Fairer Machine Learning Through Multi-objective Evolutionary Learning *

Qingquan Zhang[1,2], Jialin Liu[1,2][0000−0001−7047−8454], Zeqi Zhang[3], Junyi Wen[3], Bifei Mao[3], and Xin Yao[1,2][0000−0001−8837−4442]

[1] Research Institute of Trustworthy Autonomous System,
Southern University of Science and Technology (SUSTech), Shenzhen, China
[2] Guangdong Provincial Key Laboratory of Brain-inspired Intelligent Computation,
Department of Computer Science and Engineering,
Southern University of Science and Technology (SUSTech), Shenzhen, China
[3] Trustworthiness Theory Research Center,
Huawei Technologies Co., Ltd., Shenzhen, China
11930582@mail.sustech.edu.cn, {liujl,xiny}@sustech.edu.cn

Abstract. Dilemma between model accuracy and fairness in machine learning models has been shown theoretically and empirically. So far, dozens of fairness measures have been proposed, among which incompatibility and complementarity exist. However, no fairness measure has been universally accepted as the single fairest measure. No one has considered multiple fairness measures simultaneously. In this paper, we propose a multi-objective evolutionary learning framework for mitigating unfairness caused by considering a single measure only, in which a multi-objective evolutionary algorithm is used during training to balance accuracy and multiple fairness measures simultaneously. In our case study, besides the model accuracy, two fairness measures that are conflicting to each other are selected. Empirical results show that our proposed multi-objective evolutionary learning framework is able to find Pareto-front models efficiently and provide fairer machine learning models that consider multiple fairness measures.

Keywords: Fairness in machine learning · Discrimination in machine learning · AI ethics · Fairness measures · Multi-objective learning.

* This work was supported by the Research Institute of Trustworthy Autonomous Systems, the Guangdong Provincial Key Laboratory (Grant No. 2020B121201001), the Program for Guangdong Introducing Innovative and Entrepreneurial Teams (Grant No. 2017ZT07X386), the Guangdong Basic and Applied Basic Research Foundation (Grant No. 2021A1515011830), the Shenzhen Science and Technology Program (Grant No. KQTD2016112514355531), the Shenzhen Fundamental Research Program (Grant Nos. JCYJ20180504165652917, JCYJ20190809121403553) and Huawei project on "Fundamental Theory and Key Technologies of Trustworthy Systems". Corresponding author: X. Yao (xiny@sustech.edu.cn).

1 Introduction

Machine learning techniques are widely applied in real life, such as image recognition [9] and job application screening [21]. However, unfairness or discrimination from training data may lead to unfair data-driven models and unfair predictions. During the last decade, there has been a significantly growing research interest in measuring and mitigating unfairness in machine learning [1, 4, 14, 27, 34].

Dozens of (un)fairness measures for determining and evaluating the fairness of machine learning models trained only for accuracy have been defined [1, 4, 14, 27, 34] mainly based on the evaluated model's predicted outcomes, the predicted and actual outcomes, predicted probabilities and actual outcomes, similarity, and causal reasoning [27]. However, two dilemmas exist [1]: (i) the trade-off between model accuracy and its fairness has been theoretically and empirically shown [26]; (ii) some fairness measures have been shown to be conflicting with each other [3, 26], such as individual fairness and group fairness [26]. To tackle the former dilemma, some approaches for mitigating unfairness take into account a single fairness measure as a regularisation term or a constraint aiming at balancing between model accuracy and fairness [1]. However, no fairness measure has been universally accepted as the single fairest measure. To our best knowledge, no one has considered multiple fairness measures simultaneously during training. Motivated by this gap, we attempt to answer the following research question in this paper: *Can we make fair machine learning fairer by considering multiple fairness measures simultaneously?* This can be divided into three questions more precisely: ($\mathcal{Q}1$) Whether multi-objective learning can simultaneously optimise the model accuracy and multiple conflicting fairness measures? ($\mathcal{Q}2$) Can multi-objective learning optimise one or several fairness measures without degenerating others? In other words, whether the obtained models will become less fair according to any of the considered measures? ($\mathcal{Q}3$) Can we obtain a group of diverse models by applying multi-objective learning so that different trade-offs can be seen clearly?

To answer those questions, in this paper, we propose a multi-objective evolutionary learning framework which trains machine learning models for accuracy and multiple fairness measures simultaneously. The main contributions of this paper are as follows. (i) We propose a multi-objective evolutionary learning framework for training fairer machine learning models. Multi-objective optimisation is applied to consider model accuracy and multiple fairness measures simultaneously during training. (ii) We have implemented our framework with a concrete instantiation which uses the model error and two conflicting but complementary unfairness measures as three objectives during model training. (iii) Empirical results on three well-known benchmark sets show that our framework is able to find Pareto-front models effectively. (iv) The obtained models can act as good candidates for human decision-makers' use with different preferences.

The remainder of this paper is organised as follows. Section 2 presents related work. Our framework is proposed in Section 3.1. An effective instantiation of our framework is described in Section 3.2. Section 4 presents and discusses empirical studies. Section 5 concludes the paper.

2 Background

Concepts of (un)fairness in machine learning have been considered in many research fields for more than half a decade [14]. Different perspectives of fairness have different preferences and there is no universal definition of fairness [1, 14]. The interpretation of individual fairness and group fairness is one perspective being accepted by many studies [7, 26]. Individual fairness implies that similar individuals should be treated similarly. On the contrary, group fairness considers the fairness of different groups, which often depends on sensitive (also called protected [4]) attributes. Typically, sensitive attributes are traits considered to discriminate against by law, such as gender, race, age and so on. A number of articles [4, 14, 27, 34] have reviewed research progresses on measuring and mitigating unfairness in machine learning, in which dozens of fairness or unfairness measures have been defined.

Different (un)fairness measures have been used as a regularisation term or a constraint together with model accuracy/error during model training for balancing between model performance and fairness [1, 23]. In the work of [29], the fairness-accuracy Pareto front was estimated by training models for different weighted sums of model accuracy and one single fairness measure. However, correlation and disagreement between fairness measures have been observed [3, 26]. A model determined as fair under one measure could be unfair under another measure. Speicher *et al.* [26] illustrated the trade-off between model accuracy and fairness, as well as the trade-off between individual fairness and group fairness.

How to make fair machine learning fairer is a core research topic. In this paper, we intend to treat model accuracy and multiple fairness measures equally with multi-objective learning, which is significantly different from existing works.

3 Multi-objective Evolutionary Learning for Mitigating Unfairness

To make fair machine learning fairer, we propose a multi-objective evolutionary learning framework [2] to train models for accuracy and multiple fairness measures simultaneously. An instantiation of our framework is also provided.

3.1 Proposed Framework

Our general framework is presented in Algorithm 1. It takes as input a set of training data $\mathcal{D}_{train}$, a set of validation data $\mathcal{D}_{validation}$, a number of initial models $\mathcal{M}$, a multi-objective optimisation algorithm π and a set of model evaluation criteria $\mathcal{E}$, including an accuracy measure and multiple (un)fairness measures. Every time a new model is initialised or generated (line 1, 9), partial training [30, 31] on $\mathcal{D}_{train}$ is always applied as a kind of local search. Each model is evaluated with criteria $\mathcal{E}$ as objectives of π. In the main loop, μ promising models are selected according to some selection strategy of π (line 6), from which λ new models $\mathcal{M}'$ are generated with the aim of inheriting information from promising

Algorithm 1 Multi-objective learning framework for fairer machine learning.

Require: Initial models $\mathcal{M}_1, \ldots, \mathcal{M}_\lambda$
Require: Set of model evaluation criteria $\mathcal{E}$
Require: Training dataset $\mathcal{D}_{train}$, validation dataset $\mathcal{D}_{validation}$
Require: Multi-objective optimiser π
1: Partially train [30,31] $\mathcal{M}_1, \ldots, \mathcal{M}_\lambda$ over $\mathcal{D}_{train}$
2: **for** $i \in \{1, \ldots, \lambda\}$ **do**
3: $\quad \epsilon_i \leftarrow$ Evaluate $\mathcal{M}_i$ with criteria $\mathcal{E}$ on $\mathcal{D}_{validation}$
4: **end for**
5: **while** terminal conditions are not fulfilled **do**
6: $\quad \mathcal{P} \leftarrow$ Select μ promising models from $\mathcal{M}_1, \ldots, \mathcal{M}_\lambda$ with "best" $\epsilon_1, \ldots, \epsilon_\mu$ according to π
7: $\quad \mathcal{M}' \leftarrow$ Generate λ new models $\mathcal{M}'_1, \ldots, \mathcal{M}'_\lambda$ from $\mathcal{P}$ according to π
8: $\quad$ **for** $i \in \{1, \ldots, \lambda\}$ **do**
9: $\quad\quad \mathcal{M}'_i \leftarrow$ Partially train [30,31] $\mathcal{M}'_i$ on $\mathcal{D}_{train}$
10: $\quad\quad \epsilon'_i \leftarrow$ Evaluate $\mathcal{M}'_i$ with criteria $\mathcal{E}$ on $\mathcal{D}_{validation}$
11: $\quad$ **end for**
12: $\quad < \mathcal{M}_1, \epsilon_1 >, < \mathcal{M}_2, \epsilon_1 >, \ldots, < \mathcal{M}_\lambda, \epsilon_\lambda > \leftarrow$ Select λ promising models from $\{\mathcal{M}_1, \ldots, \mathcal{M}_\lambda\} \bigcup \{\mathcal{M}'_1, \ldots, \mathcal{M}'_\lambda\}$ by π based on $\epsilon_1, \ldots, \epsilon_\lambda$ and $\epsilon'_1, \ldots, \epsilon'_\lambda$, and then update $\mathcal{M}_1, \ldots, \mathcal{M}_\lambda$ and $\epsilon_1, \ldots, \epsilon_\lambda$ accordingly
13: **end while**
14: **Return** $\mathcal{M}_1, \ldots, \mathcal{M}_\lambda$

models (line 7). After partial training and model evaluation (line 8-11), λ models are selected from $\mathcal{M} \bigcup \mathcal{M}'$ by π as new $\mathcal{M}$ (line 12). The above steps repeat until a termination criterion is reached.

The core steps of our framework are the model evaluation based on multiple criteria and the generation of new models. Multi-objective evolutionary algorithms (MOEAs) [19] are ideal for those tasks. The output models can be further selected by human decision-makers or used as an ensemble of models [2, 31].

3.2 An Effective Instantiation of Our Framework

The choices of the model set, multi-objective optimisation algorithm and evaluation criteria in our proposed framework can vary according to the prediction tasks and actual preferences. We provide an instantiation of our framework and perform experimental studies with this instantiation in Section 4. Core ingredients of this instantiation are as follows.

Evaluation criteria (measuring between- and in-group unfairness). The mean square error (MSE), individual unfairness and group unfairness [26] are used as three evaluation criteria. In this paper, we view the conflicting but complementary in- and between- group unfairness as individual unfairness (f_I) and group unfairness (f_G), respectively, which were formulated in [26] as:

$$f_I = \sum_{g=1}^{|G|} \frac{n_g}{n} \left(\frac{\mu_g}{\mu} \right)^\alpha \epsilon^\alpha(\boldsymbol{b^g}), \tag{1}$$

$$f_G = \frac{1}{n\alpha(\alpha-1)} \sum_{g=1}^{|G|} n_g \left((\frac{\mu_g}{\mu})^\alpha - 1 \right), \quad (2)$$

where $|G|$ is the number of groups, n_g refers to the size of group g (e.g., male, female), n is the number of individuals (i.e., $n = \sum_g n_g$). The work of [26] defined the notion of *benefit vector* in group g, denoted as $\boldsymbol{b^g}$. For each individual i, its benefit is quantified as $b_i = \theta x_i - y_i + 1$ [12,26], where y_i, θ and x_i denote the true labels, parameters of models and data, respectively. The averaged benefit of a group g and the whole dataset are denoted and calculated as $\mu_g = mean(\boldsymbol{b^g})$ and $\mu = mean(\boldsymbol{b})$ [26], respectively. α is a constant $\notin \{0, 1\}$ and set as 2 in our empirical studies as in [26]. $\epsilon^\alpha(\boldsymbol{b^g})$ is a family of inequality indices from the perspective of economics [5], namely generalised entropy indices, which measures the degree of (un)fairness of the benefits obtained by algorithmic predictors and ground-truth, calculated as $\epsilon^\alpha(\boldsymbol{b^g}) = \epsilon^\alpha(b_1, b_2, \ldots, b_n) = \frac{1}{n\alpha(\alpha-1)} \sum_{i=1}^{n} \left((\frac{b_i}{\mu})^\alpha - 1 \right)$.

Model set. Various machine learning models can be used. In this work, a set of artificial neural nets (ANNs) with an identical architecture is used.

Multi-objective optimiser. The non-dominated sorting genetic algorithm-II (NSGA-II) [6], a well-known Pareto dominance-based MOEA, is used in our instantiation. The weights and bias of each ANN are encoded as a real-value vector. During selection and replacement, NSGA-II uses Pareto dominance mechanisms to select non-dominated solutions based on the evaluation criteria. When reproducing individuals, isotropic Gaussian perturbation $\delta \sim \mathcal{N}(0, \sigma^2)$ is added to weights and bias of each parent [11, 22], where σ indicates mutation strength. During partial training [30], model parameters are updated by Adam [16].

4 Experimental Studies

To answer our research questions proposed in Section 1, we perform several experiments on three benchmark datasets. On each dataset, the instantiation of our framework is used to optimise simultaneously three objectives, the model error (MSE), individual and group unfairness of models (f_I and f_G, respectively). This tri-objective case is referred to as F_{EIG} in our experiments.

To our best knowledge, there is no study that directly applied individual and group unfairness [26] to mitigate model discrimination. Therefore, in order to verify the effectiveness of our framework, we perform the following three ablation studies as baselines for comparison: a bi-objective one considering MSE and f_I (referred to as F_{EI}) during model training; another bi-objective one considering MSE and f_G (F_{EG}); and a single-objective one considering MSE alone (F_E). In each case, 15 independent trials have been performed.

4.1 Experimental Setup

Datasets. Three benchmark datasets, *German* [15], *COMPAS* [18] and *Adult* [17], have been used in our experimental study. The task of *German* is predicting if

a person has an acceptable credit risk. The sensitive attributes are gender and age. Both have two categories. The task of *COMPAS* (Correctional Offender Management Profiling for Alternative Sanctions) [18] is predicting whether an arrested offender will be rearrested within two years counting from taking the test. The sensitive attributes are race and gender, which have six and two categories, respectively. The task of *Adult* is predicting whether a person can get income higher than $50,000 per year or not. The sensitive attributes are race and gender, which have three and two categories, respectively. The pre-processing on each dataset is the same as in [8]. Each dataset is randomly split into 3 partitions, with a ratio of 6:2:2, as training, validation and test sets.

Parameter setting. Our model set is composed of 50 ANN models. In our primary experiments, all models are fully connected with one hidden layer of 64, 128 and 256 nodes for *German*, *COMPAS* and *Adult*, respectively. The weights are initialised as in [10], which is commonly used. The learning rate is set as 0.001 for *German* and 0.01 for both *COMPAS* and *Adult*. The mutation strength is set as 0.01. The μ and λ of NSGA-II are 50. In the tri-objective and two bi-objective cases, the experimental setting are the same. Since F_E considers one single objective, the top λ models considering MSE only are directly selected as the new population for the next generation. The termination condition is set as a maximum number of 500 generations of NSGA-II.

Performance measures. Three popular indicators [20], hyper volume (HV) [33], pure diversity (PD) [25] and spacing [24], are used to evaluate the solution set obtained by NSGA-II. HV is widely used as a common measure to evaluate the overall performance of a solution set in terms of convergence and diversity, while PD and spacing indicators emphasise more the diversity of solutions. In this work, due to the unknown true Pareto front, when calculating HV, all the non-dominated solutions found in all the experimental trials on the same dataset are collected as a pseudo Pareto front. Eq. (7) of the work [32] is used to calculated HV, with Nadir point $\{1.1, \ldots, 1.1\}$ after normalisation. PD is calculated with Eq. (5) of [25]. Spacing is calculated with Eq. (1) on page 136 of [25].

4.2 Experimental Results and Discussions

In this section, our research questions are answered with experimental results.

($\mathcal{Q}1$) ***Whether multi-objective learning can simultaneously optimise the model accuracy and multiple conflicting fairness measures or not?***

We answer $\mathcal{Q}1$ from two perspectives on the *validation set*: (i) visualisation of optimisation process and (ii) convergence curves of HV values.

Fig. 1 illustrates the optimisation process of arbitrarily selected trials of F_{EI}, F_{EG} and F_{EIG} on the validation set, where non-dominated solutions of each generation are drawn with colour darken as the evolution progresses. It's clearly shown that, model error and one or two unfairness measures converge simultaneously towards Pareto fronts (green stars).

In addition, Fig. 2 illustrates the convergence curves of HV values obtained by F_{EI}, F_{EG} and F_{EIG} considering their corresponding objectives in calculating HV values.

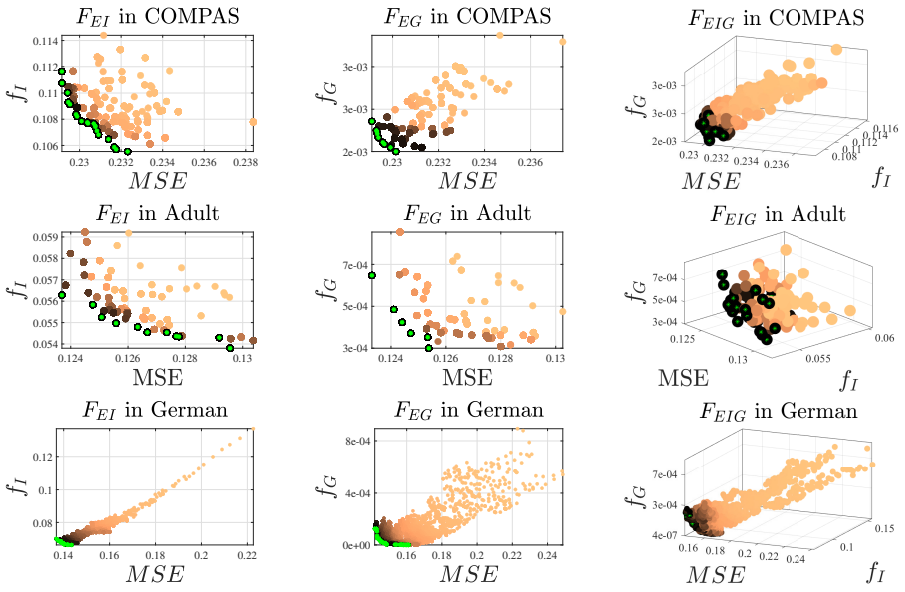

Fig. 1. Illustrative examples: evaluated values on validation set. Different colours indicate solutions at different generations. Green stars highlight the non-dominated solutions in the final generation. Left to right: F_{EI}, F_{EG} and F_{EIG}.

In all the three studies considering two or three objectives, their HV values increase along with evolution, which implies that the model error, individual and group unfairness decrease along with evolution while the diversity significantly increases.

($\mathcal{Q}2$) ***Can multi-objective learning optimise one or several fairness measures without degenerating others?***

$\mathcal{Q}2$ is answered by results on the *test set*: (i) model quality measured by HV values and (ii) model unfairness (comparing the models trained for error and unfairness to the ones trained for error only).

For the first aspect, HV values of F_{EI}, F_{EG} F_{EIG}, and F_E are compared. In order to achieve fair comparison, MSE, f_I and f_G are involved as the three objectives in the calculation of HV for all the four aforementioned cases.

Fig. 3 illustrates how the average HV value changes along with evolution. In F_{EIG} (black curve), the obtained HV value is usually higher than the ones of other cases. Especially on *Adult* dataset, this tri-objective F_{EIG} significantly outperforms the bi-objective and single-objective cases. It is worth mentioning that although only MSE is optimised in F_E, the HV values of its solutions increase with the generation number. This can be easily justified. The calculations of individual and group unfairness [26] involve the true positives, true negatives, false positives and false negatives which imply the model error. According to benefit $b_i = \theta x_i - y_i + 1$, minimising MSE means making $\theta x_i - y_i$

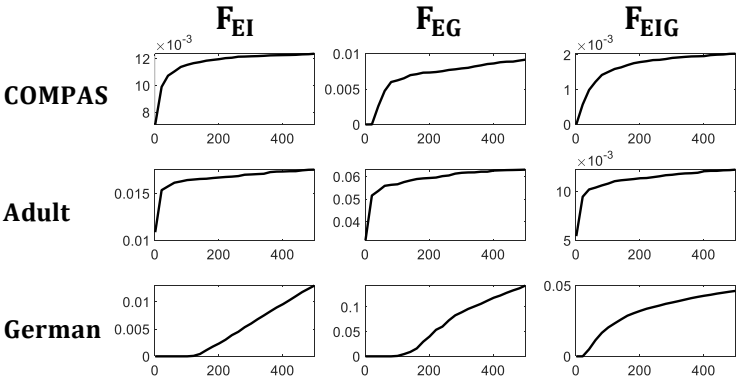

Fig. 2. HV values averaged over 15 trials considering their corresponding objectives on the *validation set*. *x-axis*: generation number; *y-axis*: HV value.

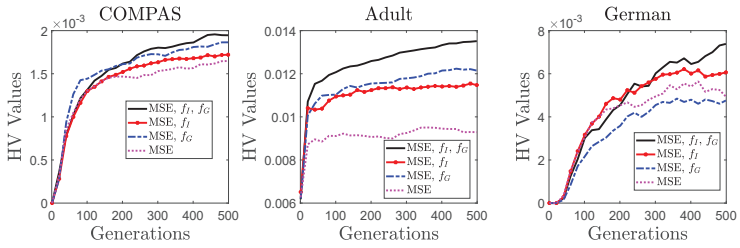

Fig. 3. HV values averaged over 15 trials considering each objectives on *test* set.

closer to 0 and b_i closer to 1. According to Eqs. (1) and (2), f_I is minimal when $b_1 = b_2 = \ldots b_n$ and f_G has the similar trend to f_I. Therefore, optimising MSE alone will decrease the individual and group unfairness but the effect is limited. This was also shown in [26]. However, on *Adult* and *German*, the contribution of solely optimising MSE to HV gradually decreases as shown in Fig. 3.

Table 1 provides the statistical analysis of HV values of the final solution sets illustrated in Fig. 3. Overall, F_{EIG}, considering simultaneously MSE, f_I and f_G, achieves statistically the best HV values in 7 out of 9 cases according to the Wilcoxon rank sum test [13] with 0.05 significance level. From the perspective of multi-objective optimisation, F_{EIG} has superior performance, in terms of convergence and diversity, in comparison with the single- and bi-objective optimisation cases.

For the second aspect, we focus on studying each unfairness measure individually for further analysis. A set of models, providing different tradeoffs among the optimised objectives, can be obtained by our proposed framework. Among the models returned at the end of each run of an F_* ("$*$" indicates the objective(s) considered during training), the models shown to be non-dominated on the test data and "best" on at least one of the metrics (MSE, f_I and f_G) are selected (in other words, extreme points are selected). The MSE, f_I and f_G val-

Table 1. HV values of final solutions averaged over 15 trials. "+/≈/-" indicates that the average HV value of corresponding algorithm (specified by column header) is statistically better/similar/worse than the one of F_{EIG} according to Wilcoxon rank sum test with 0.05 significance level.

Dataset	F_{EI}	F_{EG}	F_E	F_{EIG}
COMPAS	0.00172-	0.00186≈	0.00165-	0.00195
Adult	0.01147-	0.01214-	0.00929-	0.01351
German	0.00606≈	0.00476-	0.00494-	0.00738

Table 2. Comparing models shown to be non-dominated on the test data and best on at least one of the metrics (MSE, f_I and f_G), in terms of the three metric values averaged over 15 runs. "NDM" stands for "non-dominated models". "+/≈/-" indicates that the average objective value (specified by row header) of corresponding algorithm (specified by column header) is statistically better/similar/worse than the one of F_{EIG} according to Wilcoxon rank sum test with 0.05 significance level.

		NDM with best MSE			NDM with best f_I			NDM with best f_G		
		Avg. MSE	Avg. f_I	Avg. f_G	Avg. MSE	Avg. f_I	Avg. f_G	Avg. MSE	Avg. f_I	Avg. f_G
COMPAS	F_{EIG}	0.22268	0.10575	0.00290	0.22585	0.10203	0.00297	0.22351	0.10371	0.00287
	F_{EI}	0.22270≈	0.10567≈	0.00297-	0.22582≈	0.10206≈	0.00299≈	0.22363≈	0.10402≈	0.00291≈
	F_{EG}	0.22246≈	0.10486+	0.00289≈	0.22296+	0.10368-	0.00285+	0.22279+	0.10446-	0.00285≈
	F_E	0.22362-	0.10610≈	0.00299-	0.22415+	0.10368-	0.00292≈	0.22333≈	0.10451-	0.00290≈
Adult	F_{EIG}	0.12565	0.05738	0.00057	0.13163	0.05396	0.00065	0.12926	0.05791	0.00027
	F_{EI}	0.12552≈	0.05810-	0.00066≈	0.13133≈	0.05411≈	0.00064≈	0.12777+	0.05540+	0.00048-
	F_{EG}	0.12546≈	0.05804-	0.00065-	0.12758+	0.05643-	0.00038+	0.12933≈	0.05807≈	0.00027≈
	F_E	0.12556≈	0.05828-	0.00066≈	0.12679+	0.05608-	0.00063≈	0.12705+	0.05709+	0.00040-
German	F_{EIG}	0.16380	0.08071	0.00115	0.16813	0.07707	0.00121	0.16449	0.08224	0.00110
	F_{EI}	0.16375≈	0.07830+	0.00125-	0.16716≈	0.07537+	0.00125≈	0.16468≈	0.07748+	0.00125-
	F_{EG}	0.16314≈	0.08302-	0.00110≈	0.16341+	0.08297-	0.00110≈	0.16476≈	0.08361≈	0.00107≈
	F_E	0.16574≈	0.08367-	0.00120≈	0.16512+	0.08172-	0.00119≈	0.16489≈	0.08294≈	0.00117≈

ues of those selected models are averaged separately over 15 runs and reported in Table 2. Statistical tests assist with the determination of dominance relation [20] when comparing to the aforementioned, selected models obtained by F_{EIG} with the ones obtained by other algorithms. Highlights are summarised as follows.

F_{EIG} **versus** F_{EI}**.** In only one case (German, highlighted in grey in Table 2), F_{EIG}'s "best" model has worse f_I (0.07707) than the one (0.07537) obtained by F_{EI}, while no statistically significant difference has been observed on their MSE or f_G values; in the other cases, F_{EIG}'s "best" models are not dominated by the ones of F_{EI}, and even dominate F_{EI}'s in some cases.

F_{EIG} **versus** F_{EG}**.** Similarly, in only one case (COMPAS, highlighted in grey in Table 2), F_{EIG}'s "best" model has worse f_I (0.10575) than the one (0.10486) of F_{EG}, while no statistically significant difference has been observed on their MSE or f_I values; in the other cases, F_{EIG}'s best models are not dominated by the ones of F_{EG}, and even dominate F_{EG}'s in some cases.

F_{EIG} **versus** F_E**.** "best" models of F_{EIG} are never dominated by F_E's.

To summarise, in 25 out of 27 cases, our framework is able to optimise multiple fairness measures without degenerating others.

Table 3. PD and spacing of final solutions averaged over 15 trials. "+/≈/-" indicates that the average PD or spacing value of corresponding algorithm (specified by column header) is statistically better/similar/worse than the one of F_{EIG} according to Wilcoxon rank sum test with 0.05 significance level. Larger PD values and smaller spacing values imply better performance.

		F_{EI}	F_{EG}	F_E	F_{EIG}
PD	*COMPAS*	77509-	72794-	63791-	94144
	Adult	80454-	72202-	69703-	94345
	German	76695-	51746-	68083-	97476
Spacing	*COMPAS*	0.1614≈	0.1846-	0.2599-	0.1462
	Adult	0.1518≈	0.1716-	0.2434-	0.1308
	German	0.1086-	0.1046-	0.1775-	0.0409

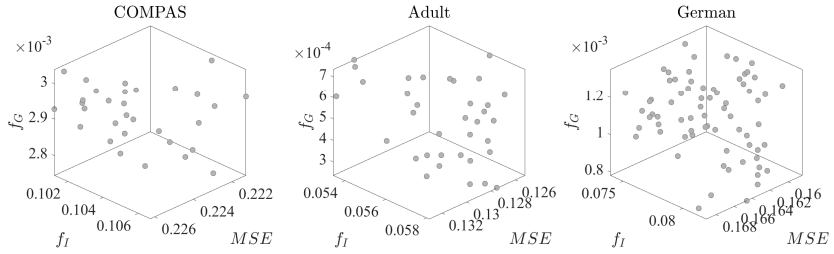

Fig. 4. Pareto fronts obtained by F_{EIG} from all the generations over 15 runs.

($\mathcal{Q}3$) ***Can we obtain a group of diverse models by applying multi-objective learning?***

To verify if multi-objective learning can generate a group of diverse models providing different tradeoffs among the concerned objectives, Fig. 4 illustrates the Pareto fronts generated by F_{EIG}, i.e., our algorithm considering MSE, f_I and f_G simultaneously. All the points in Fig. 4 are non-dominated solutions. Table 3 shows that tri-objective F_{EIG} obtains models with better diversity in 16 out of 18 cases than the other three algorithms according to PD [28] and spacing [24] indicators (cf. Section 4.1).

5 Conclusion

In this paper, we propose a novel multi-objective evolutionary learning framework for mitigating unfairness in machine learning models considering simulta-

neously multiple (un)fairness measures. An instantiation of the proposed framework is evaluated on three well-known benchmark datasets. Experimental results show that our framework is able to optimise one or several fairness measures without degenerating others and find Pareto-front models efficiently to provide human decision-makers with diverse candidate models for their choice.

In the future, we plan to evaluate our framework on a larger number of data sets. We will investigate (1) more appropriate multi-objective optimisation algorithm as the learning algorithm in our framework; (2) more efficient fitness evaluation methods for evaluating learning models; and (3) ensemble strategies for combining multiple models [2].

References

1. Caton, S., Haas, C.: Fairness in machine learning: A survey. arXiv preprint arXiv:2010.04053 (2020)
2. Chandra, A., Yao, X.: Ensemble learning using multi-objective evolutionary algorithms. Journal of Mathematical Modelling and Algorithms **5**(4), 417–445 (2006)
3. Chouldechova, A.: Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big data **5**(2), 153–163 (2017)
4. Corbett-Davies, S., Goel, S.: The measure and mismeasure of fairness: A critical review of fair machine learning. arXiv preprint arXiv:1808.00023 (2018)
5. Cowell, F.A., Kuga, K.: Additivity and the entropy concept: an axiomatic approach to inequality measurement. Journal of Economic Theory **25**(1), 131–143 (1981)
6. Deb, K., Pratap, A., Agarwal, S., Meyarivan, T.: A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE Transactions on Evolutionary Computation **6**(2), 182–197 (2002)
7. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.: Fairness through awareness. In: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference. pp. 214–226 (2012)
8. Friedler, S.A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E.P., Roth, D.: A comparative study of fairness-enhancing interventions in machine learning. In: Proceedings of the Conference on Fairness, Accountability, and Transparency. pp. 329–338 (2019)
9. Fujiyoshi, H., Hirakawa, T., Yamashita, T.: Deep learning-based image recognition for autonomous driving. IATSS research **43**(4), 244–252 (2019)
10. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the thirteenth international conference on artificial intelligence and statistics. pp. 249–256. JMLR Workshop and Conference Proceedings (2010)
11. Gong, Z., Chen, H., Yuan, B., Yao, X.: Multiobjective learning in the model space for time series classification. IEEE Transactions on Cybernetics **49**(3), 918–932 (2018)
12. Heidari, H., Ferrari, C., Gummadi, K.P., Krause, A.: Fairness behind a veil of ignorance: A welfare analysis for automated decision making. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. p. 1273–1283. NIPS'18, Curran Associates Inc., Red Hook, NY, USA (2018)
13. Hollander, M., Wolfe, D.A., Chicken, E.: Nonparametric statistical methods, vol. 751. John Wiley & Sons (2013)

14. Hutchinson, B., Mitchell, M.: 50 years of test (un) fairness: Lessons for machine learning. In: Proceedings of the Conference on Fairness, Accountability, and Transparency. pp. 49–58 (2019)
15. Kamiran, F., Calders, T.: Classifying without discriminating. In: 2nd International Conference on Computer, Control and Communication. pp. 1–6. IEEE (2009)
16. Kingma, D.P., Ba, J.L.: Adam: A method for stochastic gradient descent. In: ICLR: International Conference on Learning Representations. pp. 1–15 (2015)
17. Kohavi, R., Becker, B.: UCI machine learning repository: The adult income data set (1998), `https://archive.ics.uci.edu/ml/datasets/adult`
18. Larson, J., Mattu, S., Kirchner, L., Angwin, J.: Data and analysis for "how we analyzed the compas recidivism algorithm" (2016), `https://github.com/propublica/compas-analysis`
19. Li, B., Li, J., Tang, K., Yao, X.: Many-objective evolutionary algorithms: A survey. ACM Comput. Surv. **48**(1) (Sep 2015)
20. Li, M., Yao, X.: Quality evaluation of solution sets in multiobjective optimisation: A survey. ACM Comput. Surv. **52**(2) (Mar 2019)
21. Liem, C.C., Langer, M., Demetriou, A., Hiemstra, A.M., Wicaksana, A.S., Born, M.P., König, C.J.: Psychology meets machine learning: Interdisciplinary perspectives on algorithmic job candidate screening. In: Explainable and interpretable models in computer vision and machine learning, pp. 197–253. Springer (2018)
22. Minku, L.L., Yao, X.: Software effort estimation as a multiobjective learning problem. ACM Trans. Softw. Eng. Methodol. **22**(4) (Oct 2013)
23. Perrone, V., Donini, M., Zafar, M.B., Schmucker, R., Kenthapadi, K., Archambeau, C.: Fair bayesian optimization. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (2021)
24. Schott, J.R.: Fault tolerant design using single and multicriteria genetic algorithm optimization. Ph.D. thesis, Massachusetts Institute of Technology (1995)
25. Solow, A., Polasky, S., Broadus, J.: On the measurement of biological diversity. Journal of Environmental Economics and Management **24**(1), 60–68 (1993)
26. Speicher, T., Heidari, H., Grgic-Hlaca, N., Gummadi, K.P., Singla, A., Weller, A., Zafar, M.B.: A unified approach to quantifying algorithmic unfairness: Measuring individual &group unfairness via inequality indices. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 2239–2248 (2018)
27. Verma, S., Rubin, J.: Fairness definitions explained. In: 2018 IEEE/ACM International Workshop on Software Fairness (FairWare). pp. 1–7. IEEE (2018)
28. Wang, H., Jin, Y., Yao, X.: Diversity assessment in many-objective optimization. IEEE Transactions on Cybernetics **47**(6), 1510–1522 (2017)
29. Wei, S., Niethammer, M.: The fairness-accuracy Pareto front. arXiv preprint arXiv:2008.10797 (2020)
30. Yao, X.: Evolving artificial neural networks. Proceedings of the IEEE **87**(9), 1423–1447 (1999)
31. Yao, X., Liu, Y.: A new evolutionary system for evolving artificial neural networks. IEEE Transactions on Neural Networks **8**(3), 694–713 (1997)
32. Zhang, Q., Wu, F., Tao, Y., Pei, J., Liu, J., Yao, X.: D-MAENS2: A self-adaptive D-MAENS algorithm with better decision diversity. In: Proceedings of the 2020 IEEE Symposium Series on Computational Intelligence. pp. 2754–2761. IEEE (2020)
33. Zitzler, E., Laumanns, M., Thiele, L.: SPEA2: Improving the strength pareto evolutionary algorithm. TIK-report **103** (2001)
34. Žliobaitė, I.: Measuring discrimination in algorithmic decision making. Data Mining and Knowledge Discovery **31**(4), 1060–1089 (2017)