Fair Multiclass Classification for a Black-Box Classifier

Grigori Jasnovidov
griga1995@yandex.ru
ITMO University
Saint Petersburg, Russian Federation

Elizaveta Tarasova
el.u.tarasova@gmail.com
ITMO University
Saint Petersburg, Russian Federation

ABSTRACT

Algorithmic fairness is a widely studied area in machine learning field. The tasks of fair regression and fair binary classification are quite well explored up to the current moment. However, just few works consider a problem of fair multi-class classification despite its potential usefulness in areas like credit scoring, school and university admission, criminal jurisdiction, etc. Indeed, in all these issues, the predicted label may take more than two values. The credit liability may be estimated as 'low','medium' and 'high'; the risk of recidivism may also have several values; the future performance of a student can be evaluated as a non-binary variable. In this paper, we present a post-processing type algorithm that increases fairness in multi-class classification problems. The core of our approach is a linear programming problem that allows our algorithm to relabel some predictions of the initial classifier in order to improve fairness with a small possible loss in accuracy. We evaluate performance of our algorithm on synthetic and real datasets. As the results show, depending on the dataset, our algorithm increases fairness without statistically significant loss in accuracy.

KEYWORDS

fairness, multiclass classification, conditional use accuracy equality, post-processing, linear programming

ACM Reference Format:
Grigori Jasnovidov and Elizaveta Tarasova. 2023. Fair Multiclass Classification for a Black-Box Classifier. In *Proceedings of 2nd ACM SIGKDD Workshop on Ethical Artificial Intelligence: Methods and Applications (ACM SIGKDD).* ACM, New York, NY, USA, 6 pages.

1 INTRODUCTION

In the modern world a lot of AIs and, in particular, machine learning technologies are used in many social areas leading to various fairness issues. The concept of machine learning fairness is reduction of a bias caused by abuse of the sensitive variables during the decision-making process. Examples of sensitive variables are gender, ethnicity, sexual orientation, disability, etc.

The work [26] provides an overview of fairness in ML algorithms: a lot of research is focused on such areas as criminal justice, immigration, other public sectors and healthcare, which highlights the importance of studying ML fairness. The study [11] discusses the legislative complexities that arise when trying to reduce discrimination of legally protected groups. Another work [22] is devoted to fairness in the healthcare sector. In doing so, models learn from historical data, which in the past was based partially on human and structural biases. In this regard, discrimination arises for certain groups of patients. The paper [15] considers the problem of violation of fairness in school based on racial characteristics when predicting grades. One of the benchmarks in fair machine learning is the COMPAS dataset that considers the problem of predicting the recidivism of certain groups of prisoners. COMPAS data are used in more studies to test different definitions of algorithmic fairness, see [3, 8, 29].

As we see, in all cases above, a problem of fair classification appears naturally. Despite a lot of papers on of fair binary classification see e.g., [2, 16] and references therein just few works consider a problem of multi-class classification, see [7, 10, 19, 21]. However, in some cases the classification may be not binary, but multinomial. For example, if a bank assesses the solvency of clients, the gradation of "solvent" and "insolvent" may not be enough; it can be necessary to divide customers into more categories such as "unreliable", "medium-reliable" and "reliable". Another example where more than two classes are required is the classification of pathologies in the context of image analysis in medicine, see [23]. For COMPAS studies [3, 8, 29], a prisoner usually classified to several risk recidivism groups. A student grade prediction problem may also have more than two outcomes, see [7, 24].

In this article, we propose a Fairness Multiclass Classification Linear Programming Algorithm (later on, FMCLP algorithm) for solving the fair multiclass classification problem. FMCLP algorithm is a post-processing type algorithm that processes the results of an arbitrary classifier combined with the values of the sensitive attribute and relabels them in order to improve a chosen fairness metric. The core of our approach is the linear programming problem corresponding to the output of a given classifier, values of the sensitive attribute and a chosen fairness metric. Solution of this problem allows us to improve the chosen fairness metric, see Section 4 for detailed description of FMCLP algorithm. A merit of our algorithm is its generality and lack of prior assumptions on a given dataset and initial classifier. As we will see in Section 4, we can apply the FMCLP algorithm to a big variety of classification problems and fairness metrics. We tested FMCLP algorithm on two synthetic and three real datasets, the results are presented in Section 5. According to the results, there is a decrease in the percentage of people who are discriminated on all real datasets after implementation of FMCLP algorithm. At the same time, the task of multi-classification is applicable in various areas of life. Thus, this work has not only technical but also social significance.

ACM SIGKDD, August 7, 2023, Long Beach, California

2 RELATED WORK

Binary classification. Some papers are aimed at working with certain groups of classifiers, with the consideration of the classification method as a "black box". The article [2] presents a general approach for the black box case in the context of binary classification. The proposed approach covers a wide range of fairness definitions which can be formalized using linear inequalities in conditional moments, such as demographic equality or equalized odds. Based on this contribution, the fairlearn open source package [5] is developed by Microsoft.

Multi-class classification. The paper [28] focuses essentially on Support Vector Machine fair predictions. One of the main ideas there is the addition to the loss function of a classical ML algorithm a special term responsible for fairness improvement. The obtained optimization problem can be interpreted as a mixed integer linear problem and solved by the modern developed techniques, see, e.g, [13]. In another work [21], an approach for modifying classifier predictions by linear programming to achieve fairness in a multi-class environment is also considered. Contribution [27] is devoted to fair deep learning, particularly on image classification task. The idea there is simultaneous training of sample weights and neural network parameters trying to optimize accuracy and satisfy fairness constraints, while keeping the architecture of the neural network unchanged. We refer to study [7] for a detailed discussion on demographic parity improvement in multidimensional setup. The work [10] presents the debiasing multi-variable method. The idea behind this preprocessing-type approach is preprocessing the data in order to make all discriminated groups balanced and then apply the initially given ML classifier. The other side of discrimination can be biases in prediction that affects minority subgroups in the training data. To solve this problem, [19] proposed a post-processing multi-accuracy audit framework. The principal idea here is improving accuracy on discriminated minority subgroups using a special algorithm "multi-accuracy boost" that processes the results of a given black-box classifier and the values of protected attributes.

3 CONDITIONAL USE ACCURACY EQUALITY

In this section we discuss fairness metric *conditional use accuracy equality*, later on *cuae-metric*. This metric is presented in [4] for binary case; in this contribution we discuss its multidimensional analogous (see [21] also). We decided to use this metric to show how FMCLP algorithm performs because of its benefits against demographic parity and equalized odds, see the end of this section for more explanations. For definitions and properties of other fairness metrics, we refer to [6, 18, 20] and references therein.

Consider some dataset with a label Y, classifier R and sensitive attribute A. We say that R satisfies *conditional use accuracy equality* if $P(Y = y|R = y, A = a) = P(Y = y|R = y, A = b), \quad \forall y \in Y,\ a, b \in A$ and *cuae-metric* of R is the matrix of all probabilities $P(Y = y|R = y, A = a), y \in Y, a \in A$. To numerically evaluate a classifier in terms of cuae-metric, for any classifier R we define *cuae-difference* by $\max_{y \in Y, a,b \in A} (P(Y = y|R = y, A = a) - P(Y = y|R = y, A = b)$, *cuae-ratio* by

$$\max_{y \in Y, a,b \in A} \frac{P(Y = y|R = y, A = a)}{P(Y = y|R = y, A = b)} \tag{1}$$

and *cuae-variation* by

$$\max_{x,y \in Y, a,b \in A} (P(Y = x|R = x, A = a) - P(Y = y|R = y, A = b)) .$$

It is clear from the definition, that if cuae-difference of a classifier is close to zero or cuae-ratio is close to one, then the classifier almost satisfies cuae. The cuae-variation measures dispersion of performance of a classifier over different protected groups. The cuae-metric reminds equalized odds metric (see [12] for the definition); the difference is that we do not impose equations like

$$P(Y = y_1|R = y_2, A = a) = P(Y = y_1|R = y_2, A = b), \\ y_1 \neq y_2 \in Y, a, b \in A, \tag{2}$$

that makes the cuae-metric more flexible than equalized odds and gives more chances for a classifier to satisfy it. Observe that regardless of relations between variables, the ideal classifier $R = Y$ fulfils cuae, that in case of correlation between A and Y makes cuae more preferable than demographic parity metric (see [9]).

4 FMCLP ALGORITHM

This section provides a description of the implementation of FMCLP algorithm for cuae-metric. We show the input requirements, all steps of the algorithm and how the problem is reduced to linear programming.

Input. Input is a dataset for classification problem with an already trained ML classifier (later on *initial classifier*) that solves a multi-class classification task. We assume that the initial classifier returns prior probabilities that a certain observation belongs to one of the classes, later on we refer to these probabilities as *black-box probabilities*. Suppose that one of the features is a sensitive attribute. We do not assume any particular structure of the initial classifier, like classical ML algorithm, neural network, etc.

Step 1. For $s >> 1$ randomly chosen observations we build the matrix consisting of the corresponding black-box probabilities, values of the sensitive attribute and true labels.

Step 2. For the matrix obtained in Step 1 we solve the classification problem (features are the sensitive attribute and black-box probabilities, target is the labels) trying to maximize accuracy and satisfy cuae condition in the following way. Assume that the sensitive attribute takes values from $\mathcal{A} := \{1, 2, ..., a\}$, where $a > 1$ and $\mathcal{L} = \{1, 2, ..., l\}$ is the set of labels. We split all considered observations into a groups, say, $A_1, A_2, ..., A_a$ regarding to their value of the sensitive attribute; suppose that $A_i, i \in \mathcal{A}$ consists of K_i observations and $\mathcal{K}_i = \{1, 2, ..., K_i\}$. Let $p_i^{(k,j)}, i \in \mathcal{A}, j \in \mathcal{L}, k \in \mathcal{K}_i$ be the black-box probability that the k-th observation from A_i belongs to class j and l_i^k be the true label of this observation. The initial classifier classifies each observation to the class with the biggest value of the corresponding probability $p_i^{(k,j)}$. Our aim is to reclassify some of the observations above in order to improve cuae-metric with the smallest possible decrease in accuracy. Denote as C the set of all possible $c_i^{(k,j)} \in \{0, 1\}, i \in \mathcal{A}, j \in \mathcal{L}, k \in \mathcal{K}_i$ such that for all i, k $\sum_{j \in \mathcal{L}} c_i^{(k,j)} = 1$. Let Y be the true mapping between the features and target, and Cl be the classifier we are looking for. We maximize

$$\sum_{i \in \mathcal{A}, j \in \mathcal{L}, k \in \mathcal{K}_i} p_i^{(k,j)} c_i^{(k,j)} \tag{3}$$

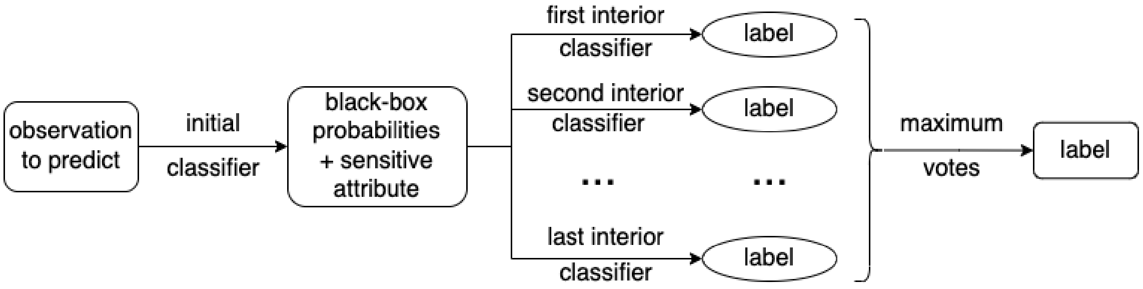

Figure 1: Predicting of FMCLP Algorithm.

over C subject to *cuae-constraint*

$$P(Cl = x|Y = x, A = a) = P(Cl = y|Y = y, A = a),$$
$$\forall a \in \mathcal{A}, \forall x, y \in \mathcal{L}. \qquad (4)$$

Let $I_{a,x}$ stands for the set of observations, such that their value of sensitive attribute is a and their label is x. The last line can be rewritten in terms of a linear programming problem. Indeed, for all $x \in \mathcal{L}$ and $a \in A$

$$P(Cl = x|Y = x, A = a) =$$
$$= \frac{P(Cl = x, Y = x, A = a)}{P(Y = x, A = a)} = \frac{\sum\limits_{k \in I_{a,x}, j \in \mathcal{L}} c_a^{(k,j)}}{|I_{a,x}|}, \qquad (5)$$

where $|\cdot|$ denotes the number of elements in a finite set. Hence, we can rewrite our problem as max $\sum\limits_{i \in \mathcal{A}, j \in \mathcal{L}, k \in \mathcal{K}_i} p_i^{(k,j)} c_i^{(k,j)}$ over C subject to

$$\frac{\sum\limits_{k \in I_{a,x}, j \in \mathcal{L}} c_a^{(k,j)}}{|I_{a,x}|} = \frac{\sum\limits_{k \in I_{a,y}, j \in \mathcal{L}} c_a^{(k,j)}}{|I_{a,y}|}, \quad \forall a \in \mathcal{A},\ x, y \in \mathcal{L}. \qquad (6)$$

This linear programming problem can be solved using the HiGHS dual simplex solver method, see [13]. The result of this solution is some set $c_i^{(k,j)}, i \in \mathcal{A}, j \in \mathcal{L}, k \in \mathcal{K}_i$. For any fixed $i \in \mathcal{A}, j \in \mathcal{L}$ $\sum\limits_{k} c_i^{(k,j)}$=1 and hence it implies to a natural way of classification k-th observation from A_i the label x, where $c_i^{(k,x)}$ is maximal among $c_i^{(k,j)}$. In this way, we have build a vector of new labels, that is the output of this step.

Step 3. Now we have a dataset, where features are the values of sensitive attribute and black-box probabilities, while target is the predicted in the previous step labels. We take a simple classical ML classifier like decision tree, random forest, k-nearest neighbours algorithm, etc., (later on *interior classifier*) and train it on this data. Thus, we obtain a classifier that predicts "fair" labels based on the value of the sensitive attribute and the black-box probabilities.

Step 4. We repeat Steps 1–3 n times and get an ensemble of classifiers. That is the output of FMCLP algorithm.

Predicting. By the initial classifier, we obtain the black-box probabilities corresponding to a given observation. Next, each of the classifiers obtained in Step 4 predicts label of the given observation. The final prediction is the label, that gets the maximal amount of votes.

If we want to apply FMCLP algorithm to optimize a different fairness metric, we need to impose the corresponding condition instead of cuae-constraint (4). This leads to changes in (6) and, consequently, to a new linear programming problem. The rest of FMCLP algorithm remains the same.

5 EXPERIMENTS AND RESULTS

We have tested our approach for 3 classes classification problems with sensitive attribute taking 2 values over real and synthetic datasets. In all problems we take as initial classifier Light Gradient Boosted Machine classifier (later on LGBM classifier, see [17]) because of its flexibility and usual good performance. For the code on Python 3 and full record of the results, see [14]. To evaluate our results, for each dataset, we run the experiment 100 times and computed cuae-metrics of initial and fair classifiers over the test group. Each table below presents the average values of characteristics of the corresponding cuae-metric. In all cases, we check statistical significance of fairness improvement by Wilcoxon-Signed Rank Test, see [25].

Synthetic datasets. We consider two types of synthetic data: with and without dependence between sensitive attribute and target. The test results for the cuae-difference, cuae-ratio, cuae-variation and accuracy are presented in Table 1.

Table 1: Synthetic dataset with / no dependence between sensitive attribute and target

	Difference		Ratio		Variation		Accuracy	
	with	no	with	no	with	no	with	no
Unfair	0.079	0.046	1.105	1.061	0.182	0.171	0.837	0.836
Fair	0.07	0.048	1.109	1.065	0.177	0.176	0.837	0.835
Impr.	11.1 %	-5.1 %	1.2 %	-0.4 %	2.7 %	-2.9 %	-0.02 %	0.1 %

For the dataset with dependence between a binary sensitive attribute and target we observe a small increase in fairness with a negligible loss in accuracy. Secondly, we evaluate behaviour of FMCLP algorithm when target and sensitive attribute are statistically independent. We observe a small decrease in fairness with statistically insignificant increase in accuracy. Indeed, since the sensitive attribute does not depend on the target, it is natural to observe a small decrease in fairness; meanwhile, FMCLP algorithm does not reduce accuracy in this case. This example shows that FMCLP algorithm does not collapses if applied to a data without dependence between sensitive attribute and target variable.

Real datasets. Based on the data provided by LSAC (see [1]), it is possible to model the problem of multiclass classification of law school students according to their GPA category, determined by their numerical GPA score. One may consider three categories instead of the binary classification "high—low", since high scores affect the further development of a student's career (e.g., they will be taken into account when applying for a job), and at the same time, middle scores also allow developing a career, unlike low scores. The task is to predict the category that a 3rd year student will fall into based on several appropriate features. We choose sensitive attributes to be gender and race, and test non-white people and non-white women on discrimination. Here we observe a decent improvement of all parameters in cuae-metric. This example shows that FMCLP algorithm can improve fairness for protected group combined of different sensitive attributes.

Table 2: LSAC dataset for non-white people (All) and non-white women (W)

	Difference		Ratio		Variation		Accuracy	
	All	W	All	W	All	W	All	W
Unfair	0.219	0.225	1.347	1.369	0.285	0.306	0.840	0.842
Fair	0.164	0.177	1.230	1.258	0.196	0.195	0.827	0.831
Impr.	25.1 %	21.3 %	8.7 %	8.1 %	31.2%	36.3%	-1.6 %	-1.3%

Our next example is a credit scoring dataset LOAN (see [14]). The target variable here is the score of the client: bad, middle or good, while sensitive attribute is one of the parameters in dataset. The implementation of FMCLP algorithm here strongly decreases cuae-difference, cuae-ratio and cuae-variation, while accuracy change is insignificant. This behaviour is consistent during all 100 experiments. Resuming, FMCLP algorithm has a decent performance over this dataset.

Table 3: LOAN and COMPAS datasets

	LOAN				COMPAS			
	Diff.	Ratio	Variat.	Acc.	Diff.	Ratio	Variat.	Acc.
Unfair	0.561	2.972	0.711	0.937	0.145	1.354	0.571	0.833
Fair	0.435	1.88	0.499	0.937	0.146	1.323	0.503	0.826
Impr.	22.4 %	36.7%	29.7 %	-0.02 %	-0.6 %	2.3 %	13.5 %	-0.8 %

Our last example is one of the COMPAS datasets (see [14]). Here the sensitive attribute is race (white or black) and the target is the potential possibility to recidivism (low, medium, or high). The results of testing on the COMPAS dataset are presented in the table 3. We see significant decrease in cuae-variation with small changes in cuae-difference, cuae-ratio and accuracy.

For all real datasets we estimate the percentage of people from sensitive groups who are unfairly assigned by the original classifier to the lower group that they actually performed, see Figure 2.

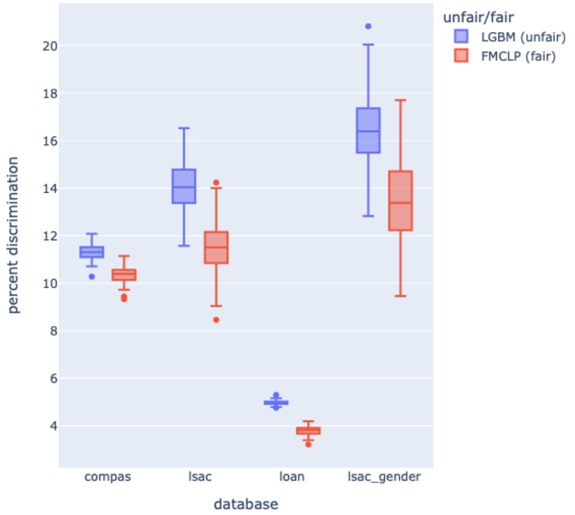

Figure 2: The dispersion of the discrimination percentage for LGBM (unfair) and FMCLP (fair) algorithms on different real datasets.

We see, that implementation of FMCLP algorithm reduces these numbers in all datasets and hence may have positive impact on the quality of life of individuals with certain sensitive parameters.

We have tested a relatively similar approach based on linear programming from [21] for all our real datasets. For each dataset we run the experiment 10 times using LGBM classifier as initial classifier and evaluated performance over test part of the data. The comparison of the results is presented in the Table 4.

Table 4: Comparison with other approach (1 - algorithm from [21], 2 - FMCLP)

	COMPAS		LOAN		LSAC		LSAC G	
	1	2	1	2	1	2	1	2
Diff.	0.079	0.146	0.763	0.435	0.161	0.164	0.160	0.177
Ratio	1.185	1.323	1.478	1.88	1.372	1.230	1.385	1.258
Variat.	0.769	0.503	0.906	0.499	0.388	0.196	0.400	0.195
Accur.	0.729	0.826	0.764	0.937	0.659	0.827	0.659	0.831
Unfair Acc.	0.833		0.937		0.840		0.842	

As we see from the results, our approach is significantly better in terms of drops in accuracy over all datasets. We have a quite similar performance over LSAC and LSAC-gender datasets in terms of fairness, while for LOAN dataset we have slightly better performance in cuae-metric. For COMPAS dataset algorithm from [21] better improves cuae-difference and cuae-ratio, but the loss in accuracy is significant.

Resuming, in all trials FMCLP has a small or statistically insignificant loss in accuracy, while fairness metric is typically improved. Moreover, for all real datasets we decreased a number of individuals that are assigned to the lower group that they deserve, that may have an application to some social real world problems.

6 DISCUSSION

An advantage of the FMCLP algorithm is its general applicability, i.e., it requires just the black-box probabilities and access to the protected groups. Thus, one can implement our approach to various classification problems without any restrictions on type of initial classifier, number of labels etc. Depending on a problem, FMCLP algorithm can be used to optimize different fairness metrics, like demographic parity, equalized odds, etc. To achieve this, we need to change the corresponding constraint in the linear programming problem in Step 2. Also, if an initial classifier returns the labels instead of the black-box probabilities, we can still use the FMCLP algorithm, considering the output as probabilities (0,0,…,1,…,0), where 1 corresponds to the predicted label. We suppose, that it makes sense to implement FMCLP algorithm in problems with at least three classes. For binary classification, there are several efficient algorithms, see, e.g., Fairlearn package [5].

As our experiments show, the loss in accuracy after implementation of FMCLP algorithm is usually small and often is statistically negligent, while fairness is improved.

7 CONCLUSION

In this paper, we introduced FMCLP algorithm for reducing the discrimination in multi-class classification problems and evaluated its performance on real and synthetic datasets. In Section 5 we see, that FMCLP algorithm improves fairness metric conditional use accuracy equality without significant loss in accuracy over three real and one synthetic datasets. Moreover, for all real datasets we observed that FMCLP algorithm decreases amount of people assigned by the initial classifier to the lower group that they belong to.

The framework presented in our paper can be applied very broadly. Indeed, FMCLP algorithm needs as input just black-box probabilities and values of sensitive attributes and, due to its linear program structure, can optimize various fairness metrics. We suppose that good performance of our approach over datasets in Section 5 may warrant FMCLP algorithm to be a subject of further investigation.

ACKNOWLEDGMENTS

This work was supported by the Analytical Center for the Government of the Russian Federation (IGK 000000D730321P5Q0002), agreement №70-2021-00141.

REFERENCES

[1] 2022. Law School Admissions Bar Passage. *Kaggle* (2022). https://www.kaggle.com/datasets/danofer/law-school-admissions-bar-passage/discussion/350765

[2] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. 2018. A Reductions Approach to Fair Classification. (2018). https://doi.org/10.48550/ARXIV.1803.02453

[3] Wael Alghamdi, Hsiang Hsu, Haewon Jeong, Hao Wang, P. Winston Michalak, Shahab Asoodeh, and Flavio P. Calmon. 2022. Beyond Adult and COMPAS: Fairness in Multi-Class Prediction. (2022). https://doi.org/10.48550/ARXIV.2206.07801

[4] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2017. Fairness in Criminal Justice Risk Assessments: The State of the Art. (2017). https://doi.org/10.48550/ARXIV.1703.09207

[5] Fairlearn contributors. 2018. fairlearn.metrics package. (2018). https://fairlearn.org/v0.5.0/api_reference/fairlearn.metrics.html

[6] Sam Corbett-Davies and Sharad Goel. 2018. The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. (2018).

[7] Christophe Denis, Romuald Elie, Mohamed Hebiri, and François Hu. 2021. Fairness guarantee in multi-class classification. (2021). https://doi.org/10.48550/ARXIV.2109.13642

[8] Julia Dressel and Hany Farid. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science Advances* 4, 1 (2018), eaao5580. https://doi.org/10.1126/sciadv.aao5580 arXiv:https://www.science.org/doi/pdf/10.1126/sciadv.aao5580

[9] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Rich Zemel. 2011. Fairness Through Awareness. (2011). https://doi.org/10.48550/ARXIV.1104.3913

[10] Giordano d'Aloisio, Giovanni Stilo, Antinisca Marco, and Andrea D'Angelo. 2022. *Enhancing Fairness in Classification Tasks with Multiple Variables: A Data- and Model-Agnostic Approach.* 117–129. https://doi.org/10.1007/978-3-031-09316-6_11

[11] Philipp Hacker. 2018. Teaching fairness to artificial intelligence: Existing and novel strategies against algorithmic discrimination under EU law. *Common Market Law Review* (2018), 1143–1185. http://www.kluwerlawonline.com/api/Product/CitationPDFURL?file=Journals%5CCOLA%5CCOLA2018095.pdf

[12] Moritz Hardt, Eric Price, Eric Price, and Nati Srebro. 2016. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems*, D Lee, M Sugiyama, U Luxburg, I Guyon, and R Garnett (Eds.), Vol. 29. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2016/file/9d2682367c3935defcb1f9e247a97c0d-Paper.pdf

[13] J. A. J. Huangfu, Q., Galabova, I., Feldmeier, M., Hall. 2020. HiGHS - high performance software for linear optimization. (2020). https://www.maths.ed.ac.uk/hall/HiGHS/%5C#guide

[14] G. Jasnovidov. 2023. https://github.com/GrigoriJasnovidov/fairness_submision

[15] Weijie Jiang and Zachary A. Pardos. 2021. Towards Equity and Algorithmic Fairness in Student Grade Prediction. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society.* ACM. https://doi.org/10.1145/3461702.3462623

[16] Faisal Kamiran and Toon Calders. 2011. Data Pre-Processing Techniques for Classification without Discrimination. *Knowledge and Information Systems* 33 (10 2011). https://doi.org/10.1007/s10115-011-0463-8

[17] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu. [n. d.]. LightGBM: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems, Long Beach, CA, USA, 3146–3154.

[18] Niki Kilbertus, Mateo Rojas-Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. 2017. Avoiding Discrimination through Causal Reasoning. (2017). https://doi.org/10.48550/ARXIV.1706.02744

[19] Michael Kim, Amirata Ghorbani, and James Zou. 2018. Multiaccuracy: Black-Box Post-Processing for Fairness in Classification. *arXiv* (2018). https://doi.org/10.48550/arXiv.1805.12317

[20] Karima Makhlouf, Sami Zhioua, and Catuscia Palamidessi. 2021. On the Applicability of ML Fairness Notions. *SIGKDD Explor. Newsl.* 23, 1 (2021), 14–23. https://doi.org/10.1145/3468507.3468511

[21] Preston Putzel and Scott Lee. 2022. Blackbox Post-Processing for Multiclass Fairness. *arXiv* (2022). https://doi.org/10.48550/arXiv.2201.0446

[22] Alvin Rajkomar, Michaela Hardt, Michael Howell, Greg Corrado, and Marshall Chin. 2018. Ensuring Fairness in Machine Learning to Advance Health Equity. *Annals of Internal Medicine* 169 (2018). https://doi.org/10.7326/M18-1990

[23] María Agustina Ricci Lara, Rodrigo Echeveste, and Enzo Ferrante. 2022. Addressing fairness in artificial intelligence for medical imaging. *Nature Communications* 13, 1 (2022), 4581. https://doi.org/10.1038/s41467-022-32186-3

[24] L. F. Wightman and H. Ramsey. 1998. LSAC national longitudinal bar passage study. Law School, Admission Council. (1998).

[25] F. Wilcoxon. 1945. Individual comparisons by ranking methods. *Biometrics* 1 (1945), 80–83.

[26] Khensani Xivuri and Hossana Twinomurinzi. 2021. A Systematic Review of Fairness in Artificial Intelligence Algorithms. In *Responsible AI and Analytics for an Ethical and Inclusive Digitized Society*, Denis Dennehy, Anastasia Griva, Nancy Pouloudi, Yogesh K Dwivedi, Ilias Pappas, and Matti Mäntymäki (Eds.). Springer International Publishing, Cham, 271–284.

[27] Bobby Yan, Skyler Seto, and Nicholas Apostoloff. 2022. FORML: Learning to Reweight Data for Fairness. *Apple Inc., Stanford University* (2022).

[28] Qing Ye and Weijun Xie. 2020. Unbiased Subdata Selection for Fair Classification: A Unified Framework and Scalable Algorithms. (2020). https://arxiv.org/abs/2012.12356

[29] Hongjun Yoon. 2018. A Machine Learning Evaluation of the COMPAS Dataset. In *2018 IEEE International Conference on Big Data (Big Data)*. 5474–5474. https://doi.org/10.1109/BigData.2018.8622343

APPENDIX

In this section, we give some details on FMCLP algorithm workflow.

Time limitations. For the training time of FMCLP algorithm we have

$$T \approx n * (t_{lp}(s) + t_{train}(s)),$$

where $t_{lp}(s)$ is the time of solving a linear programming problem on a sample of size s, $t_{train}(s)$ is the training time of the interior classifier on the sample of size s and n is the number of iterations in Step 4. For large s typically $t_{lp}(s) \gg t_{train}(s)$ and hence T is approximately $n * t_{lp}(s)$. The prediction time for each observation is approximately $n * t_{pred}$, where t_{pred} is the prediction time for one interior classifier.

Parameter tuning. The first parameter to choose is the interior classifier. Baseline is "random forest" classifier, but it is possible to use different algorithms as logistic regression, random trees, support vector machines, k-nearest neighbours algorithm, etc. The main demand for the interior classifier is its operation time: algorithms like neural networks consume a lot of time for training and this will result in low speed of FMCLP algorithm. The second parameter to adjust is s, the size of a sample in Step 1, which affects both speed and quality. Large s (comparable to number of observations in the initial dataset) may make the linear programming problem computationally complex and time-consuming; small values of s typically do not provide a good performance. As we observed from experiments on our datasets, usually medium size of s is an optimal choice. On datasets from Section 5, we took $s \in [20\sqrt{N}, 45\sqrt{N}]$, where N is a number of observations in the initial dataset. If the number of classes is greater than 3, we suppose that one should choose the biggest s such that the linear program can be solved regarding computational resources. The last parameter to customize is n, a number of iterations in Step 4. The increase of this parameter linearly increases training and predicting time, but the results are become a little more robust. A typical good choice of this parameter is between 10 and 30. Smaller values may lead to high variance and inconsistency of the results, while bigger values just increases operation time without significant improvement of the result.

Non-optimal initial classifier case. Here, we show performance of FMCLP algorithm for non-optimal initial classifier choice. We train Gaussian Naive Bayesian classifier over COMPAS and LSAC datasets and then apply FMCLP algorithm. The results are presented in Table 5:

Table 5: LOAN and COMPAS datasets for non-optimal initial classifier

	COMPAS				LSAC			
	Diff.	Ratio	Variat.	Acc.	Diff.	Ratio	Variat.	Acc.
Unfair	0.285	2.193	0.859	0.514	0.442	2.699	0.743	0.453
Fair	0.269	2.328	0.0.876	0.456	0.374	3.757	0.808	0.419
Impr.	5.6 %	-6.2%	-2.0 %	-12.7 %	15.4 %	-39.2 %	-8.7 %	-8.1 %

We see that FMCLP algorithm still improves cuae-difference, but cuae-ratio is increased and accuracy is decreased. We think, that such behaviour of FMCLP algorithm can be due to the low accuracy of the initial classifier. Indeed, low accuracy of the initial classifier implies low reliability of black-box probabilities and hence FMCLP algorithm may show this performance.