“Ignorance and Prejudice” in Software Fairness

Jie M. Zhang
University College London
London, UK
jie.zhang@ucl.ac.uk

Mark Harman
University College London
London, UK
mark.harman@ucl.ac.uk

***Abstract*—Machine learning software can be unfair when making human-related decisions, having prejudices over certain groups of people. Existing work primarily focuses on proposing fairness metrics and presenting fairness improvement approaches. It remains unclear how key aspect of any machine learning system, such as feature set and training data, affect fairness. This paper presents results from a comprehensive study that addresses this problem. We find that enlarging the feature set plays a significant role in fairness (with an average effect rate of 38%). Importantly, and contrary to widely-held beliefs that greater fairness often corresponds to lower accuracy, our findings reveal that an enlarged feature set has *both* higher accuracy *and* fairness. Perhaps also surprisingly, we find that a larger training data does *not* help to improve fairness. Our results suggest a larger training data set has more unfairness than a smaller one when feature sets are insufficient; an important cautionary finding for practising software engineers.**

***Index Terms*—software fairness, machine learning fairness**

I. Introduction

> *“Prejudice is the child of ignorance.”*
> *— William Hazlitt, On Prejudice*

Machine learning software has become an inseparable part of our daily lives. It is widely adopted to make decisions, such as to select job applicants, to evaluate employees’ performance, to predict recidivism, to predict credit risks, and to predict medical treatment.

Machine learning tends to learn what human features and data teach it. However, humans may have bias over cognition, further affecting the data collected or labelled and the algorithm designed, leading to unfairness in machine learning software. The unfairness adversely affects the benefit of people in minority groups or historically disadvantageous groups. It may also lead to consequences for software engineering, if software run afoul of laws against discrimination, such as the Civil Rights Act [1].

Recently, fairness in machine learning software has drawn substantial attention in the software engineering community. For example, Brun and Meliou [2] mentioned that “numerous software engineering challenges in the areas of requirements, specification, design, testing, and verification need to be tackled to solve this problem”. Chakraborty et al. [3] said it is the ethical duty of software engineers to strive to reduce software discrimination. Zhang et al. [4] described fairness as a non-functional property for machine learning software that deserves substantial testing effort from software developers. Much progress has been achieved in the direction of software engineering for machine learning [2], [3], such as test generation for detecting fairness violations [5]–[7], training data mutation for locating the unfairness[1] [8], and empirical studies to understand the effectiveness and efficiency of existing fairness improvement methods [3], [9].

This paper presents a large-scale study on the impact of the size of feature set and training data on Machine Learning fairness (ML fairness, which means the fairness of machine learning software). The findings will provide implications for software developers and machine learning practitioners for building fairer machine learning software.

Our study is inspired by two facts. **First**, the size of feature set and training data set are well acknowledged critical practices to optimise ML software [10]. The impact of these factors on model performance (e.g., test accuracy) has been well studied in the literature. Nevertheless, their critical role in performance improvement might be so well known that people may have ignored to study their roles in fairness improvement. **Second**, among the studies of human prejudices in the social psychology domain, it is recognised that knowledge enhancing is an effective way to reduce human prejudice [11], [12], while older adults with more experiences have a tendency to be more prejudiced than their younger counterparts [13]–[16]. Thus, we are curious whether the amount of features (analogous to the “knowledge level” of the model) and training data (analogous to the “experience level” of the model) of machine learning exhibit similar impact on its fairness.

We conduct our study with five widely-explored datasets in the fairness literature, and four widely-studied fairness metrics. We investigate the impact of the feature set size and training data size on ML fairness separately. We also check the coupling effect between these two aspects, as well as how data balance condition and fairness improvement methods affect our findings. We include a discussion of the relationship between human and ML prejudices, and provide practical suggestions to developers for building fairer machine learning software based on our findings.

Our study reveals the following interesting findings: **1)** Feature set size has a notable impact on ML fairness, with an

(This work has been accepted by ICSE 2021.)

[1] In this paper, we use unfairness and bias interactively to refer to the opposite of fairness.

average change rate of 38% across our evaluation subjects. **2)** Perhaps surprisingly, we do not observe that a larger training data has greater fairness. Indeed, we found that in 28% of the cases, a larger set of training data even has *greater* unfairness than a smaller set. **3)** The negative impact of a larger training data is more pernicious when the feature set size is small. **4)** Fairness improvement methods are effective in reducing the negative impact brought by a larger training data set.

Fairness is naturally a domain specific problem, but in this paper we show that there are crosscutting generic fairness drivers in the size of the feature set and the amount of training data. These are two key dimensions for the design of any machine learning software. Therefore, thereby, we can provide general principles to help software engineers to improve the fairness of their systems.

To conclude, this paper makes the following contributions:
1) A systematic *empirical study* on the impact of enlarging feature set and training data set when building fair machine learning software.
2) *Implications* on the impact of feature and training data size for building fairer ML models.

The rest of the paper is organised as follows. Section II introduces the preliminaries. Section III provides the details for our experimental setup. Results and analysis are presented in Section IV. Section V discusses the findings, implications, and actionable conclusions. Section VI introduces the related work. Section VII concludes.

II. Preliminaries

This section provides the preliminaries, including the definitions, terms, and metrics in ML fairness (in Section II-A), as well as the current progress in software engineering for fairness (in Section II-B).

A. Fairness Definitions and Metrics

1) Definitions on Fairness. Machine learning is a widely-adopted statistical method that aids decision making, such as income prediction and medical treatment prediction. During the process of these critical decisions, the characteristics that are sensitive and need to be protected against unfairness are called ***protected attributes*** (also called *protected characteristics* [17] or *sensitive attributes* [4]). Examples of legally recognised protected attributes include race, sex, age, pregnancy, familial status, disability status, and so on. Such protected attributes are not universal, but application specific.

Protected attributes partition a population into different sub-groups: the ***privileged group*** and ***unprivileged groups***, where unprivileged group members are often at systematic disadvantage. For example, when predicting income, *sex* is a protected attribute. The predictive model may favour male groups over female groups, where the male group is the privileged group, the female group is the unprivileged group.

Machine learning fairness is defined in terms of protected attributes and privileged/unprivileged groups. There are several types of machine learning fairness definitions proposed in the literature [18]–[20]. *Fairness Through Unawareness (FTU)* means that an algorithm is deemed to be fair if the sensitive attributes are not explicitly used in the decision-making process [21]. Another type of fairness is *group fairness*. A model has group fairness when privileged groups and unprivileged groups are treated equally (e.g., have an equal probability of decision outcomes or predictive performances [4]). There is also *individual fairness* [22]. A model with individual fairness should give similar predictive results among similar individuals.

In this paper, we focus on **group fairness** due to the following reasons. First, group fairness is more widely adopted and studied in the literature [3], [4], [23], [24]. Second, group fairness has well-defined and acknowledged mathematical fairness metrics that measure fairness quantitatively. Third, group fairness aligns better with legal regulations on fairness [25].

2) Group fairness metrics studied in this paper. This section introduces the most popular fairness metrics for group fairness. Let X be the quantified features of a sample. Let $A \in \{0,1\}$ be a binary protected attribute. For unprivileged group, $A = 0$. C is the predictive outcome. Y is the original label, with $Y = 1$ being the favourable one[2].

Statistical parity difference, also called Demographic Parity difference, is one of the most well-known criteria for fairness [26]. It is the difference between the acceptance rates of the applicants from the privileged and unprivileged groups:

$$P[C=1|A=0]-P[C=1|A=1] \tag{1}$$

Average absolute odds difference is the average of difference in false positive rate and true positive rate for unprivileged and privileged groups:

$$\begin{aligned}&\frac{1}{2}(|P[C=1|A=0,Y=0]-P[C=1|A=1,Y=0]|\\&+|P[C=1|A=0,Y=1]-P[C=1|A=1,Y=1]|)\end{aligned} \tag{2}$$

Equal opportunity difference is the true positive rate difference between unprivileged and privileged groups:

$$P[C=1|A=0,Y=1]-P[C=1|A=1,Y=1] \tag{3}$$

Disparate impact is the ratio of the acceptance rate of the unprivileged group applicants against that of the privileged group applicants:

$$\frac{P[C=1|A=0]}{P[C=1|A=1]} \tag{4}$$

Among these metrics, *disparate impact* suggests the greatest fairness when it equals 1. The remaining metrics suggest the greatest fairness when they equal 0. For ease of presentation and observation, we turn all the values into their absolute values. For *disparate impact*, we normalise it to be between 0 and 1. In this way, for all the fairness metrics, larger metric values indicate more bias.

[2] Favourable label is a label whose value corresponds to an outcome that provides an advantage to the recipient

B. Software Engineering for Fairness

Violations of fairness regulations are regarded as fairness "bugs" in the software engineering community. As early as 2009, Finkelstein et al. [27] used multi-objective search-based methods to aid optimising software fairness in requirement engineering. Brun and Meliou mentioned that software fairness is analogous to software quality, and that "numerous software engineering challenges in the areas of requirements, specification, design, testing, and verification need to be tackled to solve this problem" [2]. Chakraborty et al. [3] claim that software bias detection and mitigation should be included in the software life-cycle. In the recent survey on machine learning testing [4], fairness is classified as a non-functional property that merits significant testing effort from developers.

There have been numerous successful applications of software testing methodology and techniques for fairness improvement. For example, Galhotra et al. [5], [28] proposed Themis, which uses random test generation techniques to evaluate the degree of fairness. Udeshi et al. [6] proposed Aequitas, which first randomly samples the input space to discover the presence of discriminatory inputs, then searches the neighbourhood of these inputs to find more of them. Agarwal et al. [7] used symbolic execution together with local explainability to generate test inputs. Tramer et al. [29] presented a comprehensive testing tool, aiming to help developers test and debug fairness bugs with an easily interpretable bug report.

There are also empirical studies in software engineering seeking to understand software fairness and to get practical implications for developers. Chakraborty et al. [3] studied whether fairness improvement methods damage model prediction performance, as well as the efficiency of fairness improvement methods. Sharma and Wehrheim [8] studied the causes of unfairness via checking whether the algorithm under test is sensitive to training data mutations. Biswas and Rajan [9] conducted an extensive study on the effectiveness and efficiency of existing bias mitigation methods.

In this paper, as in previous work [3], we regard ML fairness as a type of non-functional software property. In the software life cycle, the traditional roles of feature set elicitation and training data extension are well known and reported upon. However, no previous work has studied whether the two factors have a role to play in the construction of fair software.

III. Experimental Setup

A. Research Questions

The evaluation answers the following research questions.

RQ1: How does the feature set size affect ML fairness?

RQ2: How does the training data size affect ML fairness?

To answer the first question, we use the full set of training data, but build different models with gradually-increased-size feature set. To answer the second question, we use the full set of features, but build different models with different-sized random data samples.

Under each research question, we first use visualisation to answer the top-level question. We then design sub-questions to deep dive into the results with further statistical analysis (more details in Section III-E).

RQ3: What is the coupling effect between feature set size and training data size on ML fairness?

Different from the first two research questions, this question investigates the coupling effect of the feature set size and training data size. The purpose is to investigate how these two aspects collectively impact fairness, thereby revealing any important interactions between them. To answer this question, for each of the different-sized datasets, we build models with different-sized feature sets. We use 3D surface plots to visualise the changes of fairness, in which two dimensions are feature set size and data size respectively.

We design further analysis to seek for practical solutions for building software with greater fairness and better accuracy. The analysis is presented in Section V.

B. Datasets

In this paper, we use five datasets as listed in Table I. The first four datasets are the most widely adopted in the literature of machine learning fairness research [20] and software engineering for fairness [3], [5], [24]. The fifth dataset is implemented by the IBM fairness tool AIF360 [23] in their tutorial as a representative fairness dataset.

Table I: Fairness datasets used in this paper

Name	Abbr.	#Features	ProtectedAttributes	Size
Adult Income [30]	adult	14	sex, race	45,222
Bank Marketing [31]	bank	20	age	30,488
COMPAS Score [32]	compas	10	sex, race	6,167
German Credit [33]	german	20	sex, age	1,000
Medical Survey 2015 [34]	meps	41	race	15,830

Below briefly introduces each dataset:

adult: a dataset built to predict whether income exceeds $50K/yr based on census data.

bank: a dataset related with direct marketing campaigns (phone calls) of a Portuguese banking institution. The goal is to predict whether the client will subscribe a term deposit.

german: a data set used to predict people's credit risk levels.

compas: this dataset is used to assess the likelihood that a criminal defendant will re-offend.

meps: this dataset consists of data on the cost and use of health care and health insurance coverage across the United States.

Each dataset has its protected attribute(s), which are determined by its provider depending on specific tasks. For simplicity of exposition, we consider just one protected attribute each time, following previous work [35]. This leads to eight *dataset-attribute* pairs (e.g., *adult-sex*, *adult-race*, *bank-age*). These eight pairs are the evaluation subjects for this paper.

C. Fairness Metrics

Most fairness papers adopt one or two metrics [20], [24], [36]. In this paper, we use all the four fairness metrics introduced in Section II-A2 to measure fairness. The consistency in the observations with different metrics will increase our confidence in getting answers for the research questions.

As introduced in Section II-A2, the metrics have different ranges and fairness optimal values. For ease of observation, we turn all the values into their absolute values. For *disparate impact*, we calculate its distance to one and normalise the distance to be between 0 to 1. In this way, all the fairness metrics are positive. Larger metric values indicate more bias.

D. Fairness Improvement Methods

Fairness improvement methods are used for the deep dive into our conclusions as an exploration for practical suggestions and solutions, if necessary. There are three types of fairness improvement methods in the literature [20]: 1) pre-processing: to transform the training data so that the underlying bias in the data is removed or reduced; 2) in-processing: to modify and change the learning algorithms in order to optimise fairness during the model training process; 3) post-processing: to post-process the prediction results after model training and predictions. This paper studies the impact of feature set size and training data size on building fair machine learning models, which has no intersection with post-processing methods. Thus, we only consider pre-processing and in-processing methods.

In particular, for pre-processing, we use *reweighing* [37]. This technique weights the samples in each group differently to ensure fairness before classification. For in-processing, we use *prejudice remover* [38], which adds a discrimination-aware regularisation term to the learning objective.

In addition, the imbalance of samples in privileged and unprivileged groups are also regarded as a cause for ML unfairness [20]. Thus, we also explore the impact of the training data size with balanced samples.

E. Analysis Approaches

We demonstrate the changes of fairness metric values with different feature/training sets in the following three ways.

First, we **visualise** the changes of the mean metric values as well as the standard deviation (within the 50 runs) across different feature/training set size to answer the top-level research questions. Larger values represent more bias in the models. For RQ1 and RQ2, we use line plots so as to observe the trend of changes. For RQ3, we use 3D surface heat plots.

Second, we use **one-way analysis of variance (ANOVA)** to deep dive into the statistical significance in the fairness differences among the models built with different feature/data sets. We report the *F-statistics* (the ratio of variation between sample means against the variation within the samples), the *p-values* (with a significance level of 0.05), and the *Tukey Honest Significant Differences* (TukeyHSD, it reports the statistical significance in the differences between each two group).

Third, we report the **absolute changes** (in the mean metric values) and **relative changes** (the change ratios) between the minimum and maximum feature set size/data size to further investigate the fairness changes. We only report the changes that have been determined as statistically significant by ANOVA analysis.

F. Experimental Details

By default, we show the results of fairness changes with models built from Decision Trees (which are widely used in industry due to their high efficiency and interpretability) [39]. We use another three widely-used models (i.e., Logistic Regression, Random Forests, AdaBoost. Results are on our homepage [40]) to check whether our conclusions are model-dependent.

For each dataset, we split the data into 80% training data and 20% test data. When investigating the impact of feature set size, we conduct feature sampling on both training data and test data. We start from a minimum feature set of three that contains the protected attribute, to ensure a reasonably effective machine learning model. We then gradually augment features following the default feature order in the AIF360 implementation, so as to obtain different sizes of feature sets[3]. When investigating the impact of training data size, for the training data, we randomly select different-size subsets (with a proportion of 10%, 20%, ..., 100%) to build different models. Due to the instability of machine learning fairness metric values [41], we repeat each prediction process 50 times. This amount of iterations also allows us sufficient data to conduct the one-way ANOVA analysis.

IV. Results

A. RQ1: Impact of Feature Set Size on ML Fairness

The first research question investigates the influence of feature set size on ML fairness. To answer this question, for each dataset, we leave its training data size untouched (i.e., the default 80% of the full dataset), while gradually augment its features one by one.

Figure 1 shows the visualisation of the arithmetic mean fairness values (y-axis) with different feature set size (x-axis). We observe that most of the fairness metrics exhibit notable increases in fairness when feature set size increases.

For the *german-sex* dataset, all the metric values remain unvaried when the number of features increases. This is because most metric values show good fairness (close to 0) even when there are only three features, thus there is not much bias to mitigate.

These observations reveal that a larger feature set exhibits a notable positive influence on the fairness of machine learning models. We next take a deep dive into the results with one-way ANOVA analysis and absolute/relative changes analysis.

1) RQ1.1: What is the statistical significance of the fairness changes among different feature sets? We use one-way ANOVA analysis to answer this question. For a dataset-attribute pair and a fairness metric, each feature set size corresponds to a group with 50 values (coming from the 50 runs). We checked the data and confirmed that they meet the assumptions for one-way ANOVA analysis. For the TukeyHSD results, when there are n groups, there would be all together $\binom{n}{2}$ difference significance results between each two groups.

[3] Section V-A discusses this threat with the default feature order.

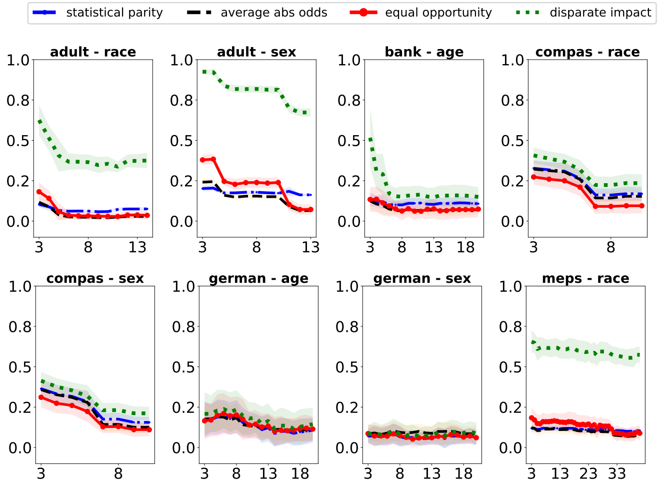

Fig. 1: **RQ1**: Visualisation of the impact of feature set size on fairness. Each line represents the arithmetic mean values for a fairness metric. The shadows represent the standard deviation across multiple runs. We observe that, when the number of features increases, the metric values tend to draw closer to 0. This observation indicates that the size of the feature set has a notable influence on fairness.

For ease of observation, we report the percentage of the results with significant p-values among all the combinations.

Table II shows the results. As shown by table, 27 of the 32 cases (84.4%) exhibit p-values smaller than 0.05. The F-statistic values (the percentage of variation between sample means against the variation within the samples) are often much larger than one (with 87.5% of cases larger than 5). These observations suggest that the fairness changes brought by a larger feature set are significant.

Table II: ***RQ1.1***: One-way ANOVA analysis results for fairness differences among different-sized feature sets. Most differences are significant, with F-statistics values larger than 1, p-values smaller than 0.05, and large percentages of significant TukeyHSD results.

	statistical parity			average abs odds			equal oppo.			disparate impact		
	F	p	HSD	F	p	HSD	F	p	HSD	F	p	HSD
adult-sex	183	***	23%	107	***	57%	108	***	57%	16	***	44%
adult-race	58	***	23%	17	***	43%	8	***	26%	1	>0.05	0%
german-sex	8	***	17%	3	***	3%	7	***	15%	8	***	17%
german-age	0	>0.05	0%	0	>0.05	0%	0	>0.05	0%	0	>0.05	0%
compas-race	316	***	83%	339	***	83%	407	***	83%	266	***	83%
compas-sex	$3e^{29}$	***	100%	$2e^{29}$	***	97%	e^{29}	***	100%	$2e^{29}$	***	100%
bank-age	29	***	31%	25	***	43%	2	**	9%	16	***	52%
meps-race	47	***	23%	28	***	51%	19	***	54%	5	***	37%

F: F-statistics; P: p-value; HSD: the percentage of group differences with significant TukeyHSD p-values.
∗ ∗ ∗: p-value smaller than 0.001; ∗∗: p-value between 0.001 and 0.01; ∗: p-value between 0.01 and 0.05.

2) RQ1.2: What are the absolute and relative fairness changes when changing the feature set size? We calculate the absolute and relative changes in fairness metric values between the smallest and the largest feature sets. Table III shows the results. The fairness for the cells marked with '–' is considered to be unchanged, because the changes are statistically insignificant in the ANOVA analysis. Light grey/dark grey cells are those whose values are increased/decreased respectively.

We observe that among the 27 significantly changed metric values, 25 (92.6%) are increased, 22 (81.5%) have a change rate larger than 20%. For the *german-sex* dataset, there are two decreased metric values, but the decreases are minor, and are likely caused by randomness.

If we treat the changes in the blank cells as zero, the average absolute change in fairness metric values is +0.120; the average change rate is +38.0%. These observations indicate that the difference of fairness between different-sized feature sets are notable.

Table III: ***RQ1.2***: Absolute and relative fairness changes (in brackets) when changing the size of feature sets. Cells marked with '–' denote statistically insignificant changes (according to the ANOVA analysis). Positive/negative values indicate increased/decreased fairness. Most changes (81.5%) have a change rate of over 20%, indicating that adding more features may considerably improve ML fairness.

dataset	statistical parity	average abs odds	equal opportunity	disparate impact
adult-sex	0.039 (20%)	0.176 (73%)	0.307 (81%)	0.252 (27%)
adult-race	0.028 (27%)	0.089 (76%)	0.146 (80%)	–
german-sex	0.006 (9%)	-0.001 (-1%)	0.025 (29%)	-0.003 (-4%)
german-age	–	–	–	–
compas-race	0.158 (48%)	0.169 (52%)	0.179 (65%)	0.172 (42%)
compas-sex	0.208 (57%)	0.233 (65%)	0.201 (64%)	0.202 (49%)
bank-age	0.029 (21%)	0.054 (43%)	0.060 (45%)	0.367 (71%)
meps-race	0.014 (12%)	0.049 (40%)	0.095 (52%)	0.081 (12%)
average	0.120 (38.0%)			

Overall, our observations lead to the following conclusion for the first research question:

> Answer to **RQ1**: Richer feature sets exhibit a notable positive influence on the fairness of machine learning models. The average fairness change rate is +38.0% across our evaluation subjects, when the feature set size increases. This highlights the importance of feature enrichment for building fair ML models.

These observations may be due to the fact that more features bring extra information for the model to make fairer decisions. The strong connection between the protected attribute and the labels is one cause for unfairness [42]. With more features, such connection will be weakened.

During the process of changing feature set, a new feature may contain information that has a strong correlation with the protected attribute [18], thereby allowing extra unfairness to creep into via correlation. From Figure 1, we only observe that for *german-age*, the bias increases a bit when augmenting the fifth and sixth feature, which are the percentage of investment against income, and the length of current residence, respectively. These two features have connections with the protected attribute *age*. However, as we observe, including other features later on eliminates the extra bias brought by these two features.

B. RQ2: Impact of Training Data Size on ML Fairness

To get the answer to RQ2, we use each dataset's full set of features, while adjusting the training data size ratio from 0.1

to 1.0 with a step size of 0.1. The other experimental settings are the same as those for RQ1.

Figure 2 shows the visualisation results. By sharp contrast to the observations for RQ1, we do not observe a unified pattern on the impact of training data size. All but one metric remain almost unvaried over different training data sizes. For *disparate impact* on *adult-sex* and *meps-race*, the metric value increases together with the training data size increases.

These observations suggest that, unlike the richness of feature set, more information in terms of training data is not able to increase ML fairness. On the contrary, with a larger set of training data, the ML fairness may even get worse.

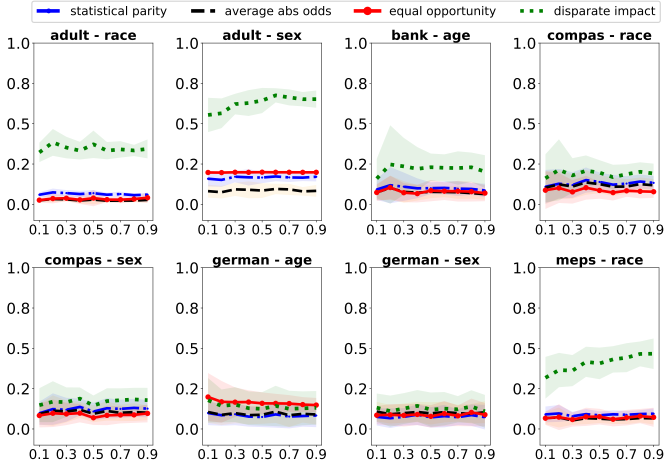

Fig. 2: **RQ2**: Visualisation of the impact of training data size on fairness. Most fairness metrics remain unchanged over different training data size, with several increase with more data. This suggests that a larger set of training data does not have better ML fairness than a smaller one.

1) RQ2.1: What is the statistical significance of the fairness changes among different training data sets? Table IV shows the one-way ANOVA analysis results for the fairness of ML models built with different training data sizes. Unlike the results in Table III, we observe 21 out of the 32 cases (65.6%) exhibit insignificant changes (i.e., with p-values larger than 0.05) for different training data sizes. In addition, the F-statistics values are much smaller than those for feature set size impact. The proportions of significant TukeyHSD values are also smaller. These observations suggest that for most of the time, with the full set of features, the fairness changes brought by a larger set of training data are insignificant.

2) RQ2.2: What are the absolute and relative fairness changes when changing the size of the training data set? The absolute and relative changes of fairness metric values are shown by Table V. We observe that 9 out of the 32 cases (28.1%) have decreased fairness. Together with the 21 cases with insignificant changes, this means that for 93.8% cases, a larger set of training data does not bring significant improvement in fairness or even decreases fairness, compared to a smaller set. Note that this conclusion is obtained on the full set of features. In Section IV-C, we will show more negative impact from a larger set of training data when there are fewer features.

Table IV: ***RQ2.1***: One-way ANOVA analysis for fairness differences among different-sized training data. Most changes are in-significant, with F-statistics values equal to or smaller than 1, p-values larger than 0.05, and small percentages of significant TukeyHSD results.

	statistical			average abs odds			equal opp.			disparate impact		
	F	p	HSD	F	p	HSD	F	p	HSD	F	p	HSD
adult-sex	0	>0.05	0%	0	>0.05	0%	0	>0.05	0%	3	**	9%
adult-race	3	**	9%	1	>0.05	0%	1	>0.05	0%	2	**	3%
german-sex	0	>0.05	0%	0	>0.05	0%	0	>0.05	0%	0	>0.05	0%
german-age	0	>0.05	0%	0	>0.05	0%	0	>0.05	0%	1	>0.05	0%
compas-race	5	***	23%	0	>0.05	0%	3	***	17%	5	***	31%
compas-sex	3	***	17%	0	>0.05	0%	2	**	6%	3	***	17%
bank-age	1	>0.05	0%	2	*	6%	0	>0.05	0%	1	>0.05	3%
meps-race	1	>0.05	0%	1	>0.05	0%	1	>0.05	0%	13	***	40%

F: F-statistics; P: p-value; HSD: the percentage of group differences with significant TukeyHSD p-values. $*\!*\!*$: p-value smaller than 0.001; $*\!*$: p-value between 0.001 and 0.01; $*$: p-value between 0.01 and 0.05.

Table V: ***RQ2.2***: Absolute and relative changes (in brackets) in fairness metrics when changing training data size. With a larger set of training data, there is a decrease of fairness for metric *statistical* and *disparate*.

dataset	statistical parity	average abs odds	equal opportunity	disparate impact
adult-sex	–	–	–	-0.098 (-18%)
adult-race	-0.0 (-1%)	–	–	-0.022 (-7%)
german-sex	–	–	–	–
german-age	–	–	–	–
compas-race	-0.024 (-22%)	–	0.009 (11%)	-0.03 (-18%)
compas-sex	-0.029 (-30%)	–	-0.013 (-16%)	-0.031 (-21%)
bank-age	–	0.01 (14%)	–	–
meps-race	–	–	–	-0.149 (-47%)
average	-0.012 (-4.8%)			

The outcome for RQ2.1 and RQ2.2 is surprising, given the intuition that insufficient data can be a source of unfairness [43]. To further explore the reason, we first check the bias in the original training data.

3) RQ2.3: What is the unfairness in the training data? Can it explain our observations? We use *statistical parity* and *disparate impact* to show data bias. The remaining two metrics require both the original and the predicted label, thus are not applicable for measuring data bias.

Table VI shows the data bias results in the full training data set[4]. Remember that from Figure 2, for *adult-sex* and *meps-race*, their disparate impact bias increases to around 0.65 and 0.5 respectively. These two values are very close to the corresponding disparate impact data bias shown by Table VI. Thus, we suspect that a larger set of training data may allow the model to learn training data bias better. For other cases, the prediction unfairness does not change much, because when the data size is small, its bias is already close to the training data bias.

We now have found evidence that with a larger set of training data there can be more unfairness. However, enlarging training data is also a critical practice to optimise model accuracy. Thus, there is a conflict between improving accuracy

[4] We use random selection to get different-sized training sets, thus, the bias measured by the two metrics in different sets is expected to be similar.

Table VI: ***RQ2.3***: Bias in the training data. Combined with Figure 2, when the training data set is larger, the model prediction bias is approaching the training data unfairness.

dataset	statistical parity	disparate impact	dataset	statistical parity	disparate impact
adult-sex	0.2	0.637	adult-race	0.104	0.394
german-sex	0.057	0.078	german-age	0.114	0.154
compas-race	0.112	0.181	compas-sex	0.136	0.208
bank-age	0.115	0.147	meps-race	0.136	0.499

and fairness through changing the size of training data. In the following, we dig deep into the influence of training data size, so as to investigate possible solutions to avoid or reduce the accuracy-fairness optimisation conflict in the practice of training data extension.

In particular, we further investigate the influence of training data size 1) when the data are balanced for privileged and unprivileged groups; and 2) after applying two popular fairness improvement methods.

4) RQ2.4: What is the impact of changing the size of training data on fairness when the data are balanced? There have been discussions that the imbalance in the data for privileged and unprivileged groups is one cause for the unfairness in ML models [20]. Indeed, intuitively, if there are more data for privileged groups than unprivileged groups, the model may have better performance when predicting the results for privileged groups, leading to larger differences in the performance metrics and more bias in terms of the widely-studied fairness metrics we explore. When we change the training data size gradually, more data will be added for privileged groups than for unprivileged groups, which may lead to a negative impact of training data size on ML fairness.

To investigate whether this intuition is true for our experiments, we first check the data balance condition for privileged groups and unprivileged groups for each dataset. The results are shown by Table VII. We find that five out of eight dataset-attribute pairs have more data in the privileged group than in the unprivileged group. However, there are three pairs with the opposite circumstances: *compas-race*, *compas-sex*, and *meps-race*. These three still suffer from unfairness on the ML models built on them, and negative impact from more training data (according to Table V). This may indicate that the imbalance in data is not the primary reason for the negative impact of training data size on ML fairness.

Table VII: ***RQ2.4***: Data size in privileged groups and unprivileged groups for each dataset. For *compas-race*, *compas-sex*, and *meps-race*, there are more data in the unprivileged group than in the privileged group.

dataset	total size	size for privileged group	size for unprivileged group
adult-sex	45,222	30,527 (67.5%)	14,695 (32.5%)
adult-race	45,222	38,903 (86.0%)	6,319 (14.0%)
german-sex	1,000	690 (69.0%)	310 (31.0%)
german-age	1,000	810 (81.0%)	190 (19.0%)
compas-race	6,167	2,100 (34.1%)	4,067 (65.9%)
compas-sex	6,167	1,173 (19.0%)	4,994 (81.0%)
bank-age	30,488	29,624 (97.2%)	864 (2.8%)
meps-race	15,830	5,656 (35.7%)	10,174 (64.3%)

We further conduct experiments on the five dataset-attribute pairs which have more privileged data than unprivileged data. When sampling training data, we ensure that privileged and unprivileged groups have equal sizes, then calculate the fairness metric values on the balanced data.

Table VII shows the results of absolute and relative changes. For ease of comparison, we also show the results of the original imbalanced data on the five dataset-attribute pairs (the bottom sub-table). Overall, when the data is balanced, the size of training set has more positive impact on fairness. Nevertheless, the improvement is rather limited with a larger data set: improved from -0.006 to +0.002 on average.

Table VIII: ***RQ2.4***: Influence of training set size with balanced/imbalanced data (top/bottom table). The average effect of a richer balanced training data on fairness is slightly improved (i.e., from -0.006 to 0.002 on average).

Influence of Training Data Size (Balanced Data)

dataset	statistical parity	average abs odds	equal opportunity	disparate impact
adult-sex	–	–	-0.012 (-12%)	–
adult-race	-0.015 (-24%)	–	0.003 (5%)	–
german-sex	–	–	0.025 (27%)	0.054 (34%)
german-age	–	–	–	–
bank-age	–	0.025 (28%)	0.006 (8%)	-0.049 (-20%)
average	0.002 (2.3%)			

Influence of Training Data Size (Imbalanced Data)

dataset	statistical parity	average abs odds	equal opportunity	disparate impact
adult-sex	–	–	–	-0.098 (-18%)
adult-race	-0.0 (-1%)	–	–	-0.022 (-7%)
german-sex	–	–	–	–
german-age	–	–	–	–
bank-age	–	0.01 (14%)	–	–
average	-0.006 (-0.6%)			

5) RQ2.5: What is the impact of training data size with fairness improvement methods applied? We also investigate whether fairness improvement methods affect our observations on the impact of training data size. In particular, we investigate reweighing for pre-processing and prejudice remover for in-processing methods (see more details in Section III-D).

Table IX shows the results. We observe that after applying the fairness improvement methods, there are more light grey cells and fewer dark grey cells when the training data set is larger. In particular, without fairness improvement methods, there are 2/9 light/dark grey cells (in Table V). With reweighing, there are 12/7 light/dark grey cells; with prejudice remover, there are 14/5 light/dark grey cells. These changes indicate that with fairness improvement methods, there is a higher probability that a larger training data would bring greater fairness.

The observations with reweighing, which is a pre-processing method, confirm our previous conjecture that the negative impact of sampling more training data might be caused by the bias in the original training data. When the reweighing reduces the bias in the original training data, the negative impact from sampling more training data is also reduced.

Overall, for the second research question, our observations lead to the following conclusion:

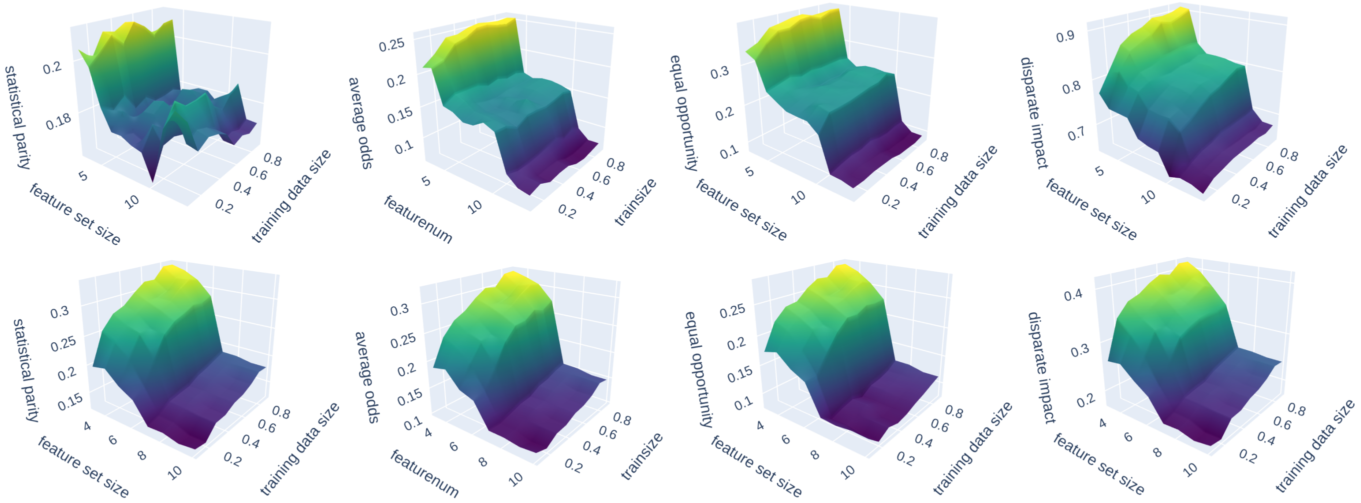

Fig. 3: RQ3: 3D surface plot showing the collective impact of feature set size and training data size. The first row is for the *adult-sex* dataset, the second row is for *compas-race*. We observe that for all the metrics, the surfaces rise more steeply when the feature set size is small. This indicates that when there are fewer features, the negative impact of larger training data on fairness is more significant.

Table IX: ***RQ2.5***: Comparison on the influence of training data size with fairness improvement methods.

Influence of Training Data Size With Reweighing				
dataset	statistical parity	average abs odds	equal opportunity	disparate impact
adult-sex	–	-0.023 (-28%)	-0.054 (-56%)	–
adult-race	-0.005 (-6%)	0.003 (6%)	-0.007 (-13%)	-0.076 (-25%)
german-sex	–	0.011 (9%)	0.015 (17%)	–
german-age	0.049 (41%)	–	0.08 (50%)	0.066 (40%)
compas-race	–	–	–	–
compas-sex	–	–	–	–
bank-age	0.051 (47%)	0.016 (26%)	-0.023 (-57%)	0.116 (44%)
meps-race	0.007 (8%)	0.017 (28%)	0.028 (42%)	-0.026 (-6%)
average	0.008 (5.2%)			
Influence of Training Data Size With Prejudice Remover				
task	statistical parity	average abs odds	equal opportunity	disparate impact
adult-sex	–	-0.015 (-9%)	-0.029 (-10%)	–
adult-race	0.012 (46%)	0.012 (17%)	0.041 (47%)	0.06 (47%)
german-sex	–	0.045 (35%)	0.07 (57%)	–
german-age	0.02 (19%)	–	0.022 (21%)	0.043 (29%)
compas-race	–	–	–	–
compas-sex	–	–	–	–
bank-age	0.032 (47%)	-0.023 (-23%)	-0.099 (-76%)	0.219 (46%)
meps-race	0.026 (50%)	0.006 (14%)	-0.006 (-13%)	0.065 (19%)
average	0.016 (11.4%)			

> Answer to **RQ2**: Perhaps surprisingly, a larger training data does not exhibit more fairness. When training data size increases, ML fairness decreases in 28% of the cases, and does not exhibit significant changes in 66% of the cases. The overall fairness change rate is -4.8%. However, fairness improvement methods can turn this change rate into 11.4%.

C. RQ3: Coupling Effect of Feature Set Size and Training Data Size on ML Fairness

This research question is designed to investigate whether there are any notable interactions between the size of feature set and training set in their collective impact on fairness. For each feature set, we investigate different training data sizes and record the fairness metric values. We then draw 3D surface plots to visualise the changes of fairness with different training data size and number of features.

Figure 3 shows the results. For brevity, we only show the results for the *adult-sex* (top row) and *compas-race* (bottom row) datasets. Each sub-figure is for one fairness metric. Different colours represent different values, with lighter colour representing larger unfairness.

Our first observation is that in each sub-graph, the largest unfairness appears at a position with the smallest feature set and the largest training set. The smallest unfairness appears at a position with the largest feature set and the smallest training data in all but one case (for the first sub-figure, the value changes are very variable). This is consistent with our previous conclusions that greater feature set size brings greater fairness, while a larger training data may have a negative affect on fairness.

Interestingly, when the feature set size is small, the surfaces rise more steeply when going from a small to a large training set. This indicates that, when there are fewer features, larger training data introduces more unfairness. However, adding more training data is a common practice to improve model accuracy. This observation highlights the importance of feature sufficiency, to reduce the negative impact brought by a larger set of training data.

> Answer to **RQ3**: When there are fewer features, the unfairness increases faster with a larger training set.

V. Discussion

In this section, we discuss the threats to validity and the trade-off between fairness and accuracy. We also derive implications and actionable conclusions from our findings for practising software engineers. Finally, we discuss the relationship between human prejudices and ML bias.

A. Threats to Validity

The primary threat to internal validity lies in the implementation of the study. To reduce this threat, the authors independently reviewed the experimental scripts to check their correctness. We also used IBM AIF360 [23], a widely adopted fairness tool in software fairness research [9], [24], to obtain the fairness and accuracy of a model and the results of bias mitigation methods.

The threats to external validity lie primarily with the subjects. We use five datasets that are widely adopted in the literature of fairness research. We use four widely-adopted fairness metrics to improve the generalisation of our conclusions.

We took several steps to address the threats to construct validity. First, we use different machine learning models (e.g., Decision Trees, Logistic Regression, Random Forests, AdaBoost) to examine whether the chosen machine learning model is a factor that would affect our conclusions. Second, for the default Decision Trees model, we use different complexity configuration (i.e., different fixed maximum depths and also grid search for each feature set and training data) to check whether complexity is a factor that would affect our conclusions. Third, we try different orders when changing the size of the feature set when answering RQ1.

We avoid adding features/data points not found in the original datasets to reduce the threat brought by unreliable data as well as to better control variables. Instead, we construct different-sized feature sets and training data sets via feature and data sampling. This is of course unrealistic in practice, where developers often have reliable resources to extend their feature set and training data. In future, we plan to explore the impact of extending feature/data sets in realistic scenarios.

This paper provides results with the default configuration with one model and one fixed complexity configuration. The full results for other configurations, together with our code, are available at our homepage [40]. All results demonstrate that the default configuration is not a threat to our conclusions.

B. Relationship Between Fairness and Accuracy

Seeking better fairness usually comes at the price of affecting the accuracy, as also reported by many previous studies [38], [44]–[47]. Figure 4 shows the test accuracy for the *adult-sex* dataset with different feature set sizes and data sizes. It is important for software engineering to be able to find a solution approach that improves both fairness and accuracy. The most immediately actionable finding of this paper is that there *does* exist such a sweet spot: to have a richer feature set.

This finding is a positive counterpoint to previous tradeoff theory between accuracy and fairness. Our observations implicate that "You CAN have your cake and eat it" – through changing the feature set.

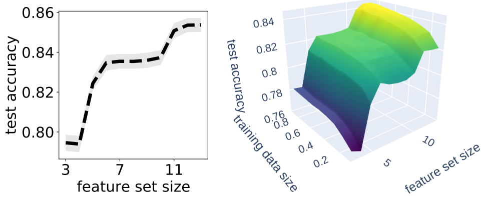

Fig. 4: Test accuracy for different feature set sizes and data sizes. Contrary to widely-held beliefs that accuracy has to be traded for fairness, our findings indicate that there is a way to improve both fairness and accuracy: to get richer features.

We also observe that there is a conflict between accuracy and fairness improvement with a larger training data. We analyse and provide practical solutions to tackle this conflict next.

C. Implications for Software Practitioners and Researchers

1) Implications on feature engineering. It is well known that the role played by the features in the success of learning algorithms is crucial: with poor features, uncorrelated with the target labels, learning could become challenging or even impossible [10]. Our findings suggest that a larger feature set not only helps improve model accuracy, but also helps to substantially improve ML fairness.

The choice of features reflect the software engineer's prior knowledge about the learning task. Our finding highlights the importance of feature engineering for building fairer software.

2) Implications on training data extension. Enriching training data is also a well acknowledged critical practice to optimise model accuracy [10]. However, from RQ2, RQ3, and Figure 4, a larger set of training data may make the model learn greater degree of bias, especially when the feature set size is small. This makes the practice of enriching training data challenging, and developers may often face a choice between accuracy and fairness.

Based on our findings, developers can have three options to alleviate the negative impact from large training data (with a fixed feature set). They can either: 1) reduce the training data size; 2) ensure the training data is balanced if the unprivileged group is a minority; or 3) apply bias mitigation methods. Our previous findings have demonstrated that each of these three options works in reducing bias, with the third one being the most effective.

We then compare the accuracy loss of each option. As shown by Table X, applying fairness improvement methods is the overall best solution that preserves the most accuracy (except for prejudice remover on the compas dataset). This indicates that applying fairness improvement methods is effective to reduce the negative impact from data extension without sacrificing too much accuracy.

For researchers, our findings imply that it is important to compare fairness improvement method effectiveness on a level playing field, which gives all techniques the same amount of training data and the same feature set for a dataset.

Table X: Accuracy and fairness changes for practical options

dataset	original data	smaller data	balanced	reweigh	prejudice remover
adult-sex	0.83	0.77	0.82	0.82	0.82
adult-race	0.84	0.80	0.79	0.82	0.83
german-sex	0.57	0.51	0.55	0.68	0.58
german-age	0.56	0.51	0.52	0.65	0.57
compas-race	0.63	0.56	–	0.65	0.34
compas-sex	0.62	0.55	–	0.66	0.34
bank-age	0.87	0.77	0.71	0.89	0.90
meps-RACE	0.77	0.69	–	0.79	0.80

D. Connection between Human and ML Prejudices

Human learning and machine learning share many similarities. This makes it highly tempting to speculate on simulation between ML fairness and human bias/prejudice. This may yield insights for both domains of study.

Humans often learn by examples and experiences. They process the information (feature engineering), then find patterns in the information (build a learning model by connecting the features to the labels) to aid decision making [48].

William Hazlitt once said: “Prejudice is the child of ignorance”. In human learning, the term *ignorance* usually refers to a lack of knowledge or information. The term *prejudice* means a prejudgement based on inadequate knowledge. In the domain of social psychology, knowledge enhancement has long been regarded as an approach to prejudice reduction [11], [12].

Interestingly, our findings about ML fairness accord well with the existing literature on human prejudice. In particular, if we regard feature sufficiency as *knowledge* of ML, our findings highlight that, for ML models, prejudgement based on inadequate features leads to greater prejudice.

Moreover, if we regard the size of training data as analogous to human *experience*, our findings reveal that when knowledge is inadequate, more experience could not reduce prejudices. Instead, the more inadequate the knowledge is, the more harmful additional experience is. Previous findings have found that older adults have a tendency to be more prejudiced than their younger counterparts [13]–[16]. However, age differences in prejudice are well documented but poorly understood [15]. Our findings on ML fairness reveal that when the experience (training data size) is richer, the prejudice for “knowledgeable data” (with many features) remain stable or slightly increase; for “unknowledgeable data” (with a few features), the prejudice increases much faster. This may make it interesting to design social psychology experiments, to compare human prejudice changes over time between knowledgeable groups and less knowledgeable groups.

Of course, there are many differences between human learning and machine learning. It might be possible that the connections we find between human prejudices and ML prejudices are coincidental. Nevertheless, we believe these feature/knowledge adequacy relationships, may shed further light on possible solutions to tackling both ML and human bias. We raise the connection here to motivate future research on this potential for successful cross-fertilisation.

VI. Related Work

Fairness has been studied in the software engineering literature since 2009 [27]. Research on fairness focuses on measuring, discovering, understanding, and coping with unfairness. This section introduces the work that is most related to ours.

A. Software Engineering for Fairness

We introduced the compelling visions on software fairness [2]–[4] in Section II. Here we discuss the progress that has been made in SE for fairness.

Galhotra et al. [5], [28] proposed Themis, which uses random test generation techniques to evaluate the degree of discrimination (based on fairness scores). Udeshi et al. [6] proposed Aequitas, focusing on test generation to uncover discriminatory inputs and those inputs essential to understand individual fairness. The generation approach first randomly samples the input space to discover the presence of discriminatory inputs, then searches the neighbourhood of these inputs to find more such inputs. Agarwal et al. [7] used symbolic execution (together with local explainability) to generate test inputs. The key idea is to use the local explanation, specifically Local Interpretable Model-agnostic Explanations to identify whether factors that drive decisions include protected attributes. Sun et al. [49] proposed to combine input mutation and metamorphic relations to automatically testing and improving the fairness of machine translations.

Tramer et al. [29] proposed a comprehensive fairness testing tool, aiming to help developers test and debug fairness bugs with an ‘easily interpretable’ bug report. The tool is available for various application areas including image classification, income prediction, and health care prediction.

Sharma and Wehrheim [8] used data mutation to locate fairness bugs by checking whether the algorithm under test is sensitive to the mutants. They mutated the training data in various ways to generate new datasets, such as changing the order of rows, columns, and shuffling feature names and values. 12 out of 14 classifiers were found to be sensitive to these changes.

B. Empirical Studies on Software Fairness

Chakraborty et al. [3] empirically studied the effectiveness and efficiency of existing fairness improvement methods. They further studied the impact of model complexity parameters on ML fairness [24]. Biswas and Rajan conducted a large-scale study on the effectiveness and efficiency of existing bias mitigation methods [9]. Kearns et al. [50] studied the effectiveness and fairness-accuracy tradeoffs of rich subgroup fairness with four datasets.

There has also been work exploring fairness with human studies. Dodge et al. [51] conducted a human study to explore how different styles of explanation impact people’s fairness judgement of machine learning systems. Harrison et al. [52] performed a survey on 502 Mechanical Turk workers that investigated their attitudes to difficult choices when faced with fairness-related trade-offs. Wang et al. [53] used human annotated data to analyse the similarity between different examples for COMPAS dataset, to facilitate individual fairness. Grgic-Hlaca et al. [54] studied humans' attitudes towards whether different attributes should be regarded as protected attributes.

As far as we know, there are no theoretical or empirical studies of the impact of feature set size and training data size on ML fairness.

VII. Conclusion

This paper presented a large study on the influence of feature set size and training data size on the fairness of machine learning software. We found that a larger feature set has 38% more fairness than a smaller one on average. In addition, a larger feature set can substantially slow down the increase of unfairness brought by larger training data. Based on our conclusion, we provided practical implications for software engineers and researchers. We also discussed the potential connection revealed by our findings between ML bias and human bias. Our findings suggest a potential for cross-fertilisation between social psychology and software engineering.

Data Availability

The five data sets that support the findings of this study were accessed following the instructions of AIF360[5]. The code and extra results for this paper are made public at figshare with identifier 10.6084/m9.figshare.12887249.

Acknowledgement

The authors are supported by the ERC advanced grant with No. 741278.

References

[1] Ruth G Blumrosen. Wage discrimination, job segregation, and the title vii of the civil rights act of 1964. *U. Mich. JL Reform*, 12:397, 1978.

[2] Yuriy Brun and Alexandra Meliou. Software fairness. In *Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pages 754–759, 2018.

[3] Joymallya Chakraborty, Tianpei Xia, Fahmid M Fahid, and Tim Menzies. Software engineering for fairness: A case study with hyperparameter optimization. *arXiv preprint arXiv:1905.05786*, 2019.

[4] J. M. Zhang, M. Harman, L. Ma, and Y. Liu. Machine learning testing: Survey, landscapes and horizons. *IEEE Transactions on Software Engineering*, pages 1–1, 2020.

[5] Sainyam Galhotra, Yuriy Brun, and Alexandra Meliou. Fairness testing: testing software for discrimination. In *Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering*, pages 498–510, 2017.

[6] Sakshi Udeshi, Pryanshu Arora, and Sudipta Chattopadhyay. Automated directed fairness testing. In *Proceedings of the 33rd ACM/IEEE International Conference on Automated Software Engineering*, pages 98–108. ACM, 2018.

[7] Aniya Aggarwal, Pranay Lohia, Seema Nagar, Kuntal Dey, and Diptikalyan Saha. Black box fairness testing of machine learning models. In *Proceedings of the 2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pages 625–635, 2019.

[8] Arnab Sharma and Heike Wehrheim. Testing machine learning algorithms for balanced data usage. In *Proc. ICST*, pages 125–135, 2019.

[9] Sumon Biswas and Hridesh Rajan. Do the machine learning models on a crowd sourced platform exhibit bias? an empirical study on model fairness. *FSE 2020 (to appear)*, 2020.

[10] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

[11] Yehuda Amir. Contact hypothesis in ethnic relations. *Psychological bulletin*, 71(5):319, 1969.

[12] Gordon Willard Allport, Kenneth Clark, and Thomas Pettigrew. The nature of prejudice. 1954.

[13] Glenn Firebaugh and Kenneth E Davis. Trends in antiblack prejudice, 1972-1984: Region and cohort effects. *American Journal of Sociology*, 94(2):251–272, 1988.

[14] Karen Gonsalkorale, Jeffrey W Sherman, and Karl Christoph Klauer. Aging and prejudice: Diminished regulation of automatic race bias among older adults. *Journal of Experimental Social Psychology*, 45(2):410–414, 2009.

[15] Brandon D Stewart, William von Hippel, and Gabriel A Radvansky. Age, race, and implicit prejudice: Using process dissociation to separate the underlying components. *Psychological Science*, 20(2):164–168, 2009.

[16] Cristina G Wilson, Amy T Nusbaum, Paul Whitney, and John M Hinson. Age-differences in cognitive flexibility when overcoming a preexisting bias through feedback. *Journal of clinical and experimental neuropsychology*, 40(6):586–594, 2018.

[17] Sam Corbett-Davies and Sharad Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*, 2018.

[18] Pratik Gajane and Mykola Pechenizkiy. On formalizing fairness in prediction with machine learning. *arXiv preprint arXiv:1710.03184*, 2017.

[19] Sahil Verma and Julia Rubin. Fairness definitions explained. In *International Workshop on Software Fairness*, 2018.

[20] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*, 2019.

[21] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pages 4066–4076, 2017.

[22] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226. ACM, 2012.

[23] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias, October 2018.

[24] Joymallya Chakraborty, Suvodeep Majumder, Zhe Yu, and Tim Menzies. Fairway: A way to build fair ml software. In *FSE (to appear)*, 2019.

[25] Solon Barocas and Andrew D Selbst. Big data's disparate impact. *Calif. L. Rev.*, 104:671, 2016.

[26] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. *arXiv preprint arXiv:1507.05259*, 2015.

[27] Anthony Finkelstein, Mark Harman, S Afshin Mansouri, Jian Ren, and Yuanyuan Zhang. A search based approach to fairness analysis in requirement assignments to aid negotiation, mediation and decision making. *Requirements engineering*, 14(4):231–245, 2009.

[28] Rico Angell, Brittany Johnson, Yuriy Brun, and Alexandra Meliou. Themis: Automatically testing software for discrimination. In *Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pages 871–875. ACM, 2018.

[29] Florian Tramer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, Jean-Pierre Hubaux, Mathias Humbert, Ari Juels, and Huang Lin. Fairtest: Discovering unwarranted associations in data-driven applications. In *2017 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 401–416. IEEE, 2017.

[30] Ronny Kohavi and Barry Becker. UCI machine learning repository, 2017.

[31] Sérgio Moro, Paulo Cortez, and Paulo Rita. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62:22–31, 2014.

[32] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias: There's software used across the country to predict future criminals. *And it's biased against blacks. ProPublica*, 23, 2016.

[5] https://github.com/Trusted-AI/AIF360/tree/master/aif360/data

[33] Hans Hofmann. UCI machine learning repository, 1994.

[34] Medical Expenditure Panel Survey. https://meps.ahrq.gov/mepsweb/. Accessed: 2020-07-06.

[35] Till Speicher, Hoda Heidari, Nina Grgic-Hlaca, Krishna P Gummadi, Adish Singla, Adrian Weller, and Muhammad Bilal Zafar. 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*, pages 2239–2248, 2018.

[36] Michael Wick, Jean-Baptiste Tristan, et al. Unlocking fairness: a trade-off revisited. In *Advances in Neural Information Processing Systems*, pages 8783–8792, 2019.

[37] Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33, 2012.

[38] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 35–50. Springer, 2012.

[39] Zachary C Lipton. The mythos of model interpretability. *Queue*, 16(3):31–57, 2018.

[40] Homepage. = "https://doi.org/10.6084/m9.figshare.12887249.v1".

[41] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, A Mojsilović, et al. Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, 63(4/5):4–1, 2019.

[42] Toshihiro Kamishima, Shotaro Akaho, and Jun Sakuma. Fairness-aware learning through regularization approach. In *2011 IEEE 11th International Conference on Data Mining Workshops*, pages 643–650. IEEE, 2011.

[43] Nicol Turner Lee, Paul Resnick, and Genie Barton. Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms. *Center for Technology Innovation, Brookings. Tillgänglig online: https://www. brookings. edu/research/algorithmic-bias-detection-and-mitigation-bestpractices-and-policies-to-reduce-consumer-harms/# footnote-7 (2019-10-01)*, 2019.

[44] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops*, pages 13–18. IEEE, 2009.

[45] L Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K Vishnoi. Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 319–328, 2019.

[46] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 259–268, 2015.

[47] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On fairness and calibration. In *Advances in Neural Information Processing Systems*, pages 5680–5689, 2017.

[48] Knud Illeris. A comprehensive understanding of human learning. *Contemporary theories of learning: Learning theorists... in their own words*, pages 7–20, 2009.

[49] Zeyu Sun, Jie M Zhang, Mark Harman, Mike Papadakis, and Lu Zhang. Automatic testing and improvement of machine translation. In *Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering*, pages 974–985, 2020.

[50] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. An empirical study of rich subgroup fairness for machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 100–109, 2019.

[51] Jonathan Dodge, Q Vera Liao, Yunfeng Zhang, Rachel KE Bellamy, and Casey Dugan. Explaining models: an empirical study of how explanations impact fairness judgment. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pages 275–285, 2019.

[52] Galen Harrison, Julia Hanson, Christine Jacinto, Julio Ramirez, and Blase Ur. An empirical study on the perceived fairness of realistic, imperfect machine learning models. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 392–402, 2020.

[53] Hanchen Wang, Nina Grgic-Hlaca, Preethi Lahoti, Krishna P Gummadi, and Adrian Weller. An empirical study on learning fairness metrics for compas data with human supervision. *arXiv preprint arXiv:1910.10255*, 2019.

[54] Nina Grgic-Hlaca, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian Weller. The case for process fairness in learning: Feature selection for fair decision making. In *NIPS Symposium on Machine Learning and the Law*, volume 1, page 2, 2016.