Supplementary material - ABCFair: an Adaptable Benchmark approach for Comparing Fairness Methods

MaryBeth Defrance
Ghent University
marybeth.defrance@ugent.be

Maarten Buyl
Ghent University
maarten.buyl@ugent.be

Tijl De Bie
Ghent University
tijl.debie@ugent.be

A Datasets

Our experiments were performed on two types of data: the dual label *SchoolPerformance* dataset, and the staple, large-scale *folktables* datasets. In this section, we provide more details on the preprocessing (not to be confused with *fairness* preprocessing) we performed for each dataset.

A.1 SchoolPerformance

The SchoolPerformance dataset was created by Lenders and Calders [4]. This dataset is based on the "Student Alcohol Consumption"-dataset [2]. The *unbiased* labels are the labels of the original dataset and they indicate whether someone succeeded in their education. The biased labels are collected through human experiments, where human subjects are given some of the student's features and they note whether they think that student would succeed or not.

We used the sex and the education of the student's parents as the sensitive attributes for this dataset.

We removed all features that are other expressions of the labels (i.e. outcomes) and we removed the ID and name of the student from the dataset.

A.2 Folktables

The following datasets are all part of the folktables [3] datasets. The following holds for all of the datasets: Age is encoded to a binary feature which encodes whether someone's age is higher or lower than the average value when calculating for intersectional groups. Smaller race categories are grouped in order to maintain statistical power.

A.2.1 ACSPublicCoverage

The goal of the ACSPublicCoverage dataset is to predict whether someone is covered by public health insurance. Note that this is the only folktables dataset on which we report results in the main paper.

Sex, age, and rage are used as sensitive features for this datasets.

The features on ancestry and specific information of disability type are omitted in our use of the dataset. We deem these features as not relevant for this use case.

A.2.2 ACSEmployment

The goal in the ACSEmployment dataset is to predict whether someone is employed or not.

Submitted to the 38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks. Do not distribute.

Sex, age, marital status, race, and disability status are used as sensitive features.

We drop the column concerning relationship status as this is encoded in a less elaborate way in the marital status attribute.

A.2.3 ACSIncome

The goal in the ACSIncome dataset is to predict whether someone earns more than $50.000 per year.

Sex, age, marital status, race and disability status are used as sensitive features.

We drop the column concerning relationship status as this is encoded in a less elaborate way in the marital status attribute.

A.2.4 ACSMobility

The goal of the ACSMobility dataset is to predict whether someone has changed their address in the previous year.

Sex, age, race, and disability are the sensitive attributes.

The features on relationship status, ancestry, and specific disability type are omitted from the dataset.

A.2.5 ACSTravelTime

The goal of the ACSTravelTime is to predict whether someone has to commute for longer than 20 minutes to work.

Sex, age, race, and disability are used as sensitive attributes.

Relationship status and employment status of parents are not included as features.

B Experiment Setup

B.1 Model Architecture and Training Hyperparameters

The underlying model in all experiments was a fully-connected neural net. All hyperparameters (including the number of hidden layers in the neural net) were chosen based on the performance on the validation set when applying no fairness method (the naive baseline). The resulting hidden layer sizes, learning rates, number of epochs, and batch sizes are reported in Table A1.

Table A1: Hidden layer size, learning rate, number of epochs, and batch size used per dataset.

	Hidden Layer Sizes	Learning rate	# Epochs	Batch size
SchoolPerformance	[64]	0.001	80	64
ACSPublicCoverage	[512,256,64,32]	0.0001	40	2048
ACSEmployment	[512,256, 64, 32]	0.0001	40	2048
ACSIncome	[512,256,64]	0.0001	40	512
ACSMobility	[512,256,64]	0.0001	45	2048
ACSTravelTime	[16,256,128,64]	0.0001	20	1024

B.2 Fairness Strengths

All fairness methods have a hyperparameter that regulates the strength of fairness. Unfortunately, the most suitable scales for these strengths varies significantly across methods. In Tab. A2, we detail which fairness strength we used for each method and the additional strengths that were used for the ACSPublicCoverage dataset. These additional strengths were selected manually to further populate Tables 4, 5, and 6 (in the main paper).

Table A2: The standard and additional strengths used for each fairness method during training.

	Standard strengths	Additional strengths
Data Repairer	[0.1, 0.5, 0.8, 0.9, 1]	[1.3, 1.5, 2, 2.5, 3, 5]
Label Flipping	[0.001, 0.01, 0.03, 0.1, 0.3]	[0.5, 0.7, 1, 1.3, 1.5, 2]
Prevalence Sampling	[0.1, 0.5, 0.8, 0.9, 1]	[2, 3]
Learning Fair Repr.	[2, 5, 25, 50, 75]	[0.1, 0.5, 1, 10, 5000]
Fairret Norm	[0.001, 0.01, 0.1, 1, 3]	[0.0001, 0.5, 0.7, 5]
Fairret KL_{proj}	[0.001, 0.01, 0.1, 1, 3]	[1e-05, 5e-05, 0.0001, 0.0005, 0.001]
LAFTR	[0.001, 0.01, 0.1, 0.3, 1]	[0.0001, 2, 3, 5, 7, 10]
Prejudice Remover	[0.001, 0.01, 0.1, 0.3, 1]	[1e-05, 0.0001, 0.0005, 2, 3, 5]
Exponentiated Gradient	[0.8, 0.9, 0.95, 0.99, 1]	[0.3, 0.5, 0.6, 0.7]
Error Parity	[0.005, 0.01, 0.05, 0.1, 0.3]	[1e-05, 5e-05, 0.0001, 0.0005, 0.001]

B.3 Computational Resources

All experiments were conducted on an internal server equipped with a 12 Core Intel(R) Xeon(R) Gold processor and 256 GB of RAM. All experiments, including preliminary and failed experiments, cost approximately 800 hours per CPU.

This large computational cost results from the breath of the possible combinations of desiderata across a large set of methods.

C Additional Fairness Notions

In the main paper, we discuss the *demographic parity* (`dem_par`) and *equalized opportunity* (`eq_opp`) fairness notions. In our full benchmark, we consider 5 more [1]:

- *predictive equality* (`pred_eq`) requires *false positive rates* (fpr) $\gamma(k;h) = \frac{\mathbb{E}[S_k(1-h(X))]}{\mathbb{E}[S_k(1-Y)]}$ to be equal. It is a natural variant of equalized opportunity, but applied to negative labels.
- *predictive parity* (`pred_par`) requires *precisions* (ppv) $\gamma(k;h) = \frac{\mathbb{E}[S_k Y h(X)]}{\mathbb{E}[S_k h(X)]}$ to be equal.
- *false omission rate parity* (`forp`) requires *false omission rates* (for) $\gamma(k;h) = \frac{\mathbb{E}[S_k Y(1-h(X))]}{\mathbb{E}[S_k(1-h(X))]}$ to be equal. It is a natural variant of predictive parity, but applied to negative labels.
- *accuracy equality* (`acc_eq`) requires *accuracy* (acc) $\gamma(k;h) = \frac{\mathbb{E}[S_k(1-Y+(2Y-1)h(X))]}{\mathbb{E}[S_k]}$ to be equal.
- *F_1-score equality* (`f1_score_eq`) requires *F_1-scores* $\gamma(k;h) = \frac{\mathbb{E}[2S_k Y h(X)]}{\mathbb{E}[S_k(Y-h(X))]}$ to be equal.

Note that the shorthand name for each notion corresponds to an option in Sec. D.1. Though we measure the violations of these notions, most methods are not designed to optimize for these lesser known notions. We refer to Tab. 3 in the main paper for an overview of which method can equalize which statistic (also shorthanded in the list above).

D Additional Results

In the main paper, we only report the results of one dataset for three possible configurations of desiderata. Many more configurations can reported for each of the datasets, as we evaluate on 6 datasets (+ 1 from the unbiased labels in SchoolPerformance), 7 fairness notions, and 2 output formats, bringing the total amount of Tables we can generate to 98. The amount of trade-off curves we can generate (like in Fig. 2) is again multiplied by the amount of sensitive feature formats (3), making 294 plots possible.

Including all these results would overly clutter the appendix. Hence, we make all our results available in our repo at `https://github.com/aida-ugent/abcfair` and provide a simple command line interface to generate the Tables and Figures as shown in the main paper.

D.1 Performance Table Generation

The performance table allows for three configuration options: the dataset, the fairness notion with respect to which violation is measured, and the output format. Here, the k values used to generate the table can either be edited into the script. If not, k values will be automatically inferred from the fairness violation k' of the naive baseline as the values $[k'/4, k'/2, k']$.

The command line options are:

```
--data_name [DATA_NAME]
    Name of the data set. Current options are
    ['ACSPublicCoverage', 'ACSEmployment', 'ACSIncome', 'ACSMobility',
    'ACSTravelTime', 'SchoolPerformanceBiased', 'SchoolPerformanceUnbiased']
--notion [NOTION]
    The fairness notion to be used. Current options are
    ['dem_par', 'eq_opp', 'forp', 'pred_par', 'acc_eq', 'f1_score_eq', 'pred_eq']
--output_type [OUTPUT_TYPE]
    The output type. Options are
    ['hard', 'soft']
```

D.2 Trade-off Figure Generation

The accuracy-fairness trade-off figure has an additional configuration option: the sensitive feature format. To express uncertainty of the mean estimator of two-dimensional variables (the accuracy and the fairness violation), the plots show confidence ellipses, based on the methodology in [1] (Appendix D.4). The ellipse radii use the covariance matrix for the standard error.

The command line options are:

```
--data_name [DATA_NAME]
    Name of the data set. Current options are
    ['ACSPublicCoverage', 'ACSEmployment', 'ACSIncome', 'ACSMobility',
    'ACSTravelTime', 'SchoolPerformanceBiased', 'SchoolPerformanceUnbiased']
--notion [NOTION]
    The fairness notion to be used. Current options are
    ['dem_par', 'eq_opp', 'forp', 'pred_par', 'acc_eq', 'f1_score_eq', 'pred_eq']
--output_type [OUTPUT_TYPE]
    The output type. Options are
    ['hard', 'soft']
--sens_attr [SENS_ATTR]
    The sensitive attribute format. Current options are
    ['binary', 'intersectional', 'parallel']
```

References

[1] Maarten Buyl, MaryBeth Defrance, and Tijl De Bie. fairret: a Framework for Differentiable Fairness Regularization Terms. In *International Conference on Learning Representations*, 2024.

[2] Paulo Cortez and Alice Silva. Using data mining to predict secondary school student performance. *EUROSIS*, 01 2008.

[3] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. *Advances in Neural Information Processing Systems*, 34, 2021.

[4] Daphne Lenders and Toon Calders. Real-life Performance of Fairness Interventions - Introducing A New Benchmarking Dataset for Fair ML. In *Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing*, pages 350–357, Tallinn Estonia, March 2023. ACM.