Data vs. Model Machine Learning Fairness Testing: An Empirical Study

Author(s)
Affiliation(s)
author(s)@example.org

Abstract

Preference has primarily been given to testing for robustness and correctness of ML systems while other non-functional properties such as fairness have been ignored. Although several fairness definitions and bias mitigation techniques exist in the literature, all existing solutions evaluate fairness after the training stage. This paper presents an empirical analysis of the relationship between model dependent and independent fairness metrics using 2 fairness metrics, 4 ML algorithms, 5 real-world datasets and 1600 training and fairness evaluation cycles. We find a linear relationship between data and model fairness metrics when the distribution and the size of the training data changes. Our results indicate that testing for fairness prior to training can be a "cheap" and effective means of catching a biased data collection process early; detecting data drifts in production systems and minimising execution of full training cycles thus reducing development time and costs.

1 Introduction

While several contributions toward testing ML systems have been made in recent years, preference has primarily been given to robustness and correctness while other non-functional properties such as security, privacy, efficiency, interpretability and fairness have been ignored [Zhang *et al.*, 2020; Zhang and Harman, 2021; Mehrabi *et al.*, 2021; Wan *et al.*, 2021]. Testing for fairness in ML systems however, is a multi-faceted problem and involves both technological and social factors. Although an abundance of definitions for fairness and consequently techniques to mitigate said bias exists in the scientific literature, all existing solutions evaluate fairness after the training stage, using the predictions of the ML model.

In addition to the underlying codebase of ML software systems, both the training data and the ML algorithms constantly evolve and change throughout the ML development lifecycle [Sculley *et al.*, 2015; Bosch *et al.*, 2021; Hutchinson *et al.*, 2021]. In contrast to prior work, we take a more holistic approach by testing for fairness at two distinct locations of the ML development lifecycle. First, prior to model training using fairness metrics that can quantify the bias in the training data (henceforth Data Fairness Metric or DFM). And second, after model training using fairness metrics that quantify the bias in the predictions of the trained model (henceforth Model Fairness Metric or MFM).

This paper presents an empirical study of the relationship between DFM and MFM. The analysis is conducted using 2 fairness metrics, 4 ML algorithms, 5 real-world tabular datasets and 1600 fairness evaluation cycles. All source code and results of the study are publicly accessible under the CC-BY 4.0 license[1]. The research questions along with the contributions of this paper are as follows.

- **RQ1.** What is the relationship between DFM and MFM as the fairness properties of the underlying training dataset changes?

  DFM and MFM convey the same information when the distribution of the underlying training dataset changes. This implies that DFM can be used as early warning systems to catch data drifts in production ML systems that may affect its fairness.

- **RQ2.** How does the training sample size affect the relationship between DFM and MFM?

  Our analysis of the training sample size and how it influences the relationship between DFM and MFM reveals the presence of a trade-off between fairness, efficiency and performance. In Section 5.1 we provide some practical guidelines on how to best navigate this trade-off.

- **RQ3.** What is the relationship between DFM and MFM across various training and feature sample sizes?

  DFM and MFM convey the same information when the training sample size changes. This implies that DFM can help practitioners catch fairness issues upstream and avoid execution costs of a full training cycle.

The remainder of the paper is structured as follows. Section 2 summarises related concepts and prior work done in the field of ML fairness testing. In Section 3, the experimental design and fairness evaluation strategy is presented. The results of this study are presented in Section 4. The implications of the results along with their threats to validity are discussed in Section 5.

[1] https://figshare.com/s/67206f7c219b12885a6f

2 Preliminaries

This section provides a summary of the relevant concepts and prior work done in ML fairness testing.

2.1 Algorithmic Bias, Bias Mitigation and Group Fairness

Manually validating the fairness of the labels is often an expensive and time consuming process which is still prone to cognitive and social biases of the human auditors. Significant efforts have therefore been made to quantify the bias present in a ML model using fairness metrics. Existing fairness metrics are restricted to supervised binary classification problems where one of the outcomes is more favourable than the other and the dataset contains one or more *protected attributes* such as *race, sex, age, colour, religion or disability status*. An ML model is said to make unfair decisions if it favours a certain group or individual pertaining to one or more protected attributes in the dataset.

Fairness metrics can be broadly classified into two categories—namely, *group fairness* and *individual fairness*. Individual fairness dictates that the predictions of an ML model should not differ for two individuals who only differ in the value of the protected attribute. While group fairness dictates that the predictions of an ML model should be similar for both privileged and unprivileged groups present in the dataset [Castelnovo *et al.*, 2022; Hellman, 2020; Mitchell *et al.*, 2021; Kusner *et al.*, 2017; Grgic-Hlaca *et al.*, 2016; Dwork *et al.*, 2012; Barocas *et al.*, 2019; Barocas and Selbst, 2016; Hardt *et al.*, 2016; Binns, 2018; Verma and Rubin, 2018; Saxena *et al.*, 2019].

Once an appropriate definition of fairness is identified, the relevant techniques to mitigate said bias must be found. Bias mitigation techniques can be classified into three groups based on the location where the mitigation technique is applied: *pre-processing, in-processing and post-processing*. Several in-processing techniques or novel ML algorithms have been proposed that take fairness into consideration while training from biased data [Zhang *et al.*, 2018; Agarwal *et al.*, 2018; Kearns *et al.*, 2018; Kamishima *et al.*, 2012]. In contrast to in-processing techniques which target the ML model, pre-processing [Feldman *et al.*, 2015; Zemel *et al.*, 2013; Calmon *et al.*, 2017; Kamiran and Calders, 2012] and post-processing [Pleiss *et al.*, 2017; Hardt *et al.*, 2016; Kamiran *et al.*, 2012] techniques are applied to the training data and the predictions of the ML model respectively.

This study uses group fairness metrics due to their popularity in existing empirical studies on ML fairness testing and ease of understandability [Zhang and Harman, 2021; Biswas and Rajan, 2020; Biswas and Rajan, 2021; Hort *et al.*, 2021; Chakraborty *et al.*, 2021]. Our analysis of the relationship between the DFM and MFM (see Section 5.1) presents practical guidelines for practitioners to pick the appropriate bias mitigation strategy based on the particular fairness issue they are facing.

2.2 Prior Work in ML Fairness Testing

There is a growing consensus amongst academics that not all fairness metrics can be satisfied simultaneously. There is also a consensus that fairness and performance of ML systems are orthogonal to one another and involve a trade-off. In fact, identifying the correct fairness metric is the primary challenge which typically depends on the domain and problem at hand. It is therefore recommended to consider fairness as early as the requirements engineering and software design phase [Zhang *et al.*, 2020; Chen *et al.*, 2022; Mehrabi *et al.*, 2021; Zhang and Harman, 2021].

Several literature surveys have been conducted to classify and catalogue various ML fairness testing and bias mitigation techniques. Wan *et al.* conducted a large scale survey on in-processing bias mitigation strategies and their effectiveness. Chen *et al.* and Mehrabi *et al.* conducted a survey of existing literature on fairness testing in ML systems. Mehrabi *et al.* present a comprehensive survey on the state-of-the-art research on fairness in ML with an emphasise on fairness issues arising from both the data and the model. Chen *et al.* survey 113 papers addressing fairness testing and provide a formal definition of fairness bugs and fairness testing in ML from a software engineering perspective. Authors also reflect on how fairness testing differs from traditional software testing and provide practical guidelines on how and where to test for fairness within the entire Software Development Lifecycle [Wan *et al.*, 2021; Chen *et al.*, 2022; Mehrabi *et al.*, 2021].

Prior work has also focused on conducting empirical analysis of bias mitigation techniques. Biswas and Rajan take a holistic view of the entire ML pipeline and analyse the effect of common data pre-processing techniques such as standardisation, feature selection, encoding and sampling on the fairness of ML models. The analysis is conducted using 37 real-world ML pipelines from Kaggle notebooks which operate on five datasets. Typically data for the unprivileged group tends to be limited which make pre and post processing bias mitigation techniques less effective. Feffer *et al.* thus conduct an empirical analysis of bias mitigation in combination with popular ensemble techniques to understand the effectiveness of such combinations. Zhang and Harman studied the effect of training sample and feature sample size on the fairness of ML models. Authors observe that a large feature sample combined with a small training set helps reduce bias [Biswas and Rajan, 2021; Feffer *et al.*, 2022; Zhang and Harman, 2021].

Prior work discussed so far operate under the assumption that the training data and learning algorithm are accessible to the ML practitioner—often referred to as *white-box testing*. In contrast, *black-box testing* makes no such assumptions and treats the entire ML pipeline as a black-box where we can only control the input to the system and see the corresponding output. For such situations, several test input generation techniques have been proposed. Galhotra *et al.* proposed *Themis*, a tool which automatically generates a testing suite to measure discrimination in software using causal fairness testing. Udeshi *et al.* propose *Aequitas*, an automated tool that accepts a model and the protected attributes as input and explores the input space to detect specific examples that may produce discriminatory behaviour in the model. Aggarwal *et al.* propose a new technique for generating test input using symbolic execution which

DFM	
DI	$\frac{P(Y=1\mid D=0)}{P(Y=1\mid D=1)}$
SPD	$P(Y=1\mid D=0)-P(Y=1\mid D=1)$
MFM	
DI	$\frac{P(\hat{Y}=1\mid D=0)}{P(\hat{Y}=1\mid D=1)}$
SPD	$P(\hat{Y}=1\mid D=0)-P(\hat{Y}=1\mid D=1)$

Table 1: Fairness metrics used in this study

Name	**Prot.**	**#Eg.**
German [Hofmann, 1994]	age, sex	1000
Compas[Angwin *et al.*, 2016]	race, sex	6167
MEPS [mep,]	race	15675
Bank[Moro *et al.*, 2014]	age	30488
Adult[Kohavi and others, 1996]	race, sex	45222

Table 2: Datasets used in the study

Parameter	**Count**
Fairness metrics	2
ML models	4
Datasets	8
Total cases	$8 \times 4 = 32$
Iterations	50
Total fairness evaluation cycles	$32 \times 50 = 1600$

Table 3: Parameters of the study

accounts for correlation amongst the protected and unprotected attributes [Aggarwal *et al.*, 2019; Udeshi *et al.*, 2018; Galhotra *et al.*, 2017].

This study conducts an empirical analysis of the relationship between DFM and MFM, and as such, operates under white-box testing assumptions. The experimental design of this study is similar in spirit to that proposed by Zhang and Harman (2021). However our objective, results and implications are entirely different. While Zhang and Harman (2021) study the effect of training and feature sample size on the fairness of the model, this study aims to understand the relationship between DFM and MFM. We analyse how change in the distribution, sample size and number of features in the training set affects this relationship.

3 Experimental Design

This section presents the datasets, ML models and fairness metrics used in this study followed by the methodology used to evaluate DFM and MFM.

3.1 Datasets, ML Models and Fairness Metrics

Table 1 shows the group fairness metrics along with their mathematical formulas used in this study. We include all group fairness metrics—namely *Disparate Impact (DI)* and *Statistical Parity Difference (SPD)*—for which both model dependent and independent variants are available. The DFM use the labels of the data (Y) where as the MFM use the predictions of the trained ML models ($\hat{Y}$). Favourable and unfavourable outcomes are represented by 1 and 0 respectively. Similarly, privileged and unprivileged groups of the protected attribute (D) are represented by 1 and 0 respectively. All fairness metrics and datasets used in this study are obtained from the *AIF360* python library [Bellamy *et al.*, 2019].

Table 2 presents the datasets used in this study. We consider tabular datasets which have been extensively used in prior scientific contributions on ML fairness testing [Zhang and Harman, 2021; Biswas and Rajan, 2020; Biswas and Rajan, 2021; Chen *et al.*, 2022]. Based on prior work, we only consider one protected attribute at any given time thus giving us eight independent datasets. We follow the default pre-processing steps implemented in the AIF360 library—missing values are dropped and categorical features are label encoded. Prior to training, the features in the training and testing subsets are standardised by removing the mean of the sample and scaling to unit variance.

We use the scikit-learn [Pedregosa *et al.*, 2011] python library for creating the train-test splits and training the ML models. We use four ML models of varying complexity namely, *Logistic Regression*, *Decision Trees*, *Random Forest* and *Ada boost* based on their popularity in practise and in prior scientific publications [Zhang and Harman, 2021; Biswas and Rajan, 2021; Biswas and Rajan, 2020].

3.2 Fairness Evaluation

Figure 1 presents the methodology used in this study for evaluating the fairness of ML models and datasets. A 75–25 split with shuffling is used to create the training and testing splits. DFMs and MFMs are used to quantify the bias in the underlying distribution of the training set and the predictions of the models respectively. We adopted the transformation steps from prior work to scale all fairness metric values between 0 and 1 such that higher values indicate more bias [Zhang and Harman, 2021; Hort *et al.*, 2021].

We extend the above experiment further in two ways. First, we experiment with different number of examples and second with different number of features in the training set. For both experiments, we shuffle the order of the examples in the training and testing sets. Additionally, for the feature sample size experiment we shuffle the order of the features.

For the training sample size experiment, we generate different training samples of varying sizes starting from 10% of the original training data, and increase in steps of 10% until the full quantity is reached. For the feature sample size experiment, we start with a minimum of three features (in addition to the protected attribute and target) and introduce one new feature until all the features are utilised. Both the training and testing sets undergo the same feature sampling procedure in the feature sample size experiment. No such sampling is done in the testing set for the training sample size experiment.

We use Spearman Rank Correlation to quantify the linear relationship between the DFM and MFM. We repeat all experiments 50 times and report the statistical significance of our results. We consider cases where $pvalue \leq 0.05$ to be

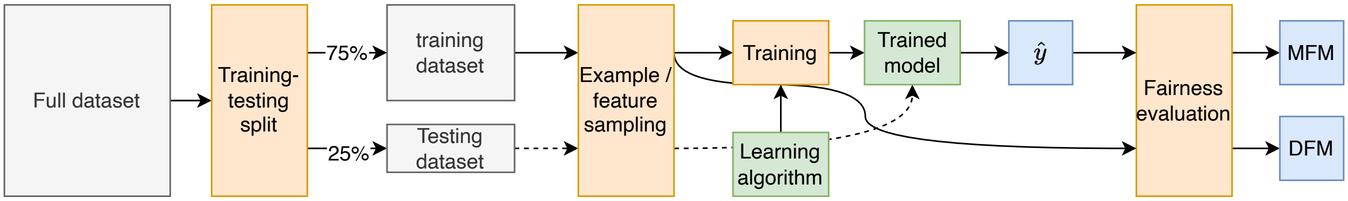

Figure 1: Methodology for evaluating fairness of datasets and ML models using DFM and MFM.

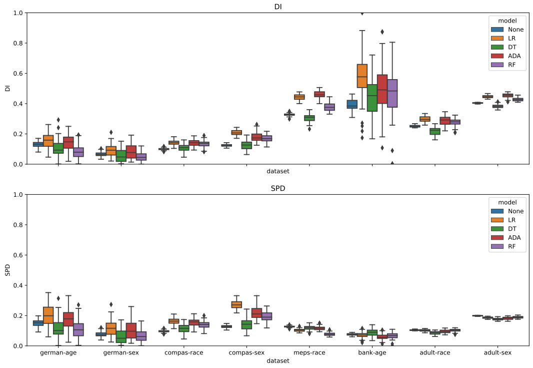

Figure 2: Distribution of DFM and MFM across all datasets and models.

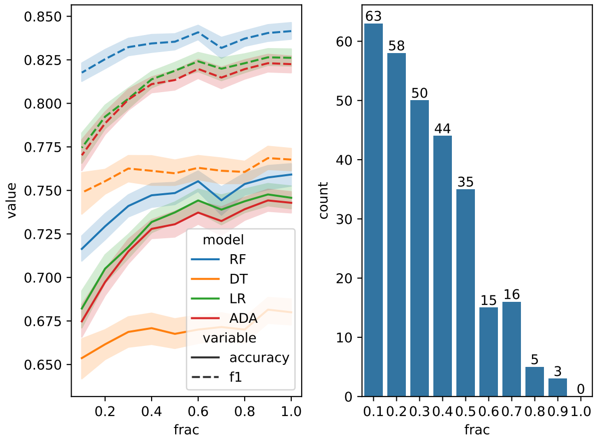

Figure 3: ***(left)*** Accuracy and f1 across various training sample size in the german-age dataset. ***(right)*** Count of cases with significant change in accuracy and f1.

statistically significant in our evaluation. Table 3 summarises the parameters of the study. We train 4 ML models on 8 datasets producing 32 total cases. The fairness for each case is evaluated 50 times using two fairness metrics, thus producing a total of 1600 training and fairness evaluation cycles.

4 Results

This section presents the analysis of the relationship between DFM and MFM. The section is broken into two subsections. Section 4.1 presents the relationship between DFM and MFM as the distribution of the training data changes while Section 4.2 presents their relationship across varying training and feature sample sizes.

4.1 Full Training Set Experiments

RQ1. What is the relationship between DFM and MFM as the fairness properties of the underlying training dataset changes?

Figure 2 presents a boxplot with the distribution of the fairness metrics across the datasets. The x-axis represents the datasets used in this study while the y-axis presents the value of the fairness metric. The models used in this study are represented using different colors—note that the model "None" represented in blue refers to the DFM. Both the fairness metrics DI (top) and SPD (bottom) are presented in separate plots. We observe that distribution of the DFM and MFM are similar in all cases indicating that they convey similar information. The variability of the DFM is less than the MFM. This is because in addition to the randomness from the data shuffling in the training set, the models are assigned random initial states in every iteration. Finally in several cases the tree-based classifiers (DT and RF) make fairer decisions compared to the other classifiers, sometimes even better than the baseline provided by the DFM.

To analyse the relationship between the DFM and MFM across various data distributions, we calculate the DFM and MFM across training samples of varying size. The data distribution in smaller training samples will change more frequently in the 50 iterations at the loss of data quality. To identify a sample size that captures a variety of data distribution changes while also being a realistic training dataset, we analyse the *accuracy* and *f1 score* of the models across the training sample sizes. Next, we conduct *student t-test* to identify the smallest sample size where the performance of the models is similar to that obtained when trained using the full training set.

Figure 3 (right) presents a histogram of the number of cases where there was a significant difference between the two populations. We note that there is a significant difference in the performance of the models in the majority of the cases when the training size is reduced to 50% while it remains consistent when using a training size of 60% or higher. This is also corroborated by the lineplot in Figure 3 (left) which shows the accuracy and f1 of all models across various training sample sizes in the *german-age* dataset. We observe that the performance stabilises starting from 60% training sample size. Thus for the majority of the cases, a training sample of 60%

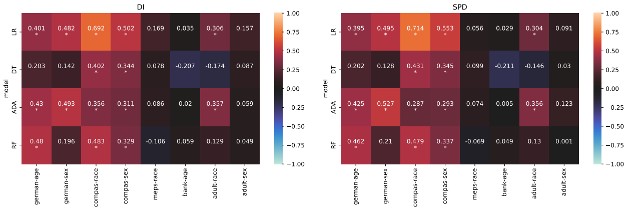

Figure 4: Correlation between DFM and MFM across all models and datasets using 60% training data. The statistical significance is reported using asterisks at three α levels. ***: $p \leq 0.01$; **: $0.01 > p \leq 0.05$; *: $0.05 > p \leq 0.1$

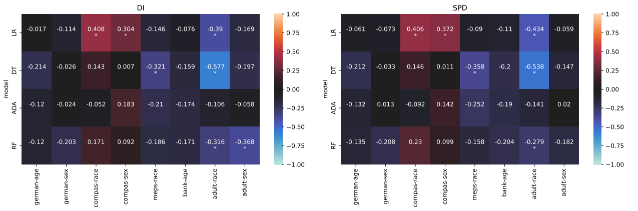

Figure 5: Correlation between DFM and MFM using the full training data. ***: $p \leq 0.01$; **: $0.01 > p \leq 0.05$; *: $0.05 > p \leq 0.1$

allows us to train models with acceptable performance, while also capturing a wide variety of fairness issues in the underlying training data within the 50 iterations.

Figure 4 shows the correlation between the DFM and MFM across all models and datasets when trained using 60% of the original training set. The models used in this study are represented along the y-axis and the datasets along the x-axis. Darker colours indicate weaker correlation whereas brighter colours indicate stronger correlation. Positive correlation is indicated using hues of red while negative correlation is indicated using hues of blue. The correlation between DFM and MFM for both fairness metrics are shown separately. We primarily observe a positive correlation between the DFM and MFM. This indicates that the DFM and the MFM convey the same information as the distribution—and consequently the fairness properties—of the underlying training dataset changes.

In contrast to Figure 4, Figure 5 shows the correlation between DFM and MFM when the full training set is used. Due to lack of significant change in the distribution of the training data, we primarily observe darker colours indicating that the DFM and MFM are not linearly related to one another anymore.

> ***Answer to RQ1:*** *DFM and MFM are positively correlated and thus convey the same information as the distribution—and consequently the fairness properties—of the underlying training dataset changes.*

RQ2. How does the training sample size affect the correlation between DFM and MFM?

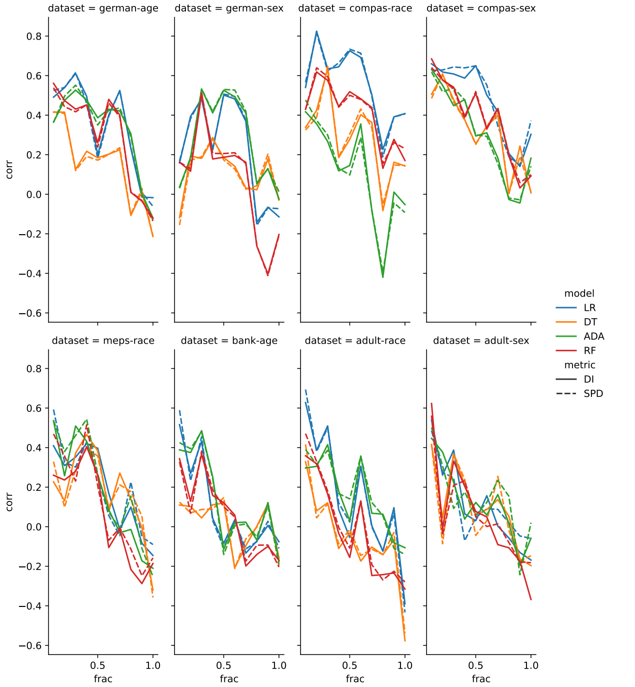

Figure 6: Distribution of correlation between DFM and MFM within various training sample sizes across all datasets and models.

In Figure 4, the correlation in the smaller datasets are more positive compared to the larger datasets when trained using 60% of the original training data. When the training data size is increased, the correlation in the datasets decrease as observed in Figure 5. Based on these observations, we hypothesise that the quantity of training data influences the relationship between the DFM and MFM. Our hypothesis is corroborated by Figure 6 which shows the distribution of the correlation between DFM and MFM within the various training sample sizes, across all datasets and models. The x-axis presents the training sample size and the y-axis presents the correlation between the DFM and MFM. The colours represent the models while the style of the line represents the two fairness metrics. Each dataset is shown as a separate subplot. The overwhelming majority shows that the correlation between the DFM and MFM decreases as we increase the training sample size.

> ***Answer to RQ2:*** *The training sample size has a profound effect on the relationship between the DFM and MFM. The correlation between the DFM and MFM decreases as we increase the training sample size.*

4.2 Training and Feature Sample Size Experiments

RQ3. What is the relationship between DFM and MFM across various training and feature sample sizes?

Training Sample Size. In this section we analyse the relationship between DFM and MFM across varying training sample sizes. In contrast to Section 4.1 where we calculated the correlation between DFM and MFM within each training sample size, here we calculate the correlation across all training

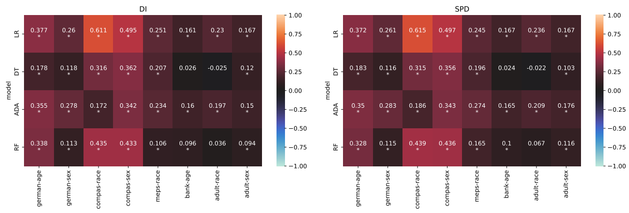

Figure 7: Correlation between DFM and MFM acoss various training sample sizes. ***: $p \leq 0.01$; **: $0.01 > p \leq 0.05$; *: $0.05 > p \leq 0.1$

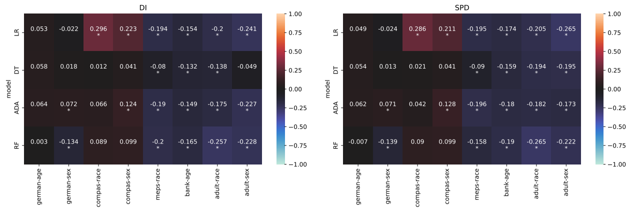

Figure 8: Correlation between DFM and MFM across various feature sample sizes. ***: $p \leq 0.01$; **: $0.01 > p \leq 0.05$; *: $0.05 > p \leq 0.1$

sample sizes. The correlation between the DFM and MFM is shown in Figure 7. We primarily observe colours indicating that the DFM and MFM convey similar information as the training sample size changes.

Feature Sample Size. In this section we analyse the relationship between the DFM and MFM across varying feature sample sizes. In contrast to the training sample size experiment above, we change the number of features in the training set and randomise the feature order in each iteration. Figure 8 presents the correlation between the DFM and MFM across all feature sample sizes. We primarily notice darker colours indicating that there is no significant correlation between the DFM and MFM as the number of features in the training dataset changes.

From Table 1, we note that the feature sample size does not affect the DFM thus explaining the lack of significant correlation between the DFM and MFM. The larger datasets show a more negative correlation. This however is due to the change in training set distribution caused by the training-testing split within the 50 iterations as explained in Section 4.1. This can also be verified in Figure 9 which shows the relationship between the correlation and the feature sample sizes across all datasets and models. There is no discernible relationship between the correlation and the feature sample size in the top row containing datasets with a small training sample size and varying feature sample size. A slight relationship can only be observed in the bottom right subplots which contain datasets with a large training sample size but small feature sample size.

> ***Answer to RQ3:*** *DFM and MFM convey similar information as the training sample size changes but not when the feature sample size changes.*

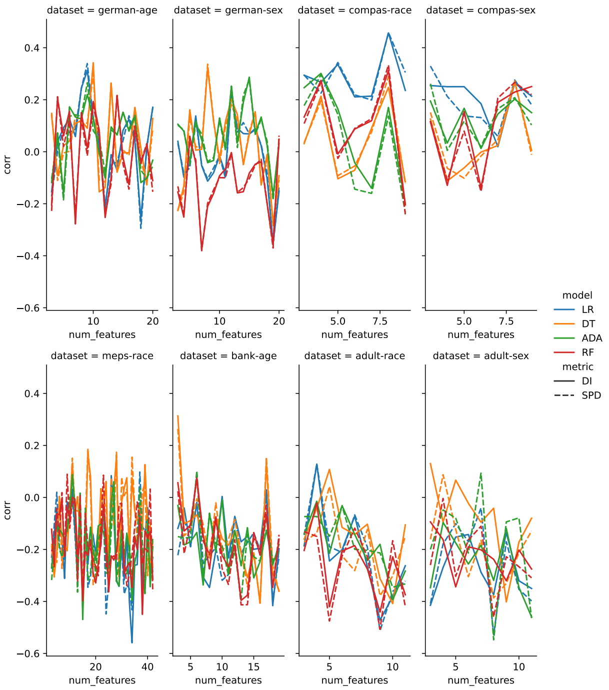

Figure 9: Distribution of correlation between DFM and MFM across various features sample sizes.

5 Implications

The implications of the results presented in Section 4 along with their threats to validity are discussed in this section.

5.1 Discussion

Data Drift

Results from RQ1 indicate that the DFM and MFM convey the same information when the distribution and consequently the fairness properties of the training data changes. ML systems running in a production environment are often monitored to detect degradation in model performance. A standard practise is to combine the data encountered by the model in the production environment with its predictions to create the training set for the next iteration [Biessmann *et al.*, 2021]. Since data reflects the real world, change in its underlying distribution over time is eminent. Our results indicate that DFM can be used as a early warning system to identify fairness related data drifts in automated ML pipelines.

Fairness, efficiency and performance trade-off

Results from RQ2 indicate that the quantity of training data profoundly influences the relationship between DFM and MFM. We primarily see a positive correlation between the DFM and MFM in smaller training sample sizes which indicates that the DFM and MFM convey the same information. With sufficient training data, the correlation starts to drop and eventually becomes negative. This indicates that the models learn to make fairer predictions and are able to circumvent the bias in the training data to a certain extent.

Thus a positive correlation between the DFM and MFM may indicate lack of sufficient training data. Under such circumstances, practitioners can either choose to collect more data if possible or use bias mitigation techniques to address the fairness issue.

A negative correlation between the DFM and MFM implies that the MFM reported a lower value compared to the DFM. This does not however guarantee the absence of bias in the model. Zhang and Harman (2021) showed that in addition to the quantity of training data, the quality itself affects the fairness of ML systems. Introducing more data does not fix the bias in the model if the final distribution remains biased.

Lower MFM compared to DFM also presents a trade-off between efficiency and performance. A slight reduction in the training sample size, in combination with appropriate bias mitigation techniques, can allow practitioners to build fair ML systems with quicker training cycles at the cost of negligible predictive performance [Verdecchia *et al.*, 2022]. Engineering efficient, high quality training data can reduce training cycles, development time and ultimately project costs. Compounded over the entire duration that an ML system stays in production along with the human effort required to keep such a system operational, the benefits can be more than substantial.

No correlation between DFM and MFM presents a trade-off between fairness, efficiency and performance. Practitioners may opt for a more efficient system by reducing the training sample size. However this may reduce the predictive performance of the model and require more engineering effort to mitigate the fairness issues in the system. Alternatively, they may opt for more accurate predictions by using a larger training sample size. A larger training size may mitigate some of the bias in the training set at the cost of more compute.

Test Reduction

Results from RQ3 indicate that the DFM and MFM are positively correlated and thus convey the same information when the training sample size changes. DFM can therefore aid practitioners catch fairness issues upstream and avoid execution costs of a full training cycle. Considering the entire duration that an ML system is operational, along with the multiple iterations it takes to test an ML system, avoiding a full training cycle while evaluating its fairness can be energy efficient and sustainable in the long run.

However the same test reduction cannot be made when experimenting with the feature sample size of the training set. Zhang and Harman (2021) showed that a larger feature sample size typically improves the fairness of the model. To the best of our knowledge, there are no fairness metrics that consider the influence of other features on the fairness at the data level. Thus when experimenting with the feature sample size, it is recommended that ML practitioners evaluate the fairness both before and after training.

Locating root cause of bias

Testing for fairness after training the model makes it very difficult to identify where exactly the bias was introduced. Testing for fairness both at the data (using DFM) and model (using MFM) level provides a holistic view on the fairness of the entire system and can aid practitioners identify the root cause of bias. If the DFM indicates presence of bias, it can be an early indication of flaws in the data collection process or flaws in the initial design of the system itself. In the event that the DFM does not indicate bias but the MFM does, practitioners can narrow down the cause of bias to the learning algorithm itself and opt for in-processing or post-processing bias mitigation techniques.

Explaining fairness in decision trees

We consistently observe that Decision Trees (DTs) are able to make fairer predictions with minimal effort. For instance, in Figure 2 we observe that DTs report lower values for both fairness metrics across all datasets. In Figure 6 we observe that DTs consistently report lower correlation between DFM and MFM compared to other models. This indicates that DTs are able to mitigate the bias present in the training data with a smaller training sample size and continue to do so as the training same size changes. The above observation pose an interesting line of query into examining why DTs are able to produce fairer predictions using explainable AI techniques.

5.2 Threats to Validity

We do not apply the *Bonferroni correction* to the correlation analysis results. Although we report the $pvalue$ for completeness, we do not base our implications only on the statistically significant results. But rather on general trends observed in our analysis. For all experiments, we additionally conduct linear regression analysis using ordinary-least squares and check the coefficient of determination (R^2) and the mean squared error (MSE) in the residuals to evaluate the goodness of fit.

For the experiment conducted in Section 4.1 the selection of the training sample size of 60% is a gross approximation and may not be a good fit for all datasets used in this study. An alternative albeit computationally more expensive solution would be to identify this threshold individually for each dataset and update it dynamically as its distribution changes.

6 Conclusion

Prior work in ML fairness testing evaluate fairness after the training stage, using the predictions of the ML model. In contrast, this study presents a more holistic approach by testing for fairness at two distinct locations of the ML development lifecycle. We analyse the relationship between model dependent and independent fairness metrics empirically and find a linear relationship between data and model fairness metrics when the distribution and the size of the training data changes. Our results indicate that testing for fairness prior to training can be a "cheap" and effective means of catching fairness issues in the upstream stages of automated ML pipelines and aid practitioners navigate the complex landscape of fairness testing. As an extension of this study, we wish to evaluate the effectiveness of DFM in real-world ML systems.

References

[Agarwal *et al.*, 2018] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *International Conference on Machine Learning*, pages 60–69. PMLR, 2018.

[Aggarwal *et al.*, 2019] 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.

[Angwin *et al.*, 2016] 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, 2016.

[Barocas and Selbst, 2016] Solon Barocas and Andrew D Selbst. Big data's disparate impact. *California law review*, pages 671–732, 2016.

[Barocas *et al.*, 2019] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019. http://www.fairmlbook.org.

[Bellamy *et al.*, 2019] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra 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.

[Biessmann *et al.*, 2021] Felix Biessmann, Jacek Golebiowski, Tammo Rukat, Dustin Lange, and Philipp Schmidt. Automated data validation in machine learning systems. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering.[Google Scholar]*, 2021.

[Binns, 2018] Reuben Binns. Fairness in machine learning: Lessons from political philosophy. In *Conference on Fairness, Accountability and Transparency*, pages 149–159. PMLR, 2018.

[Biswas and Rajan, 2020] Sumon Biswas and Hridesh Rajan. Do the machine learning models on a crowd sourced platform exhibit bias? an empirical study on model fairness. In *Proceedings of the 28th ACM joint meeting on European software engineering conference and symposium on the foundations of software engineering*, pages 642–653, 2020.

[Biswas and Rajan, 2021] Sumon Biswas and Hridesh Rajan. Fair preprocessing: towards understanding compositional fairness of data transformers in machine learning pipeline. In *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pages 981–993, 2021.

[Bosch *et al.*, 2021] Jan Bosch, Helena Holmström Olsson, and Ivica Crnkovic. Engineering ai systems: A research agenda. In *Artificial Intelligence Paradigms for Smart Cyber-Physical Systems*, pages 1–19. IGI Global, 2021.

[Calmon *et al.*, 2017] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. Optimized pre-processing for discrimination prevention. *Advances in neural information processing systems*, 30, 2017.

[Castelnovo *et al.*, 2022] Alessandro Castelnovo, Riccardo Crupi, Greta Greco, Daniele Regoli, Ilaria Giuseppina Penco, and Andrea Claudio Cosentini. A clarification of the nuances in the fairness metrics landscape. *Scientific Reports*, 12(1):1–21, 2022.

[Chakraborty *et al.*, 2021] Joymallya Chakraborty, Suvodeep Majumder, and Tim Menzies. Bias in machine learning software: why? how? what to do? In *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pages 429–440, 2021.

[Chen *et al.*, 2022] Zhenpeng Chen, Jie M Zhang, Max Hort, Federica Sarro, and Mark Harman. Fairness testing: A comprehensive survey and analysis of trends. *arXiv e-prints*, pages arXiv–2207, 2022.

[Dwork *et al.*, 2012] 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, 2012.

[Feffer *et al.*, 2022] Michael Feffer, Martin Hirzel, Samuel C Hoffman, Kiran Kate, Parikshit Ram, and Avraham Shinnar. An empirical study of modular bias mitigators and ensembles. *arXiv preprint arXiv:2202.00751*, 2022.

[Feldman *et al.*, 2015] 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.

[Galhotra *et al.*, 2017] 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.

[Grgic-Hlaca *et al.*, 2016] 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. Barcelona, Spain, 2016.

[Hardt *et al.*, 2016] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

[Hellman, 2020] Deborah Hellman. Measuring algorithmic fairness. *Virginia Law Review*, 106(4):811–866, 2020.

[Hofmann, 1994] Hans Hofmann. German credit data, 1994.

[Hort *et al.*, 2021] Max Hort, Jie M Zhang, Federica Sarro, and Mark Harman. Fairea: A model behaviour mutation approach to benchmarking bias mitigation methods. In *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pages 994–1006, 2021.

[Hutchinson *et al.*, 2021] Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 560–575, 2021.

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

[Kamiran *et al.*, 2012] Faisal Kamiran, Asim Karim, and Xiangliang Zhang. Decision theory for discrimination-aware classification. In *2012 IEEE 12th International Conference on Data Mining*, pages 924–929. IEEE, 2012.

[Kamishima *et al.*, 2012] 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.

[Kearns *et al.*, 2018] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning*, pages 2564–2572. PMLR, 2018.

[Kohavi and others, 1996] Ron Kohavi et al. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *Kdd*, volume 96, pages 202–207, 1996.

[Kusner *et al.*, 2017] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *Advances in neural information processing systems*, 30, 2017.

[Mehrabi *et al.*, 2021] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.

[mep,] Medical expenditure panel survey. Accessed on 2022-10-05.

[Mitchell *et al.*, 2021] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application*, 8:141–163, 2021.

[Moro *et al.*, 2014] 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.

[Pedregosa *et al.*, 2011] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[Pleiss *et al.*, 2017] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On fairness and calibration. *Advances in neural information processing systems*, 30, 2017.

[Saxena *et al.*, 2019] Nripsuta Ani Saxena, Karen Huang, Evan DeFilippis, Goran Radanovic, David C Parkes, and Yang Liu. How do fairness definitions fare? examining public attitudes towards algorithmic definitions of fairness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 99–106, 2019.

[Sculley *et al.*, 2015] David Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-Francois Crespo, and Dan Dennison. Hidden technical debt in machine learning systems. *Advances in neural information processing systems*, 28:2503–2511, 2015.

[Udeshi *et al.*, 2018] 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, 2018.

[Verdecchia *et al.*, 2022] Roberto Verdecchia, Luís Cruz, June Sallou, Michelle Lin, James Wickenden, and Estelle Hotellier. Data-centric green ai an exploratory empirical study. In *2022 International Conference on ICT for Sustainability (ICT4S)*, 2022.

[Verma and Rubin, 2018] Sahil Verma and Julia Rubin. Fairness definitions explained. In *2018 ieee/acm international workshop on software fairness (fairware)*, pages 1–7. IEEE, 2018.

[Wan *et al.*, 2021] Mingyang Wan, Daochen Zha, Ninghao Liu, and Na Zou. Modeling techniques for machine learning fairness: A survey. *arXiv preprint arXiv:2111.03015*, 2021.

[Zemel *et al.*, 2013] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International conference on machine learning*, pages 325–333. PMLR, 2013.

[Zhang and Harman, 2021] Jie M Zhang and Mark Harman. "Ignorance and prejudice" in software fairness. In *2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE)*, pages 1436–1447. IEEE, 2021.

[Zhang *et al.*, 2018] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340, 2018.

[Zhang *et al.*, 2020] Jie M Zhang, Mark Harman, Lei Ma, and Yang Liu. Machine learning testing: Survey, landscapes and horizons. *IEEE Transactions on Software Engineering*, 2020.