Assessing Fair Machine Learning Strategies Through a Fairness-Utility Trade-off Metric

Luiz Fernando F. P. de Lima[1],
Danielle Rousy D. Ricarte[1], **Clauirton A. Siebra**[1]

[1]Centro de Informática (CI) – Universidade Federal da Paraíba (UFPB)
João Pessoa, PB – Brazil

luizfernando@ppgi.ci.ufpb.br, {danielle, clauirton}@ci.ufpb.br

Abstract. *Due to the increasing use of artificial intelligence for decision making and the observation of biased decisions in many applications, researchers are investigating solutions that attempt to build fairer models that do not reproduce discrimination. Some of the explored strategies 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 comparing current approaches a complex task. In that sense, we defined a utility and fairness trade-off metric. We assessed 15 fair model implementations and a baseline model using this metric, providing a systemically comparative ruler for other approaches.*

1. Introduction

Due to the increase of available data, machine learning (ML) algorithms have become a standard approach for building decision making software in the most diverse areas such as health, finances, security, and education. However, this increasing importance of ML algorithms as decision making resources, mainly in critical areas, brought about some problems embedded in such algorithms. 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].

The fairness subarea concerns building models that mitigate bias and discrimination problems in algorithms. 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 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 was biased with racial prejudice. [GARCIA 2016] demonstrated how applications to determine online advertisement delivery had a sexist bias. [BOLUKBASI et al. 2016] also demonstrated the 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].

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.

Adversarial learning has been used in representation learning tasks and showed to be helpful to 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.

In that sense, we can encode fairness constraints into our models through an adversary, for example, in the works of [BEUTEL et al. 2017], [ZHANG et al. 2018] and [MADRAS et al. 2018]. Both [BEUTEL et al. 2017] and [MADRAS et al. 2018] worked in fair models focusing on learning fair representation. [ZHANG et al. 2018], on the other hand, worked in structuring a model-agnostic adversarial debiasing architecture. In general, they use an adversary and a classifier, where the adversary aims to correctly predict the protected attribute from a fair representation or the classifier's outcomes.

Moreover, it is common to works of the fair ML area to define a specific metric to evaluate their models in relation to the model fairness. In that sense, there is no standard benchmark for comparing the models and evaluating new proposals in the current literature. 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 that we discuss in Section 2.

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, it is essential to develop a benchmark that includes the adversarial learning approaches to systematically evaluate these proposals, proposals with other strategies, and new proposals that emerge like those proposed in this work.

In that sense, this work presents the development of a benchmark for the adversarial based fair models. We present a fairness-utility trade-off metric and assess 15 fair model implementations based on [MADRAS et al. 2018] and [ZHANG et al. 2018] papers and a baseline model using this metric.

The remainder of this paper is organized as follows. Section 2 presents the related concepts to understand the fairness area, the adversarial approaches and the found benchmark. Section 3 presents the methods and details behind the models' implementation. Section 4 presents and discusses the results for the chosen approaches. Finally, Section 5 concludes the work.

2. Related Concepts

2.1. Discrimination and Fairness Definitions

Discrimination can be understood, in general, as the fact of having prejudice or harm against an individual or a group in decision making [MEHRABI et al. 2019]. 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, these attributes should be considered as protected attributes in the learning process. 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. Different definitions have been formulated and presented in the literature regarding ML fairness, so there is no universal definition. Some commonly used fairness definitions are:

- **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 only perform well 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 admission to a university, promotion receipt, and credit release. In this case, the true positive rates must be equal for the two groups of the protected attribute, mathematically, $P(\hat{Y} = 1|A = 0, Y = 1) = P(\hat{Y} = 1|A = 1, Y = 1)$.

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

2.2. Fair Adversarial Strategies

As pointed in Section 1, some proposals are taking advantage on the use of adversarial learning to build fairer models. The works of [ZHANG et al. 2018] and [MADRAS et al. 2018] are examples of works that include an adversary into the ML model to encode a fairness constraint and also the base for our work. We better review them over the remainder of this Section.

In [MADRAS et al. 2018] work, the authors present the Learning Adversarially Fair and Transferable Representations (LAFTR) model. LAFTR (Figure 1) uses an encoder ($f(X)$) in order 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.

In the LAFTR model, the adversary aims to maximize its objective, while 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 descent and ascent steps to optimize

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

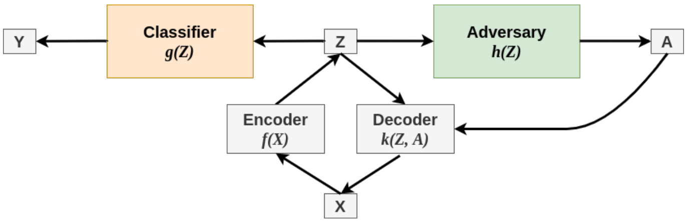

Figure 1. LAFTR model from [MADRAS et al. 2018] (adapted)

their parameters according to Equation 1. 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. 1 respectively specify a desired balance between utility, reconstruction of the inputs, and fairness.

$$L(f,g,h,k) = \alpha L_C(g(f(X,A)),Y) \ + \beta L_{Dec}(k(f(X,A),A,X) - \\ \gamma L_{Adv}(h(f(X,A)),A) \tag{1}$$

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.

On the other hand, the study of [ZHANG et al. 2018] presents a general architecture for achieving fairness through the adversarial process. The model (Figure 2) consists of training a predictor, with the objective of predicting Y from X, and an adversary, with the objective of predicting A from $\hat{Y}$. For each fairness definition to be achieved, different input data is used for the adversary.

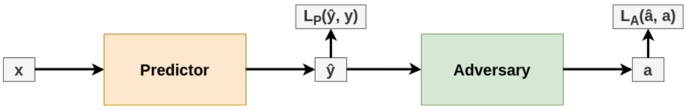

Figure 2. [ZHANG et al. 2018] general architecture (adapted)

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. For equalized odds, in addition to $\hat{Y}$, 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. Each training step U is updated to minimize L_A according to the gradient $\nabla_U L_A$. W is updated according to Equation 2. The term $proj_{\nabla_U L_A} \nabla_W L_P$ prevents the predictor from moving in a direction that helps the adversary decreasing its loss. Furthermore, the last term, $\alpha\nabla_U L_A$, attempts to increase the adversary's loss, α is a tunable hyperparameter to balance this attempt.

$$W = W - \nabla_W L_P - proj_{\nabla_U L_A} \nabla_W L_P - \alpha \nabla_W L_A \tag{2}$$

2.3. Evaluating Fair Models

In 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. Then, it is hard to compare the approaches presented in the literature to other literature approaches or to new proposals. That was the motivation for the [JONES et al. 2020] work.

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 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 in 0.5. The PPR policy, in which the threshold is determined so that the positive predictive rate matches a pre-determined value of 20% within a fixed tolerance. Finally, the Policy Free evaluation considers all possible values in a range for the threshold. For this latter aspect, they define and apply the fair efficiency metric (Equation 3).

$$\Theta_{p,f} = 2\frac{K_p K_f}{K_p + K_f} \tag{3}$$

$$K_m = \int_0^1 \int_0^1 m(\lambda, \tau)\, d\tau\, d\lambda \tag{4}$$

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 4) is a additional integral that considers all possible values for m, i.e., the full range for all combinations of τ and λ. 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 when the model is maximally fair and useful $\Theta = 1$ and the model is optimal.

The weakness of their work, which we intend to address, is that any evaluated model is an adversarial strategy. Moreover, the benchmark evaluation is limiting because [JONES et al. 2020] considers 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 by [MADRAS et al. 2018], 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, most of adversarial works pointed in Section 1 consider the trade-off coefficients as tunable hyperparameters. Thus, it is essential to any comparative proposal 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.

3. Research Method

This Section presents the methodological aspects used to develop the benchmark to assess the adversarial based fair models. To benchmark all models, we used the dataset presented in Section 3.3 following the detailed train/validation/test split and pre-processing. For each model training, we evaluated the utility and fairness metrics presented in Section 3.2 and computed the $FU\text{-}score$ (Section 3.1).

3.1. 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 $FU\text{-}score$ (Equation 5). $FU\text{-}score$ is a fairness-utility trade-off metric inspired on the $F1\text{-}score$[2], but is also a simplification from the fair efficiency metric proposed by [JONES et al. 2020].

$$FU\text{-}score = 2\frac{pf}{p+f} \tag{5}$$

Similar to the fair efficiency metric, $FU\text{-}score$ 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, $FU\text{-}score$ penalizes models that score highly for fairness but do not present a good utility and vice versa. In addition, $FU\text{-}score$ takes into accounts the fairness and utility metrics that we want to maximize, i.e., achieve results near to 1. Then, $FU\text{-}score = 0$ means that the model is whether maximally unfair or non-predictive. Whereas when the model is optimal, i.e., the model is maximally fair and useful, $FU\text{-}score = 1$.

$FU\text{-}score$ do not consider the additional integral K_m proposed by [JONES et al. 2020]. Thus, we can use this metric to compare the same model but changing its fair hyperparameter that is tunable. With this more general form, $FU\text{-}score$ can assist in the model's tune process where one could compare the same model to find the 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].

[2] $F1\text{-}score$ is a utility metric commonly used to access machine learning models. $F1\text{-}score$ takes the harmonic mean from two other performance metrics, $Precision$ and $Recall$.

3.2. Fairness and Performance Metrics

To assess the models' predictive performance, we used the overall accuracy defined by Equation 6. The accuracy measures the overall model utility by looking at the prediction's hit rate over the total number of classifications.

$$Acc = \frac{TN + TP}{TN + FP + FN + TP} \tag{6}$$

In order to evaluate the model's fairness, we considered the three commonly used fairness definitions, demographic parity, equalized odds and equal opportunity. Thus, we can measure this by the demographic disparity (Eq. 7), disparity in equal odds (Eq. 8) and disparity in equal opportunity (Eq. 9).

$$DemDisp = |P(\hat{Y} = 1|A = 0) - P(\hat{Y} = 1|A = 1)| \tag{7}$$

$$DispEqOdds = |P(\hat{Y} = 1|A = 0, Y = y) - P(\hat{Y} = 1|A = 1, Y = y)| \tag{8}$$

$$DispEqOpp = |P(\hat{Y} = 1|A = 0, Y = 1) - P(\hat{Y} = 1|A = 1, Y = 1)| \tag{9}$$

But there is a problem with these disparity definitions (Equations 7, 8 and 9). The $FU\text{-}score$ 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 by 1 in those metrics. Thus, we rewrite the fairness metrics as in Equations 10, 11 and 12. We can apply this modification to any fair definition or metric when necessary.

$$DemDisp = 1 - |P(\hat{Y} = 1|A = 0) - P(\hat{Y} = 1|A = 1)| \tag{10}$$

$$DispEqOdds = 1 - |P(\hat{Y} = 1|A = 0, Y = y) - P(\hat{Y} = 1|A = 1, Y = y)| \tag{11}$$

$$DispEqOpp = 1 - |P(\hat{Y} = 1|A = 0, Y = 1) - P(\hat{Y} = 1|A = 1, Y = 1)| \tag{12}$$

3.3. Dataset

To assess the models on the benchmark procedure, we used the UCI Adult Income dataset[3]. The original Adult dataset is separated into two sets, training with 32561 examples and test with 16281 examples.

We use the continuous attributes as real-valued. The attribute "fnlwgt" was disregarded. Age was the only continuous attribute bucketized at boundaries [18, 25, 30, 35, 40, 45, 50, 55, 60, 65]. The categorical features were converted to one-hot encoding vectors. Finally, the binary categorical attributes, i.e., the protected attribute sex and the target attribute income, were binarized. We discharged examples with missing values and then remained 30162 examples from the training set and 15060 from the test set.

Training data was split into training and validation sets. To this, we randomly shuffled data and took 20% from this to the validation set. The test set was not changed from the original split.

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

3.4. Implemented Models

In this work, we implemented and benchmarked some adversarial strategies present in the literature, which are the LAFTR model [MADRAS et al. 2018] and the model proposed by [ZHANG et al. 2018]. We also implemented a baseline model without any fairness constraint. Like in [ZHANG et al. 2018], our baseline model is a logistic regression that computes the predictions through Equation 13.

$$\hat{y} = \sigma(wx + b) \tag{13}$$

For the approach proposed by [ZHANG et al. 2018] we provided three implementations. Each implementation aims to incorporate in the model a fairness constraint. We followed their implementation for a model that enforces equal odds. This model has a predictor model like in Equation 13 and an adversarial model to predict the protected attribute defined by Equations 14a and 14b.

$$s = \sigma[(1 + |c|)\sigma^{-1}(\hat{y})] \tag{14a}$$

$$\hat{a} = u[s, sy, s(1 - y)] + b \tag{14b}$$

The model that enforces equal opportunity is similar to the last one but differs by using only data examples where y = 1. Finally, the implementation to add the demographic parity constraint has an adversarial model that is a simplified model from the last one, we define this adversary by Equations 15a and 15b.

$$s = \sigma[(1 + |c|)\sigma^{-1}(\hat{y})] \tag{15a}$$

$$\hat{a} = us + b \tag{15b}$$

In these Equations, σ is the sigmoid function and σ^{-1} is its inverse function, 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 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 between all implementations, we use a single hidden layer for each of our encoder, classifier and adversary, with 8 hidden units and a latent space of dimension = 8. We applied the Leaky ReLU function as activation function for all layers.

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

All models were trained with 100 epochs and a minibatch equal to 64 examples, using the Adam optimizer. Also, using a learning rate = 0.001, except the baseline model that seemed to take advantage in the use of the approach adopted by [ZHANG et al. 2018] to avoid local minimum problems, decrease the learning rate (lr) for each epoch setting it to $lr = 0.001/t$, where t is the step/epoch counter. The weights u and w in Equations 13, 14b, and 15b and the weights for the layers in the neural networks were initialized with zeros. On the other hand, the b's in Equations 13, 14b, and 15b and c's in Equations 14a, 15a and the bias parameters for the neural networks were initialized with ones.

We also defined the model specific parameters as follows. For all models based on [ZHANG et al. 2018] work we used $\alpha = 1/t$, where t is the step counter. This worked better than the $\alpha = \sqrt{t}$ used in the original work and keep the guarantee that $\alpha lr \to 0$. 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 $\gamma = [0.2, 0.5, 0.7, 1]$.

As technological resources for this implementation, we used the programming language Python (version 3.8.5) and the packages TensorFlow (version 2.2.0) and NumPy (version 1.19.2). The parameters related to these packages and not specified here were used as default. For reproducibility, we provide our code as well the training, validation, and test split at this work GitHub repository[4].

4. Results and Discussion

This Section presents and discusses the results for the proposed assessment in Chapter 3. Firstly, we present the understanding of the models' behaviors for utility and fairness. Thereon, we discuss the models' results for the FU-$score$ metric.

4.1. Models Utility and Fairness

Figure 3 presents the accuracy and fairness results for each implemented model and for each chosen fairness metric. The expected behavior when building fair models is to face a decrease in its accuracy while the fairness increases. In that sense, our baseline model presents a higher accuracy than the fair models. On the other hand, for all fairness metric, the fair models present a better result.

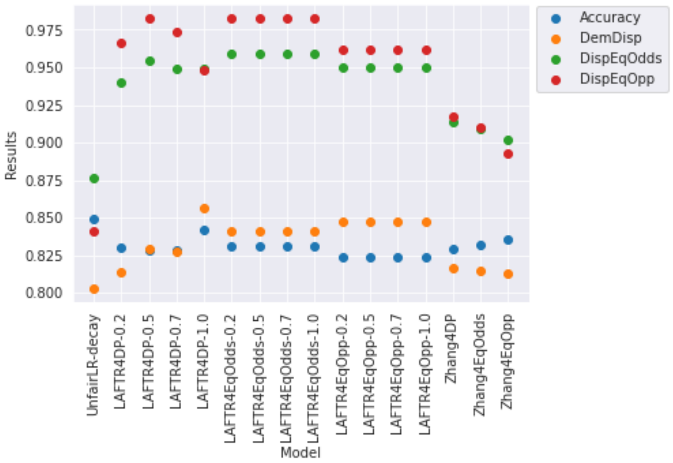

Figure 3. Accuracy and fairness results for each model

The LAFTR implementation for demographic parity ($\gamma = 1$) presented better utility performance over the other fair implementations. The LAFTR implementations

[4] https://github.com/limafernando/falsb

for equal odds and equal opportunity presented a consistent behavior independent of the chosen value for the fair coefficient.

The implementations based on [ZHANG et al. 2018] work presented a utility performance slightly similar to the other fair models. Furthermore, this models' implementations presented a slightly worse fair performance, however still better than the baseline model.

4.2. Models Results for FU-$score$

Figure 4 presents the trade-off between utility and fairness, measured by the FU-$score$ metric, for each model and selected fairness metric. How LAFTR implementations for equal odds and equal opportunity presented a consistent performance in accuracy and fairness, the trade-offs measured for these models are also consistent independent of the value for the fair coefficient.

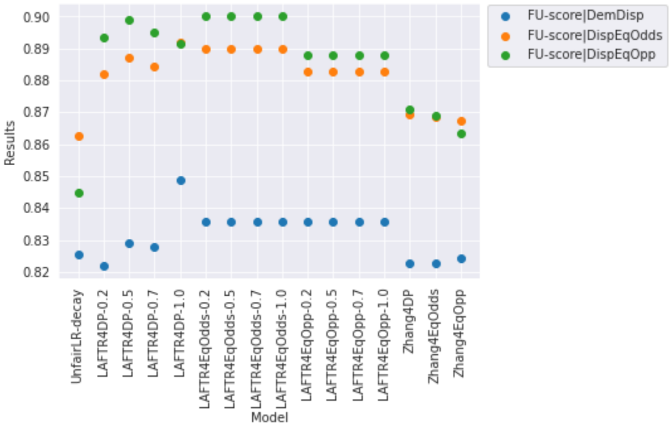

Figure 4. FU-$score$ **for each model and each fairness definition**

When we look at the FU-$score$ results considering the demographic disparity, we observe that ZHANGs implementations and LAFTR implementation for demographic parity ($\gamma = 0.2$) are penalized due to its lower fairness results. On the other hand, LAFTR implementation for demographic parity ($\gamma = 1$) presents a better trade-off result.

LAFTR model for demographic parity ($\gamma = 1$) also presents the better FU-$score$ when we look at the disparity in equal odds. However, the trade-off results are slightly worse for both LAFTR implementation for equal odds and LAFTR implementation for demographic parity ($\gamma = 0.5$).

Furthermore, when we look at the disparity in equal opportunity, we observe that the LAFTR model for equal opportunity presents the better FU-$score$ result, followed by the LAFTR implementation for demographic parity ($\gamma = 0.5$). However, the FU-$score$ result of LAFTR implementation for demographic parity ($\gamma = 1$) considering the disparity in equal opportunity metric is just 0.01 lower than the better model's result for this fairness definition. Then, we could consider using just this implementation for our classification task that presents a good enough trade-off performance for all selected fairness metrics.

Finally, besides being penalized when considering demographic disparity, implementations based on [ZHANG et al. 2018] work present the inferior FU-$score$ results for the other fairness metrics among the assessed fair models.

5. Conclusions

The growing use of machine learning solutions, including in critical areas, has raised serious concerns about the socio-algorithmic problems in AI solutions such as fairness, transparency, accountability, explainability and privacy.

To build fairer models that mitigate bias and discrimination in its decisions, some approaches are based on the use of an adversary to ensure a fairness constraint in the model. Furthermore, it is common in the fair ML area to works 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 propose the $FU\text{-}score$, a fairness-utility trade-off metric and initiate a systematic benchmark to provide a comparative ruler for the adversarial based fair models.

The $FU\text{-}score$ metric jointly evaluates the model fairness and predictive performance given chosen utility and fairness metrics. In this sense, this trade-off metric penalizes models that score highly for fairness but do not present a good utility and vice versa. Through this metric, we assessed 9 implementations for the LAFTR model [MADRAS et al. 2018] and 3 implementations for the model proposed by [ZHANG et al. 2018]. We also implemented and assessed a logistic regression model as a baseline without any fairness constraint.

Our implemented fair models behaved as expects, increased fairness at the cost of decreasing accuracy. Looking at the $FU\text{-}score$ metric, we found that the ZHANGs models presented the lower trade-off results. While the LAFTR implementation for demographic parity ($\gamma = 1$) worked well for all fairness definitions, despite not performing the best trade-off result for the disparity in equal opportunity, it is just 0.01 inferior to the best model for this fairness metric.

Acknowledgements

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001.

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.

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.

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.

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.

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.

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.

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.

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

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*.

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.

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.