLR-decay and BEUTEL4DP (with the presented concerns). However, the t-test result failed to reject the null hypothesis when comparing the LAFTR4DP-0.7 model to the other models.

We can apply a similar interpretation to the LAFTR4DP-0.5 model. This model presents a worse result for the demographic disparity when compared with the UnfairLR-decay and BEUTEL4DP (with the presented concerns). However, the t-test result failed to reject the null hypothesis when comparing the LAFTR4DP-0.5 model to the other models.

With this individually observation of the demographic disparity metric, we only can say UnfairLR-decay, and BEUTEL4DP models outperform all the other models.

For the disparity in equal odds, UnfairLR-decay reached the optimal value (DispEqOdds = 1). Moreover, the BEUTEL4DP reached a DispEqOdds = $\approx$ 0.99. However, we keep the previous understanding that this could mean that these models miss the correct prediction for most data points equally.

ZHANG et al. (2018) based implementations reached lower results for DispEqOdds ($\approx$ 0.85, 0.85, and 0.86). This result is worse than the UnfairLR ($\approx$ 0.88). Excluding the UnfairLR-decay and BEUTEL4DP models, the LAFTR4DP-0.7 model presented the better result for the disparity in equal odds, followed by the LAFTR4EqOdds (regardless of the fair coefficient). Both fair models outperformed the UnfairLR implementation by $\approx$ 0.03.

When we compare the models statistically, we see that only the comparisons between the UnfairLR model and UnfairLR-decay, Zhang4DP, and BEUTEL4DP models reject the null hypothesis. This result means that the Zhang4DP model performs worst than the baseline model, and the UnfairLR-decay and BEUTEL4DP models outperform the baseline (with the presented concerns).

The statistical comparisons also show that the LAFTR4DP-0.7 presents a worse result for the disparity in equal odds when compared with the UnfairLR-decay and BEUTEL4DP models (with the presented concerns). On the other hand, the t-test result failed to reject the null hypothesis when comparing the LAFTR4DP-0.7 with the other models.

Again, observing the disparity in equal odds metric individually, we only can say UnfairLR-decay, and BEUTEL4DP models outperform all the other models.

Finally, when we look at the models' results for the disparity in equal opportunity, we see that all models almost reached the optimal result, presenting results between 0.93 and 1. We also see that the models that better performed for utility presented the lower results for this metric (UnfairLR and ZHANG et al. (2018) based implementations). The UnfairLR-decay and BEUTEL4DP outperformed all other models reaching DispEqOpp $= 1$ and DispEqOpp ≈ 0.99, respectively. However, we keep the previous understanding that this could mean that these models miss the correct prediction for most data points.

Excluding the UnfairLR-decay and BEUTEL4DP models, the LAFTR4DP-0.7 model presented the better result for the disparity in equal opportunity, followed by LAFTR4DP-0.5. These fair models outperformed the UnfairLR implementation by ≈ 0.02 and 0.01, respectively.

When we look statistical comparisons for this metric, we see that only the comparisons between the UnfairLR model and UnfairLR-decay and BEUTEL4DP models reject the null hypothesis, which the UnfairLR-decay and BEUTEL4DP models outperform the UnfairLR baseline (with the presented concerns).

The statistical comparisons also show that the LAFTR4DP-0.7 presents a worse result for the disparity in equal opportunity compared with the UnfairLR-decay and BEUTEL4DP models (with the presented concerns). On the other hand, the t-test result failed to reject the null hypothesis when comparing the LAFTR4DP-0.7 with the other models.

We can apply a similar interpretation for the LAFTR4DP-0.5. This model presents a worse result for the disparity in equal opportunity compared with the UnfairLR-decay (with the presented concerns). However, the t-test result failed to reject the null hypothesis when comparing the LAFTR4DP-0.5 with the other models.

Again, observing the disparity in equal odds metric individually, we only can say

UnfairLR-decay, and BEUTEL4DP models outperform all the other models.

6.2.3 *FU-score*

For the trade-off results between accuracy and demographic disparity, the UnfairLR and ZHANG et al. (2018) based models presented the highest accuracies for the german dataset and performed well for the demographic disparity metric. Their trade-off performances were $\approx$ 0.79, 0.78, 0.79, and 0.79. On the other hand, the UnfairLR-decay and BEUTEL4DP models presented the lowest accuracies but the highest fairness results. The *FU-score* balances the trade-off result for these models, and both models performed this trade-off as $\approx$ 0.82.

Excluding the UnfairLR-decay and BEUTEL4DP models, the highest trade-off performance was achieved by the LAFTR4DP-0.7 ($\approx$ 0.8024. This model is followed by the LAFTR4EqOdds (regardless of the fair coefficient) that also performed $\approx$ 0.802. When we look at the statistical experiments for these trade-off results, we observe that all experiments failed to reject the null hypothesis.

Then, for the german dataset, we can not say that any model performs better or worst than others when looking at the *FU-score* between accuracy and demographic disparity.

In the case of the trade-off results between accuracy and disparity in equal odds, the best performances were also demonstrated by UnfairLR-decay and BEUTEL4DP models ($\approx$ 0.82 for both). All the other models reached results near 0.78 and 0.79 for this evaluation.

The statistical tests for these trade-offs showed that almost no model significantly differs in performance for this trade-off. However, the t-test demonstrated that the BEUTEL4DP model performs better when compared with the LAFTR4DP-0.5, LAFTR4DP-1.0, and LAFTR4EqOdds (regardless of the fair coefficient) models.

We have a similar understanding of the trade-off results between accuracy and disparity in equal opportunity. The UnfairLR-decay and BEUTEL4DP models also demonstrated the best performances ($\approx$ 0.82 for both). However, for this trade-off, all the other models reached results near 0.80 and 0.81.

Again, when we look at the statistical experiments for these trade-off results, we observe that all experiments failed to reject the null hypothesis. Then, for the german dataset, we can not say that any model performs better or worst than others when looking at the *FU-score* between accuracy and disparity in equal opportunity.

6.2.4 Discussion

The t-test results of the models' assessments for the german dataset showed that the UnfairLR does not present a statistically significant difference in accuracy. When looking for the trade-off results, almost all t-test results failed to reject the null hypothesis.

Furthermore, when looking at the trade-off between accuracy and demographic disparity, all t-test results failed to reject the null hypothesis. The same understanding occurs for the trade-off between accuracy and disparity in equal opportunity. For the trade-off accuracy and disparity in equal odds, we understand that the BEUTEL4DP model performs better when compared with the LAFTR4DP-0.5, LAFTR4DP-1.0, and LAFTR4EqOdds (regardless of the fair coefficient) models.

6.3 Results for Adult (sex) Dataset

In this section, we present and discuss the models' results for the adult dataset considering sex as the protected attribute. This assessment brings the original papers' reproduction results. Table 8 presents the models' accuracies, fairness and $fu\text{-}score$ results for the adult (sex) dataset. In the following subsections, we discuss each of these metrics.

Table 8: Models' results for Adult (sex) dataset

Model	Accuracy	DemDisp	DispEqOdds	DispEqOpp	*FU-score* (DemDisp)	*FU-score* (DispEqOdds)	*FU-score* (DispEqOpp)
UnfairLR	0.8502 ± 0.0025	0.8098 ± 0.0059	0.9013 ± 0.0087	0.8813 ± 0.0138	0.8295 ± 0.0038	0.875 ± 0.0052	0.8654 ± 0.0078
UnfairLR-decay	0.8355 ± 0.0015	0.806 ± 0.0051	0.8598 ± 0.0088	0.8048 ± 0.0153	0.8205 ± 0.0025	0.8474 ± 0.0037	0.8198 ± 0.0074
Zhang4DP	0.8507 ± 0.0029	0.8081 ± 0.0057	0.8994 ± 0.0071	0.879 ± 0.0114	0.8289 ± 0.0039	0.8743 ± 0.0048	0.8646 ± 0.0069
Zhang4EqOdds	0.8507 ± 0.0029	0.8075 ± 0.0058	0.898 ± 0.0078	0.8767 ± 0.0134	0.8285 ± 0.004	0.8737 ± 0.0052	0.8635 ± 0.0079
Zhang4EqOpp	0.85 ± 0.0022	0.8087 ± 0.0055	0.9001 ± 0.008	0.8801 ± 0.0141	0.8288 ± 0.0035	0.8743 ± 0.0047	0.8648 ± 0.0077
LAFTR4DP-0.2	0.8498 ± 0.0019	0.8139 ± 0.0146	0.9189 ± 0.0087	0.9158 ± 0.0102	0.8314 ± 0.0074	0.883 ± 0.0039	0.8815 ± 0.0049
LAFTR4DP-0.5	0.8492 ± 0.0016	0.8335 ± 0.0126	0.952 ± 0.0107	0.9685 ± 0.0184	0.8412 ± 0.0063	0.8976 ± 0.0043	0.9049 ± 0.0076
LAFTR4DP-0.7	0.8485 ± 0.0019	0.8554 ± 0.0136	0.9648 ± 0.0052	0.9791 ± 0.0097	0.8519 ± 0.0063	0.9029 ± 0.0022	0.9091 ± 0.0047
LAFTR4DP-1.0	0.8479 ± 0.0026	0.871 ± 0.0162	0.9567 ± 0.008	0.9529 ± 0.0234	0.8592 ± 0.0071	0.899 ± 0.0037	0.8972 ± 0.0109
LAFTR4EqOdds-0.2	0.8492 ± 0.0021	0.8485 ± 0.0147	0.9671 ± 0.0051	0.9889 ± 0.0082	0.8488 ± 0.0068	0.9043 ± 0.002	0.9137 ± 0.004
LAFTR4EqOdds-0.5	0.8492 ± 0.0021	0.8485 ± 0.0147	0.9671 ± 0.0051	0.9889 ± 0.0082	0.8488 ± 0.0068	0.9043 ± 0.002	0.9137 ± 0.004
LAFTR4EqOdds-0.7	0.8492 ± 0.0021	0.8485 ± 0.0147	0.9671 ± 0.0051	0.9889 ± 0.0082	0.8488 ± 0.0068	0.9043 ± 0.002	0.9137 ± 0.004
LAFTR4EqOdds-1.0	0.8492 ± 0.0021	0.8485 ± 0.0147	0.9671 ± 0.0051	0.9889 ± 0.0082	0.8488 ± 0.0068	0.9043 ± 0.002	0.9137 ± 0.004
LAFTR4EqOpp-0.2	0.8474 ± 0.002	0.8694 ± 0.0171	0.9528 ± 0.0135	0.9465 ± 0.0332	0.8582 ± 0.0075	0.897 ± 0.0062	0.894 ± 0.0155
LAFTR4EqOpp-0.5	0.8474 ± 0.002	0.8694 ± 0.0171	0.9528 ± 0.0135	0.9465 ± 0.0332	0.8582 ± 0.0075	0.897 ± 0.0062	0.894 ± 0.0155
LAFTR4EqOpp-0.7	0.8474 ± 0.002	0.8694 ± 0.0171	0.9528 ± 0.0135	0.9465 ± 0.0332	0.8582 ± 0.0075	0.897 ± 0.0062	0.894 ± 0.0155
LAFTR4EqOpp-1.0	0.8474 ± 0.002	0.8694 ± 0.0171	0.9528 ± 0.0135	0.9465 ± 0.0332	0.8582 ± 0.0075	0.897 ± 0.0062	0.894 ± 0.0155
BEUTEL4DP	0.6599 ± 0.1351	0.8872 ± 0.2468	0.8834 ± 0.2442	0.8804 ± 0.2432	0.7553 ± 0.1808	0.754 ± 0.1798	0.7529 ± 0.1794

6.3.1 Utility

When training for this dataset, the baseline models, Unfair-LR and UnfairLR-decay, achieved the utility performance of 85% and 83%, respectively. ZHANG et al. (2018) based implementations also presented an accuracy of 85%, performing as good as the baseline. All MADRAS et al. (2018) based implementations presented an accuracy near 84%, and the BEUTEL4DP model achieved the lower result of $\approx 65\%$.

From the statistical comparisons, we understand that the UnfairLR model outperforms, with a significance level, the UnfairLR-decay, LAFTR4DP-1.0, LAFTR4EqOdds

(regardless of the fair coefficient), LAFTR4EqOpp (regardless the fair coefficient), and BEUTEL4DP models.

The ZHANG et al. (2018) base implementations present a better accuracy, with statically confidence, than the UnfairLR-decay, LAFTR4DP-1.0, LAFTR4EqOdds (regardless of the value for the fairness coefficient), LAFTR4EqOdds (regardless of the value for the fairness coefficient), and BEUTEL4DP models.

6.3.2 Fairness

As opposed to the accuracies results, for the demographic disparity metric, the models that better performed for utility presented lower demographic disparities (UnfairLR and ZHANG et al. (2018) based implementations). Nevertheless, these models achieved high results ($\approx$ 0.80). For this task, the models with the worst accuracies (UnfairLR-decay and BEUTEL4DP) presented different values for the demographic disparity. UnfairLR-decay presented a DemDisp $\approx$ 0.80, and BEUTEL4DP presented the higher result for this metric (DemDisp $\approx$ 0.88). However, the BEUTEL4DP model also presented a high value for the standard deviation ($\approx$ 0.24, while all other models reached a stdev < 0.02), which means the result is less stable than the other models.

Excluding the BEUTEL4DP model, the LAFTR4DP-1.0 model presented the better result for DemDisp, followed by the LAFTR4EqOpp (regardless of the value for the fairness coefficient). These fair models outperformed the UnfairLR implementation by $\approx$ 0.07 and 0.06, respectively.

When we look at the statistically comparisons between the UnfairLR model and the other implementations, we see that almost all tests reject the null hypothesis. This result means that the models are worse or better than the UnfairLR model for this metric. The exceptions are the Zhang4DP, Zhang4EqOpp, LAFTR4DP-0.2, and BEUTEL4DP models.

The paired t-test results also show that the LAFTR4DP-1.0 outperforms the other models for this metric, but the t-test failed to reject the null hypothesis when comparing the LAFTR4DP-1.0 with the BEUTEL4DP model. The same occurs for the LAFTR4EqOpp (regardless of the value for the fairness coefficient). This model outperforms all other models, but the t-test failed to reject the null hypothesis when comparing it with the BEUTEL4DP model.

For the disparity in equal odds, our baseline models reached high values ($\approx$ 0.9 and $\approx$ 0.85). Analyzing the fair models, the BEUTEL4DP model presented the lower result (DispEqOdds $\approx$ 0.88). ZHANG et al. (2018) based implementations reached results for DispEqOdds $\approx$ 0.89, 0.89, and 0.90. The LAFTR4EqOdds (regardless of the fair

coefficient) presented the better result for this fair metric, followed by the LAFTR4DP-0.7 model. These both fair models outperformed the UnfairLR implementation by ≈ 0.06.

When we looking the statistically comparisons between the UnfairLR model and the other implementation, we see that almost all results reject the null hypothesis. This result means that the UnfairLR model outperforms, with a statistical significance, the UnfairLR-decay and presents an underperformance compared to almost all other models. The exceptions are ZHANG et al. (2018) based implementations and the BEUTEL4DP model, for which the t-test failed to reject the null hypothesis.

The paired t-test results also show that the LAFTR4EqOdds (regardless of the fair coefficient) outperform almost all other models. The t-test result failed to reject the null hypothesis when comparing the LAFTR4EqOdds (regardless of the fair coefficient) with the LAFTR4DP-0.7, LAFTR4DP-1.0, LAFTR4EqOpp (regardless of the fair coefficient), and BEUTEL4DP models.

Finally, when we look at the models' results for the disparity in equal opportunity, we see that all models present high performances with results between 0.80 and 0.98. For this metric, our baseline models reached high values (≈ 0.88 and ≈ 0.80). Analyzing the fair models, the BEUTEL4DP model presented the lower result DispEqOdds ≈ 0.88. ZHANG et al. (2018) based implementations reached results for DispEqOdds (≈ 0.87, 0.87, and 0.88). The LAFTR4EqOdds (regardless of the fair coefficient) presented the better result for this fair metric, followed by the LAFTR4DP-0.7 model. These both fair models outperformed the UnfairLR implementation by ≈ 0.1 and 0.09, respectively.

When statistically comparing the UnfairLR model and the other implementation result's for the disparity in equal opportunity, we see that almost all results reject the null hypothesis. This result means that the UnfairLR model outperforms, with a statistical significance, the UnfairLR-decay model and presents an underperformance compared to almost all other models. The exceptions are ZHANG et al. (2018) based implementations and the BEUTEL4DP model, for which the t-test failed to reject the null hypothesis.

The paired t-test results also show that the LAFTR4EqOdds (regardless of the fair coefficient) outperform almost all other models. The t-test result failed to reject the null hypothesis when comparing the LAFTR4EqOdds (regardless of the fair coefficient) with the LAFTR4DP-0.7, LAFTR4DP-1.0, LAFTR4EqOpp (regardless of the fair coefficient), and BEUTEL4DP models.

We can apply a similar interpretation for the LAFTR4DP-0.5. This model presents a worse result for the disparity in equal opportunity compared with the UnfairLR-decay (with the presented concerns). However, the t-test result failed to reject the null hypothesis when the LAFTR4DP-0.5 was compared with the other models.

6.3.3 *FU-score*

For the trade-offs results between accuracy and demographic disparity, the UnfairLR, UnfairLR-decay, and ZHANG et al. (2018) based models presented the highest accuracies for the adult dataset and performed well for the demographic disparity metric. Their trade-off performances were all ≈ 0.82. On the other hand, the BEUTEL4DP model presented the lowest accuracy but the highest fairness result. The *FU-score* balances the trade-off result for this model, and it performed a trade-off ≈ 0.75, but with a stdev ≈ 0.18.

The highest trade-off performance was achieved by the LAFTR4DP-1.0 model ($\approx$ 0.859), followed by the LAFTR4EqOpp (regardless of the fair coefficient) that performed a trade-off ≈ 0.858. When we look at the statistical experiments for these trade-off results, we observe that the LAFTR4DP-1.0 model outperforms almost all models for this trade-off assessment. The exception is only the BEUTEL4DP model, in which the t-test failed to reject the null hypothesis.

For the trade-off results between accuracy and disparity in equal odds, the UnfairLR and ZHANG et al. (2018) based models presented the highest accuracies for the adult dataset and performed well for the disparity in equal odds metric. Their trade-off performances were all ≈ 0.87. Moreover, the UnfairLR-decay and BEUTEL4DP presented a slightly lower result for this trade-off (≈ 0.84 and ≈ 0.75, respectively). The *FU-score* balances the trade-off result for the BEUTEL4DP model, and it performed a trade-off ≈ 0.75, but with a stdev ≈ 0.17.

The higher trade-off performance was achieved by the LAFTR4EqOdds (regardless of the fair coefficient) model (≈ 0.904), followed by the LAFTR4DP-0.7 that performed a trade-off ≈ 0.902. When we statistically comparing these trade-off results, we observe that the LAFTR4EqOdds (regardless of the fair coefficient) model outperforms almost all models for this trade-off assessment. The exceptions are the LAFTR4DP-0.7, LAFTR4DP-1.0, LAFTR4EqOpp (regardless of the fair coefficient), and BEUTEL4DP model, in which the t-test failed to reject the null hypothesis.

Finally, we have a similar understanding of the trade-off results between accuracy and disparity in equal opportunity. The UnfairLR and ZHANG et al. (2018) based models presented the highest accuracies for the adult dataset and performed well for the disparity in equal opportunity metric. Their trade-off performances were all ≈ 0.86. Moreover, the UnfairLR-decay and BEUTEL4DP presented a slightly lower result for this trade-off (≈ 0.81 and ≈ 0.75, respectively). The *FU-score* balances the trade-off result for the BEUTEL4DP model, and it performed a trade-off ≈ 0.75, but with a stdev ≈ 0.17.

The higher trade-off performance was achieved by the LAFTR4EqOdds (regardless of the fair coefficient) model (≈ 0.91), followed by the LAFTR4DP-0.7 which performed

a trade-off ≈ 0.90. When we statistically comparing these trade-off results, we observe that the LAFTR4EqOdds (regardless of the fair coefficient) model outperforms almost all models for this trade-off assessment. The exceptions are the LAFTR4DP-0.7, and BEUTEL4DP model, in which the t-test failed to reject the null hypothesis.

6.3.4 Discussion

The t-test results of the models' assessments for the adult dataset (having sex as the protected attribute) showed that the UnfairLR outperforms only a set of fair models with a statistically significant difference.

When looking at the trade-off between accuracy and demographic disparity, the LAFTR4DP-1.0 outperforms all models but BEUTEL4DP. However, this late model has a trade-off value of 0.75 with a higher stdev ≈ 0.18. On the other hand, the LAFTR4DP-1.0 has a trade-off value of 0.85 with a stdev ≈ 0.007, which makes the LAFTR4DP-1.0 present a more consistent result.

Moreover, for the trade-off accuracy and disparity in equal odds, the t-test results showed that LAFTR4EqOdds outperforms, with statistical significance, all other models but LAFTR4DP-0.7, LAFTR4DP-1.0, LAFTR4EqOpp (regardless the fair coefficient), and BEUTEL4DP models. A similar understanding occurs for the trade-off between accuracy and disparity in equal opportunity. The t-test results showed that LAFTR4EqOdds outperforms, with statistical significance, all other models but LAFTR4DP-0.7 and BEUTEL4DP models.

When choosing a model for considering these both trade-offs, one could look to the utility and fair metrics individually to identify the most suitable model. For example, the LAFTR4EqOdds model presents a higher accuracy, with statistical significance, than the LAFTR4DP-1.0, LAFTR4EqOpp, and BEUTEL4DP models. For the DispEqOdds metric, the LAFTR4EqOdds model does not present a statistically significant difference in its performance compared to the LAFTR4DP-0.7, LAFTR4DP-1.0, LAFTR4EqOpp, and BEUTEL4DP models. Finally, for the DispEqOpp metric, the LAFTR4EqOdds model presents a higher result, with statistical significance, than the LAFTR4DP-1.0 and LAFTR4EqOpp models.

6.4 Results for Adult (race) Dataset

In this section, we present and discuss the models' results for the adult dataset considering race as the protected attribute. This assessment brings the first view on how these models perform for a non-binary protected attribute. Table 9 presents the models' accuracies, fairness and *fu-score* results for the adult (race) dataset. In the following subsections, we discuss each of these metrics.

Table 9: Models' results for Adult (race) dataset

Model	Accuracy	DemDisp	DispEqOdds	DispEqOpp	*FU-score* (DemDisp)	*FU-score* (DispEqOdds)	*FU-score* (DispEqOpp)
UnfairLR	0.8502 ± 0.0025	0.1628 ± 0.0119	0.759 ± 0.032	0.5766 ± 0.0618	0.2731 ± 0.0168	0.8017 ± 0.0175	0.6856 ± 0.044
UnfairLR-decay	0.8355 ± 0.0015	0.2431 ± 0.0377	0.7573 ± 0.0305	0.5786 ± 0.06	0.3754 ± 0.0439	0.7942 ± 0.0171	0.6824 ± 0.0414
Zhang4DP	0.8499 ± 0.0024	0.1676 ± 0.0124	0.7619 ± 0.0331	0.5873 ± 0.067	0.2798 ± 0.0174	0.8032 ± 0.0182	0.6928 ± 0.0478
Zhang4EqOdds	0.8501 ± 0.0021	0.16 ± 0.013	0.7523 ± 0.0305	0.5627 ± 0.0629	0.2691 ± 0.0185	0.7979 ± 0.0171	0.6755 ± 0.0463
Zhang4EqOpp	0.85 ± 0.0021	0.1645 ± 0.0137	0.7573 ± 0.0294	0.5754 ± 0.0596	0.2754 ± 0.0192	0.8008 ± 0.0165	0.6848 ± 0.0437
LAFTR4DP-0.2	0.8501 ± 0.002	0.18 ± 0.0188	0.79 ± 0.0253	0.6466 ± 0.0577	0.2967 ± 0.0253	0.8187 ± 0.0135	0.7333 ± 0.038
LAFTR4DP-0.5	0.8499 ± 0.0018	0.1823 ± 0.0232	0.7996 ± 0.0129	0.6669 ± 0.038	0.2996 ± 0.0306	0.8239 ± 0.0069	0.7469 ± 0.0234
LAFTR4DP-0.7	0.8497 ± 0.0015	0.1919 ± 0.0121	0.811 ± 0.0177	0.6964 ± 0.0391	0.3129 ± 0.016	0.8298 ± 0.0088	0.7649 ± 0.0238
LAFTR4DP-1.0	0.8496 ± 0.0019	0.1977 ± 0.0187	0.8122 ± 0.0171	0.7039 ± 0.0438	0.3205 ± 0.0243	0.8304 ± 0.0091	0.7694 ± 0.0263
LAFTR4EqOdds-0.2	0.8494 ± 0.0016	0.1816 ± 0.0189	0.7932 ± 0.0235	0.6559 ± 0.0531	0.2988 ± 0.0253	0.8202 ± 0.0125	0.7392 ± 0.035
LAFTR4EqOdds-0.5	0.8494 ± 0.0016	0.1816 ± 0.0189	0.7932 ± 0.0235	0.6559 ± 0.0531	0.2988 ± 0.0253	0.8202 ± 0.0125	0.7392 ± 0.035
LAFTR4EqOdds-0.7	0.8494 ± 0.0016	0.1816 ± 0.0189	0.7932 ± 0.0235	0.6559 ± 0.0531	0.2988 ± 0.0253	0.8202 ± 0.0125	0.7392 ± 0.035
LAFTR4EqOdds-1.0	0.8494 ± 0.0016	0.1816 ± 0.0189	0.7932 ± 0.0235	0.6559 ± 0.0531	0.2988 ± 0.0253	0.8202 ± 0.0125	0.7392 ± 0.035
LAFTR4EqOpp-0.2	0.8488 ± 0.002	0.2063 ± 0.018	0.823 ± 0.0182	0.7311 ± 0.0413	0.3316 ± 0.0231	0.8356 ± 0.0091	0.785 ± 0.0243
LAFTR4EqOpp-0.5	0.8488 ± 0.002	0.2063 ± 0.018	0.823 ± 0.0182	0.7311 ± 0.0413	0.3316 ± 0.0231	0.8356 ± 0.0091	0.785 ± 0.0243
LAFTR4EqOpp-0.7	0.8488 ± 0.002	0.2063 ± 0.018	0.823 ± 0.0182	0.7311 ± 0.0413	0.3316 ± 0.0231	0.8356 ± 0.0091	0.785 ± 0.0243
LAFTR4EqOpp-1.0	0.8488 ± 0.002	0.2063 ± 0.018	0.823 ± 0.0182	0.7311 ± 0.0413	0.3316 ± 0.0231	0.8356 ± 0.0091	0.785 ± 0.0243

6.4.1 Utility

When training for this dataset, the baseline models, Unfair-LR and UnfairLR-decay, achieved utility performance of 85% and 83%, respectively. ZHANG et al. (2018) based implementations also presented an accuracy of $\approx$ 84%, 85%, and 85%. The LAFTR4DP-0.2 model reached an accuracy ≈ 0.85, and all other MADRAS et al. (2018) based implementations presented an accuracy near 84%. The fair models performed as well as the baseline.

From the statistical experiments, we understand that the UnfairLR model only outperforms, with a significance level, the UnfairLR-decay, and LAFTR4EqOpp (regardless of the fair coefficient).

Almost all comparisons between the fair models failed to reject the null hypothesis. However, the t-test experiments rejects the null hypothesis indicating that the LAFTR4EqOpp presents an underperformance for the accuracy when compared with the Zhang4EqOdds, Zhang4EqOpp, LAFTR4DP-0.2, LAFTR4DP-1.0 models.

6.4.2 Fairness

For the demographic disparity, all models presented low fairness performances. The UnfairLR baselines presented a DemDisp ≈ 0.16, and the UnfairLR-decay baseline presented a DemDisp ≈ 0.24 (which is the higher result). All ZHANG et al. (2018) based implementations presented similar results for the demographic disparity of ≈ 0.16. The LAFTR4EqOpp (regardless of the value for the fairness coefficient) model presented the second best result for DemDisp, followed by the LAFTR4DP-1.0 model. These fair models outperformed the UnfairLR implementation by ≈ 0.04 and 0.03, respectively.

When we look at the statistically comparisons between the UnfairLR model and the other implementation, we see that almost all tests reject the null hypothesis. This result means that the models are better than the UnfairLR model for this metric. The exceptions are the Zhang4EqOdds and Zhang4EqOpp models.

On the other hand, when we look at the t-test results from the comparison between the UnfairLR-decay model and the other implementation, we see that almost all tests reject the null hypothesis. This result means that the models are worse than the UnfairLR-decay model for this metric. The exceptions are the LAFTR4DP-0.7, LAFTR4DP-1.0, and LAFTR4EqOpp (regardless of the fair coefficient) models.

The paired t-test results also show that the LAFTR4EqOpp (regardless of the fair coefficient) outperforms the other models for this metric. However, the t-test failed to reject the null hypothesis compared with the UnfairLR-decay model.

For the disparity in equal odds, our baseline models reached high values (≈ 0.75). Analyzing the fair models, the three ZHANG et al. (2018) based implementations reached results for DispEqOdds $\approx$ 0.76, 0.75, and 0.75. The LAFTR4EqOpp (regardless of the fair coefficient) presented the best result for this fair metric, followed by the LAFTR4DP-1.0 model. These fair models outperformed the UnfairLR implementation by ≈ 0.07 and 0.06, respectively.

When we look at the statistical experiments' results from the comparison between the UnfairLR model and the other implementation, we see that almost all results reject the null hypothesis. This result means that the UnfairLR model presents an underperformance compared to almost all other models. The exceptions are the UnfairLR-decay model and ZHANG et al. (2018) based implementations, for which the t-test failed to reject the null hypothesis.

We have a similar understanding when we look at the t-test results from comparing the UnfairLR-decay model and the other implementation. We see that almost all tests reject the null hypothesis, which means that the models are better than the UnfairLR-decay model for this metric. The exceptions are the UnfairLR, ZHANG et al. (2018) based models, and the LAFTR4DP-0.2 model.

The paired t-test results also show that the LAFTR4EqOdds (regardless of the fair coefficient) outperform almost all other models. The t-test result failed to reject the null hypothesis only when the LAFTR4EqOdds (regardless of the fair coefficient) was compared with the LAFTR4DP-1.0 model.

Finally, when we look at the models' results for disparity in equal opportunity we also see a decrease in the models' fairness performance. The UnfairLR and UnfairLR-decay baselines both presented a DemDisp ≈ 0.57. ZHANG et al. (2018) based implementations reached results for DispEqOdds ($\approx$ 0.58, 0.56, and 0.57). The LAFTR4EqOpp (regardless of the value for the fairness coefficient) model presented the best result for DispEqOpp, followed by the LAFTR4DP-1.0 model. These fair models outperformed the UnfairLR implementation by ≈ 0.16 and 0.13, respectively.

When we look at the statistical experiments' results from the comparison between

the UnfairLR model and the other implementation, we see that all results from comparisons between this model and MADRAS et al. (2018) based models reject the null hypothesis. This result means that the UnfairLR presents an underperformance compared to almost models. The exceptions are the UnfairLR-decay model and ZHANG et al. (2018) based implementations, for which the t-test failed to reject the null hypothesis.

The paired t-test results also show that the LAFTR4EqOpp (regardless of the fair coefficient) outperforms almost all other models. The t-test result failed to reject the null hypothesis only when the LAFTR4EqOpp (regardless of the fair coefficient) was compared with the LAFTR4DP-1.0.

6.4.3 *FU-score*

Although all models presented a good accuracy performance, they performed badly for the DemDisp, which makes them penalized by the *FU-score* (when looking for this fair metric). The UnfairLR baselines presented a trade-off ≈ 0.27, and the UnfairLR-decay baseline presented a trade-off ≈ 0.37 (which is the higher result). ZHANG et al. (2018) based implementations reached results for this metric (≈ 0.27, 0.26, and 0.27). The LAFTR4EqOpp (regardless of the value for the fairness coefficient) model presented the second best result for this trade-off metric, followed by the LAFTR4DP-1.0 model. These fair models outperformed the UnfairLR implementation by ≈ 0.06 and 0.05, respectively.

When we look at the statistical experiments' results from comparing the UnfairLR model and the other implementation, we see that almost all tests reject the null hypothesis. This result means that the models are better than the UnfairLR model for this metric. The exceptions are the Zhang4EqOdds, Zhang4EqOpp, and LAFTR4DP-0.5 models.

On the other hand, when we look at the t-test results from the comparison between the UnfairLR-decay model and the other implementation, we see that almost all tests reject the null hypothesis, which means that the models are worse than the UnfairLR-decay model for this metric. The exceptions are the LAFTR4DP-1.0 and LAFTR4EqOpp (regardless of the fair coefficient) models.

The paired t-test results also show that the LAFTR4EqOpp (regardless of the fair coefficient) outperforms the other models for this metric. However, the t-test failed to reject the null hypothesis when comparing it with the UnfairLR-decay model.

For the trade-off between accuracy and disparity in equal odds assessment, almost all models presented a performance higher than 0.8. The UnfairLR-decay baseline model reached the lower trade-off value (≈ 0.794). The UnfairLR baseline model reached a trade-off value ≈ 0.8. Analyzing the fair models, ZHANG et al. (2018) based implementations reached results for DispEqOdds (≈ 0.8, 0.797, and 0.8). The LAFTR4EqOpp (regardless of the fair coefficient) presented the best result for this trade-off metric, followed by the

LAFTR4DP-1.0 model. These both models outperformed the UnfairLR implementation by ≈ 0.03.

When we look at the statistically comparisons between the UnfairLR model and the other implementation, we see that almost all tests reject the null hypothesis. This result means that the models are better than the UnfairLR model for this metric. The exceptions are the UnfairLR-decay, Zhang4DP, Zhang4EqOdds, Zhang4EqOpp, and LAFTR4DP-0.2 models.

The paired t-test results also show that the LAFTR4EqOpp (regardless of the fair coefficient) outperforms the other models for this metric, but the t-test failed to reject the null hypothesis when comparing it with the LAFTR4DP-0.7, and LAFTR4DP-1.0 model.

For the trade-off between accuracy and disparity in equal opportunity assessment, the models presented performances between 0.67 and 0.78. The unfair baseline models reached low trade-off values (≈ 0.68). Analyzing the fair models, ZHANG et al. (2018) based implementations reached results for DispEqOdds ($\approx$ 0.69, 0.67, and 0.68). The LAFTR4EqOpp (regardless of the fair coefficient) presented the best result for this trade-off metric, followed by the LAFTR4DP-1.0 model. These fair models outperformed the UnfairLR implementation by ≈ 0.1 and 0.08, respectively.

When we look at the statistical experiments for this trade-off from comparing the UnfairLR model and the other implementation, we see that almost all tests reject the null hypothesis. This result means that the models are better than the UnfairLR model for this metric. The exceptions are the UnfairLR-decay, Zhang4DP, Zhang4EqOdds, and Zhang4EqOpp models.

The paired t-test results also show that the LAFTR4EqOpp (regardless of the fair coefficient) outperforms the other models for this metric, but the t-test failed to reject the null hypothesis when comparing it with the LAFTR4DP-1.0 model.

6.4.4 Discussion

These results showed how the fair adversarial approaches perform when considering a non-binary protected attribute. This change did not harm the utility of these models. The t-test results of the models' assessments for the adult dataset showed that the UnfairLR outperforms, with a statistically significant difference, only the UnfairLR-decay, and LAFTR4EqOpp models.

Any model presented a good performance for the DemDisp metric. Which made the *FU-score* penalizes their accuracies. For this assessment, we understand that none of the models could encode the demographic parity for a non-binary protected attribute. For the other fair metrics, the models achieved better results. However, comparing their

performance for the other datasets, we observe that they achieved lower results for DispEqOdds and DispEqOpp.

Looking at the trade-off between accuracy and DispEqOdds, the LAFTR4EqOpp (regardless of the fair coefficient) had the highest result. A similar result occurs when looking at the trade-offs between accuracy and DispEqOpp. Also, the LAFTR4EqOpp had the highest result.

The LAFTR4EqOpp model's result only does not differ, with a statistical significance, compared to the LAFTR4DP-0.7 and LAFTR4DP-1.0 models' results for the trade-off between accuracy and DispEqOdds. Moreover, for the trade-off between accuracy and DispEqOpp, the LAFTR4EqOpp model's result only does not differ, with a statistical significance, compared to the LAFTR4DP-1.0 model's result.

When choosing a model for considering these both trade-offs, one could look to the utility and fair metrics individually to identify the most suitable model. For example, the LAFTR4EqOpp model presents a lower accuracy, with statistical significance, than the LAFTR4DP-0.7. On the other hand, for the DispEqOdds metric, the LAFTR4EqOpp model presents a higher result, with statistical significance, than the LAFTR4DP-0.7 model. Finally, for the DispEqOpp metric, the LAFTR4EqOpp model presents a higher result, with statistical significance, than the LAFTR4DP-0.7 model.

7 Conclusions

The ML fairness area is concerned with building models that mitigate bias and discrimination problems in algorithms. Many works are exploring this area and bringing strategies to build fairer models. Some of these are adversarial-based approaches.

Some of these approaches are in-processing strategies and based on an adversary's use to ensure a fairness constraint in the model. In addition to these strategies, we find other pre-processing proposals based on generative adversarial networks to generate fair data.

Furthermore, it is common in the fair ML area works to evaluate their models in a specific way, making it difficult to make a systematic assessment between the literature approaches and/or new strategies. Therefore, we mainly aimed to develop a benchmark to assess fair machine learning strategies using a performance-fairness trade-off metric, helping in the fairness area maturity. To achieve this goal, we:

- Proposed the *FU-score*, a fairness-utility trade-off metric to evaluate the fair strategies systemically;
- Defined a benchmark procedure, presenting the utility and fairness metrics, statistical tests, datasets, models and its implementation details;
- Applied the benchmark procedure the non-generative adversarial strategies to provide a comparative ruler for the fair ML area.

Following this procedure, we assessed the works of MADRAS et al. (2018), ZHANG et al. (2018), and BEUTEL et al. (2017) for the titanic, german, and adult datasets. We demonstrated how these approaches behave on these data, exploring the utility metric (accuracy), fairness metric (demographic disparity, disparity in equal odds, and disparity in equal opportunity), and the *FU-score* that computes the trade-off between accuracy and each of the fair metrics.

7.1 Final Remarks

We evaluated the non-generative adversarial models for the Titanic, German and Adult datasets over the utility, fairness, and *FU-score* perspectives. Our assessment brings the reproduction of the non-generative adversarial models' implementations for the Adult dataset with sex as the protected attribute. The assessment also brings the first view of the non-generative adversarial models' implementations and results for the following datasets: Titanic (with sex as the protected attribute), German (also with race

as the protected attribute), and Adult (with race as the protected attribute, which is a non-binary attribute).

The models' accuracy, fairness, and trade-off results vary over the assessed datasets. Thus, looking over all datasets and *FU-score* assessments, we can not determine a unique model that better performs the trade-off between accuracy and fairness for all cases (datasets and fairness metrics).

On the other hand, if we look individually for the accuracy or fairness metric, we can observe some behavior patterns. For example, the ZHANG et al. (2018) based models present better accuracies than the other fair models for all datasets and the unfair models in some cases. However, this is not the only aim of these models, and they do not present a good performance for the fairness metrics. This result shows the importance of evaluating the models from utility and fairness perspectives and, ideally, by a trade-off perspective.

Looking individually at the DemDisp, DispEqOdds, and DispEqOpp metrics for the Titanic and German datasets, we observe that the LAFTR4DP-0.7 model presents a higher performance than the other fair models. The BEUTEL4DP model presents a similar behavior, except for the DispEqOdds metrics and training the models for the Titanic dataset. However, when applying the paired t-test, we can make statements with statistical significance. For example, we observed that for Titanic, the LAFTR4DP-0.7 model did not outperform the fairness performance (for any metric) of the LAFTR4DP-1.0 and LAFTR4EqOpp models with statistical significance.

Moreover, we have different behaviors over the fairness metrics for the Adult dataset (sex). For DemDisp, the BEUTEL4DP model presents a higher result, followed by LAFTR4DP-1.0 and LAFTR4EqOpp models. For the other fair metrics, the LAFTR4EqOdds model presents a higher performance. However, looking at the paired t-test results for the DispEqOdds, we understand that the LAFTR4EqOdds model does not present a statistically significant difference in its performance compared to the LAFTR4DP-0.7, LAFTR4DP-1.0, LAFTR4EqOpp, and BEUTEL4DP models. On the other hand, for the DispEqOpp metric, the LAFTR4EqOdds model presents a better result, with statistical significance, than the LAFTR4DP-1.0 and LAFTR4EqOpp models.

Finally, we also have different behaviors over the metrics for the Adult dataset (race). For the DemDisp, any model could present a good performance, which made the *FU-score* metric penalize their utility performances. Again, this shows the importance of evaluating the models from a trade-off perspective. Furthermore, for the other fair metrics, the LAFTR4EqOpp model presents a higher performance, followed by the LAFTR4DP-1.0 model. The t-test showed that both models do not significantly differ in performance in these metrics.

All these trade-offs and individual assessments also show the importance of employing a statistical test in the assessment. When applying the statistical test, we ensure a certain confidence level in the results, statements, and analyses.

With this assessment and results, we could analyze the literature models from the same metric perspective and with statistical confidence in comparisons. From the *FU-score* perspective, we could observe if any model performed better for each dataset. On the other hand, given the variation of models' results (accuracy and fairness) for each dataset, we could not determine a unique model that better performs the trade-off between accuracy and fairness for all cases (datasets and fairness metrics). Nevertheless, new fairness works can use these results as a ruler to evaluate how its model proposal performs concerning the non-generative adversarial works.

7.2 Limitations and Future Work

Many of this work's limitations and difficulties were related to the models' implementations. For each model, we spent a considerable amount of time optimizing the model as much as the original paper describes.

A limiting factor is related to the reproducibility of the works. We did not find implementation details such as the number of hidden layers of a part of the model, applied data preparation, and the used activation and weight initialization functions.

This lack of reproducibility details caused us a change of work objectives moment. We firstly were attempting to reproduce the results presented by XU et al. (2018) to validate the hypothesis that by combining this approach with others, such as the MADRAS et al. (2018), we could increase accuracy simultaneously we would be able to increase fairness. However, many of these reproducibility problems were faced when attempting to implement the XU et al. (2018) work. We also tried to contact the authors to clarify some questions, however, we were not successful in getting an answer. Thus, due to time limitations, we decided to pivot the work to build the presented benchmark considering the non-generative adversarial approaches.

Time limitations were also a faced as a challenge. After this decision point, we only had 6 more months to conclude the work. Moreover, at this point, we still did not have the data preparation (for titanic and german datasets), statistical experiment defined, nor the BEUTEL et al. (2017) model implemented (and we would face a lack of implementation details for this work too). This is the main reason we could not provide the BEUTEL model implementation for equal odds, nor its implementation to a non-binary protected attribute.

The machine learning area presents many opportunities and gaps for fairness research. Here we point out some possible paths for futures works when considering the

adversarial approaches and our proposed benchmark (we pointed out some of these opportunities in our previous work LIMA et al. (2022)):

- In our work, we assessed only the non-generative adversarial works. One could consider applying the same systematical evaluation to other fair work or including the generative adversarial models;
- Many datasets present multiple attributes that we could consider protected. One could explore the intersectionality perspective on the adversarial models;
- One could explore the rich set of fairness definitions present in the literature, expanding the adversarial strategies to consider other fairness definitions and expanding the benchmark of these models;
- The *FU-score* metric was thought for classification problems. However, there are many datasets for the regression tasks used in fairness research. Thus, one could consider expanding the trade-off metric for this kind of task.

REFERENCES

ANGWIN, J., LARSON, J., MATTU, S., and KIRCHNER, L. (2016). Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. *ProPublica*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

BERK, R., HEIDARI, H., JABBARI, S., KEARNS, M., and ROTH, A. (2018). Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*.

BEUTEL, A., CHEN, J., ZHAO, Z., and CHI, E. H. (2017). Data decisions and theoretical implications when adversarially learning fair representations. *arXiv preprint arXiv:1707.00075*.

BOLUKBASI, T., CHANG, K.-W., ZOU, J. Y., SALIGRAMA, V., and KALAI, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*, pages 4349–4357.

BOUSMALIS, K., TRIGEORGIS, G., SILBERMAN, N., KRISHNAN, D., and ERHAN, D. (2016). Domain separation networks. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 343–351.

BROWNLEE, J. (2019). Statistical significance tests for comparing machine learning algorithms. https://machinelearningmastery.com/statistical-significance-tests-for-comparing-machine-learning-algorithms/.

CALDERS, T., KAMIRAN, F., and PECHENIZKIY, M. (2009). Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops*, pages 13–18. IEEE.

CALDERS, T. and VERWER, S. (2010). Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21(2):277–292.

COVER, T. M. and THOMAS, J. A. (2012). *Elements of information theory*. John Wiley & Sons.

DAL'EVEDOVE, P. R. and FUJITA, M. S. L. (2009). A abordagem sociológica em ciência da informação: um novo olhar investigativo. *A ciência da informação criadora de conhecimento*, 2:147–156.

DWORK, C., HARDT, M., PITASSI, T., REINGOLD, O., and ZEMEL, R. (2012). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226.

FACELI, K., LORENA, A. C., GAMA, J., ALMEIDA, T., and CARVALHO, A. (2021). *Inteligência Artificial: Uma Abordagem de Aprendizado de Máquina*. LTC, 2nd edition.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., et al. (2018). Ai4people—an ethical framework for a good ai society: Opportunities, risks, principles, and recommendations. *Minds and machines*, 28(4):689–707.

GANIN, Y., USTINOVA, E., AJAKAN, H., GERMAIN, P., LAROCHELLE, H., LAVIOLETTE, F., MARCHAND, M., and LEMPITSKY, V. (2016). Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030.

GARCIA, M. (2016). Racist in the machine: The disturbing implications of algorithmic bias. *World Policy Journal*, 33(4):111–117.

GOODFELLOW, I., POUGET-ABADIE, J., MIRZA, M., XU, B., WARDE-FARLEY, D., OZAIR, S., COURVILLE, A., and BENGIO, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.

HAO, K. (2020). The uk exam debacle reminds us that algorithms can't fix broken systems. *MIT Technology Review*. https://www.technologyreview.com/2020/08/20/1007502/uk-exam-algorithm-cant-fix-broken-system/.

HARDT, M., PRICE, E., and SREBRO, N. (2016). Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323.

HUTCHINSON, B. and MITCHELL, M. (2019). 50 years of test (un)fairness: Lessons for machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 49–58, New York, NY, USA. Association for Computing Machinery.

Jobin, A., Ienca, M., and Vayena, E. (2019). The global landscape of ai ethics guidelines. *Nature Machine Intelligence*, 1(9):389–399.

JONES, G. P., HICKEY, J. M., DI STEFANO, P. G., DHANJAL, C., STODDART, L. C., and VASILEIOU, V. (2020). Metrics and methods for a systematic comparison of fairness-aware machine learning algorithms. *arXiv preprint arXiv:2010.03986*.

KEARNS, M. and ROTH, A. (2019). *The ethical algorithm: The science of socially aware algorithm design*. Oxford University Press.

LEAVY, S. (2018). Gender bias in artificial intelligence: The need for diversity and gender theory in machine learning. In *Proceedings of the 1st international workshop on gender equality in software engineering*, pages 14–16.

LIMA, L., RICARTE, D., and SIEBRA, C. (2021). Assessing fair machine learning strategies through a fairness-utility trade-off metric. In *Anais do XVIII Encontro Nacional de Inteligência Artificial e Computacional*, pages 607–618, Porto Alegre, RS, Brasil. SBC.

LIMA, L. F. F. P., RICARTE, D. R. D., and SIEBRA, C. A. (2022). An overview on the use of adversarial learning strategies to ensure fairness in machine learning models. In *XVIII Brazilian Symposium on Information Systems*, SBSI, New York, NY, USA. Association for Computing Machinery.

LUM, K. and JOHNDROW, J. (2016). A statistical framework for fair predictive algorithms. *arXiv preprint arXiv:1610.08077*.

MAAS, A. L., HANNUN, A. Y., NG, A. Y., et al. (2013). Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, volume 30, page 3. Citeseer.

MADRAS, D., CREAGER, E., PITASSI, T., and ZEMEL, R. (2018). Learning adversarially fair and transferable representations. In *Proceedings of the 35th International Conference on Machine Learning*, pages 3384–3393.

MEHRABI, N., MORSTATTER, F., SAXENA, N., LERMAN, K., and GALSTYAN, A. (2019). A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*.

ODENA, A., OLAH, C., and SHLENS, J. (2017). Conditional image synthesis with auxiliary classifier gans. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2642–2651. JMLR. org.

RUSSEL, S. and NORVIG, P. (2021). *Artificial intelligence: a modern approach*. Person, 4rd edition.

VERMA, S. and RUBIN, J. (2018). Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*, pages 1–7. IEEE.

WAZLAWICK, R. S. (2020). *Metodologia de pesquisa para ciência da computação*. GEN LTC, 3 edition.

XU, D., YUAN, S., ZHANG, L., and WU, X. (2018). Fairgan: Fairness-aware generative adversarial networks. In *IEEE International Conference on Big Data (Big Data)*, pages 570–575. IEEE.

XU, D., YUAN, S., ZHANG, L., and WU, X. (2019). Fairgan+: Achieving fair data generation and classification through generative adversarial nets. In *IEEE International Conference on Big Data (Big Data)*, pages 1401–1406. IEEE.

ZHANG, B. H., LEMOINE, B., and MITCHELL, M. (2018). Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340.

A HYPOTHESIS TESTS' RESULTS

This Appendix presents the hypotheses test results for each task and metric. We present the t-test results in a NxN table, where N is the number of assessed models. We dashed the table's diagonal to indicate we did not make a paired comparison of the same model. When the means results are equal, the t-test returns an undefined value. We present these cases when the cells under the diagonal have no value filled. We bolted all results that reject the null hypothesis are bolted.

Table 10: Accuracy t-test results for Titanic dataset

	UnfairLR	UnfairLR-decay	Zhang4DP	Zhang4EqOdds	Zhang4EqOpp	LAFTR4DP-0.2	LAFTR4DP-0.5	LAFTR4DP-0.7	LAFTR4DP-1.0	LAFTR4EqOdds-0.2	LAFTR4EqOdds-0.5	LAFTR4EqOdds-0.7	LAFTR4EqOdds-1.0	LAFTR4EqOpp-0.2	LAFTR4EqOpp-0.5	LAFTR4EqOpp-0.7	LAFTR4EqOpp-1.0	BEUTEL4DP
UnfairLR	-																	
UnfairLR-decay	**0.00023026**	-																
Zhang4DP	0.0890093	**0.000199572**	-															
Zhang4EqOdds	0.304559	**0.000207585**	0.177808	-														
Zhang4EqOpp	0.512225	**0.000183753**	0.86053	0.871219	-													
LAFTR4DP-0.2	0.431377	**0.000126099**	0.183567	0.215116	0.256287	-												
LAFTR4DP-0.5	**0.0446534**	**0.00012734**	**0.0229836**	**0.0230504**	**0.0239335**	**0.00206001**	-											
LAFTR4DP-0.7	**0.0364125**	**7.59292e-05**	**0.0242062**	**0.0255526**	**0.017326**	**0.0122172**	0.105464	-										
LAFTR4DP-1.0	**0.0233994**	**0.000147266**	**0.017641**	**0.0189077**	**0.00863963**	**0.0216988**	0.33206	0.654148	-									
LAFTR4EqOdds-0.2	0.572441	**0.000176996**	0.433569	0.476621	0.402829	0.905727	0.108196	**0.0199788**	**0.00321042**	-								
LAFTR4EqOdds-0.5	0.572441	**0.000176996**	0.433569	0.476621	0.402829	0.905727	0.108196	**0.0199788**	**0.00321042**		-							
LAFTR4EqOdds-0.7	0.572441	**0.000176996**	0.433569	0.476621	0.402829	0.905727	0.108196	**0.0199788**	**0.00321042**			-						
LAFTR4EqOdds-1.0	0.572441	**0.000176996**	0.433569	0.476621	0.402829	0.905727	0.108196	**0.0199788**	**0.00321042**				-					
LAFTR4EqOpp-0.2	**0.00375982**	**0.000277892**	**0.00418107**	**0.005024**	**0.00212713**	**0.0176028**	0.175269	0.497338	0.182696	**0.0105444**	**0.0105444**	**0.0105444**	**0.0105444**	-				
LAFTR4EqOpp-0.5	**0.00375982**	**0.000277892**	**0.00418107**	**0.005024**	**0.00212713**	**0.0176028**	0.175269	0.497338	0.182696	**0.0105444**	**0.0105444**	**0.0105444**	**0.0105444**		-			
LAFTR4EqOpp-0.7	**0.00375982**	**0.000277892**	**0.00418107**	**0.005024**	**0.00212713**	**0.0176028**	0.175269	0.497338	0.182696	**0.0105444**	**0.0105444**	**0.0105444**	**0.0105444**			-		
LAFTR4EqOpp-1.0	**0.00375982**	**0.000277892**	**0.00418107**	**0.005024**	**0.00212713**	**0.0176028**	0.175269	0.497338	0.182696	**0.0105444**	**0.0105444**	**0.0105444**	**0.0105444**				-	
BEUTEL4DP	**1.09728e-05**	**0.0359656**	**5.84568e-06**	**6.57382e-06**	**1.37817e-05**	**1.33334e-05**	**4.5945e-05**	**0.000200407**	**0.000210828**	**0.000222224**	**0.000222224**	**0.000222224**	**0.000222224**	**0.000240537**	**0.000240537**	**0.000240537**	**0.000240537**	-

Table 11: DemDisp t-test results for Titanic dataset

	UnfairLR	UnfairLR-decay	Zhang4DP	Zhang4EqOdds	Zhang4EqOpp	LAFTR4DP-0.2	LAFTR4DP-0.5	LAFTR4DP-0.7	LAFTR4DP-1.0	LAFTR4EqOdds-0.2	LAFTR4EqOdds-0.5	LAFTR4EqOdds-0.7	LAFTR4EqOdds-1.0	LAFTR4EqOpp-0.2	LAFTR4EqOpp-0.5	LAFTR4EqOpp-0.7	LAFTR4EqOpp-1.0	BEUTEL4DP
UnfairLR	-																	
UnfairLR-decay	**2.22005e-05**	-																
Zhang4DP	**0.0434741**	**3.51775e-05**	-															
Zhang4EqOdds	0.0526358	**3.52586e-05**	0.72806	-														
Zhang4EqOpp	**0.0116988**	**2.82907e-05**	0.0908577	0.124094	-													
LAFTR4DP-0.2	0.423335	**3.02235e-05**	0.0799851	0.0680194	0.0564408	-												
LAFTR4DP-0.5	**0.0026986**	**0.000215864**	**0.000891085**	**0.000579785**	**0.00121755**	**0.000337263**	-											
LAFTR4DP-0.7	**0.000917337**	**0.00084282**	**0.000908349**	**0.000814917**	**0.00135976**	**9.47001e-05**	**0.0382887**	-										
LAFTR4DP-1.0	**0.00122218**	**0.00203897**	**0.00223834**	**0.00246287**	**0.00228778**	**0.00580311**	0.369332	0.50097	-									
LAFTR4EqOdds-0.2	0.623217	**0.000809744**	0.144536	0.163679	0.0925939	0.955923	0.058028	**0.0177**	**0.0143353**	-								
LAFTR4EqOdds-0.5	0.623217	**0.000809744**	0.144536	0.163679	0.0925939	0.955923	0.058028	**0.0177**	**0.0143353**		-							
LAFTR4EqOdds-0.7	0.623217	**0.000809744**	0.144536	0.163679	0.0925939	0.955923	0.058028	**0.0177**	**0.0143353**			-						
LAFTR4EqOdds-1.0	0.623217	**0.000809744**	0.144536	0.163679	0.0925939	0.955923	0.058028	**0.0177**	**0.0143353**				-					
LAFTR4EqOpp-0.2	**0.00349597**	**0.00797577**	**0.00249455**	**0.00264824**	**0.00230751**	**0.018038**	0.346547	0.903711	0.675781	**0.0123241**	**0.0123241**	**0.0123241**	**0.0123241**	-				
LAFTR4EqOpp-0.5	**0.00349597**	**0.00797577**	**0.00249455**	**0.00264824**	**0.00230751**	**0.018038**	0.346547	0.903711	0.675781	**0.0123241**	**0.0123241**	**0.0123241**	**0.0123241**		-			
LAFTR4EqOpp-0.7	**0.00349597**	**0.00797577**	**0.00249455**	**0.00264824**	**0.00230751**	**0.018038**	0.346547	0.903711	0.675781	**0.0123241**	**0.0123241**	**0.0123241**	**0.0123241**			-		
LAFTR4EqOpp-1.0	**0.00349597**	**0.00797577**	**0.00249455**	**0.00264824**	**0.00230751**	**0.018038**	0.346547	0.903711	0.675781	**0.0123241**	**0.0123241**	**0.0123241**	**0.0123241**				-	
BEUTEL4DP	**0.0329164**	0.215613	**0.0323446**	**0.0334259**	**0.027792**	**0.0415468**	0.219932	0.319916	0.214703	**0.0480956**	**0.0480956**	**0.0480956**	**0.0480956**	0.341307	0.341307	0.341307	0.341307	-

Table 12: DispEqOdds t-test results for Titanic dataset

	UnfairLR	UnfairLR-decay	Zhang4DP	Zhang4EqOdds	Zhang4EqOpp	LAFTR4DP-0.2	LAFTR4DP-0.5	LAFTR4DP-0.7	LAFTR4DP-1.0	LAFTR4EqOdds-0.2	LAFTR4EqOdds-0.5	LAFTR4EqOdds-0.7	LAFTR4EqOdds-1.0	LAFTR4EqOpp-0.2	LAFTR4EqOpp-0.5	LAFTR4EqOpp-0.7	LAFTR4EqOpp-1.0	BEUTEL4DP
UnfairLR	-																	
UnfairLR-decay	**6.15438e-05**	-																
Zhang4DP	**0.0231616**	**0.000106632**	-															
Zhang4EqOdds	**0.0199864**	**6.5997e-05**	0.463645	-														
Zhang4EqOpp	**0.023807**	**0.000237936**	0.151706	0.184829	-													
LAFTR4DP-0.2	0.316532	**9.00272e-05**	0.0542516	**0.0367923**	0.0504345	-												
LAFTR4DP-0.5	**0.0055359**	**0.00566202**	**0.00154637**	**0.000686592**	**0.00116126**	**0.00424402**	-											
LAFTR4DP-0.7	**0.00128758**	**0.00988101**	**0.000536242**	**0.000249998**	**0.000823147**	**0.000455799**	**0.023922**	-										
LAFTR4DP-1.0	**0.000640762**	**0.0304422**	**0.000776635**	**0.00054694**	**0.00110759**	**0.0051634**	0.278914	0.884872	-									
LAFTR4EqOdds-0.2	0.306602	**0.00495946**	**0.0140258**	**0.0186956**	**0.0162008**	0.898853	0.0527645	**0.017369**	**0.0164048**	-								
LAFTR4EqOdds-0.5	0.306602	**0.00495946**	**0.0140258**	**0.0186956**	**0.0162008**	0.898853	0.0527645	**0.017369**	**0.0164048**		-							
LAFTR4EqOdds-0.7	0.306602	**0.00495946**	**0.0140258**	**0.0186956**	**0.0162008**	0.898853	0.0527645	**0.017369**	**0.0164048**			-						
LAFTR4EqOdds-1.0	0.306602	**0.00495946**	**0.0140258**	**0.0186956**	**0.0162008**	0.898853	0.0527645	**0.017369**	**0.0164048**				-					
LAFTR4EqOpp-0.2	**0.0045572**	**0.0395546**	**0.00163824**	**0.00127736**	**0.00259813**	**0.0144813**	0.702581	0.369134	0.262159	**0.0153754**	**0.0153754**	**0.0153754**	**0.0153754**	-				
LAFTR4EqOpp-0.5	**0.0045572**	**0.0395546**	**0.00163824**	**0.00127736**	**0.00259813**	**0.0144813**	0.702581	0.369134	0.262159	**0.0153754**	**0.0153754**	**0.0153754**	**0.0153754**		-			
LAFTR4EqOpp-0.7	**0.0045572**	**0.0395546**	**0.00163824**	**0.00127736**	**0.00259813**	**0.0144813**	0.702581	0.369134	0.262159	**0.0153754**	**0.0153754**	**0.0153754**	**0.0153754**			-		
LAFTR4EqOpp-1.0	**0.0045572**	**0.0395546**	**0.00163824**	**0.00127736**	**0.00259813**	**0.0144813**	0.702581	0.369134	0.262159	**0.0153754**	**0.0153754**	**0.0153754**	**0.0153754**				-	
BEUTEL4DP	0.487826	0.110734	0.299867	0.294581	0.26477	0.704641	0.493117	0.308879	0.27745	0.776709	0.776709	0.776709	0.776709	0.376528	0.376528	0.376528	0.376528	-

Table 13: DispEqOpp t-test results for Titanic dataset

	UnfairLR	UnfairLR-decay	Zhang4DP	Zhang4EqOdds	Zhang4EqOpp	LAFTR4DP-0.2	LAFTR4DP-0.5	LAFTR4DP-0.7	LAFTR4DP-1.0	LAFTR4EqOdds-0.2	LAFTR4EqOdds-0.5	LAFTR4EqOdds-0.7	LAFTR4EqOdds-1.0	LAFTR4EqOpp-0.2	LAFTR4EqOpp-0.5	LAFTR4EqOpp-0.7	LAFTR4EqOpp-1.0	BEUTEL4DP
UnfairLR	-																	
UnfairLR-decay	**1.98178e-05**	-																
Zhang4DP	**0.0609167**	**0.000115342**	-															
Zhang4EqOdds	**0.0609167**	**0.000115342**		-														
Zhang4EqOpp	**0.0274399**	**5.76391e-05**	0.329804	0.329804	-													
LAFTR4DP-0.2	**0.00135477**	**0.000739217**	**0.00178284**	**0.00178284**	**0.00326366**	-												
LAFTR4DP-0.5	**0.000268499**	**0.0132267**	**0.000374593**	**0.000374593**	**0.000813915**	**4.40063e-05**	-											
LAFTR4DP-0.7	**4.10506e-05**	**0.0228214**	**0.000456132**	**0.000456132**	**0.000472613**	**0.000244552**	**0.0382108**	-										
LAFTR4DP-1.0	**0.000398081**	**0.023049**	**0.00150696**	**0.00150696**	**0.000692389**	**0.0126628**	0.472019	0.479552	-									
LAFTR4EqOdds-0.2	0.0556269	**0.00542025**	**0.00610717**	**0.00610717**	**0.00445207**	0.750956	**0.0491236**	**0.0258225**	**0.0437574**	-								
LAFTR4EqOdds-0.5	0.0556269	**0.00542025**	**0.00610717**	**0.00610717**	**0.00445207**	0.750956	**0.0491236**	**0.0258225**	**0.0437574**		-							
LAFTR4EqOdds-0.7	0.0556269	**0.00542025**	**0.00610717**	**0.00610717**	**0.00445207**	0.750956	**0.0491236**	**0.0258225**	**0.0437574**			-						
LAFTR4EqOdds-1.0	0.0556269	**0.00542025**	**0.00610717**	**0.00610717**	**0.00445207**	0.750956	**0.0491236**	**0.0258225**	**0.0437574**				-					
LAFTR4EqOpp-0.2	**0.00112641**	**0.0414516**	**0.00147207**	**0.00147207**	**0.000604776**	**0.0294963**	0.698792	0.428973	0.607071	**0.0248686**	**0.0248686**	**0.0248686**	**0.0248686**	-				
LAFTR4EqOpp-0.5	**0.00112641**	**0.0414516**	**0.00147207**	**0.00147207**	**0.000604776**	**0.0294963**	0.698792	0.428973	0.607071	**0.0248686**	**0.0248686**	**0.0248686**	**0.0248686**		-			
LAFTR4EqOpp-0.7	**0.00112641**	**0.0414516**	**0.00147207**	**0.00147207**	**0.000604776**	**0.0294963**	0.698792	0.428973	0.607071	**0.0248686**	**0.0248686**	**0.0248686**	**0.0248686**			-		
LAFTR4EqOpp-1.0	**0.00112641**	**0.0414516**	**0.00147207**	**0.00147207**	**0.000604776**	**0.0294963**	0.698792	0.428973	0.607071	**0.0248686**	**0.0248686**	**0.0248686**	**0.0248686**				-	
BEUTEL4DP	0.0572937	0.0725939	**0.0449842**	**0.0449842**	**0.02884**	0.39683	0.795759	0.453735	0.536472	0.340328	0.340328	0.340328	0.340328	0.653725	0.653725	0.653725	0.653725	-

Table 14: FU-score (DemDisp) t-test results for Titanic dataset

	UnfairLR	UnfairLR-decay	Zhang4DP	Zhang4EqOdds	Zhang4EqOpp	LAFTR4DP-0.2	LAFTR4DP-0.5	LAFTR4DP-0.7	LAFTR4DP-1.0	LAFTR4EqOdds-0.2	LAFTR4EqOdds-0.5	LAFTR4EqOdds-0.7	LAFTR4EqOdds-1.0	LAFTR4EqOpp-0.2	LAFTR4EqOpp-0.5	LAFTR4EqOpp-0.7	LAFTR4EqOpp-1.0	BEUTEL4DP
UnfairLR	-																	
UnfairLR-decay	**0.00835088**	-																
Zhang4DP	**0.0673071**	**0.0478992**	-															
Zhang4EqOdds	**0.0661591**	0.0506239	0.577423	-														
Zhang4EqOpp	**0.0271712**	0.050179	0.0885953	0.117071	-													
LAFTR4DP-0.2	0.395658	**0.0156854**	0.0921053	0.0759643	0.0594083	-												
LAFTR4DP-0.5	**0.00328988**	**0.00180787**	**0.00302254**	**0.00212766**	**0.00293349**	**0.00115225**	-											
LAFTR4DP-0.7	**0.00152019**	**0.00197509**	**0.00263361**	**0.002116**	**0.0030538**	**0.000321792**	**0.0238783**	-										
LAFTR4DP-1.0	**0.00226055**	**0.00309871**	**0.00521835**	**0.00481473**	0.0053207	**0.00494618**	0.380344	0.458721	-									
LAFTR4EqOdds-0.2	0.851859	**0.0472673**	0.0906212	0.107157	0.0618697	0.775206	0.0558418	**0.028282**	**0.0316763**	-								
LAFTR4EqOdds-0.5	0.851859	**0.0472673**	0.0906212	0.107157	0.0618697	0.775206	0.0558418	**0.028282**	**0.0316763**		-							
LAFTR4EqOdds-0.7	0.851859	**0.0472673**	0.0906212	0.107157	0.0618697	0.775206	0.0558418	**0.028282**	**0.0316763**			-						
LAFTR4EqOdds-1.0	0.851859	**0.0472673**	0.0906212	0.107157	0.0618697	0.775206	0.0558418	**0.028282**	**0.0316763**				-					
LAFTR4EqOpp-0.2	**0.00423564**	**0.00428387**	**0.00391803**	**0.00352328**	**0.00454858**	**0.0146623**	0.56893	0.437987	0.696705	**0.0218584**	**0.0218584**	**0.0218584**	**0.0218584**	-				
LAFTR4EqOpp-0.5	**0.00423564**	**0.00428387**	**0.00391803**	**0.00352328**	**0.00454858**	**0.0146623**	0.56893	0.437987	0.696705	**0.0218584**	**0.0218584**	**0.0218584**	**0.0218584**		-			
LAFTR4EqOpp-0.7	**0.00423564**	**0.00428387**	**0.00391803**	**0.00352328**	**0.00454858**	**0.0146623**	0.56893	0.437987	0.696705	**0.0218584**	**0.0218584**	**0.0218584**	**0.0218584**			-		
LAFTR4EqOpp-1.0	**0.00423564**	**0.00428387**	**0.00391803**	**0.00352328**	**0.00454858**	**0.0146623**	0.56893	0.437987	0.696705	**0.0218584**	**0.0218584**	**0.0218584**	**0.0218584**				-	
BEUTEL4DP	0.524731	0.284946	0.961611	0.941774	0.719948	0.358291	0.0651472	**0.0281051**	**0.0201632**	0.552308	0.552308	0.552308	0.552308	**0.0380442**	**0.0380442**	**0.0380442**	**0.0380442**	-

Table 15: FU-score (DispEqOdds) for Titanic dataset

	UnfairLR	UnfairLR-decay	Zhang4DP	Zhang4EqOdds	Zhang4EqOpp	LAFTR4DP-0.2	LAFTR4DP-0.5	LAFTR4DP-0.7	LAFTR4DP-1.0	LAFTR4EqOdds-0.2	LAFTR4EqOdds-0.5	LAFTR4EqOdds-0.7	LAFTR4EqOdds-1.0	LAFTR4EqOpp-0.2	LAFTR4EqOpp-0.5	LAFTR4EqOpp-0.7	LAFTR4EqOpp-1.0	BEUTEL4DP
UnfairLR	-																	
UnfairLR-decay	**0.0118624**	-																
Zhang4DP	**0.0419506**	**0.00180849**	-															
Zhang4EqOdds	**0.02897**	**0.00180645**	0.394363	-														
Zhang4EqOpp	**0.043306**	**0.000769722**	0.146556	0.184801	-													
LAFTR4DP-0.2	0.303166	**0.00234744**	0.0588631	**0.0355293**	0.050867	-												
LAFTR4DP-0.5	**0.0128048**	**0.000485385**	**0.00426188**	**0.00174208**	**0.0024235**	**0.0174375**	-											
LAFTR4DP-0.7	**0.00696413**	**0.000438254**	**0.00244555**	**0.000933293**	**0.00168681**	**0.007955**	**0.00472428**	-										
LAFTR4DP-1.0	**0.00454481**	**0.0006751**	**0.00346742**	**0.00169848**	**0.00324326**	**0.00872891**	0.261582	0.962152	-									
LAFTR4EqOdds-0.2	0.50863	**0.00189517**	**0.0146965**	**0.0115082**	**0.00779967**	0.967473	0.0609396	**0.0341508**	**0.0443463**	-								
LAFTR4EqOdds-0.5	0.50863	**0.00189517**	**0.0146965**	**0.0115082**	**0.00779967**	0.967473	0.0609396	**0.0341508**	**0.0443463**		-							
LAFTR4EqOdds-0.7	0.50863	**0.00189517**	**0.0146965**	**0.0115082**	**0.00779967**	0.967473	0.0609396	**0.0341508**	**0.0443463**			-						
LAFTR4EqOdds-1.0	0.50863	**0.00189517**	**0.0146965**	**0.0115082**	**0.00779967**	0.967473	0.0609396	**0.0341508**	**0.0443463**				-					
LAFTR4EqOpp-0.2	**0.0172711**	**0.00145062**	**0.00486995**	**0.00244093**	**0.00719349**	**0.0155922**	0.408239	0.123616	0.0982226	0.0803652	0.0803652	0.0803652	0.0803652	-				
LAFTR4EqOpp-0.5	**0.0172711**	**0.00145062**	**0.00486995**	**0.00244093**	**0.00719349**	**0.0155922**	0.408239	0.123616	0.0982226	0.0803652	0.0803652	0.0803652	0.0803652		-			
LAFTR4EqOpp-0.7	**0.0172711**	**0.00145062**	**0.00486995**	**0.00244093**	**0.00719349**	**0.0155922**	0.408239	0.123616	0.0982226	0.0803652	0.0803652	0.0803652	0.0803652			-		
LAFTR4EqOpp-1.0	**0.0172711**	**0.00145062**	**0.00486995**	**0.00244093**	**0.00719349**	**0.0155922**	0.408239	0.123616	0.0982226	0.0803652	0.0803652	0.0803652	0.0803652				-	
BEUTEL4DP	**0.0164483**	0.517014	0.0584049	0.0666543	0.133978	**0.0150864**	**0.0119839**	**0.00993473**	**0.00696571**	**0.0328527**	**0.0328527**	**0.0328527**	**0.0328527**	**0.00842418**	**0.00842418**	**0.00842418**	**0.00842418**	-

Table 16: FU-score (DispEqOpp) for Titanic dataset

	UnfairLR	UnfairLR-decay	Zhang4DP	Zhang4EqOdds	Zhang4EqOpp	LAFTR4DP-0.2	LAFTR4DP-0.5	LAFTR4DP-0.7	LAFTR4DP-1.0	LAFTR4EqOdds-0.2	LAFTR4EqOdds-0.5	LAFTR4EqOdds-0.7	LAFTR4EqOdds-1.0	LAFTR4EqOpp-0.2	LAFTR4EqOpp-0.5	LAFTR4EqOpp-0.7	LAFTR4EqOpp-1.0	BEUTEL4DP
UnfairLR	-																	
UnfairLR-decay	**0.0150854**	-																
Zhang4DP	0.0813774	**0.0276992**	-															
Zhang4EqOdds	0.076856	**0.0278067**	0.178939	-														
Zhang4EqOpp	**0.0390229**	0.0501385	0.29625	0.308761	-													
LAFTR4DP-0.2	**0.000579089**	**0.0020498**	**0.00271022**	**0.0025574**	**0.00279908**	-												
LAFTR4DP-0.5	**0.000131342**	**0.000944257**	**0.000736207**	**0.000690203**	**0.000693483**	**0.000133216**	-											
LAFTR4DP-0.7	**0.000130614**	**0.00086072**	**0.0010502**	**0.000990174**	**0.000782315**	**0.000452565**	**0.0253197**	-										
LAFTR4DP-1.0	**0.000621049**	**0.00181324**	**0.00300891**	**0.00286089**	**0.00153837**	**0.0153791**	0.518112	0.534362	-									
LAFTR4EqOdds-0.2	0.0675969	**0.00822472**	**0.00269088**	**0.00267609**	**0.0014867**	0.622986	0.0779957	0.0600882	0.0928945	-								
LAFTR4EqOdds-0.5	0.0675969	**0.00822472**	**0.00269088**	**0.00267609**	**0.0014867**	0.622986	0.0779957	0.0600882	0.0928945		-							
LAFTR4EqOdds-0.7	0.0675969	**0.00822472**	**0.00269088**	**0.00267609**	**0.0014867**	0.622986	0.0779957	0.0600882	0.0928945			-						
LAFTR4EqOdds-1.0	0.0675969	**0.00822472**	**0.00269088**	**0.00267609**	**0.0014867**	0.622986	0.0779957	0.0600882	0.0928945				-					
LAFTR4EqOpp-0.2	**0.000440463**	**0.00272119**	**0.00217741**	**0.0020697**	**0.000934142**	**0.0424351**	0.622681	0.151301	0.190456	0.0811994	0.0811994	0.0811994	0.0811994	-				
LAFTR4EqOpp-0.5	**0.000440463**	**0.00272119**	**0.00217741**	**0.0020697**	**0.000934142**	**0.0424351**	0.622681	0.151301	0.190456	0.0811994	0.0811994	0.0811994	0.0811994	0.0811994	-			
LAFTR4EqOpp-0.7	**0.000440463**	**0.00272119**	**0.00217741**	**0.0020697**	**0.000934142**	**0.0424351**	0.622681	0.151301	0.190456	0.0811994	0.0811994	0.0811994	0.0811994	0.0811994	0.0811994	-		
LAFTR4EqOpp-1.0	**0.000440463**	**0.00272119**	**0.00217741**	**0.0020697**	**0.000934142**	**0.0424351**	0.622681	0.151301	0.190456	0.0811994	0.0811994	0.0811994	0.0811994	0.0811994	0.0811994	0.0811994	-	
BEUTEL4DP	0.137496	0.312254	0.645216	0.652484	0.787255	0.0274702	**0.0111974**	**0.00765775**	**0.0067257**	0.0625868	0.0625868	0.0625868	0.0625868	**0.00897153**	**0.00897153**	**0.00897153**	**0.00897153**	-

Table 17: Accuracy t-test results for German dataset

	UnfairLR	UnfairLR-decay	Zhang4DP	Zhang4EqOdds	Zhang4EqOpp	LAFTR4DP-0.2	LAFTR4DP-0.5	LAFTR4DP-0.7	LAFTR4DP-1.0	LAFTR4EqOdds-0.2	LAFTR4EqOdds-0.5	LAFTR4EqOdds-0.7	LAFTR4EqOdds-1.0	LAFTR4EqOpp-0.2	LAFTR4EqOpp-0.5	LAFTR4EqOpp-0.7	LAFTR4EqOpp-1.0	BEUTEL4DP
UnfairLR	-																	
UnfairLR-decay	0.306791	-																
Zhang4DP	0.183567	0.183676	-															
Zhang4EqOdds	0.136945	0.167103	0.373901	-														
Zhang4EqOpp	0.362662	0.298015	1	0.748868	-													
LAFTR4DP-0.2	0.438199	0.601832	0.238628	0.15858	0.258651	-												
LAFTR4DP-0.5	0.135965	0.730399	0.0756321	**0.036352**	0.113539	0.337502	-											
LAFTR4DP-0.7	0.119055	0.850086	0.054217	**0.030466**	0.0971852	0.134702	0.426317	-										
LAFTR4DP-1.0	0.234719	0.775076	0.106083	0.0628076	0.141927	0.189004	0.827565	0.373901	-									
LAFTR4EqOdds-0.2	0.31943	0.611529	0.128353	0.079488	0.201555	0.688457	0.704	**0.0341094**	0.208	-								
LAFTR4EqOdds-0.5	0.31943	0.611529	0.128353	0.079488	0.201555	0.688457	0.704	**0.0341094**	0.208		-							
LAFTR4EqOdds-0.7	0.31943	0.611529	0.128353	0.079488	0.201555	0.688457	0.704	**0.0341094**	0.208			-						
LAFTR4EqOdds-1.0	0.31943	0.611529	0.128353	0.079488	0.201555	0.688457	0.704	**0.0341094**	0.208				-					
LAFTR4EqOpp-0.2	0.163608	0.897595	0.0858104	0.053604	**0.0320899**	0.350815	0.674518	1	0.803333	0.600424	0.600424	0.600424	0.600424	-				
LAFTR4EqOpp-0.5	0.163608	0.897595	0.0858104	0.053604	**0.0320899**	0.350815	0.674518	1	0.803333	0.600424	0.600424	0.600424	0.600424	0.600424	-			
LAFTR4EqOpp-0.7	0.163608	0.897595	0.0858104	0.053604	**0.0320899**	0.350815	0.674518	1	0.803333	0.600424	0.600424	0.600424	0.600424	0.600424	0.600424	-		
LAFTR4EqOpp-1.0	0.163608	0.897595	0.0858104	0.053604	**0.0320899**	0.350815	0.674518	1	0.803333	0.600424	0.600424	0.600424	0.600424	0.600424	0.600424	0.600424	-	
BEUTEL4DP	0.0625058	0.568982	**0.0272489**	**0.0256154**	0.105366	0.344976	0.387222	0.51157	0.481934	0.313323	0.313323	0.313323	0.313323	0.659131	0.659131	0.659131	0.659131	-

Table 18: DemDisp t-test results for German dataset

	UnfairLR	UnfairLR-decay	Zhang4DP	Zhang4EqOdds	Zhang4EqOpp	LAFTR4DP-0.2	LAFTR4DP-0.5	LAFTR4DP-0.7	LAFTR4DP-1.0	LAFTR4EqOdds-0.2	LAFTR4EqOdds-0.5	LAFTR4EqOdds-0.7	LAFTR4EqOdds-1.0	LAFTR4EqOpp-0.2	LAFTR4EqOpp-0.5	LAFTR4EqOpp-0.7	LAFTR4EqOpp-1.0	BEUTEL4DP
UnfairLR	-																	
UnfairLR-decay	0.0501535	-																
Zhang4DP	0.0501878	**0.029407**	-															
Zhang4EqOdds	0.194388	**0.0314463**	0.914128	-														
Zhang4EqOpp	0.201009	**0.0345959**	0.28447	0.307761	-													
LAFTR4DP-0.2	0.760115	**0.0152871**	0.302858	0.296848	0.397866	-												
LAFTR4DP-0.5	0.541042	**0.00160958**	0.245374	0.247676	0.303698	0.455189	-											
LAFTR4DP-0.7	0.471448	**0.0173506**	0.245877	0.263108	0.297871	0.327221	0.655237	-										
LAFTR4DP-1.0	0.550153	**0.00563292**	0.283083	0.299585	0.344437	0.560847	0.919533	0.646896	-									
LAFTR4EqOdds-0.2	0.458494	**0.0027926**	0.211052	0.23164	0.268919	0.440744	0.701552	0.811744	0.783898	-								
LAFTR4EqOdds-0.5	0.458494	**0.0027926**	0.211052	0.23164	0.268919	0.440744	0.701552	0.811744	0.783898		-							
LAFTR4EqOdds-0.7	0.458494	**0.0027926**	0.211052	0.23164	0.268919	0.440744	0.701552	0.811744	0.783898			-						
LAFTR4EqOdds-1.0	0.458494	**0.0027926**	0.211052	0.23164	0.268919	0.440744	0.701552	0.811744	0.783898				-					
LAFTR4EqOpp-0.2	0.768199	**0.00736089**	0.428271	0.45611	0.538021	0.929576	0.685053	0.545498	0.557592	0.490548	0.490548	0.490548	0.490548	-				
LAFTR4EqOpp-0.5	0.768199	**0.00736089**	0.428271	0.45611	0.538021	0.929576	0.685053	0.545498	0.557592	0.490548	0.490548	0.490548	0.490548		-			
LAFTR4EqOpp-0.7	0.768199	**0.00736089**	0.428271	0.45611	0.538021	0.929576	0.685053	0.545498	0.557592	0.490548	0.490548	0.490548	0.490548			-		
LAFTR4EqOpp-1.0	0.768199	**0.00736089**	0.428271	0.45611	0.538021	0.929576	0.685053	0.545498	0.557592	0.490548	0.490548	0.490548	0.490548				-	
BEUTEL4DP	**0.0479632**	0.373901	**0.0276554**	**0.0293546**	**0.0321525**	**0.013688**	**0.00111499**	**0.0170566**	**0.00542071**	**0.00236269**	**0.00236269**	**0.00236269**	**0.00236269**	**0.0088721**	**0.0088721**	**0.0088721**	**0.0088721**	-

Table 19: DispEqOdds t-test results for German dataset

	UnfairLR	UnfairLR-decay	Zhang4DP	Zhang4EqOdds	Zhang4EqOpp	LAFTR4DP-0.2	LAFTR4DP-0.5	LAFTR4DP-0.7	LAFTR4DP-1.0	LAFTR4EqOdds-0.2	LAFTR4EqOdds-0.5	LAFTR4EqOdds-0.7	LAFTR4EqOdds-1.0	LAFTR4EqOpp-0.2	LAFTR4EqOpp-0.5	LAFTR4EqOpp-0.7	LAFTR4EqOpp-1.0	BEUTEL4DP
UnfairLR	-																	
UnfairLR-decay	**0.00129466**	-																
Zhang4DP	0.14304	**0.00455159**	-															
Zhang4EqOdds	0.363574	**0.00530081**	0.592547	-														
Zhang4EqOpp	0.763703	**0.018249**	0.477311	0.480589	-													
LAFTR4DP-0.2	0.697123	**0.0204406**	0.805845	0.923645	0.860366	-												
LAFTR4DP-0.5	0.487764	**0.028559**	0.315506	0.374278	0.525424	0.212976	-											
LAFTR4DP-0.7	0.292439	0.0583224	0.207157	0.2297	0.348037	0.0871335	0.0960264	-										
LAFTR4DP-1.0	0.860962	**0.00873719**	0.410477	0.53513	0.779792	0.426602	0.355206	0.135798	-									
LAFTR4EqOdds-0.2	0.533043	**0.00289786**	0.239639	0.239293	0.478598	0.231589	0.715604	0.240055	0.599856	-								
LAFTR4EqOdds-0.5	0.533043	**0.00289786**	0.239639	0.239293	0.478598	0.231589	0.715604	0.240055	0.599856		-							
LAFTR4EqOdds-0.7	0.533043	**0.00289786**	0.239639	0.239293	0.478598	0.231589	0.715604	0.240055	0.599856			-						
LAFTR4EqOdds-1.0	0.533043	**0.00289786**	0.239639	0.239293	0.478598	0.231589	0.715604	0.240055	0.599856				-					
LAFTR4EqOpp-0.2	0.980417	**0.0112619**	0.580021	0.686297	0.875937	0.665289	0.0916445	0.0705306	0.838049	0.541676	0.541676	0.541676	0.541676	-				
LAFTR4EqOpp-0.5	0.980417	**0.0112619**	0.580021	0.686297	0.875937	0.665289	0.0916445	0.0705306	0.838049	0.541676	0.541676	0.541676	0.541676		-			
LAFTR4EqOpp-0.7	0.980417	**0.0112619**	0.580021	0.686297	0.875937	0.665289	0.0916445	0.0705306	0.838049	0.541676	0.541676	0.541676	0.541676			-		
LAFTR4EqOpp-1.0	0.980417	**0.0112619**	0.580021	0.686297	0.875937	0.665289	0.0916445	0.0705306	0.838049	0.541676	0.541676	0.541676	0.541676				-	
BEUTEL4DP	**0.0012675**	0.373901	**0.00398905**	**0.00452844**	**0.017586**	**0.02166**	**0.0350064**	0.0688271	**0.00980604**	**0.00302007**	**0.00302007**	**0.00302007**	**0.00302007**	**0.0142295**	**0.0142295**	**0.0142295**	**0.0142295**	-

Table 20: DispEqOpp t-test results for German dataset

	UnfairLR	UnfairLR-decay	Zhang4DP	Zhang4EqOdds	Zhang4EqOpp	LAFTR4DP-0.2	LAFTR4DP-0.5	LAFTR4DP-0.7	LAFTR4DP-1.0	LAFTR4EqOdds-0.2	LAFTR4EqOdds-0.5	LAFTR4EqOdds-0.7	LAFTR4EqOdds-1.0	LAFTR4EqOpp-0.2	LAFTR4EqOpp-0.5	LAFTR4EqOpp-0.7	LAFTR4EqOpp-1.0	BEUTEL4DP
UnfairLR	-																	
UnfairLR-decay	**1.98178e-05**	-																
Zhang4DP	**0.0609167**	**0.000115342**	-															
Zhang4EqOdds	**0.0609167**	**0.000115342**		-														
Zhang4EqOpp	**0.0274399**	**5.76391e-05**	0.329804	0.329804	-													
LAFTR4DP-0.2	**0.00135477**	**0.000739217**	**0.00178284**	**0.00178284**	**0.00326366**	-												
LAFTR4DP-0.5	**0.000268499**	**0.0132267**	**0.000374593**	**0.000374593**	**0.000813915**	**4.40063e-05**	-											
LAFTR4DP-0.7	**4.10506e-05**	**0.0228214**	**0.000456132**	**0.000456132**	**0.000472613**	**0.000244552**	**0.0382108**	-										
LAFTR4DP-1.0	**0.000398081**	**0.023049**	**0.00150696**	**0.00150696**	**0.000692389**	**0.0126628**	0.472019	0.479552	-									
LAFTR4EqOdds-0.2	0.0556269	**0.00542025**	**0.00610717**	**0.00610717**	**0.00445207**	0.750956	**0.0491236**	**0.0258225**	**0.0437574**	-								
LAFTR4EqOdds-0.5	0.0556269	**0.00542025**	**0.00610717**	**0.00610717**	**0.00445207**	0.750956	**0.0491236**	**0.0258225**	**0.0437574**		-							
LAFTR4EqOdds-0.7	0.0556269	**0.00542025**	**0.00610717**	**0.00610717**	**0.00445207**	0.750956	**0.0491236**	**0.0258225**	**0.0437574**			-						
LAFTR4EqOdds-1.0	0.0556269	**0.00542025**	**0.00610717**	**0.00610717**	**0.00445207**	0.750956	**0.0491236**	**0.0258225**	**0.0437574**				-					
LAFTR4EqOpp-0.2	**0.00112641**	**0.0414516**	**0.00147207**	**0.00147207**	**0.000604776**	**0.0294963**	0.698792	0.428973	0.607071	**0.0248686**	**0.0248686**	**0.0248686**	**0.0248686**	-				
LAFTR4EqOpp-0.5	**0.00112641**	**0.0414516**	**0.00147207**	**0.00147207**	**0.000604776**	**0.0294963**	0.698792	0.428973	0.607071	**0.0248686**	**0.0248686**	**0.0248686**	**0.0248686**		-			
LAFTR4EqOpp-0.7	**0.00112641**	**0.0414516**	**0.00147207**	**0.00147207**	**0.000604776**	**0.0294963**	0.698792	0.428973	0.607071	**0.0248686**	**0.0248686**	**0.0248686**	**0.0248686**			-		
LAFTR4EqOpp-1.0	**0.00112641**	**0.0414516**	**0.00147207**	**0.00147207**	**0.000604776**	**0.0294963**	0.698792	0.428973	0.607071	**0.0248686**	**0.0248686**	**0.0248686**	**0.0248686**				-	
BEUTEL4DP	0.0572937	0.0725939	**0.0449842**	**0.0449842**	**0.02884**	0.39683	0.795759	0.453735	0.536472	0.340328	0.340328	0.340328	0.340328	0.653725	0.653725	0.653725	0.653725	-

Table 21: FU-score (DemDisp) t-test results for German dataset

	UnfairLR	UnfairLR-decay	Zhang4DP	Zhang4EqOdds	Zhang4EqOpp	LAFTR4DP-0.2	LAFTR4DP-0.5	LAFTR4DP-0.7	LAFTR4DP-1.0	LAFTR4EqOdds-0.2	LAFTR4EqOdds-0.5	LAFTR4EqOdds-0.7	LAFTR4EqOdds-1.0	LAFTR4EqOpp-0.2	LAFTR4EqOpp-0.5	LAFTR4EqOpp-0.7	LAFTR4EqOpp-1.0	BEUTEL4DP
UnfairLR	-																	
UnfairLR-decay	0.170368	-																
Zhang4DP	0.128194	0.124019	-															
Zhang4EqOdds	0.321819	0.132528	0.713404	-														
Zhang4EqOpp	0.918716	0.245869	0.093312	0.154076	-													
LAFTR4DP-0.2	0.949295	0.170636	0.552285	0.612092	0.922681	-												
LAFTR4DP-0.5	0.630883	0.0888806	0.36152	0.421506	0.722981	0.584109	-											
LAFTR4DP-0.7	0.584814	0.119981	0.36509	0.421791	0.687235	0.475305	0.501533	-										
LAFTR4DP-1.0	0.768148	0.0957443	0.438965	0.521417	0.824199	0.813001	0.526141	0.310442	-									
LAFTR4EqOdds-0.2	0.520338	0.102861	0.279881	0.350074	0.654601	0.553848	0.747654	0.903817	0.371996	-								
LAFTR4EqOdds-0.5	0.520338	0.102861	0.279881	0.350074	0.654601	0.553848	0.747654	0.903817	0.371996		-							
LAFTR4EqOdds-0.7	0.520338	0.102861	0.279881	0.350074	0.654601	0.553848	0.747654	0.903817	0.371996			-						
LAFTR4EqOdds-1.0	0.520338	0.102861	0.279881	0.350074	0.654601	0.553848	0.747654	0.903817	0.371996				-					
LAFTR4EqOpp-0.2	0.933995	0.102408	0.542668	0.635545	0.922693	0.989751	0.601217	0.503668	0.703681	0.413955	0.413955	0.413955	0.413955	-				
LAFTR4EqOpp-0.5	0.933995	0.102408	0.542668	0.635545	0.922693	0.989751	0.601217	0.503668	0.703681	0.413955	0.413955	0.413955	0.413955		-			
LAFTR4EqOpp-0.7	0.933995	0.102408	0.542668	0.635545	0.922693	0.989751	0.601217	0.503668	0.703681	0.413955	0.413955	0.413955	0.413955			-		
LAFTR4EqOpp-1.0	0.933995	0.102408	0.542668	0.635545	0.922693	0.989751	0.601217	0.503668	0.703681	0.413955	0.413955	0.413955	0.413955				-	
BEUTEL4DP	0.126979	0.587542	0.0920229	0.101739	0.218339	0.172327	0.0699114	0.1084	0.0717649	0.077985	0.077985	0.077985	0.077985	0.0865135	0.0865135	0.0865135	0.0865135	-

Table 22: FU-score (DispEqOdds) for German dataset

	UnfairLR	UnfairLR-decay	Zhang4DP	Zhang4EqOdds	Zhang4EqOpp	LAFTR4DP-0.2	LAFTR4DP-0.5	LAFTR4DP-0.7	LAFTR4DP-1.0	LAFTR4EqOdds-0.2	LAFTR4EqOdds-0.5	LAFTR4EqOdds-0.7	LAFTR4EqOdds-1.0	LAFTR4EqOpp-0.2	LAFTR4EqOpp-0.5	LAFTR4EqOpp-0.7	LAFTR4EqOpp-1.0	BEUTEL4DP
UnfairLR	-																	
UnfairLR-decay	0.240511	-																
Zhang4DP	0.161773	0.17472	-															
Zhang4EqOdds	0.330638	0.181763	0.956188	-														
Zhang4EqOpp	0.545205	0.24778	0.264196	0.31359	-													
LAFTR4DP-0.2	0.942657	0.121257	0.56589	0.598528	0.779049	-												
LAFTR4DP-0.5	0.766861	0.0653596	0.484521	0.51493	0.66803	0.646649	-											
LAFTR4DP-0.7	0.761068	0.094979	0.513233	0.547327	0.679944	0.608995	0.889974	-										
LAFTR4DP-1.0	0.807738	0.0733402	0.539739	0.575583	0.710956	0.767214	0.935734	0.822818	-									
LAFTR4EqOdds-0.2	0.678702	0.079691	0.433313	0.474946	0.613121	0.570789	0.620959	0.75733	0.424752	-								
LAFTR4EqOdds-0.5	0.678702	0.079691	0.433313	0.474946	0.613121	0.570789	0.620959	0.75733	0.424752		-							
LAFTR4EqOdds-0.7	0.678702	0.079691	0.433313	0.474946	0.613121	0.570789	0.620959	0.75733	0.424752			-						
LAFTR4EqOdds-1.0	0.678702	0.079691	0.433313	0.474946	0.613121	0.570789	0.620959	0.75733	0.424752				-					
LAFTR4EqOpp-0.2	0.949119	0.0905733	0.709419	0.745685	0.890363	0.89069	0.573546	0.516228	0.41542	0.257009	0.257009	0.257009	0.257009	-				
LAFTR4EqOpp-0.5	0.949119	0.0905733	0.709419	0.745685	0.890363	0.89069	0.573546	0.516228	0.41542	0.257009	0.257009	0.257009	0.257009		-			
LAFTR4EqOpp-0.7	0.949119	0.0905733	0.709419	0.745685	0.890363	0.89069	0.573546	0.516228	0.41542	0.257009	0.257009	0.257009	0.257009			-		
LAFTR4EqOpp-1.0	0.949119	0.0905733	0.709419	0.745685	0.890363	0.89069	0.573546	0.516228	0.41542	0.257009	0.257009	0.257009	0.257009				-	
BEUTEL4DP	0.208957	0.550229	0.145934	0.15376	0.218994	0.0971376	**0.0279495**	0.0649025	**0.0414791**	**0.0439668**	**0.0439668**	**0.0439668**	**0.0439668**	0.0689468	0.0689468	0.0689468	0.0689468	-

Table 23: FU-score (DispEqOpp) for German dataset

	UnfairLR	UnfairLR-decay	Zhang4DP	Zhang4EqOdds	Zhang4EqOpp	LAFTR4DP-0.2	LAFTR4DP-0.5	LAFTR4DP-0.7	LAFTR4DP-1.0	LAFTR4EqOdds-0.2	LAFTR4EqOdds-0.5	LAFTR4EqOdds-0.7	LAFTR4EqOdds-1.0	LAFTR4EqOpp-0.2	LAFTR4EqOpp-0.5	LAFTR4EqOpp-0.7	LAFTR4EqOpp-1.0	BEUTEL4DP
UnfairLR	-																	
UnfairLR-decay	0.372113	-																
Zhang4DP	0.646827	0.402464	-															
Zhang4EqOdds	0.850734	0.463396	0.276611	-														
Zhang4EqOpp	0.734964	0.580419	0.424061	0.766963	-													
LAFTR4DP-0.2	0.432749	0.162834	0.672349	0.516909	0.503992	-												
LAFTR4DP-0.5	0.847884	0.205831	0.984216	0.8401	0.801092	0.367115	-											
LAFTR4DP-0.7	0.990817	0.182457	0.880147	0.945951	0.884192	0.309855	0.35855	-										
LAFTR4DP-1.0	0.453289	0.194053	0.709107	0.548817	0.556102	0.798341	0.251232	0.251666	-									
LAFTR4EqOdds-0.2	0.658251	0.126251	0.961225	0.713487	0.687567	0.449795	0.86947	0.53382	0.471835	-								
LAFTR4EqOdds-0.5	0.658251	0.126251	0.961225	0.713487	0.687567	0.449795	0.86947	0.53382	0.471835		-							
LAFTR4EqOdds-0.7	0.658251	0.126251	0.961225	0.713487	0.687567	0.449795	0.86947	0.53382	0.471835			-						
LAFTR4EqOdds-1.0	0.658251	0.126251	0.961225	0.713487	0.687567	0.449795	0.86947	0.53382	0.471835				-					
LAFTR4EqOpp-0.2	0.199	0.120964	0.488272	0.351519	0.37427	0.900047	0.218766	0.166049	0.48301	0.163	0.163	0.163	0.163	-				
LAFTR4EqOpp-0.5	0.199	0.120964	0.488272	0.351519	0.37427	0.900047	0.218766	0.166049	0.48301	0.163	0.163	0.163	0.163		-			
LAFTR4EqOpp-0.7	0.199	0.120964	0.488272	0.351519	0.37427	0.900047	0.218766	0.166049	0.48301	0.163	0.163	0.163	0.163			-		
LAFTR4EqOpp-1.0	0.199	0.120964	0.488272	0.351519	0.37427	0.900047	0.218766	0.166049	0.48301	0.163	0.163	0.163	0.163				-	
BEUTEL4DP	0.338439	0.573436	0.363427	0.439576	0.636264	0.195737	0.253256	0.25945	0.197561	0.0941543	0.0941543	0.0941543	0.0941543	0.110777	0.110777	0.110777	0.110777	-

Table 24: Accuracy t-test results for Adult dataset

	UnfairLR	UnfairLR-decay	Zhang4DP	Zhang4EqOdds	Zhang4EqOpp	LAFTR4DP-0.2	LAFTR4DP-0.5	LAFTR4DP-0.7	LAFTR4DP-1.0	LAFTR4EqOdds-0.2	LAFTR4EqOdds-0.5	LAFTR4EqOdds-0.7	LAFTR4EqOdds-1.0	LAFTR4EqOpp-0.2	LAFTR4EqOpp-0.5	LAFTR4EqOpp-0.7	LAFTR4EqOpp-1.0	BEUTEL4DP
UnfairLR	-																	
UnfairLR-decay	**0.000301503**	-																
Zhang4DP	0.132777	**0.00034997**	-															
Zhang4EqOdds	0.143306	**0.00030937**	0.936951	-														
Zhang4EqOpp	0.581648	**0.000130591**	0.145197	0.172379	-													
LAFTR4DP-0.2	0.31907	**0.000207414**	0.201786	0.192542	0.583122	-												
LAFTR4DP-0.5	0.257908	7.32451e-05	0.175545	0.160477	0.259088	0.296019	-											
LAFTR4DP-0.7	0.188425	7.90819e-06	0.119081	0.118603	0.122603	0.299682	0.446976	-										
LAFTR4DP-1.0	**0.0190628**	**0.000206233**	**0.00872649**	**0.00684868**	**0.00899122**	0.0842712	0.203912	0.3922	-									
LAFTR4EqOdds-0.2	0.104996	**8.96846e-05**	**0.0447931**	**0.0398621**	**0.0249743**	0.420812	1	0.315354	**0.00692789**	-								
LAFTR4EqOdds-0.5	0.104996	**8.96846e-05**	**0.0447931**	**0.0398621**	**0.0249743**	0.420812	1	0.315354	**0.00692789**		-							
LAFTR4EqOdds-0.7	0.104996	**8.96846e-05**	**0.0447931**	**0.0398621**	**0.0249743**	0.420812	1	0.315354	**0.00692789**			-						
LAFTR4EqOdds-1.0	0.104996	**8.96846e-05**	**0.0447931**	**0.0398621**	**0.0249743**	0.420812	1	0.315354	**0.00692789**				-					
LAFTR4EqOpp-0.2	**0.0324334**	**6.12996e-05**	**0.0244408**	**0.022689**	**0.0176008**	0.0678381	0.123311	0.0887613	0.356467	**0.0183996**	**0.0183996**	**0.0183996**	**0.0183996**	-				
LAFTR4EqOpp-0.5	**0.0324334**	**6.12996e-05**	**0.0244408**	**0.022689**	**0.0176008**	0.0678381	0.123311	0.0887613	0.356467	**0.0183996**	**0.0183996**	**0.0183996**	**0.0183996**		-			
LAFTR4EqOpp-0.7	**0.0324334**	**6.12996e-05**	**0.0244408**	**0.022689**	**0.0176008**	0.0678381	0.123311	0.0887613	0.356467	**0.0183996**	**0.0183996**	**0.0183996**	**0.0183996**			-		
LAFTR4EqOpp-1.0	**0.0324334**	**6.12996e-05**	**0.0244408**	**0.022689**	**0.0176008**	0.0678381	0.123311	0.0887613	0.356467	**0.0183996**	**0.0183996**	**0.0183996**	**0.0183996**				-	
BEUTEL4DP	**0.0339578**	**0.0442474**	**0.0339444**	**0.0337697**	**0.0345178**	**0.0342045**	**0.0348654**	**0.0361078**	**0.0357127**	**0.0349721**	**0.0349721**	**0.0349721**	**0.0349721**	**0.0360691**	**0.0360691**	**0.0360691**	**0.0360691**	-

Table 25: DemDisp t-test results for Adult dataset

	UnfairLR	UnfairLR-decay	Zhang4DP	Zhang4EqOdds	Zhang4EqOpp	LAFTR4DP-0.2	LAFTR4DP-0.5	LAFTR4DP-0.7	LAFTR4DP-1.0	LAFTR4EqOdds-0.2	LAFTR4EqOdds-0.5	LAFTR4EqOdds-0.7	LAFTR4EqOdds-1.0	LAFTR4EqOpp-0.2	LAFTR4EqOpp-0.5	LAFTR4EqOpp-0.7	LAFTR4EqOpp-1.0	BEUTEL4DP
UnfairLR	-																	
UnfairLR-decay	0.0887488	-																
Zhang4DP	0.128857	0.36039	-															
Zhang4EqOdds	**0.0375136**	0.509698	**0.0290636**	-														
Zhang4EqOpp	0.0978039	0.233847	0.260882	**0.0251998**	-													
LAFTR4DP-0.2	0.551137	0.348043	0.409343	0.358677	0.441604	-												
LAFTR4DP-0.5	**0.0131745**	**0.0129121**	**0.0123043**	**0.0108069**	**0.0110973**	**0.00163237**	-											
LAFTR4DP-0.7	**0.0012404**	**0.00144438**	**0.00120802**	**0.00105074**	**0.00109011**	**0.000130648**	**0.00174721**	-										
LAFTR4DP-1.0	**0.000825955**	**0.000885449**	**0.000753091**	**0.000668863**	**0.000724075**	**0.000156555**	**0.00159693**	**0.00161048**	-									
LAFTR4EqOdds-0.2	**0.00378434**	**0.00366727**	**0.00320062**	**0.00281524**	**0.00312117**	**0.000487783**	**0.0229644**	**0.0140474**	**3.5166e-05**	-								
LAFTR4EqOdds-0.5	**0.00378434**	**0.00366727**	**0.00320062**	**0.00281524**	**0.00312117**	**0.000487783**	**0.0229644**	**0.0140474**	**3.5166e-05**		-							
LAFTR4EqOdds-0.7	**0.00378434**	**0.00366727**	**0.00320062**	**0.00281524**	**0.00312117**	**0.000487783**	**0.0229644**	**0.0140474**	**3.5166e-05**			-						
LAFTR4EqOdds-1.0	**0.00378434**	**0.00366727**	**0.00320062**	**0.00281524**	**0.00312117**	**0.000487783**	**0.0229644**	**0.0140474**	**3.5166e-05**				-					
LAFTR4EqOpp-0.2	**0.00110945**	**0.00117479**	**0.00100632**	**0.000896221**	**0.00097546**	**0.000215802**	**0.002385**	**0.00418964**	**0.0199857**	**0.000138426**	**0.000138426**	**0.000138426**	**0.000138426**	-				
LAFTR4EqOpp-0.5	**0.00110945**	**0.00117479**	**0.00100632**	**0.000896221**	**0.00097546**	**0.000215802**	**0.002385**	**0.00418964**	**0.0199857**	**0.000138426**	**0.000138426**	**0.000138426**	**0.000138426**		-			
LAFTR4EqOpp-0.7	**0.00110945**	**0.00117479**	**0.00100632**	**0.000896221**	**0.00097546**	**0.000215802**	**0.002385**	**0.00418964**	**0.0199857**	**0.000138426**	**0.000138426**	**0.000138426**	**0.000138426**			-		
LAFTR4EqOpp-1.0	**0.00110945**	**0.00117479**	**0.00100632**	**0.000896221**	**0.00097546**	**0.000215802**	**0.002385**	**0.00418964**	**0.0199857**	**0.000138426**	**0.000138426**	**0.000138426**	**0.000138426**				-	
BEUTEL4DP	0.516506	0.496116	0.510726	0.50681	0.512633	0.548259	0.651313	0.786726	0.890221	0.745043	0.745043	0.745043	0.745043	0.879551	0.879551	0.879551	0.879551	-

Table 26: DispEqOdds t-test results for Adult dataset

	UnfairLR	UnfairLR-decay	Zhang4DP	Zhang4EqOdds	Zhang4EqOpp	LAFTR4DP-0.2	LAFTR4DP-0.5	LAFTR4DP-0.7	LAFTR4DP-1.0	LAFTR4EqOdds-0.2	LAFTR4EqOdds-0.5	LAFTR4EqOdds-0.7	LAFTR4EqOdds-1.0	LAFTR4EqOpp-0.2	LAFTR4EqOpp-0.5	LAFTR4EqOpp-0.7	LAFTR4EqOpp-1.0	BEUTEL4DP
UnfairLR	-																	
UnfairLR-decay	**0.000282958**	-																
Zhang4DP	0.302304	**0.000626422**	-															
Zhang4EqOdds	0.112095	**0.000863193**	**0.0205415**	-														
Zhang4EqOpp	0.261813	**0.000562228**	0.491375	0.0816226	-													
LAFTR4DP-0.2	**0.036364**	**0.000259975**	**0.0296758**	**0.0285519**	**0.0332979**	-												
LAFTR4DP-0.5	**0.000844861**	**1.18933e-05**	**0.0011189**	**0.00118616**	**0.0010178**	**0.00124668**	-											
LAFTR4DP-0.7	**0.000439075**	**3.49875e-05**	**0.000273489**	**0.000322328**	**0.00035619**	**0.000196314**	**0.0492506**	-										
LAFTR4DP-1.0	**0.0012817**	**0.000128477**	**0.000665723**	**0.000719869**	**0.000893296**	**0.00389541**	0.495197	**0.0492839**	-									
LAFTR4EqOdds-0.2	**0.000101185**	**4.68349e-06**	**6.85322e-05**	**8.76373e-05**	**0.000103827**	**0.000104355**	**0.0310549**	0.527449	0.130242	-								
LAFTR4EqOdds-0.5	**0.000101185**	**4.68349e-06**	**6.85322e-05**	**8.76373e-05**	**0.000103827**	**0.000104355**	**0.0310549**	0.527449	0.130242		-							
LAFTR4EqOdds-0.7	**0.000101185**	**4.68349e-06**	**6.85322e-05**	**8.76373e-05**	**0.000103827**	**0.000104355**	**0.0310549**	0.527449	0.130242			-						
LAFTR4EqOdds-1.0	**0.000101185**	**4.68349e-06**	**6.85322e-05**	**8.76373e-05**	**0.000103827**	**0.000104355**	**0.0310549**	0.527449	0.130242				-					
LAFTR4EqOpp-0.2	**0.00457303**	**0.000303969**	**0.0027426**1	**0.00276722**	**0.00356088**	**0.0176659**	0.921742	0.0871158	0.294073	0.12373	0.12373	0.12373	0.12373	-				
LAFTR4EqOpp-0.5	**0.00457303**	**0.000303969**	**0.0027426**1	**0.00276722**	**0.00356088**	**0.0176659**	0.921742	0.0871158	0.294073	0.12373	0.12373	0.12373	0.12373		-			
LAFTR4EqOpp-0.7	**0.00457303**	**0.000303969**	**0.0027426**1	**0.00276722**	**0.00356088**	**0.0176659**	0.921742	0.0871158	0.294073	0.12373	0.12373	0.12373	0.12373			-		
LAFTR4EqOpp-1.0	**0.00457303**	**0.000303969**	**0.0027426**1	**0.00276722**	**0.00356088**	**0.0176659**	0.921742	0.0871158	0.294073	0.12373	0.12373	0.12373	0.12373				-	
BEUTEL4DP	0.874566	0.834495	0.889296	0.8985	0.883707	0.757351	0.551342	0.499898	0.54679	0.482322	0.482322	0.482322	0.482322	0.570915	0.570915	0.570915	0.570915	-

Table 27: DispEqOpp t-test results for Adult dataset

	UnfairLR	UnfairLR-decay	Zhang4DP	Zhang4EqOdds	Zhang4EqOpp	LAFTR4DP-0.2	LAFTR4DP-0.5	LAFTR4DP-0.7	LAFTR4DP-1.0	LAFTR4EqOdds-0.2	LAFTR4EqOdds-0.5	LAFTR4EqOdds-0.7	LAFTR4EqOdds-1.0	LAFTR4EqOpp-0.2	LAFTR4EqOpp-0.5	LAFTR4EqOpp-0.7	LAFTR4EqOpp-1.0	BEUTEL4DP
UnfairLR	-																	
UnfairLR-decay	**0.000429223**	-																
Zhang4DP	0.427411	**0.000691938**	-															
Zhang4EqOdds	0.180239	**0.00105838**	0.0793817	-														
Zhang4EqOpp	0.450819	**0.0007683**	0.588762	0.14609	-													
LAFTR4DP-0.2	**0.00785656**	**2.16918e-05**	**0.00825064**	**0.00990919**	**0.0106078**	-												
LAFTR4DP-0.5	**0.00039411**	**1.90098e-06**	**0.000560706**	**0.00063794**	**0.000550319**	**0.00120586**	-											
LAFTR4DP-0.7	**0.000387995**	**4.14775e-05**	**0.000167582**	**0.000217203**	**0.00032517**	**0.00125622**	0.307127	-										
LAFTR4DP-1.0	**0.00518168**	**0.000424428**	**0.00276141**	**0.00264016**	**0.00396977**	**0.0550082**	0.285871	**0.0154498**	-									
LAFTR4EqOdds-0.2	**3.02127e-05**	**2.17103e-06**	**1.38488e-05**	**2.45665e-05**	**4.11317e-05**	**0.000157399**	**0.0400053**	0.192571	**0.0336427**	-								
LAFTR4EqOdds-0.5	**3.02127e-05**	**2.17103e-06**	**1.38488e-05**	**2.45665e-05**	**4.11317e-05**	**0.000157399**	**0.0400053**	0.192571	**0.0336427**		-							
LAFTR4EqOdds-0.7	**3.02127e-05**	**2.17103e-06**	**1.38488e-05**	**2.45665e-05**	**4.11317e-05**	**0.000157399**	**0.0400053**	0.192571	**0.0336427**			-						
LAFTR4EqOdds-1.0	**3.02127e-05**	**2.17103e-06**	**1.38488e-05**	**2.45665e-05**	**4.11317e-05**	**0.000157399**	**0.0400053**	0.192571	**0.0336427**				-					
LAFTR4EqOpp-0.2	**0.0191525**	**0.00103982**	**0.0119188**	**0.0109276**	**0.0161129**	0.162285	0.233962	**0.0421432**	0.348698	**0.0450488**	**0.0450488**	**0.0450488**	**0.0450488**	-				
LAFTR4EqOpp-0.5	**0.0191525**	**0.00103982**	**0.0119188**	**0.0109276**	**0.0161129**	0.162285	0.233962	**0.0421432**	0.348698	**0.0450488**	**0.0450488**	**0.0450488**	**0.0450488**		-			
LAFTR4EqOpp-0.7	**0.0191525**	**0.00103982**	**0.0119188**	**0.0109276**	**0.0161129**	0.162285	0.233962	**0.0421432**	0.348698	**0.0450488**	**0.0450488**	**0.0450488**	**0.0450488**			-		
LAFTR4EqOpp-1.0	**0.0191525**	**0.00103982**	**0.0119188**	**0.0109276**	**0.0161129**	0.162285	0.233962	**0.0421432**	0.348698	**0.0450488**	**0.0450488**	**0.0450488**	**0.0450488**				-	
BEUTEL4DP	0.993643	0.505026	0.990027	0.974472	0.998201	0.752579	0.438507	0.424828	0.560426	0.368688	0.368688	0.368688	0.368688	0.600774	0.600774	0.600774	0.600774	-

Table 28: FU-score (DemDisp) t-test results for Adult dataset

	UnfairLR	UnfairLR-decay	Zhang4DP	Zhang4EqOdds	Zhang4EqOpp	LAFTR4DP-0.2	LAFTR4DP-0.5	LAFTR4DP-0.7	LAFTR4DP-1.0	LAFTR4EqOdds-0.2	LAFTR4EqOdds-0.5	LAFTR4EqOdds-0.7	LAFTR4EqOdds-1.0	LAFTR4EqOpp-0.2	LAFTR4EqOpp-0.5	LAFTR4EqOpp-0.7	LAFTR4EqOpp-1.0	BEUTEL4DP
UnfairLR	-																	
UnfairLR-decay	**0.000237571**	-																
Zhang4DP	0.307002	**0.0012299**	-															
Zhang4EqOdds	0.0943987	**0.00143003**	0.0753423	-														
Zhang4EqOpp	0.163173	**0.000552393**	0.963239	0.335354	-													
LAFTR4DP-0.2	0.619479	**0.0387404**	0.512756	0.452538	0.482396	-												
LAFTR4DP-0.5	**0.0215629**	**0.00273278**	**0.0221829**	**0.0191351**	**0.0161524**	**0.00134925**	-											
LAFTR4DP-0.7	**0.00163725**	**0.000465405**	**0.00173446**	**0.0015079**	**0.00125685**	**3.85312e-05**	**0.000739665**	-										
LAFTR4DP-1.0	**0.000683311**	**0.000276518**	**0.00067456**	**0.000591646**	**0.00054814**	**9.14283e-05**	**0.00118907**	**0.00216988**	-									
LAFTR4EqOdds-0.2	**0.00389917**	**0.000959167**	**0.00360119**	**0.00317155**	**0.00290272**	**0.000310774**	**0.0180726**	**0.0231758**	**2.60458e-05**	-								
LAFTR4EqOdds-0.5	**0.00389917**	**0.000959167**	**0.00360119**	**0.00317155**	**0.00290272**	**0.000310774**	**0.0180726**	**0.0231758**	**2.60458e-05**		-							
LAFTR4EqOdds-0.7	**0.00389917**	**0.000959167**	**0.00360119**	**0.00317155**	**0.00290272**	**0.000310774**	**0.0180726**	**0.0231758**	**2.60458e-05**			-						
LAFTR4EqOdds-1.0	**0.00389917**	**0.000959167**	**0.00360119**	**0.00317155**	**0.00290272**	**0.000310774**	**0.0180726**	**0.0231758**	**2.60458e-05**				-					
LAFTR4EqOpp-0.2	**0.00100016**	**0.000385471**	**0.000997052**	**0.000877575**	**0.000811694**	**9.20829e-05**	**0.00137113**	**0.0034603**	**0.014876**	**3.83508e-05**	**3.83508e-05**	**3.83508e-05**	**3.83508e-05**	-				
LAFTR4EqOpp-0.5	**0.00100016**	**0.000385471**	**0.000997052**	**0.000877575**	**0.000811694**	**9.20829e-05**	**0.00137113**	**0.0034603**	**0.014876**	**3.83508e-05**	**3.83508e-05**	**3.83508e-05**	**3.83508e-05**		-			
LAFTR4EqOpp-0.7	**0.00100016**	**0.000385471**	**0.000997052**	**0.000877575**	**0.000811694**	**9.20829e-05**	**0.00137113**	**0.0034603**	**0.014876**	**3.83508e-05**	**3.83508e-05**	**3.83508e-05**	**3.83508e-05**			-		
LAFTR4EqOpp-1.0	**0.00100016**	**0.000385471**	**0.000997052**	**0.000877575**	**0.000811694**	**9.20829e-05**	**0.00137113**	**0.0034603**	**0.014876**	**3.83508e-05**	**3.83508e-05**	**3.83508e-05**	**3.83508e-05**				-	
BEUTEL4DP	0.405442	0.461005	0.411706	0.4132	0.411448	0.402648	0.346719	0.297502	0.267543	0.312966	0.312966	0.312966	0.312966	0.271594	0.271594	0.271594	0.271594	-

Table 29: FU-score (DispEqOdds) for Adult dataset

	UnfairLR	UnfairLR-decay	Zhang4DP	Zhang4EqOdds	Zhang4EqOpp	LAFTR4DP-0.2	LAFTR4DP-0.5	LAFTR4DP-0.7	LAFTR4DP-1.0	LAFTR4EqOdds-0.2	LAFTR4EqOdds-0.5	LAFTR4EqOdds-0.7	LAFTR4EqOdds-1.0	LAFTR4EqOpp-0.2	LAFTR4EqOpp-0.5	LAFTR4EqOpp-0.7	LAFTR4EqOpp-1.0	BEUTEL4DP
UnfairLR	-																	
UnfairLR-decay	**0.000128288**	-																
Zhang4DP	0.492943	**0.000267622**	-															
Zhang4EqOdds	0.189135	**0.000364167**	**0.0324044**	-														
Zhang4EqOpp	0.309726	**0.000204291**	0.968243	0.221168	-													
LAFTR4DP-0.2	**0.0495543**	**6.84351e-05**	**0.0480658**	**0.0440484**	**0.0411541**	-												
LAFTR4DP-0.5	**0.00123468**	**2.58889e-06**	**0.00180284**	**0.00185829**	**0.00131691**	**0.000971542**	-											
LAFTR4DP-0.7	**0.000970081**	**2.23027e-05**	**0.000666834**	**0.000751759**	**0.000626521**	**0.000672873**	0.079408	-										
LAFTR4DP-1.0	**0.00198798**	**6.0118e-05**	**0.00110517**	**0.00119067**	**0.00110053**	**0.0078192**	0.677508	**0.0394154**	-									
LAFTR4EqOdds-0.2	**0.000127488**	**2.52676e-06**	**9.31742e-05**	**0.000123296**	**0.000100104**	**0.00017276**	**0.0310786**	0.475859	0.0988299	-								
LAFTR4EqOdds-0.5	**0.000127488**	**2.52676e-06**	**9.31742e-05**	**0.000123296**	**0.000100104**	**0.00017276**	**0.0310786**	0.475859	0.0988299		-							
LAFTR4EqOdds-0.7	**0.000127488**	**2.52676e-06**	**9.31742e-05**	**0.000123296**	**0.000100104**	**0.00017276**	**0.0310786**	0.475859	0.0988299			-						
LAFTR4EqOdds-1.0	**0.000127488**	**2.52676e-06**	**9.31742e-05**	**0.000123296**	**0.000100104**	**0.00017276**	**0.0310786**	0.475859	0.0988299				-					
LAFTR4EqOpp-0.2	**0.00759075**	**0.000158787**	**0.00494627**	**0.00507505**	**0.00517932**	**0.0315034**	0.867503	0.0687424	0.27123	0.100922	0.100922	0.100922	0.100922	-				
LAFTR4EqOpp-0.5	**0.00759075**	**0.000158787**	**0.00494627**	**0.00507505**	**0.00517932**	**0.0315034**	0.867503	0.0687424	0.27123	0.100922	0.100922	0.100922	0.100922		-			
LAFTR4EqOpp-0.7	**0.00759075**	**0.000158787**	**0.00494627**	**0.00507505**	**0.00517932**	**0.0315034**	0.867503	0.0687424	0.27123	0.100922	0.100922	0.100922	0.100922			-		
LAFTR4EqOpp-1.0	**0.00759075**	**0.000158787**	**0.00494627**	**0.00507505**	**0.00517932**	**0.0315034**	0.867503	0.0687424	0.27123	0.100922	0.100922	0.100922	0.100922				-	
BEUTEL4DP	0.19978	0.302229	0.20548	0.207004	0.204137	0.179642	0.142401	0.1397	0.149624	0.133395	0.133395	0.133395	0.133395	0.156209	0.156209	0.156209	0.156209	-

Table 30: FU-score (DispEqOpp) for Adult dataset

	UnfairLR	UnfairLR-decay	Zhang4DP	Zhang4EqOdds	Zhang4EqOpp	LAFTR4DP-0.2	LAFTR4DP-0.5	LAFTR4DP-0.7	LAFTR4DP-1.0	LAFTR4EqOdds-0.2	LAFTR4EqOdds-0.5	LAFTR4EqOdds-0.7	LAFTR4EqOdds-1.0	LAFTR4EqOpp-0.2	LAFTR4EqOpp-0.5	LAFTR4EqOpp-0.7	LAFTR4EqOpp-1.0	BEUTEL4DP
UnfairLR	-																	
UnfairLR-decay	**0.000256083**	-																
Zhang4DP	0.555658	**0.000423537**	-															
Zhang4EqOdds	0.235763	**0.000630314**	0.0765191	-														
Zhang4EqOpp	0.449043	**0.000436926**	0.826021	0.199853	-													
LAFTR4DP-0.2	**0.010151**	**1.30192e-05**	**0.0120416**	**0.0136883**	**0.0127845**	-												
LAFTR4DP-0.5	**0.000413441**	**4.27811e-07**	**0.000644486**	**0.000752333**	**0.000572306**	**0.000929592**	-											
LAFTR4DP-0.7	**0.000690735**	**3.88867e-05**	**0.000333096**	**0.000416403**	**0.000496184**	**0.00239513**	0.369334	-										
LAFTR4DP-1.0	**0.00664757**	**0.000278707**	**0.00360347**	**0.00345549**	**0.00453714**	0.0778258	0.260413	**0.0152745**	-									
LAFTR4EqOdds-0.2	**4.04573e-05**	**3.47043e-06**	**1.83187e-05**	**3.54986e-05**	**4.51749e-05**	**0.000239595**	**0.0408934**	0.203345	**0.0322438**	-								
LAFTR4EqOdds-0.5	**4.04573e-05**	**3.47043e-06**	**1.83187e-05**	**3.54986e-05**	**4.51749e-05**	**0.000239595**	**0.0408934**	0.203345	**0.0322438**		-							
LAFTR4EqOdds-0.7	**4.04573e-05**	**3.47043e-06**	**1.83187e-05**	**3.54986e-05**	**4.51749e-05**	**0.000239595**	**0.0408934**	0.203345	**0.0322438**			-						
LAFTR4EqOdds-1.0	**4.04573e-05**	**3.47043e-06**	**1.83187e-05**	**3.54986e-05**	**4.51749e-05**	**0.000239595**	**0.0408934**	0.203345	**0.0322438**				-					
LAFTR4EqOpp-0.2	**0.0253525**	**0.000682339**	**0.0162636**	**0.0150708**	**0.01998**	0.220219	0.215172	**0.0420567**	0.321647	**0.0453037**	**0.0453037**	**0.0453037**	**0.0453037**	-				
LAFTR4EqOpp-0.5	**0.0253525**	**0.000682339**	**0.0162636**	**0.0150708**	**0.01998**	0.220219	0.215172	**0.0420567**	0.321647	**0.0453037**	**0.0453037**	**0.0453037**	**0.0453037**		-			
LAFTR4EqOpp-0.7	**0.0253525**	**0.000682339**	**0.0162636**	**0.0150708**	**0.01998**	0.220219	0.215172	**0.0420567**	0.321647	**0.0453037**	**0.0453037**	**0.0453037**	**0.0453037**			-		
LAFTR4EqOpp-1.0	**0.0253525**	**0.000682339**	**0.0162636**	**0.0150708**	**0.01998**	0.220219	0.215172	**0.0420567**	0.321647	**0.0453037**	**0.0453037**	**0.0453037**	**0.0453037**				-	
BEUTEL4DP	0.223749	0.437849	0.233228	0.236862	0.229055	0.175483	0.121538	0.127855	0.155908	0.113453	0.113453	0.113453	0.113453	0.167669	0.167669	0.167669	0.167669	-

Table 31: Accuracy t-test results for Adult (race) dataset

	UnfairLR	UnfairLR-decay	Zhang4DP	Zhang4EqOdds	Zhang4EqOpp	LAFTR4DP-0.2	LAFTR4DP-0.5	LAFTR4DP-0.7	LAFTR4DP-1.0	LAFTR4EqOdds-0.2	LAFTR4EqOdds-0.5	LAFTR4EqOdds-0.7	LAFTR4EqOdds-1.0	LAFTR4EqOpp-0.2	LAFTR4EqOpp-0.5	LAFTR4EqOpp-0.7	LAFTR4EqOpp-1.0
UnfairLR	-																
UnfairLR-decay	**0.000301503**	-															
Zhang4DP	0.512231	**0.000203606**	-														
Zhang4EqOdds	0.726169	**0.00017923**	0.484494	-													
Zhang4EqOpp	0.329139	**0.000202069**	0.818755	0.802395	-												
LAFTR4DP-0.2	0.844944	**0.000272957**	0.732488	0.972208	0.915907	-											
LAFTR4DP-0.5	0.819807	6.16017e-05	0.953025	0.865334	0.949974	0.854204	-										
LAFTR4DP-0.7	0.635871	7.37986e-05	0.84691	0.607379	0.757533	0.51692	0.453799	-									
LAFTR4DP-1.0	0.225705	**0.000238436**	0.579645	0.0526402	0.338849	0.196196	0.735954	0.882672	-								
LAFTR4EqOdds-0.2	0.249057	**0.000107301**	0.367407	0.0861819	0.300406	0.103184	0.417927	0.447589	0.512939	-							
LAFTR4EqOdds-0.5	0.249057	**0.000107301**	0.367407	0.0861819	0.300406	0.103184	0.417927	0.447589	0.512939		-						
LAFTR4EqOdds-0.7	0.249057	**0.000107301**	0.367407	0.0861819	0.300406	0.103184	0.417927	0.447589	0.512939			-					
LAFTR4EqOdds-1.0	0.249057	**0.000107301**	0.367407	0.0861819	0.300406	0.103184	0.417927	0.447589	0.512939				-				
LAFTR4EqOpp-0.2	**0.0424116**	**0.000353764**	0.096868	**0.0061463**	**0.0466602**	**0.00602057**	0.251639	0.210078	**0.00531119**	0.110006	0.110006	0.110006	0.110006	-			
LAFTR4EqOpp-0.5	**0.0424116**	**0.000353764**	0.096868	**0.0061463**	**0.0466602**	**0.00602057**	0.251639	0.210078	**0.00531119**	0.110006	0.110006	0.110006	0.110006		-		
LAFTR4EqOpp-0.7	**0.0424116**	**0.000353764**	0.096868	**0.0061463**	**0.0466602**	**0.00602057**	0.251639	0.210078	**0.00531119**	0.110006	0.110006	0.110006	0.110006			-	
LAFTR4EqOpp-1.0	**0.0424116**	**0.000353764**	0.096868	**0.0061463**	**0.0466602**	**0.00602057**	0.251639	0.210078	**0.00531119**	0.110006	0.110006	0.110006	0.110006				-

Table 32: DemDisp t-test results for Titanic dataset

	UnfairLR	UnfairLR-decay	Zhang4DP	Zhang4EqOdds	Zhang4EqOpp	LAFTR4DP-0.2	LAFTR4DP-0.5	LAFTR4DP-0.7	LAFTR4DP-1.0	LAFTR4EqOdds-0.2	LAFTR4EqOdds-0.5	LAFTR4EqOdds-0.7	LAFTR4EqOdds-1.0	LAFTR4EqOpp-0.2	LAFTR4EqOpp-0.5	LAFTR4EqOpp-0.7	LAFTR4EqOpp-1.0
UnfairLR	-																
UnfairLR-decay	**0.00503195**	-															
Zhang4DP	**0.0354938**	**0.0070848**	-														
Zhang4EqOdds	0.0990337	**0.00462702**	**0.000264738**	-													
Zhang4EqOpp	0.139573	**0.00561368**	0.0907771	**0.0135056**	-												
LAFTR4DP-0.2	**0.0178578**	**0.0187615**	0.0606367	**0.014099**	**0.0174894**	-											
LAFTR4DP-0.5	0.0672663	**0.0379025**	0.155954	0.0575348	0.0784625	0.621294	-										
LAFTR4DP-0.7	**0.00235751**	0.0502795	**0.00345537**	**0.00174946**	**0.00297609**	0.0648421	0.21367	-									
LAFTR4DP-1.0	**0.00428768**	0.0839521	**0.00817118**	**0.00373098**	**0.00429872**	**0.0168582**	**0.02219**	0.186152	-								
LAFTR4EqOdds-0.2	**0.0276006**	**0.0312324**	0.0615902	**0.0185746**	**0.0288709**	0.577786	0.877917	**0.0411657**	**0.00550455**	-							
LAFTR4EqOdds-0.5	**0.0276006**	**0.0312324**	0.0615902	**0.0185746**	**0.0288709**	0.577786	0.877917	**0.0411657**	**0.00550455**		-						
LAFTR4EqOdds-0.7	**0.0276006**	**0.0312324**	0.0615902	**0.0185746**	**0.0288709**	0.577786	0.877917	**0.0411657**	**0.00550455**			-					
LAFTR4EqOdds-1.0	**0.0276006**	**0.0312324**	0.0615902	**0.0185746**	**0.0288709**	0.577786	0.877917	**0.0411657**	**0.00550455**				-				
LAFTR4EqOpp-0.2	**0.0017796**	0.137007	**0.00306058**	**0.00160525**	**0.00176212**	**0.00573337**	**0.00765081**	**0.0145814**	**0.00167922**	**0.00184411**	**0.00184411**	**0.00184411**	**0.00184411**	-			
LAFTR4EqOpp-0.5	**0.0017796**	0.137007	**0.00306058**	**0.00160525**	**0.00176212**	**0.00573337**	**0.00765081**	**0.0145814**	**0.00167922**	**0.00184411**	**0.00184411**	**0.00184411**	**0.00184411**		-		
LAFTR4EqOpp-0.7	**0.0017796**	0.137007	**0.00306058**	**0.00160525**	**0.00176212**	**0.00573337**	**0.00765081**	**0.0145814**	**0.00167922**	**0.00184411**	**0.00184411**	**0.00184411**	**0.00184411**			-	
LAFTR4EqOpp-1.0	**0.0017796**	0.137007	**0.00306058**	**0.00160525**	**0.00176212**	**0.00573337**	**0.00765081**	**0.0145814**	**0.00167922**	**0.00184411**	**0.00184411**	**0.00184411**	**0.00184411**				-

Table 33: DispEqOdds t-test results for Adult (race) dataset

	UnfairLR	UnfairLR-decay	Zhang4DP	Zhang4EqOdds	Zhang4EqOpp	LAFTR4DP-0.2	LAFTR4DP-0.5	LAFTR4DP-0.7	LAFTR4DP-1.0	LAFTR4EqOdds-0.2	LAFTR4EqOdds-0.5	LAFTR4EqOdds-0.7	LAFTR4EqOdds-1.0	LAFTR4EqOpp-0.2	LAFTR4EqOpp-0.5	LAFTR4EqOpp-0.7	LAFTR4EqOpp-1.0
UnfairLR	-																
UnfairLR-decay	0.775357	-															
Zhang4DP	0.341565	0.599578	-														
Zhang4EqOdds	0.192387	0.610521	**0.0354778**	-													
Zhang4EqOpp	0.712561	0.999498	0.135315	0.123073	-												
LAFTR4DP-0.2	0.0557349	0.0909786	0.0537446	**0.0145673**	**0.0146371**	-											
LAFTR4DP-0.5	**0.0348649**	**0.0426936**	**0.0405559**	**0.0112062**	**0.0159976**	0.259956	-										
LAFTR4DP-0.7	**0.00414862**	**0.00564629**	**0.00467805**	**0.00257605**	**0.00175767**	0.0688766	0.22926	-									
LAFTR4DP-1.0	**0.0135409**	**0.0214422**	**0.012879**	**0.00371825**	**0.00467448**	**0.0283283**	**0.0130689**	0.87494	-								
LAFTR4EqOdds-0.2	**0.0182451**	**0.0424394**	**0.0134277**	**0.00397468**	**0.00238569**	0.568729	0.448484	**0.0202516**	**0.0327802**	-							
LAFTR4EqOdds-0.5	**0.0182451**	**0.0424394**	**0.0134277**	**0.00397468**	**0.00238569**	0.568729	0.448484	**0.0202516**	**0.0327802**		-						
LAFTR4EqOdds-0.7	**0.0182451**	**0.0424394**	**0.0134277**	**0.00397468**	**0.00238569**	0.568729	0.448484	**0.0202516**	**0.0327802**			-					
LAFTR4EqOdds-1.0	**0.0182451**	**0.0424394**	**0.0134277**	**0.00397468**	**0.00238569**	0.568729	0.448484	**0.0202516**	**0.0327802**				-				
LAFTR4EqOpp-0.2	**0.0049984**	**0.00701866**	**0.0040959**	**0.002143**	**0.00140124**	**0.00562735**	**0.0290028**	**0.0291239**	0.15093	**0.00156597**	**0.00156597**	**0.00156597**	**0.00156597**	-			
LAFTR4EqOpp-0.5	**0.0049984**	**0.00701866**	**0.0040959**	**0.002143**	**0.00140124**	**0.00562735**	**0.0290028**	**0.0291239**	0.15093	**0.00156597**	**0.00156597**	**0.00156597**	**0.00156597**		-		
LAFTR4EqOpp-0.7	**0.0049984**	**0.00701866**	**0.0040959**	**0.002143**	**0.00140124**	**0.00562735**	**0.0290028**	**0.0291239**	0.15093	**0.00156597**	**0.00156597**	**0.00156597**	**0.00156597**			-	
LAFTR4EqOpp-1.0	**0.0049984**	**0.00701866**	**0.0040959**	**0.002143**	**0.00140124**	**0.00562735**	**0.0290028**	**0.0291239**	0.15093	**0.00156597**	**0.00156597**	**0.00156597**	**0.00156597**				-

Table 34: DispEqOpp t-test results for Adult (race) dataset

	UnfairLR	UnfairLR-decay	Zhang4DP	Zhang4EqOdds	Zhang4EqOpp	LAFTR4DP-0.2	LAFTR4DP-0.5	LAFTR4DP-0.7	LAFTR4DP-1.0	LAFTR4EqOdds-0.2	LAFTR4EqOdds-0.5	LAFTR4EqOdds-0.7	LAFTR4EqOdds-1.0	LAFTR4EqOpp-0.2	LAFTR4EqOpp-0.5	LAFTR4EqOpp-0.7	LAFTR4EqOpp-1.0
UnfairLR	-																
UnfairLR-decay	0.878457	-															
Zhang4DP	0.123108	0.650066	-														
Zhang4EqOdds	0.135134	0.454641	**0.0140638**	-													
Zhang4EqOpp	0.878879	0.87334	0.0945864	0.0504974	-												
LAFTR4DP-0.2	**0.0410475**	0.101659	0.0559021	**0.013581**	**0.0138135**	-											
LAFTR4DP-0.5	**0.0336257**	0.0605247	**0.0493465**	**0.0136916**	**0.0174824**	0.240301	-										
LAFTR4DP-0.7	**0.00264097**	**0.00678877**	**0.00342322**	**0.00180684**	**0.00129913**	0.0573276	0.209198	-									
LAFTR4DP-1.0	**0.00808679**	**0.0198759**	**0.00953636**	**0.0027693**	**0.00320676**	**0.0195869**	**0.00942733**	0.674697	-								
LAFTR4EqOdds-0.2	**0.0162631**	0.056397	**0.016925**	**0.00468283**	**0.00357319**	0.433656	0.543755	**0.0211621**	**0.0175665**	-							
LAFTR4EqOdds-0.5	**0.0162631**	0.056397	**0.016925**	**0.00468283**	**0.00357319**	0.433656	0.543755	**0.0211621**	**0.0175665**		-						
LAFTR4EqOdds-0.7	**0.0162631**	0.056397	**0.016925**	**0.00468283**	**0.00357319**	0.433656	0.543755	**0.0211621**	**0.0175665**			-					
LAFTR4EqOdds-1.0	**0.0162631**	0.056397	**0.016925**	**0.00468283**	**0.00357319**	0.433656	0.543755	**0.0211621**	**0.0175665**				-				
LAFTR4EqOpp-0.2	**0.0024061 8**	**0.0065554**	**0.00255753**	**0.00116133**	**0.000827329**	**0.00296851**	**0.00956844**	**0.0149868**	0.0645493	**0.000342034**	**0.000342034**	**0.000342034**	**0.000342034**	-			
LAFTR4EqOpp-0.5	**0.0024061 8**	**0.0065554**	**0.00255753**	**0.00116133**	**0.000827329**	**0.00296851**	**0.00956844**	**0.0149868**	0.0645493	**0.000342034**	**0.000342034**	**0.000342034**	**0.000342034**		-		
LAFTR4EqOpp-0.7	**0.0024061 8**	**0.0065554**	**0.00255753**	**0.00116133**	**0.000827329**	**0.00296851**	**0.00956844**	**0.0149868**	0.0645493	**0.000342034**	**0.000342034**	**0.000342034**	**0.000342034**			-	
LAFTR4EqOpp-1.0	**0.0024061 8**	**0.0065554**	**0.00255753**	**0.00116133**	**0.000827329**	**0.00296851**	**0.00956844**	**0.0149868**	0.0645493	**0.000342034**	**0.000342034**	**0.000342034**	**0.000342034**				-

Table 35: FU-score (DemDisp) t-test results for Adult (race) dataset

	UnfairLR	UnfairLR-decay	Zhang4DP	Zhang4EqOdds	Zhang4EqOpp	LAFTR4DP-0.2	LAFTR4DP-0.5	LAFTR4DP-0.7	LAFTR4DP-1.0	LAFTR4EqOdds-0.2	LAFTR4EqOdds-0.5	LAFTR4EqOdds-0.7	LAFTR4EqOdds-1.0	LAFTR4EqOpp-0.2	LAFTR4EqOpp-0.5	LAFTR4EqOpp-0.7	LAFTR4EqOpp-1.0
UnfairLR	-																
UnfairLR-decay	**0.0033580 4**	-															
Zhang4DP	**0.0374543**	**0.00512602**	-														
Zhang4EqOdds	0.102025	**0.00314429**	**0.000341367**	-													
Zhang4EqOpp	0.141889	**0.00389014**	0.0918333	**0.0128526**	-												
LAFTR4DP-0.2	**0.0147193**	**0.0155266**	0.0546378	**0.0118773**	**0.0146844**	-											
LAFTR4DP-0.5	0.0596097	**0.0338926**	0.14753	0.0515094	0.0711686	0.635651	-										
LAFTR4DP-0.7	**0.0023166 8**	**0.0462542**	**0.00337426**	**0.00176656**	**0.00308495**	0.0647818	0.196936	-									
LAFTR4DP-1.0	**0.0035930 6**	0.0811194	**0.00712256**	**0.00319802**	**0.00374831**	**0.017454**	**0.0219762**	0.184973	-								
LAFTR4EqOdds-0.2	**0.0246206**	**0.0277613**	0.0560989	**0.0162519**	**0.0261084**	0.587916	0.897373	**0.040481**	**0.00591781**	-							
LAFTR4EqOdds-0.5	**0.0246206**	**0.0277613**	0.0560989	**0.0162519**	**0.0261084**	0.587916	0.897373	**0.040481**	**0.00591781**		-						
LAFTR4EqOdds-0.7	**0.0246206**	**0.0277613**	0.0560989	**0.0162519**	**0.0261084**	0.587916	0.897373	**0.040481**	**0.00591781**			-					
LAFTR4EqOdds-1.0	**0.0246206**	**0.0277613**	0.0560989	**0.0162519**	**0.0261084**	0.587916	0.897373	**0.040481**	**0.00591781**				-				
LAFTR4EqOpp-0.2	**0.00147536**	0.136928	**0.00264229**	**0.00137637**	**0.00154047**	**0.00606233**	**0.00792742**	**0.0131316**	**0.000234248**	**0.00206468**	**0.00206468**	**0.00206468**	**0.00206468**	-			
LAFTR4EqOpp-0.5	**0.00147536**	0.136928	**0.00264229**	**0.00137637**	**0.00154047**	**0.00606233**	**0.00792742**	**0.0131316**	**0.000234248**	**0.00206468**	**0.00206468**	**0.00206468**	**0.00206468**		-		
LAFTR4EqOpp-0.7	**0.00147536**	0.136928	**0.00264229**	**0.00137637**	**0.00154047**	**0.00606233**	**0.00792742**	**0.0131316**	**0.000234248**	**0.00206468**	**0.00206468**	**0.00206468**	**0.00206468**			-	
LAFTR4EqOpp-1.0	**0.00147536**	0.136928	**0.00264229**	**0.00137637**	**0.00154047**	**0.00606233**	**0.00792742**	**0.0131316**	**0.000234248**	**0.00206468**	**0.00206468**	**0.00206468**	**0.00206468**				-

Table 36: FU-score (DispEqOdds) for Adult (race) dataset

	UnfairLR	UnfairLR-decay	Zhang4DP	Zhang4EqOdds	Zhang4EqOpp	LAFTR4DP-0.2	LAFTR4DP-0.5	LAFTR4DP-0.7	LAFTR4DP-1.0	LAFTR4EqOdds-0.2	LAFTR4EqOdds-0.5	LAFTR4EqOdds-0.7	LAFTR4EqOdds-1.0	LAFTR4EqOpp-0.2	LAFTR4EqOpp-0.5	LAFTR4EqOpp-0.7	LAFTR4EqOpp-1.0
UnfairLR	-																
UnfairLR-decay	0.0834519	-															
Zhang4DP	0.381145	0.126547	-														
Zhang4EqOdds	0.174998	0.526012	**0.033964**	-													
Zhang4EqOpp	0.690343	0.285154	0.155392	0.145641	-												
LAFTR4DP-0.2	0.0515099	**0.0411798**	0.0523091	**0.014468**	**0.0139761**	-											
LAFTR4DP-0.5	**0.0323891**	**0.0200106**	**0.039045**	**0.0113172**	**0.0171268**	0.270817	-										
LAFTR4DP-0.7	**0.00446431**	**0.00291974**	**0.00562798**	**0.00315628**	**0.00248997**	0.0774618	0.236407	-									
LAFTR4DP-1.0	**0.0146327**	**0.0129261**	**0.0144589**	**0.0045504**	**0.00588847**	**0.0302538**	**0.0174316**	0.886567	-								
LAFTR4EqOdds-0.2	**0.0173202**	**0.0187407**	**0.013764**	**0.00428837**	**0.00263056**	0.634769	0.414435	**0.0248936**	**0.0365842**	-							
LAFTR4EqOdds-0.5	**0.0173202**	**0.0187407**	**0.013764**	**0.00428837**	**0.00263056**	0.634769	0.414435	**0.0248936**	**0.0365842**		-						
LAFTR4EqOdds-0.7	**0.0173202**	**0.0187407**	**0.013764**	**0.00428837**	**0.00263056**	0.634769	0.414435	**0.0248936**	**0.0365842**			-					
LAFTR4EqOdds-1.0	**0.0173202**	**0.0187407**	**0.013764**	**0.00428837**	**0.00263056**	0.634769	0.414435	**0.0248936**	**0.0365842**				-				
LAFTR4EqOpp-0.2	**0.00526278**	**0.0048111**	**0.0052804**	**0.00273459**	**0.00194124**	**0.00628247**	**0.0332915**	0.0559358	0.174041	**0.00239446**	**0.00239446**	**0.00239446**	**0.00239446**	-			
LAFTR4EqOpp-0.5	**0.00526278**	**0.0048111**	**0.0052804**	**0.00273459**	**0.00194124**	**0.00628247**	**0.0332915**	0.0559358	0.174041	**0.00239446**	**0.00239446**	**0.00239446**	**0.00239446**		-		
LAFTR4EqOpp-0.7	**0.00526278**	**0.0048111**	**0.0052804**	**0.00273459**	**0.00194124**	**0.00628247**	**0.0332915**	0.0559358	0.174041	**0.00239446**	**0.00239446**	**0.00239446**	**0.00239446**			-	
LAFTR4EqOpp-1.0	**0.00526278**	**0.0048111**	**0.0052804**	**0.00273459**	**0.00194124**	**0.00628247**	**0.0332915**	0.0559358	0.174041	**0.00239446**	**0.00239446**	**0.00239446**	**0.00239446**				-

Table 37: FU-score (DispEqOpp) for Adult (race) dataset

	UnfairLR	UnfairLR-decay	Zhang4DP	Zhang4EqOdds	Zhang4EqOpp	LAFTR4DP-0.2	LAFTR4DP-0.5	LAFTR4DP-0.7	LAFTR4DP-1.0	LAFTR4EqOdds-0.2	LAFTR4EqOdds-0.5	LAFTR4EqOdds-0.7	LAFTR4EqOdds-1.0	LAFTR4EqOpp-0.2	LAFTR4EqOpp-0.5	LAFTR4EqOpp-0.7	LAFTR4EqOpp-1.0
UnfairLR	-																
UnfairLR-decay	0.746211	-															
Zhang4DP	0.137818	0.466233	-														
Zhang4EqOdds	0.128375	0.653783	**0.0138287**	-													
Zhang4EqOpp	0.879091	0.874487	0.0910141	0.0561702	-												
LAFTR4DP-0.2	**0.0366802**	0.076994	0.0515793	**0.0129038**	**0.0125025**	-											
LAFTR4DP-0.5	**0.0330692**	**0.0458074**	0.0507705	**0.0153584**	**0.0198209**	0.250898	-										
LAFTR4DP-0.7	**0.0034916**	**0.0058074**	**0.00532102**	**0.00287567**	**0.00229851**	0.0585609	0.205732	-									
LAFTR4DP-1.0	**0.00898292**	**0.0167136**	**0.0117269**	**0.0038329**	**0.00440837**	**0.0234536**	**0.00748779**	0.68334	-								
LAFTR4EqOdds-0.2	**0.0154039**	**0.043207**	**0.0173637**	**0.00526546**	**0.0038939**	0.426227	0.522791	**0.026403**	**0.0232134**	-							
LAFTR4EqOdds-0.5	**0.0154039**	**0.043207**	**0.0173637**	**0.00526546**	**0.0038939**	0.426227	0.522791	**0.026403**	**0.0232134**		-						
LAFTR4EqOdds-0.7	**0.0154039**	**0.043207**	**0.0173637**	**0.00526546**	**0.0038939**	0.426227	0.522791	**0.026403**	**0.0232134**			-					
LAFTR4EqOdds-1.0	**0.0154039**	**0.043207**	**0.0173637**	**0.00526546**	**0.0038939**	0.426227	0.522791	**0.026403**	**0.0232134**				-				
LAFTR4EqOpp-0.2	**0.0031026**	**0.00613727**	**0.00387184**	**0.00187187**	**0.00144118**	**0.00413226**	**0.00934543**	**0.0175768**	0.0690899	**0.00106586**	**0.00106586**	**0.00106586**	**0.00106586**	-			
LAFTR4EqOpp-0.5	**0.0031026**	**0.00613727**	**0.00387184**	**0.00187187**	**0.00144118**	**0.00413226**	**0.00934543**	**0.0175768**	0.0690899	**0.00106586**	**0.00106586**	**0.00106586**	**0.00106586**		-		
LAFTR4EqOpp-0.7	**0.0031026**	**0.00613727**	**0.00387184**	**0.00187187**	**0.00144118**	**0.00413226**	**0.00934543**	**0.0175768**	0.0690899	**0.00106586**	**0.00106586**	**0.00106586**	**0.00106586**			-	
LAFTR4EqOpp-1.0	**0.0031026**	**0.00613727**	**0.00387184**	**0.00187187**	**0.00144118**	**0.00413226**	**0.00934543**	**0.0175768**	0.0690899	**0.00106586**	**0.00106586**	**0.00106586**	**0.00106586**				-