Measuring the Burden of (Un)fairness Using Counterfactuals

Alejandro Kuratomi[1] ✉, Evaggelia Pitoura[2] ✉, Panagiotis Papapetrou[1] ✉,
Tony Lindgren[1] ✉, and Panayiotis Tsaparas[2] ✉

[1] Department of Computer and Systems Sciences, Stockholm University,
Borgarfjordsgatan 12, 16455 Kista, Sweden,
{alejandro.kuratomi,tony,panagiotis}@dsv.su.se

[2] Department of Computer Science and Engineering, University of Ioannina, Ípeiros
45110, Ioannina, Greece,
{pitoura,tsap}@uoi.gr

Abstract. In this paper, we use counterfactual explanations to offer a new perspective on fairness, that, besides accuracy, accounts also for the difficulty or burden to achieve fairness. We first gather a set of fairness-related datasets and implement a classifier to extract the set of false negative test instances to generate different counterfactual explanations on them. We subsequently calculate two measures: the false negative ratio of the set of test instances, and the distance (also called *burden*) from these instances to their corresponding counterfactuals, aggregated by sensitive feature groups. The first measure is an accuracy-based estimation of the classifier biases against sensitive groups, whilst the second is a counterfactual-based assessment of the difficulty each of these groups has of reaching their corresponding desired ground truth label. We promote the idea that a counterfactual and an accuracy-based fairness measure may assess fairness in a more holistic manner, whilst also providing interpretability. We then propose and evaluate, on these datasets, a measure called Normalized Accuracy Weighted Burden, which is more consistent than only its accuracy or its counterfactual components alone, considering both false negative ratios and counterfactual distance per sensitive feature. We believe this measure would be more adequate to assess classifier fairness and promote the design of better performing algorithms in both accuracy and fairness.

Keywords: algorithmic fairness · counterfactual explanations · bias

1 Introduction

Machine Learning (ML) models assist decision-making in different applications, such as recommender systems [16,18], vehicle localization [7], student grading [6], credit assessment [1], disease diagnoses [9] and recidivism prediction [3]. These decisions should be taken impartially across sensitive features, such as religion, gender, ethnicity and age [20,27]. In order to achieve fair outcomes, the ML models must avoid making decisions based on these qualities. There are several challenges in attaining these unbiased model decisions, and we hereby describe and focus on three of them, namely **fairness evaluation**, **interpretability** and **fairness accuracy trade-off**:

1. **Fairness evaluation:** The first challenge refers to the fact that the difficulty of defining a measure for model fairness assessment lies on its selection. While there exist at least 20 such measures [8,20], none of them is perfectly suitable for all situations. More importantly, Kusner et al. [8] argue that some measures might exacerbate the perceived discrimination, and may not eliminate the biases entirely even after optimizing for them [2].
2. **Interpretability:** The second challenge is knowing the models' features weighting. The increase in model complexity and capacity to represent highly nonlinear functions to achieve superior prediction performance has raised a new challenge, that of providing trustable model explanations to understand how different features are prioritized [5,14,21,22,26]. Given that highly complex and opaque models may focus on sensitive features to elaborate a decision (even when the sensitive features are omitted from the data due to correlations with other nonsensitive, proxy features [8,20,18]), it is important to obtain model explanations to understand whether this is occurring or not. A subfield of ML, called ML Interpretability, aims to provide these model explanations. Specifically, an interpretability technique known as Counterfactual Explanations (CE) answers the following question: *how should an instance change its feature values so as to switch a model's predicted label from an undesired to a desired label?* An analogous nontrivial problem to the fairness evaluation challenge exists for CE generation: there are several different CE algorithms, each minimizing a distinct cost function and producing fairly contrasting CEs [27].
3. **Fairness-accuracy trade-off:** The third challenge refers to the fact that altering the model to deter biases naturally found in the datasets, due to highly correlated sensitive features and labels, may reduce the models performance, leading to a fairness-accuracy trade-off [13,18,20].

In this paper, we address these challenges by combining two fairness measures: one accuracy-based and one counterfactual-based.

In particular, we assume that for each sensitive feature, there are at least two sensitive groups, e.g. the sensitive feature *Sex* has two sensitive groups *Male* and *Female*, and that we have a binary classification task. To measure accuracy, we use *predictive equality* [27], which states that the False Negative Ratio (*FNR*), i.e., the fraction of false negative predictions, should be the same across sensitive groups. Other accuracy-based fairness definitions, such as predictive parity, are left for future work.

The CE x' of an item x is a similar item to x for which the classifier produces an outcome different than the outcome of x. Let x be an item in a sensitive group that was falsely predicted to belong to the negative class. Intuitively, the distance between x and its counterfactual x' measures the amount of change that is needed to counteract unfairness in accuracy, that is, to correctly classify x in the positive class. We call *Burden* the average such distance for all items in the sensitive group that were falsely assigned to the negative class. In a sense, Burden captures the cost of achieving fairness.

The main advantages of counterfactual-based fairness are three-fold: first, it aligns with a fair treatment intuition, since the difficulty of achieving a desired output among sensitive groups should be similar [8]. This similar difficulty may be seen as a similar burden value among different groups; second, burden is calculated using a generated CE (x'), which inherently indicates the models features relevances, providing important information to tackle the models opacity; and third, it may provide both individual and group fairness assessment [23], while other metrics, like statistical parity and equalized odds, focus on group fairness.

Hence, the first contribution of this paper is a study between the FNR and the measure of burden, where the set of CEs are generated by minimizing different cost functions. The study uses 11 fairness-related, binary classification datasets from four different fields. We analyze the differences in burden among different CE methods and their relation to FNR. Moreover, the second contribution of the paper is a new measure, *Normalized Accuracy Weighted Burden* (NAWB), that assesses fairness holistically and may be used to optimize classifiers training and address the accuracy-fairness trade-off challenge.

2 Related work

In this work, two areas converge: machine learning fairness and counterfactual explainability. From the perspective of machine learning fairness, different approaches have been taken to both measure and correct biases in different applications [18,19,20,25,29].

2.1 Fairness and bias measurement

To avoid model discrimination biases, the biases must be first detected [8]. Quy et al. compiled 15 datasets from different fields that are frequently used for fairness-related research in ML and use statistical parity, equalized odds and Absolute Between-ROC Area (ABROCA) to detect biases among a set of sensitive features in each dataset [20]. Machine learning models may amplify the users input biases according to common user preferences [25]. Zafar et al. relate the recommendation bias increase to stereotypical-based biases, and highlight the strong relation of false positive rates with sensitive groups in recidivism prediction biases against african americans, and in less-paid jobs for women [29].

2.2 Counterfactual explainability

Verma et al. propose a rubric to compare different CE generation algorithms, reviewing 39 papers where methods and metrics are discussed [27]. They highlight the existence of linear and mixed-integer programming CE methods, such as the Actionable Recourse algorithm by Ustun et al. [26], that provides actionability (actionable decisions) with low computational demand, at the cost of using low-accuracy, linear classifiers.

Among the metrics discussed by Verma et al. are likelihood (the closeness of the CE to the data distribution) and sparsity (the number of changed features) [27]. Related to actionability is the property of feasibility, which considers the feature direction of change and the plausibility of the obtained feature values. Linked to sparsity is proximity, which is the inverse of the distance between the Instance of Interest (IOI) and its CE [12,26]. Finally, faithfulness may also be prioritized, as it indicates how likely (through likelihood) or justified [11] a CE is according to the data. Different algorithms prioritize different metrics.

The Nearest Neighbor Tweaking (NN) method selects the *closest* positive ground truth label instance in the training set to the IOI. The Minimum Observable (MO) method selects the closest counterfactual instance from the whole dataset (including the test instances with their predicted labels). These two methods minimize the euclidean distance function and preserve the plausibility of the feature values [12,28]. The Random Forest Tweaking (RT) method selects the *most frequent* counterfactual training instance inside the same leaves that the IOI falls in, in a Random Forest (RF) classifier, providing plausibility and faithfulness. The Counterfactual Conditional Heterogeneous Autoencoder (CCH-VAE) CE method prioritizes likelihood, outputting counterfactual instances that are likely according to the data. The method uses a variational autoencoder and creates random perturbations in its latent space. These perturbations are brought to the original space and become the generated counterfactuals [17].

Other notable more complex methods exist. Model-Agnostic CE (MACE) [4] delivers best-in-class proximity performance but with the longest computational times; Growing Spheres (GS) [10] attempts to obtain close counterfactuals by growing spheres from the IOI; Diverse CE (DiCE) [15] allows users to obtain a set of CEs instead of a single one, where the set is chosen to provide diverse feature changes. Local Rule-based Explainability (LORE) is able to provide feature relevances and CEs through the training of a local rule generation model.

2.3 Counterfactual fairness

At the intersection of these two areas lies counterfactual fairness: a characteristic of decision processes treating individuals equally in the as-is situation, and in a world where their sensitive features are different [8]. Currently, CEs provide insights on why a decision was taken and potential actionability, but cannot indicate whether these decisions are fair [13]. On the other hand, fairness measures lack the actionability and feature relevance that CEs ellicit.

Ustun et al. propose an interesting measure between the classifier model and the instances attributes, and use this to design a fair model. This measure uses the covariance between the sensitive features values and the distance between the subjects and the decision boundary. If this covariance is high, that means the distance between the instances and the decision boundary are highly related, indicating that the model may be biased according to those features [26]. This measure is however intended for linear classifiers and assumes a linear relation between classifiers and features. The authors also present an interesting evaluation of the relation between the cost of achieving a given counterfactual (cost of recourse) and split it by false and true negative prediction groups.

Coston et al. argue that traditional measures of fairness, like parity, may not necessarily lead to fairness in counterfactual scenarios. Therefore, they indicate that counterfactual reasoning must be implemented to measure fairness, and apply a set of methods to achieve fairness in a policy design framework [2]. Finally, Sharma et al. define the counterfactual-based fairness metric called burden, and indicate its usage for both individual and group fairness assessment. The authors use this metric as part of the fitness function in a genetic algorithm that generates counterfactuals [23].

3 Methodology

Given a dataset X, with labels $Y \in \{-,+\}$, $+$ being the desired, positive label, a classification function f, such that $f : X \rightarrow Y$ and a set of sensitive features S_i, $i \in \{1, 2, ..., M\}$, where M is the number of sensitive features, the accuracy-based metric of False Negative Ratio (FNR) per sensitive group is defined as follows:

$$\mathrm{FNR_s} = P(f(x) = -|S = s, Y = +), \tag{1}$$

where $\mathrm{FNR_s}$ is the false negative ratio of the sensitive group s.

For the counterfactual-based measure, we first formally define the counterfactual search as [24,12]:

$$x^* = \operatorname*{argmin}_{x'} c(x, x')|f(x) = y \wedge f(x') = y' \ , \tag{2}$$

where $c(x, x')$ is a distance-based cost function, and y' is the opposite label to y. The counterfactual reasoning is mainly applied by analyzing whether it is equally *difficult* to change the model outcome, from an undesired label $f(x) = y = -$, to a desired predicted label $f(x') = y' = +$, among sensitive groups or individuals [23,26]. Hence, the counterfactual-based measure may be obtained by calculating the average cost function $c(x, x')$ with $x \in X^s$, where X^s is the set of instances belonging to the sensitive feature group s, and the counterfactuals found x' for each x. This measure is defined as *Burden* and is formulated as follows:

$$Burden_s = \frac{1}{|X^s|} \sum_{x_i \in X^s} c(x_i, x'_i), \tag{3}$$

where $\mathrm{Burden_s}$ is the average value of the cost function $c(.)$, which may be defined as the euclidean distance, based on the concept defined by [23].

We propose and examine a combined measure based on $\mathrm{Burden_s}$ and $\mathrm{FNR_s}$ that could potentially be used to design a fair and accurate classifier. The proposed measure is called the *Accuracy Weighted Burden* or AWB. To derive it, we define the set of false negative instances per sensitive group s as:

$X^s_{FN} = \{x \in X | f(x) = -, S = s, Y = +\}$ and multiply $\mathrm{Burden_s}$ and $\mathrm{FNR_s}$ as shown:

$$\mathrm{AWB_s} = P(f(x) = - | S = s, Y = +) \frac{1}{|X^s_{FN}|} \sum_{x_i \in X^s_{FN}} d(x_i, x'_i) \tag{4}$$

$$\mathrm{AWB_s} = \frac{|X^s_{FN}|}{|\{x \in X | S = s, Y = +\}|} \frac{1}{|X^s_{FN}|} \sum_{x_i \in X^s_{FN}} d(x_i, x'_i) \tag{5}$$

$$\mathrm{AWB_s} = \frac{\sum_{x_i \in X^s_{FN}} d(x_i, x'_i)}{|\{x \in X | S = s, Y = +\}|} \tag{6}$$

where x'_i represents the CE of the x_i instance, and function $d(.)$ is the euclidean distance or burden. If we plot $\mathrm{Burden_s}$ versus $\mathrm{FNR_s}$, and locate each sensitive group as a point in this plane, a point located in the upper-right corner would present a higher general bias than one located in the lower-left corner. The $\mathrm{FNR_s}$ is the ratio of instances falsely classified as belonging to the negative class, whilst the $\mathrm{Burden_s}$ measures how far the IOI is from an existing, desired counterfactual instance, per group. In this sense, a high $\mathrm{FNR_s}$ and a high $\mathrm{Burden_s}$ indicates a high number of difficult-to-correctly classify points for a given group and classifier f. This may be translated to the *area* of the box formed between the location of the dots and the origin. This area is calculated by multiplying these variables, leading to Eq. 6.

By normalizing each of the L features in the dataset inside the [0,1] range, the range of values for $d(.)$ is [0,L], so we divide Eq. 6 by L to obtain the *Normalized Accuracy Weighted Burden* or NAWB:

$$\mathrm{NAWB_s} = \frac{\sum_{i \in X^s_{FN}} d(x_i, x'_i)}{L|\{x \in X | S = s, Y = +\}|} \tag{7}$$

After defining the basic metrics, let us outline the steps of our methodology. In order to study classifier fairness, for each dataset and classification task, we test several classifiers and search for the model parameters that provide the best performance in each case (see section 4). A single classifier (the one with the best F1 score) is used per dataset. We then execute a four-step process: (1) Calculate the FNR per sensitive group, (2) obtain CEs for the false negative instances using different CE methods (different ways and cost functions in solving Eq. 2, (3) estimate the aggregated Burden per sensitive group, per CE method, and (4) study the relation of $\mathrm{Burden_s}$ and $\mathrm{FNR_s}$ to provide a holistic view on the classifier fairness and evaluate AWB, our new combined measure.

The first step of the process is carried out using Eq. 1, where s is the sensitive group of a feature (Male, Female or White, Non-white, etc.). Ultimately, a fair classifier would have a similar $\mathrm{FNR_s}$ among the different s values belonging to each S sensitive feature.

The second step is using NN, MO, RT and CCHVAE to generate the CEs. We concentrate on the four mentioned algorithms, as they represent a set of relevant objectives currently prioritized in CE algorithms, namely proximity, feasibility and faithfulness (through likelihood), whilst maintaining relatively low complexity and computational times. These methods are applied to the false negative instances (X_{FN}), i.e, obtaining a set of four CEs for each of them.

The third step is calculating the aggregated burden by sensitive feature Burden_{s} using Eq. 3. A higher burden for a given group of subjects, in comparison to another, would mean that the individuals belonging to that group have a higher difficulty, in terms of the distance, to achieve the positive class, according to the model f.

In the fourth and final step, we discuss these metrics, presenting their evaluation on the fairness-related datasets. We analyze the FNR_{s} per dataset and evaluate the Burden_{s} per dataset and CE method. We then relate both measures and study their correlation and finally examine the combined measure AWB_{s} and its normalized version NAWB_{s}.

4 Empirical Evaluation

We describe here the datasets based on [20] and discuss the obtained results. The datasets, their main sources, and codes are available at the GitHub[3].

4.1 Datasets

The datasets and relevant characteristics are shown in Table 1. Preprocessing is carried out according to [4,20], reducing the number of features and instances by removing duplicates, missing values and low-importance features. Further details may be observed in the repository. The test group and true positive distributions are obtained after preprocessing.

4.2 Results and discussion

In this section we show the classification performance, analyze the FNR_{s} per dataset, discuss the Burden_{s} measure per dataset and CE method, and finally present and analyze an accuracy-counterfactual combined fairness measure.

Model selection. We implemented four different types of classifiers and used grid search with 5-fold validation to identify the optimal parameters according to the F1 score. The implemented classifiers are Support Vector Machines (SVM), Decision Trees (DT), Multilayer Perceptrons (MLP) and Random Forests (RF). The RF classifier achieved the best performance for 5 out of the 11 datasets (shown in Table 2 along with the model parameters), while the MLP classifier achieved the best performance for the other 6 datasets (shown in Table 3 along with the model parameters). We used the best classifier for each dataset.

FNR_{s} evaluation The dataset is split into 70% train and 30% test. The models are used to predict the label of positive ground truth test instances.

[3] `https://github.com/alku7660/counterfactual-fairness`

Table 1: Datasets instances, features, labels and sensitive groups distributions

Dataset	**Items (Feat.)**	**Classes**	**Sensitive Groups**	**Test Group Distribution**	**True Positive Distribution**
Adult	48842 (15)	+:>50kUSD -:≤50kUSD	Male/Female	9112/4455	2848/499
			White/Nonwhite	11666/1901	3032/315
			<25/25-60/>60	2214/10500/853	31/3102/214
KDD Census	299285 (41)	+:>50kUSD -:≤50kUSD	Male/Female	43193/46593	4374/1180
			White/Nonwhite	75268/14518	5052/502
German	1000 (21)	+:low risk -:high risk	Male/Female	208/92	55/28
Dutch	60420 (12)	+:low risk -:high risk	Male/Female	9090/9036	6024/3401
Bank	45211 (17)	+:deposits -:no deposit	Sing./Marr./Divor.	3785/8171/1608	574/849/196
			<25/25-60/>60	244/12939/381	68/1384/167
Credit	30000 (24)	+:no default -:defaults	Male/Female	3547/5350	854/1118
			Marr./NotMarr.	4131/4756	980/992
			Oth./HS/Uni./Gra.	143/1435/4122/3187	10/355/1006/601
Compas	7214 (52)	+:improved -:recidivist	Male/Female	1276/308	624/209
			Caucasian/African	619/965	379/454
Diabetes	101766 (50)	+:recovered -:readmitted	Male/Female	6326/7527	4743/5767
Student	395 (33)	+:high grade -:low grade	Male/Female	55/64	40/45
			<18/≥18	93/26	66/19
Oulad	32593 (12)	+:pass exam -:fail exam	Male/Female	5142/4303	2393/2061
Law	20798 (12)	+:pass bar -:fail bar	Male/Female	3426/2703	3274/2557
			White/Nonwhite	5148/981	4990/841

The $\mathrm{FNR_s}$ are shown in Fig. 1. In the Adult dataset, the highest $\mathrm{FNR_s}$ corresponds to the *<25* age group. This indicates that younger adults are expected to earn less than those with longer careers and higher education, both correlated to age. Additionally, *Females* present a considerable unfavorable bias, relative to *Males*. *Non-whites* are unfavored, though not as *Females* and young people. A similar behavior is observed in the KDD Census dataset $\mathrm{FNR_s}$ with respect to the unfavored *Female* and *Non-white* groups.

An inverted bias behavior is observed in the German and Dutch datasets, where *Males* are more likely to be incorrectly classified with bad credit or low-level occupation, respectively, than *Females.* The $\mathrm{FNR_s}$ is double for *Males* in the German dataset (similar in the Oulad dataset), while it is close to 5 times in the Dutch dataset, compared to *Females.*

In the Bank and Credit datasets, all $\mathrm{FNR_s}$ are considerable. In the Bank dataset the *>60* age group has the lowest $\mathrm{FNR_s}$ ($< 1\%$), while in the Credit dataset the *Other* education group has the highest $\mathrm{FNR_s}$ ($> 80\%$).

In the Compas dataset, the *African-Americans* and *Females* are more than twice as likely to be incorrectly classified as recidivist as *Caucasians* and *Males*,

Table 2: Datasets with RF as best classifier and F1 score

Dataset	Adult	KDD Census	Dutch	Bank	Student
F1	0.83	0.87	0.84	0.86	0.70
Max. Depth	10	10	10	10	2
Min. Samples/Leaf	1	5	3	1	5
Min. Samples/Split	5	5	5	2	2
Num. Trees	100	100	50	200	200

Table 3: Datasets with MLP as best classifier and F1 score

Dataset	Credit	German	Diabetes	Oulad	Law	Compas
F1	0.72	0.70	0.61	0.67	0.82	0.66
Activation	Tanh	ReLU	Logistic	Logistic	Tanh	Tanh
Hidden Layers	(50, 1)	(100, 10)	(100, 2)	(100, 10)	(50, 1)	(100, 10)
Solver	Adam	SGD	SGD	SGD	Adam	Adam

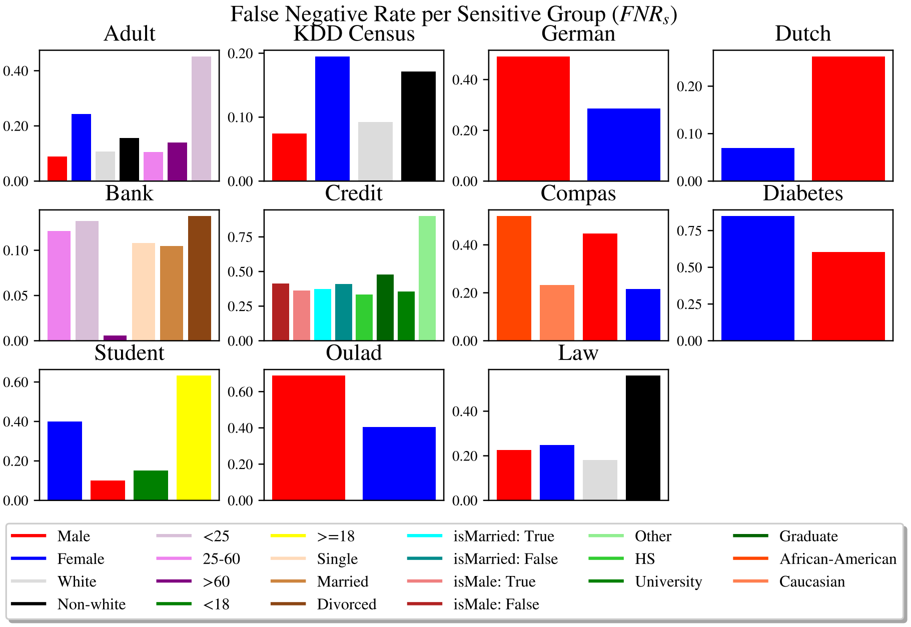

Fig. 1: FNR_s for each sensitive group

respectively. The Diabetes dataset shows the highest *Females* FNR_s caused by the low classifier performance. In the Student dataset, the highest FNR_s is observed in the *>=18* age group ($> 60\%$), in comparison with the lower FNR_s for the *Female*, *Male* and *<18* groups with 40% or lower. Finally, in the Law dataset, all FNR_s are close to 20%, except for the *Non-white* group with 50%.

$\textbf{Burden}_\text{s}$ evaluation Fig. 2 present datasets in rows and CE methods in columns. In general, all datasets show a similar relative burden among sensitive groups for NN and MO methods, since they prioritze distance and pick the counterfactual from the pool of observations (MO's Burden_s measure is lower because it also considers test instances). RT and CCHVAE present a relative different Burden_s behavior in both magnitude and relative position among sensitive groups, since these two prioritize frequency and likelihood, respectively, over proximity. The CEs obtained through CCHVAE are particularly further from their respective IOIs because they are closer to the data distribution centers to maximize likelihood. Specifically, for the Adult dataset, in the age feature, we may see that *>60* has a high Burden_s, compared to *<25* and *25-60*, specially in the NN, MO and RT methods. This could indicate a bias against older people who may have a higher difficulty of achieving a high income.

In the KDD Census dataset,The relative Burden_s magnitude is the same for all methods: higher for *Females* than *Males* and higher for *Non-whites* than *Whites*. The KDD Census dataset presents a similar behavior in terms of relative burden with the Law dataset.

In the German dataset, the correlation of burden with FNR_s is inverted in CCHVAE, while NN, MO and RT preserve the same higher bias for *Males* than *Females*. In the Dutch dataset, Burden_s is higher for *Females*, while the FNR_s ratio was higher for the *Males* (Fig. 1).

In the Bank dataset, there is a higher burden for the *<25* and *Divorced* groups relative to their counterparts in the NN, MO and RT methods, however, it is the *>60* group that has a higher burden according to CCHVAE. These behaviors are contrasting with the FNR_s in the age groups, because the *>60* has a significantly lower FNR_s.

In the Credit dataset, the behavior among groups is similar to the FNR_s relative behavior in the NN, MO and RT methods. However, it drastically changes in the CCHVAE, where the burden is high and similar across groups.

In the Compas dataset note that the RT FNR_s shows a different relative magnitude: the *Males* and *African-Americans* FNR_s is higher, whilst the burden is higher for *Females* and *Caucasians*.

In the Diabetes dataset the FNR_s of *Females* is higher than that of *Males* (even though the data is balanced among genders) but Burden_s shows a relative similar behavior for both *Females* and *Males*.

In the Student dataset, all methods showed a similar relative Burden_s behavior, which is a strong contrast with the highly unfavored age group of *>18* according to FNR_s. Finally, in the Oulad dataset, *Females* present a slightly higher Burden_s than *Males* in all methods except RT.

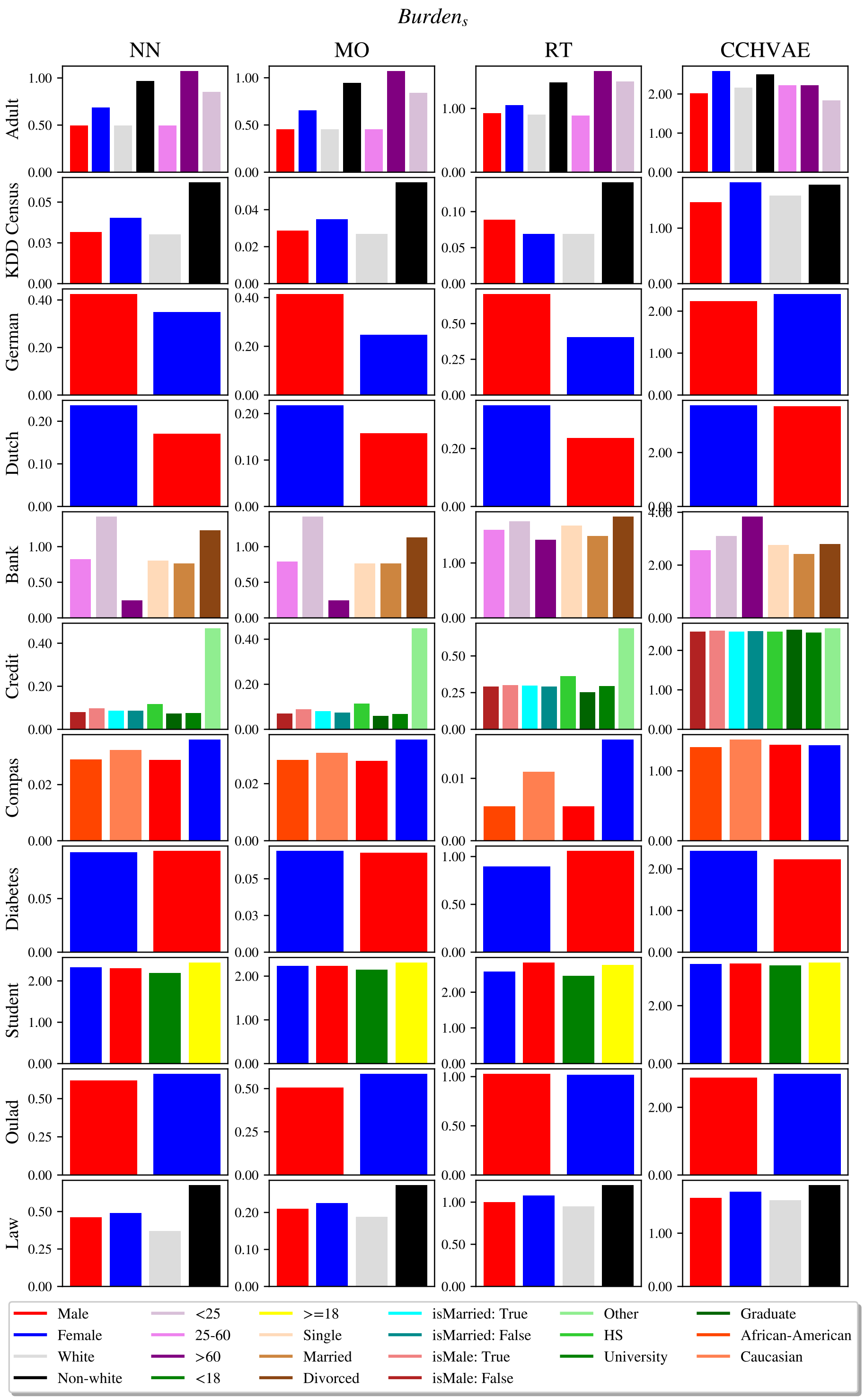

Fig. 2: Burden$_s$ for each sensitive group

FNR$_s$ and Burden$_s$ The relation between FNR$_s$ and each CE method's Burden$_s$ is observed in Fig. 3 for some of the datasets. Each scatter plot shows the FNR$_s$ in the x-axis and Burden$_s$ in the y-axis. The dots represent the sensitive groups location in the $Burden-FNR$ plane. Each color indicates a sensitive feature and each dot has its group name. Positively correlated Burden$_s$ and FNR$_s$ measures show dots of the same color (belonging to the same feature) scattered across the positive diagonal, whilst a negative correlation shows these dots closer to the negative diagonal. For example, in the Diabetes dataset, *Male* and *Female* dots are located in the negative diagonal in NN, and RT, whilst in the positive one in MO and CCHVAE.

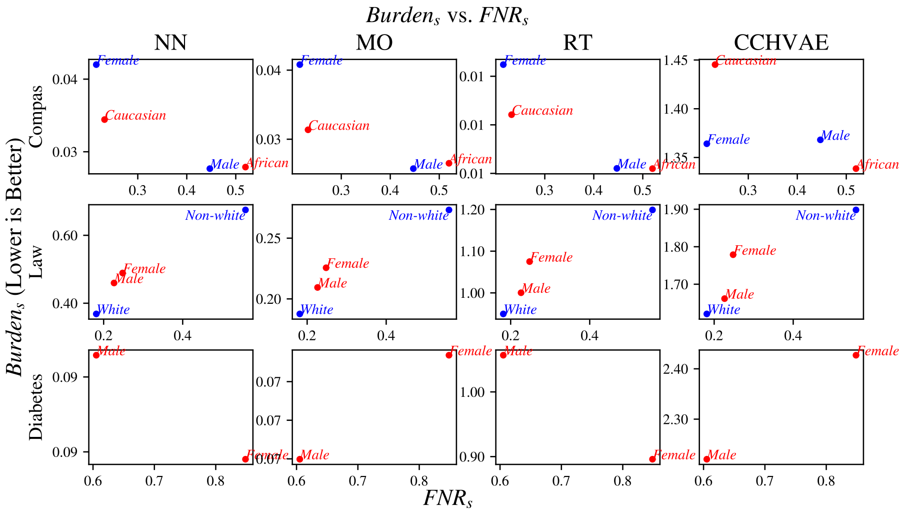

Fig. 3: False Negative Ratio (FNR$_s$) versus Burden (Burden$_s$)

A dot located in the upper-right corner of Fig. 3 has the highest area and therefore the highest general bias. This area measure is the *Accuracy Weighted Burden* or AWB, shown in Eq. 6. We then calculate its normalized version, NAWB$_s$, for all the datasets and models and show it in Fig. 4.

Normalized Accuracy Weighted Burden (NAWB) The NAWB$_s$ measure is not as sensitive to the CE method used, due to the FNR factor, however, the magnitude may still change significantly. This is observed throughout all the datasets. In the Adult dataset, *Females*, *Non-whites* and *<25* are the most unfavored in terms of bias, and the ordering of the age groups is the same across methods. This was not true for the Adult Burden$_s$ measure alone, in which (see Fig. 2) the Burden$_s$ was higher for the *>60* group. In the German dataset the NAWB$_s$ measure shows a higher bias against *Males* than *Females* than FNR$_s$ or Burden$_s$ alone indicating that the difficulty of each IOI to change its label brings an added level of bias against *Males* to the already higher ratio of false negatives in that group, compared to *Females*. This indicates that it is important to consider both metrics and in that way improve the overall perspective on the relative biases among groups. Additionally, the consistency between these relative measures is greatly improved. For example, the German and Dutch datasets Burden_s measure showed a different behavior among groups, compared to the more holistic NAWB_s measure. However, in terms of magnitude, the NAWB_s measure indicates a higher value for the RT and CCHVAE methods, which is still justified by the objectives they prioritize, as mentioned before. This may indicate that, although relative NAWB_s among the groups is more consistent across diverse CE methods, the magnitude is still dependent on the CE method. Further improvements may be done on normalization of the measures with respect to each sensitive feature, or across the features, for example, considering the NAWB_s fraction over the sum of all NAWB_s to make this metric less dependent on which CE method is applied.

5 Conclusions and future work

In this study, we performed an evaluation of four different CE methods to assess the burden on different sensitive groups due to a classifier model. The distance between the CEs and the instances are seen as a measure of fairness through counterfactual reasoning. We compared this measure with an accuracy-based fairness measure, the False Negative Ratio per sensitive group, and propose a combined product of these measures that attempts to more consistently measure the (un)fairness of classifiers. Hence, we proposed NAWB_s as a normalized, accuracy and counterfactual-based measure to determine the existence of classifier bias, proving that it may enhance the evaluation of biases among sensitive groups. We assessed the difference among the groups burden identified by different CE methods, and that future work may deal with a further normalization process to make this measure independent of the CE method used.

Additionally future work should also consider other methods, such as MACE, GS, DiCE and LORE. Finally, an extension to multi-class tasks and the application of the combined measure in the design of a classifier may be done, in order to make a generalized system that optimizes for both fairness and accuracy.

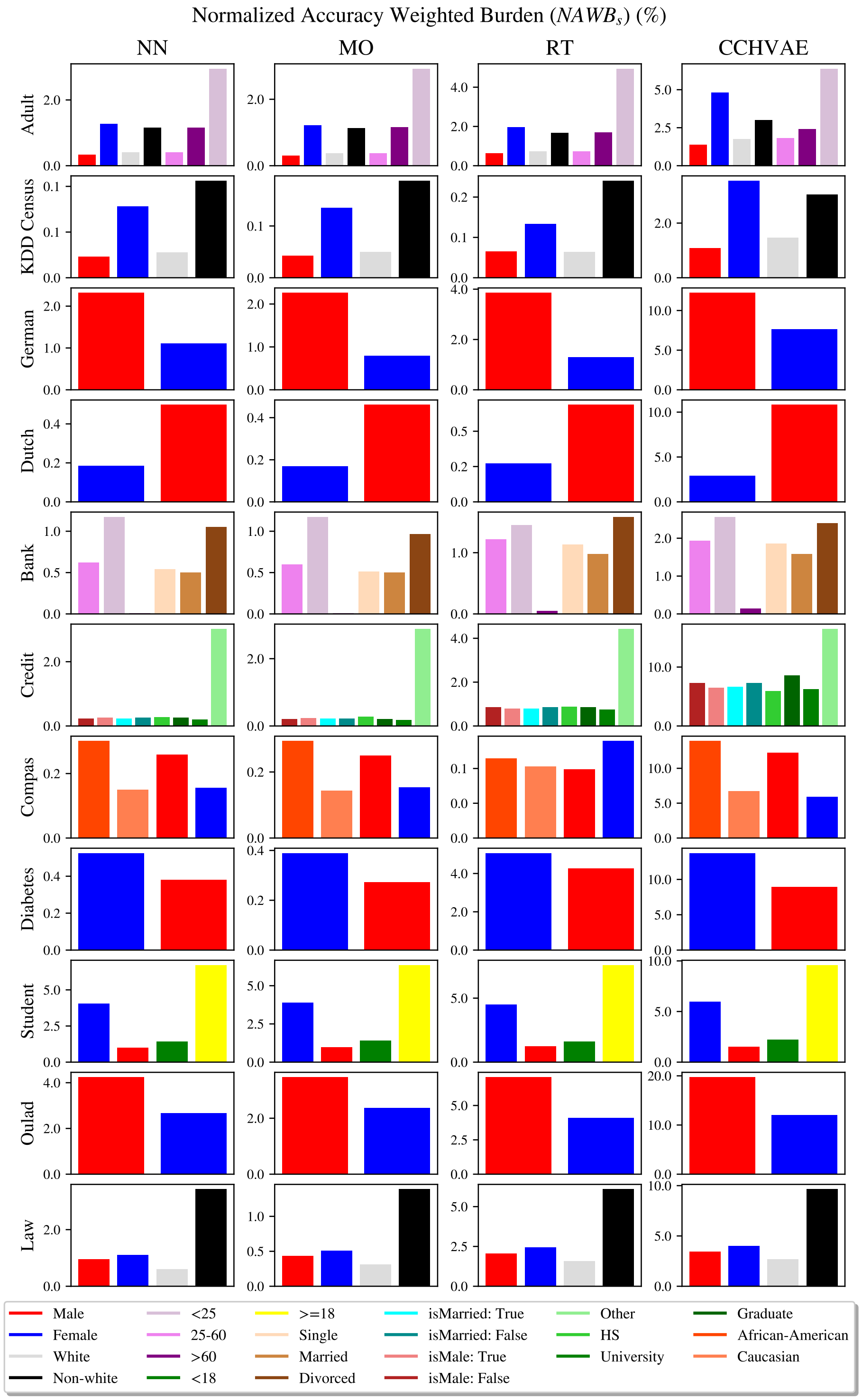

Fig. 4: Normalized Accuracy Weighted Burden (AWB) for each dataset and sensitive group

References

1. Boer, N., Deutch, D., Frost, N., Milo, T.: Just in time: Personal temporal insights for altering model decisions. In: 2019 IEEE 35th International Conference on Data Engineering (ICDE). pp. 1988–1991. IEEE (2019)
2. Coston, A., Mishler, A., Kennedy, E.H., Chouldechova, A.: Counterfactual risk assessments, evaluation, and fairness. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. pp. 582–593. ACM, Barcelona Spain (Jan 2020). `https://doi.org/10.1145/3351095.3372851`, `https://dl.acm.org/doi/10.1145/3351095.3372851`
3. Dodge, J., Liao, Q.V., Zhang, Y., Bellamy, R.K., Dugan, C.: Explaining models: an empirical study of how explanations impact fairness judgment. In: Proceedings of the 24th international conference on intelligent user interfaces. pp. 275–285 (2019)
4. Karimi, A.H., Barthe, G., Balle, B., Valera, I.: Model-agnostic counterfactual explanations for consequential decisions. In: International Conference on Artificial Intelligence and Statistics. pp. 895–905. PMLR (2020)
5. Karlsson, I., Rebane, J., Papapetrou, P., Gionis, A.: Locally and globally explainable time series tweaking. Knowledge and Information Systems **62**(5), 1671–1700 (2020)
6. Kearns, M., Neel, S., Roth, A., Wu, Z.S.: An empirical study of rich subgroup fairness for machine learning. In: Proceedings of the conference on fairness, accountability, and transparency. pp. 100–109 (2019)
7. Kuratomi, A., Lindgren, T., Papapetrou, P.: Prediction of global navigation satellite system positioning errors with guarantees. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 562–578. Springer (2020)
8. Kusner, M.J., Loftus, J.R., Russell, C., Silva, R.: Counterfactual Fairness. arXiv:1703.06856 [cs, stat] (Mar 2018), `http://arxiv.org/abs/1703.06856`, arXiv: 1703.06856
9. Kyrimi, E., Neves, M.R., McLachlan, S., Neil, M., Marsh, W., Fenton, N.: Medical idioms for clinical bayesian network development. Journal of Biomedical Informatics **108**, 103495 (2020)
10. Laugel, T., Lesot, M.J., Marsala, C., Renard, X., Detyniecki, M.: Inverse classification for comparison-based interpretability in machine learning. arXiv preprint arXiv:1712.08443 (2017)
11. Laugel, T., Lesot, M.J., Marsala, C., Renard, X., Detyniecki, M.: Unjustified classification regions and counterfactual explanations in machine learning. In: Joint European conference on machine learning and knowledge discovery in databases. pp. 37–54. Springer (2019)
12. Lindgren, T., Papapetrou, P., Samsten, I., Asker, L.: Example-based feature tweaking using random forests. In: 2019 IEEE 20th International Conference on Information Reuse and Integration for Data Science (IRI). pp. 53–60. IEEE (2019)
13. Loi, M., Ferrario, A., Viganò, E.: Transparency as design publicity: explaining and justifying inscrutable algorithms. Ethics and Information Technology **23**(3), 253–263 (Sep 2021). `https://doi.org/10.1007/s10676-020-09564-w`, `https://link.springer.com/10.1007/s10676-020-09564-w`
14. Molnar, C.: Interpretable machine learning: A guide for making black-box models explainable (2021), `https://christophm.github.io/interpretable-ml-book/limo.html`

15. Mothilal, R.K., Sharma, A., Tan, C.: Explaining machine learning classifiers through diverse counterfactual explanations. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. pp. 607–617 (2020)
16. Nobrega, C., Marinho, L.: Towards explaining recommendations through local surrogate models. In: Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing. p. 1671–1678. SAC '19, Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3297280.3297443
17. Pawelczyk, M., Broelemann, K., Kasneci, G.: Learning model-agnostic counterfactual explanations for tabular data. In: Proceedings of The Web Conference 2020. pp. 3126–3132 (2020)
18. Pitoura, E., Stefanidis, K., Koutrika, G.: Fairness in rankings and recommendations: an overview. The VLDB Journal (Oct 2021). https://doi.org/10.1007/s00778-021-00697-y, https://link.springer.com/10.1007/s00778-021-00697-y
19. Pitoura, E., Tsaparas, P., Flouris, G., Fundulaki, I., Papadakos, P., Abiteboul, S., Weikum, G.: On Measuring Bias in Online Information. ACM SIGMOD Record **46**(4), 16–21 (Feb 2018). https://doi.org/10.1145/3186549.3186553, https://dl.acm.org/doi/10.1145/3186549.3186553
20. Quy, T.L., Roy, A., Iosifidis, V., Zhang, W., Ntoutsi, E.: A survey on datasets for fairness-aware machine learning. arXiv:2110.00530 [cs] (Jan 2022), http://arxiv.org/abs/2110.00530, arXiv: 2110.00530
21. Ribeiro, M.T., Singh, S., Guestrin, C.: "why should i trust you?" explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. pp. 1135–1144 (2016)
22. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence **1**(5), 206–215 (2019)
23. Sharma, S., Henderson, J., Ghosh, J.: CERTIFAI: Counterfactual Explanations for Robustness, Transparency, Interpretability, and Fairness of Artificial Intelligence models. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society pp. 166–172 (Feb 2020). https://doi.org/10.1145/3375627.3375812, http://arxiv.org/abs/1905.07857, arXiv: 1905.07857
24. Tolomei, G., Silvestri, F., Haines, A., Lalmas, M.: Interpretable predictions of tree-based ensembles via actionable feature tweaking. In: Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining. pp. 465–474 (2017)
25. Tsintzou, V., Pitoura, E., Tsaparas, P.: Bias Disparity in Recommendation Systems. arXiv:1811.01461 [cs] (Nov 2018), http://arxiv.org/abs/1811.01461, arXiv: 1811.01461
26. Ustun, B., Spangher, A., Liu, Y.: Actionable recourse in linear classification. In: Proceedings of the conference on fairness, accountability, and transparency. pp. 10–19 (2019)
27. Verma, S., Dickerson, J., Hines, K.: Counterfactual Explanations for Machine Learning: A Review. arXiv:2010.10596 [cs, stat] (Oct 2020), http://arxiv.org/abs/2010.10596, arXiv: 2010.10596
28. Wexler, J., Pushkarna, M., Bolukbasi, T., Wattenberg, M., Viégas, F., Wilson, J.: The what-if tool: Interactive probing of machine learning models. IEEE transactions on visualization and computer graphics **26**(1), 56–65 (2019)
29. Zafar, M.B., Valera, I., Rodriguez, M.G., Gummadi, K.P.: Fairness Constraints: Mechanisms for Fair Classification. arXiv:1507.05259 [cs, stat] (Mar 2017), http://arxiv.org/abs/1507.05259, arXiv: 1507.05259