Assessing the Risk of Discriminatory Bias in Classification Datasets

Kejun Dai[1*], Jonathan Kim[1,2], Sašo Džeroski[3], Jörg Wicker[1], Gillian Dobbie[1], Katharina Dost[1,3*]

[1]Computer Science, University of Auckland, Auckland, New Zealand.
[2]AI+, Callaghan Innovation, Auckland, New Zealand.
[3]Knowledge Technologies, Jožef Stefan Institute, Ljubljana, Slovenia.

*Corresponding author(s). E-mail(s): kdai332@aucklanduni.ac.nz; katharina.dost@ijs.si;
Contributing authors: j.kim@callaghaninnovation.govt.nz; saso.dzeroski@ijs.si; j.wicker@auckland.ac.nz; g.dobbie@auckland.ac.nz;

Abstract

Bias in machine learning models remains a critical challenge, particularly in datasets with numeric features where discrimination may be subtle and hard to detect. Existing fairness frameworks rely on expert knowledge of marginalized groups, such as specific racial groups, and categorical features defining them. Furthermore, most frameworks evaluate bias in models rather than datasets, despite the fact that model bias can often be traced back to dataset shortcomings. Our research aims to remedy this gap by capturing dataset flaws in a set of meta-features at the dataset level, and to warn practitioners of bias risk when using such datasets for model training. We neither restrict the feature type nor expect domain knowledge. To this end, we develop methods to synthesize biased datasets and extend current fairness metrics to continuous features in order to quantify dataset-level discrimination risks. Our approach constructs a meta-database of diverse datasets, from which we derive transferable meta-features that capture dataset properties indicative of bias risk. Our findings demonstrate that dataset-level characteristics can serve as cost-effective indicators of bias risk, providing a novel method for data auditing that does not rely on expert knowledge. This work lays the foundation for early-warning systems, moving beyond model-focused assessments toward a data-centric approach.

Keywords: Discriminatory bias; Meta-learning; Fairness

1 Introduction

Machine learning is increasingly used as a foundation for decision-making due to its ability to quickly and accurately find patterns in large datasets. Expecting that machine learning models are incapable of prejudice and discrimination, practitioners employ them even in critical decision-making domains such as loan approval [1], job recruitments [2], and credit card risk prediction [3]. However, studies have found evidence that this expectation is a fallacy – machine learning models often make discriminatory predictions against marginalized populations that, if left unchecked, create unfairness [4, 5].

In response, many approaches to mitigating discriminatory bias in machine learning models and ensuring their predictions are fair to all populations have been proposed [6]. However, they typically assume that (i) discrimination occurs with respect to categorical demographic attributes, (ii) practitioners are aware of sensitive attributes or bias itself, and (iii) marginalized groups can be expressed using observable attributes. As a result, current frameworks for describing and assessing bias rely heavily on predefined, observable demographic categories such as race or gender. For example, bias auditing or evaluation toolkits request practitioners to name the attributes they suspect to be affected by certain biases in order to evaluate said bias in their models [7, 8]. This approach is particularly problematic if the practitioner is unaware of the bias; even more so in numeric datasets such as in healthcare data, where patients are defined by their age and numeric lab results rather than their gender or race [9]. Bias here may be more subtle, more complex to express, harder to test for, and it is impossible to exhaustively search all subspaces of the feature space.

An additional drawback is that existing tools primarily assess unfair behavior of models. However, model bias is often rooted in dataset shortcomings such as underrepresented subspaces within the feature space. In that case, *every* model trained on this dataset is likely to exhibit a bias against this subspace unless explicitly taught to prevent this. We therefore advocate for auditing datasets rather than models so that we can warn the practitioner *before* the potentially costly model training that bias mitigation is essential.

In our previous work [10], we have shown that some sources of bias can be detected in datasets with numeric, non-discrete, continuous features without any prior expert suspicion or knowledge via violations of smoothness assumptions in the data distribution. In this paper, we are searching for further dataset characteristics or indicators that increase the risk of bias in models trained on such datasets. The ultimate goal is to develop an early-warning system creating awareness in practitioners.

To this end, in this paper, we propose to test a wide range of potential dataset characteristics in a meta-learning approach. Particularly, we will derive meta-features, i.e., transferable high-level descriptors of datasets that represent a dataset in a vector. We will then train a meta-model on these meta-features to predict what type and what severity of bias we find in the dataset. Validating this meta-model indirectly validates the suitability of the meta-features to capture bias risk information in the dataset. We will focus on datasets with numeric, non-discrete, continuous features and bias that is expressed through these continuous features. To achieve this goal, we make the following contributions:

1. We provide a framework to extend the concept of marginalized subgroups to datasets with continuous values. This framework allows for intricate and complex definitions of marginalized subspaces while remaining efficient to query the membership of samples.
2. We extend three popular bias metrics to continuous datasets in accordance with the proposed framework, and propose corresponding dataset-level bias scores that quantify the risk of these three types of bias for models trained on such datasets. These scores allow us to disentangle bias risk from model bias.
3. We propose an algorithm to generate synthetic biased datasets with continuous values based on the formerly defined framework. Although particularly crucial for meta-learning approaches, this generator is a standalone contribution that can support future bias research.
4. We provide our meta-database to the community in our repository for further collaborative extension.
5. Using meta-models trained to quantify bias risks in datasets without any expert knowledge, we identify relevant dataset indicators that can warn practitioners of bias risks.

The remainder of this paper is organized as follows: Section 2 provides a brief overview of related research. Section 3 explains our methodology, including the data/bias generation, meta-features, bias scores, meta-database population, and meta-model training. Section 4 presents experimental findings. Finally, Section 5 concludes the paper.

2 Related Research

This section first defines discriminatory bias and generalizes it beyond categorical, personal data. Then, it reviews bias detection, quantification, and generation approaches – artifacts crucial for our meta-learning approach.

Discriminatory Bias

In machine learning, we denote a classification dataset as $D = \{(x_i, y_i)\}^n$, where $x_i \in X$ is the feature vector including attributes $A_1, ..., A_n$ and $y_i \in Y$ is the label. Our goal is to train a model $\mathcal{M} : x \mapsto \hat{y}$ mapping previously unseen data points x to a label prediction $\hat{y}$.

Discriminatory bias is defined with respect to sensitive attributes like ethnicity or gender that best describe the marginalized population, such as Black men [11]. The marginalized population is often termed the "protected group" in contrast to the "privileged group." Discriminatory bias is a disparity between models' predictions for protected groups and privileged groups.

Sensitive attributes can be any categorical attributes of X, but in most cases, they are either race, gender, marital status, or disability. A limitation of this definition of bias is that it heavily relies on our awareness of socioeconomic bias in the real world. If no protected groups are known in advance, practitioners can probe the model for bias against all combinations of sensitive attribute values. However, this probing approach is only feasible for categorical attributes.

In this paper, we extend the notion of discriminatory bias beyond categorical attributes. While the literature on fair machine learning focuses on discrimination due to human prejudice, bias can occur in all kinds of datasets due to biased data collection, sampling, or measurement [6]. Consequently, subspaces (rather than protected groups) can be disadvantaged. For example, models for predicting drug responses may be biased against patients with plasma cell counts below a certain threshold [12].

When investigating discriminatory bias, we are interested in whether models are giving fewer beneficial predictions to individuals from marginalized populations, leading to unfair model behavior. How to define fairness is debated, and many criteria have been proposed [6, 13]. For example, *Disparate Impact* [14] expects models to make equal numbers of preferable predictions for both protected and privileged groups. *Equal Opportunity* [15], on the other hand, expects the models to have equal accuracy of their preferable predictions for both protected and privileged groups; a concept that extends well to non-personal data. In other problem domains, such as word embeddings, there are metrics like *WEAT* (Word Embedding Association Test) [16] that expect models to provide a balanced representation for sensitive attributes in relation to other concepts. Note that these metrics are model-specific, while we strive to quantify the risk of discriminatory bias based on the model alone.

Bias Detection

In the field of fair ML, there is an abundance of model auditing tools that detect bias. AI Fairness 360 [7] and Aequitas [8] are comprehensive suites implementing 70 and 12 of these methods, respectively. However, they test models rather than datasets and rely on the definition of sensitive attributes or groups.

Alelyani [17] proposes alternation, a data-centric approach to detect bias in machine learning models. Machine learning models are trained on two datasets – one of them is the original dataset, while the other one is derived from it by alternating sensitive attribute values. The goal of the test is to see whether the performance of machine learning models is significantly impacted by the alternation of sensitive attribute values. As sensitivity to feature values is generally expected and desired, this approach is infeasible if no sensitive attributes have been defined.

The literature on dataset bias is generally focused on bias between a sample and a ground-truth data distribution [18]. One exception is our previous work [10] that investigates only the sample for distributional imperfections that can lead to specific regions in the feature space being underrepresented, which, in turn, will create unfair behavior of models towards these regions. However, this approach makes strong assumptions regarding the ground truth, and it can neither categorize nor quantify the risk of model unfairness.

Bias Quantification

There are several attempts to quantify bias in datasets. Azzalini et al. [19] propose FairDB, a framework that evaluates discriminatory bias by quantifying functional dependencies among attributes, that is, attributes (such as race) whose values often determine the value of another attribute (such as income). However, such dependencies require categorical features and the definition of sensitive attributes.

Jiang et al. [20] train a linear regression model as a surrogate for other models on the target dataset and measure the following bias metrics for the model: (i) the difference between true and predicted probabilities of a certain number of instances within pre-determined groups, (ii) the difference in accuracy between each pre-determined group, and (iii) the difference in the proportion of preferable predictions being made towards individuals from each pre-determined group. We use this approach as an inspiration for our landmarking dataset-level bias metrics.

Simonetta et al. [21] expect datasets to have samples of every possible combination of all sensitive attributes for machine learning models to rule out "blind spots" that can lead to unfair behavior. Based on this definition, the study develops metrics for datasets that calculate the ratio between the number of unique combinations of sensitive attribute values in the dataset and the number of possible combinations of all sensitive attribute values to assess the level of completeness in the dataset. Using frame theory, they extend this notion to continuous features. However, this approach addresses imbalance rather than more subtle reasons for bias.

Bias Generation

Most studies on fair ML use real-world datasets. Several historically biased open-access datasets are well-curated and frequently used [22]. We use some of these datasets to evaluate our approach, but need further datasets and an additional level of control for our meta-learning approach. Therefore, we opt for synthetic datasets and bias.

Jiang et al. [20] use genetic algorithms to generate synthetic biased datasets in the education sector. Starting with a number of real-world datasets, they apply a series of mutations and crossovers to obtain new synthetic datasets. While this approach allows for the generation of arbitrarily many biased datasets, it does not exhibit the control and diversity we aim for.

Jesus et al. [23] use disproportionate sampling techniques to create multiple biased datasets with various degrees of bias from one singular large dataset. They induce bias in their new dataset by oversampling entries belonging to privileged groups and undersampling entries belonging to protected groups. We adopt this approach as different levels of disproportionality grant us control over the strength of the induced bias.

Barbierato et al. [24] synthesize biased datasets by constructing probabilistic networks. They use directed acyclic graphs to model feature dependencies, where nodes are features and edges indicate the strength of dependency between features. The discriminatory bias can be induced by tuning the severity of dependency of sensitive attributes with other features. Despite of the degree of control over synthesized datasets, this approach is not efficient enough for the number of datasets required for meta-learning.

3 Methodology

The reasons for unfair behavior of models generated by machine learning can often be traced back to shortcomings of the dataset they were learned from. For example, population groups that are disadvantaged by a model are typically underrepresented in the training set and could therefore be identified using only the training set.

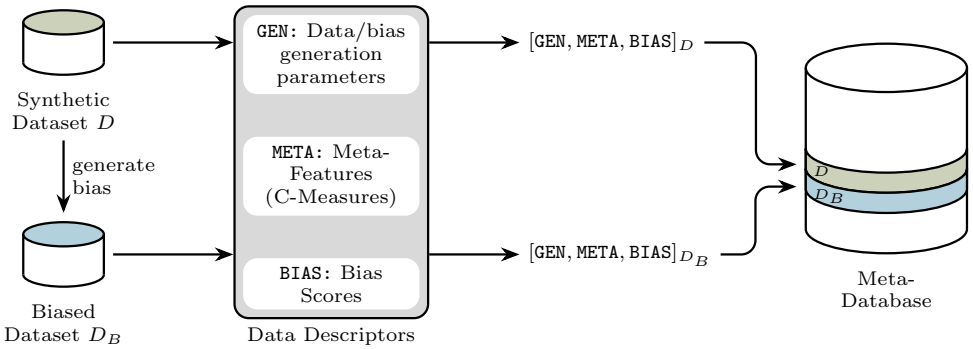


Fig. 1: Overview of the meta-database generation: We repeatedly generate a synthetic dataset and bias, extract data and bias descriptors, obtain a vector representation of the datasets, and enter them as rows into the meta-database.

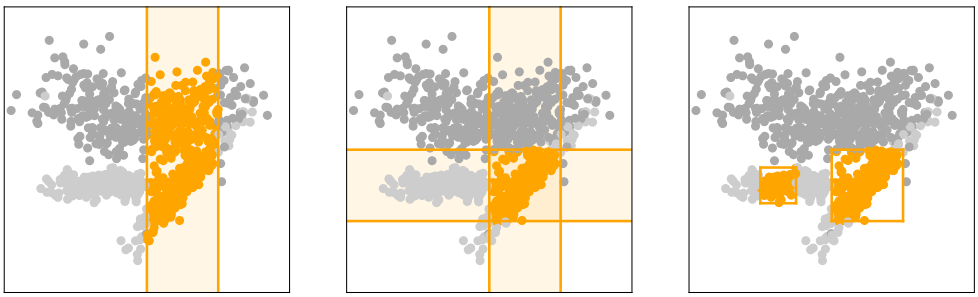
To gain a better understanding of what constitutes a bias and to identify features that capture it, in this paper, we propose a meta-learning approach. Our goal is to train a machine learning model that takes a dataset as input and decides whether there is a bias in the dataset, what type of bias it is, and how strong it is.

To obtain such a model, we need to populate a meta-database that the model can be trained on. This meta-database shall represent each dataset and its bias (of different types) by a single vector. More precisely, we follow the procedure outlined in Figure 1 to populate a meta-database: First, we generate a number of synthetic datasets with (D_B) and without (D) synthetic bias. For both the unbiased (D) and the biased (D_B) dataset, we capture parameters describing the data and bias generation (`GEN`), extract meta-features describing the dataset characteristics (`META`), and quantify different types of bias through bias scores (`BIAS`). Concatenating the obtained features into vectors [`GEN`, `META`, `BIAS`] produces two rows for the meta-database, one for the unbiased dataset and one for the biased dataset. Finally, a meta-model will be learned to take the meta-features as input and output the corresponding bias scores. We discuss each element of this pipeline in more depth subsequently.

3.1 Dataset Generation

We aim to train our meta-model on a wide spectrum of biased and unbiased datasets to best support the generality of the model. However, real-world datasets may or may not be biased to an unknown degree: In using them for meta-learning, we would lack the level of control that the use of synthetic datasets can offer. We therefore decide to generate synthetic datasets of varying characteristics to populate our meta-database.

Particularly, we generate datasets with 2k-20k rows, 2-10 classes, 2-50 features, varying numbers of dependent, independent, and repeated features, and 1-3 multivariate Gaussian clusters per class. See our supplementary materials for a detailed breakdown of the generation procedure and parameter ranges. The generated datasets will then artificially be biased, before we extract meta-features from both the original and biased versions of the data.



(a) Sensitive Slice (b) Sensitive Box (c) Sensitive Complex

Fig. 2: An example of the SBC Hierarchy: Orange regions represent marginalized regions of increasing complexity (left to right).

3.2 Bias Generation

The research field of fairness in machine learning typically assumes expert-defined marginalized subpopulation groups before assessing whether these groups are treated fairly by the model [7]. These groups are defined by categorical feature values, such as a specific race, or combinations thereof, such as the subpopulation associating with a specific race and gender. To extend this notion to more general datasets with numeric, non-discrete, continuous features, we first need to define marginalized regions within our datasets. For the sake of brevity, we refer to such datasets as "continuous". Inspired by Jesus et al. [23], we will then sample disproportionately from marginalized regions to create an underrepresentation effect.

We propose the *SBC Hierarchy (Slice, Box, and Complex)* for a streamlined definition of complex marginalized regions in continuous classification datasets. Such a dataset is a set $D = \{(x_i, y_i)\}^n$ of n samples $x_i \in \mathbb{R}^m$ capturing the sample's expression of m attributes $A_1, A_2, ...A_m$ and an associated discrete-valued label y_i.

Sensitive slices, denoted as $s^{A_i}_{n:m}$, describe marginalized slices of the attribute space defined by only one attribute A_i, such as home owners between the age of 20 and 30. An example of a sensitive slice can be seen in Figure 2a. Similar to sensitive attributes, sensitive slices are functions $s^{A_i}_{n:m} : x \mapsto$ [true,false] evaluating to True if x's value of attribute A_i falls within the interval $[n, m]$ and False otherwise.

Sensitive boxes $b = s^{A_j}_{n_j:m_j} \wedge \cdots \wedge s^{A_k}_{n_k:m_k}$ are intersections of multiple sensitive slices on different attributes as exemplified in Figure 2b. Sensitive boxes are functions that evaluate to True if all involved sensitive slices evaluate to True. Sensitive boxes describe hyperrectangular areas in the data space that represent marginalized populations, such as all home owners of age 20 to 30 with an annual income below 30,000$. Sensitive boxes are equivalent to protected groups in the fairness literature.

Sensitive complexes $c = \bigcup_i b_i$ are unions of sensitive boxes, as can be seen in Figure 2c, and represent the case where there are multiple marginalized groups within the same dataset. However, they also provide flexibility to describe more complex shapes than hyperrectangles as the involved boxes b_i can overlap.

The **insensitive subspace** is the collection of all entries that do not belong to any sensitive complex c and can be expressed as the complement $(\cup c)^{\mathrm{C}}$. The insensitive subspace is equivalent to privileged groups. As is common in fairness questions, when assessing bias with SBC representation, we will consider all sensitive complexes and contrast them to the insensitive subspace.

We use the SBC hierarchy to generate artificial bias for the previously generated datasets. To exhibit control, we introduce a parameter to the bias generation process: the number of samples that fall within the marginalized regions. We introduce and use two algorithms to find marginalized subgroups following SBC that meet the desired group size and prevalence: *Patch sensitive subspace generation* partitions the feature space into a large number of small boxes and then greedily combines them into a sensitive complex that meets the target group size. This results in a fragmented set of hyperrectangles. In contrast, *continuous sensitive subspace generation* searches for a single sensitive box. It randomly chooses the number of involved attributes as well as the attributes, and then adjusts the intervals until the target size is met.

Additional details for both bias generation procedures are provided in the supplementary materials alongside examples. During the bias generation for our meta-database, we alternate between both procedures.

3.3 Meta-Features

To be able to assess bias in datasets of varying shapes, sizes and characteristics, we need a fixed-length transferable feature representation of the datasets. We are particularly interested in understanding which kinds of meta-features capture information on bias to inform future research. Therefore, we include a broad range of meta-features:

- **Complexity measures** [25] are a set of data descriptors capturing the complexity of the classification task. They include feature-based, linearity, neighborhood, network, dimensionality and class imbalance measures.
- **Landmarking features** [26] observe the performance of simple, unoptimized models ("landmarkers") to probe the nature of the dataset.
- **Statistical and information-theoretic features** [26] describe the dataset itself. They include the number of features and classes, the average class entropy, and features capturing the data and class distribution.
- **Model-based features** [26] train an unpruned decision tree model and, similarly to landmarking, use the characteristics of the trained tree to characterize the dataset.

A full list of our meta-features is provided in our supplementary materials.

3.4 Bias Scores

In order to have a consistent categorization and quantification of bias that we can use as prediction targets for the meta-model, we propose three novel dataset-level bias scores for real-valued datasets. These scores inform about how likely machine learning models trained on such datasets will exhibit discriminatory behavior.

Inspired by Jiang et al. [20], we borrow the concept of landmarkers – simple, fast-to-train, and fairness-unaware machine learning models that have shown to be able to

reveal the existence of bias in the dataset they are trained on [27]. If we observe any discriminatory behavior, we can safely conclude that the dataset has considerable levels of bias. Landmarking bias metrics are designed for categorical features and require the definition of protected and privileged groups, which we replace with sensitive and insensitive complexes in the SBC framework. We choose decision nodes[1] as our landmarkers and calculate the following bias scores.

Disparate Parity evaluates the disparity between the number of the landmarker's positive predictions (PP) for samples in a sensitive and insensitive subspace. For each sensitive complex $c \in \mathcal{C}$ in a collection $\mathcal{C}$ of all sensitive complexes defined over this dataset and the corresponding insensitive subspace c^{C}, we define the Disparate Parity for one complex (DP_c) and the entire dataset (DP) as follows:

$$\mathrm{DP}_c = \left| \frac{\mathrm{PP}_c}{|c|} - \frac{\mathrm{PP}_{c^{\mathrm{C}}}}{|c^{\mathrm{C}}|} \right|; \qquad \mathrm{DP} = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \mathrm{DP}_c,$$

where $|c|$ counts the number of samples falling within c. In the case of multiple classes, we split the class labels into separate sets of positive and negative labels or evaluate DP individually for all labels in a one-vs-rest fashion.

Similarly, **Equal Opportunity** evaluates the disparity between the number of the landmarker's true positive predictions (TP) on sensitive and insensitive subspaces:

$$\mathrm{EQ}_c = \left| \frac{\mathrm{TP}_c}{|c|} - \frac{\mathrm{TP}_{c^{\mathrm{C}}}}{|c^{\mathrm{C}}|} \right|; \qquad \mathrm{EQ} = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \mathrm{EQ}_c.$$

Group Fairness, on the other hand, adopts the idea of fairness in terms of equal benefits of model outcomes as in the general entropy index proposed by Speicher et al. [28][2]. It assumes that larger label values represent larger benefits and treats unfairness as inequality in benefits gained from the models' prediction values. A deficit in benefits is defined as $\Delta b_i := \hat{y}_i - y_i + 1$ for sample (x_i, y_i) with predicted label $\hat{y}_i$. We denote the average benefit deficit within a sensitive complex and the entire dataset with μ_c and μ, respectively. As before, we define Group Fairness on a complex-level (GF_c) and for the entire dataset (GF) as follows:

$$\mathrm{GF}_c = \frac{1}{2|c|} \sum_{(x_i, y_i) \in c} \left(\frac{\Delta b_i}{\mu_c} \right)^2$$

$$\mathrm{GF} = \frac{1}{2|D|} \sum_{c \in \mathcal{C} \cup \mathcal{C}^{\mathrm{C}}} |c| \cdot (2\theta_c \mathrm{GF}_c + \theta_c - 1); \qquad \theta_c = \left(\frac{\mu_c}{\mu} \right)^2,$$

where D denotes the dataset. Note that for Group Fairness, we sum over all complexes in the dataset, including the insensitive one.

[1] For classification datasets with more than two classes, suitable landmarkers are decision trees with the lowest maximum depth such that they can predict all possible label values.

[2] Our definition uses an α value of 2

All of the above bias scores are calculated for each synthetic dataset using the generated sensitive subspaces. These scores will be used as the prediction targets for a meta-model.

3.5 Meta-Model

For each generated dataset, we generate one or more biased datasets and represent them as a vector each using the data and bias generation parameters (`GEN`), the meta-features (`META`), and the bias scores (`BIAS`). A meta-model will be trained on this meta-database to take the meta-features as input and predict the bias scores, i.e.,

$$[\texttt{META}] \rightarrow [\texttt{BIAS}],$$

which is formulated as a multi-target regression problem. The generation parameters are, of course, unavailable for real-world application cases and will only be used in our experiments to further investigate the strengths and weaknesses of our meta-model.

To predict the bias of a previously unseen dataset, we need to extract its meta-features and apply the meta-model to obtain bias score predictions.

One crucial aspect when creating a meta-database is to achieve diversity to best enhance the generalizability of the meta-model. This is important in the context of computational costs – as is typical for meta-learning, the highest computational effort lies in creating the meta-database. Making a prediction regarding the bias of a previously unseen dataset is comparatively inexpensive. Therefore, if the meta-database is diverse and the meta-model generalizes well, these costs occur only once.

4 Experiments and Results

To thoroughly investigate the ability of our meta-features to inform about bias, we divide our experiments according to the following research questions: (i) Do meta-features capture information regarding bias? (ii) Can meta-features be used to quantify bias? (iii) Can interactions between bias types help enhance predictive performance? (iv) Which dataset characteristics or meta-features are particularly relevant to quantify bias? (v) Case study: Can our meta-features confirm expert-identified bias in well-researched datasets? Lastly, we quantify the compute time required to populate the meta-database and to use it to predict a dataset's bias risk. All materials required to reproduce the presented results, our populated meta-database, raw result files, and additional results are provided in our repository[3]. The experiments are set up as follows.

Meta-Database. For our meta-database, we have generated 2500 smaller and 2500 larger synthetic datasets with randomly drawn parameters as specified in Section 3.1. For each synthetic dataset, we generate four sets of sensitive subspaces and create a corresponding bias as described in Section 3.2. Overall, our meta-database contains a total of 20.000 datasets with a right-skewed distribution of bias scores, as shown in Figure 3.

[3] Our repository: github.com/Hydracerynitis/Assessing-Risk-of-Bias-in-Datasets

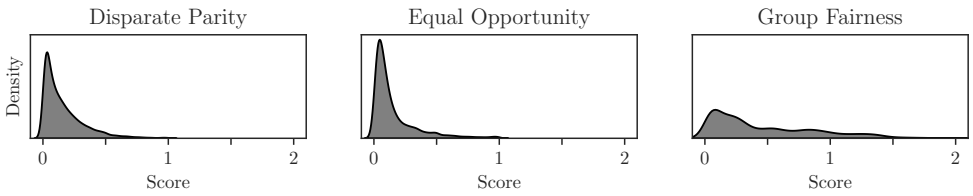


Fig. 3: Distribution of bias scores in our meta-database

Real-World Datasets. We further validate our approach on a range of well-studied real-world datasets, i.e., we include Adult[4], OULAD[5], COMPAS [4], Diabetes[6] and German[7]. We additionally include Materials[8], a dataset from the domain of materials science that has not been studied in the context of fairness before.

Single-Target Models. We include Linear Regression or Linear Support Vector Classification, LASSO Regression, Random Forest, XGBoost, LightGBM, Multi-Layered Perceptron (MLP) and a naïve model predicting the average as a baseline. Hyperparameters are tuned via grid search[9].

Multi-Target Models. For each distinct sequence of bias metrics, we train a RegressorChain model [29], where all predictors are LightGBM models (since we identify them as the strongest single-target models) with the hyperparameters tuned for every target individually. Additionally, we include an ensemble model predicting the average of all RegressorChains ("Ensemble"), and multi-target regression with CLUS [30, 31], which uses predictive clustering trees (PCTs) to simultaneously predict multiple continuous target variables.

Evaluation. We use 10-fold cross-validation to measure the models' performance on the meta-dataset. Hyperparameter tuning takes place individually per fold on a subset of the respective training set.

Metrics. To compare models, we mainly use the *Accuracy* for classification tasks and *Root Mean Squared Error (RMSE)* for regression tasks. Results with additional error metrics can be found in our supplementary materials; however, we reach the same conclusions.

4.1 Do meta-features capture information regarding bias?

To answer the core question of our research, we binarize the bias scores (prediction targets) into low/high using their mean as a splitting point. We train the single-target classifiers listed above to predict whether a bias score is high or low, given the meta-features. We include a simple baseline that predicts the majority class found in the training set. The results in Figure 4 (left) demonstrate that LightGBM, XGBoost and Random Forest are statistically significantly better at detecting bias than the baseline, validating our meta-features. Figure 4 (right) reveals that detecting Group Fairness

[4]Adult dataset: archive.ics.uci.edu/dataset/2/adult
[5]OULAD dataset: analyse.kmi.open.ac.uk/open_dataset
[6]Diabetes dataset: archive.ics.uci.edu/dataset/296/diabetes+130-us+hospitals+for+years+1999-2008
[7]German credit score dataset: archive.ics.uci.edu/dataset/144/statlog+german+credit+data
[8]Materials dataset: doi.org/10.34740/kaggle/dsv/5411884
[9]See our supplementary materials for the specifications of the search grids.

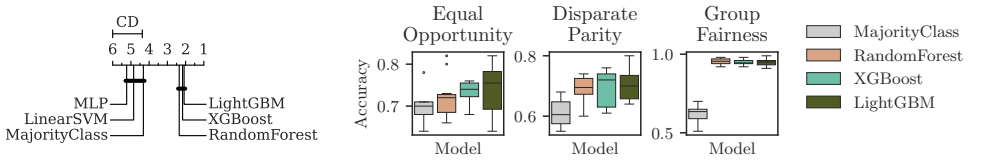


Fig. 4: Classifier performance on binary low/high bias score prediction task. **Left:** Critical Distance plot over all bias scores for non-parametric Friedman with posthoc Nemenyi test [32] under significance level $\alpha = 0.01$. **Right:** Prediction accuracy for individual bias scores.

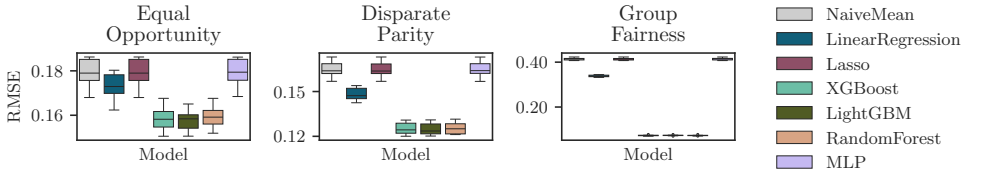


Fig. 5: Root Mean Squared Error (RMSE) for individual models trained and evaluated per bias metric. NaïveMean (gray) is the baseline predicting the average value.

is fairly simple. For Disparate Parity, we obtain moderate improvements in detection accuracy; however, Equal Opportunity violations seem to be harder to detect given our meta-features. Overall, we conclude that the meta-features capture some, but not all, information regarding bias.

4.2 Can meta-features be used to quantify bias?

To quantify bias, we train the regressors specified above to predict individual bias scores from the meta-features.

Figure 5 shows the models' RMSE. We observe that all scores can be predicted better than the baseline (NaïveMean in grey), indicating that there is information regarding fairness violations captured by our meta-features. As before, Group Fairness seems particularly simple to predict. We observe that for all bias metrics, the error values are relatively low in comparison to the score distributions in Figure 3. Overall, XGBoost, LightGBM, and Random Forest show the strongest performance; however, LightGBM has a slightly lower RMSE overall and is less computationally demanding. Therefore, we focus on LightGBM in the remainder of our analysis.

4.3 Can interactions between bias types help enhance predictive performance?

Pearson correlation analysis reveals moderate correlations between some of the bias metrics, as shown in Figure 6. Particularly, Equal Opportunity and Disparate Parity exhibit the strongest correlation, which we explain by the similarity of their definitions:

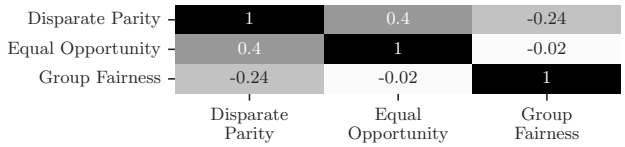


Fig. 6: Pearson correlation between bias scores.

Fig. 7: **Left:** RMSE of all RegressorChains averaged over all bias scores, sorted by increasing RMSE. **Right:** Critical Distance plot comparing single- and multi-target regression on all bias scores for non-parametric Friedman with posthoc Nemenyi test [32] under significance level $\alpha = 0.01$.

Both are concerned with the positive predictions of landmarkers with Equal Opportunity focusing only on the correct positive predictions. Group Fairness and Equal Opportunity show very little correlation, while Group Fairness and Disparate Parity are weakly negatively correlated. This may be because Group Fairness additionally takes the landmarkers' negative predictions while other bias metrics do not.

To capitalize on these correlations, we use RegressorChains [29] as multi-target models. A RegressorChain is a sequence of regression models $M_0, \ldots, M_n$, where each model M_i in the chain predicts one bias score s_i from both the meta-features and its predecessors' predicted bias scores $s_0, \ldots, s_{i-1}$.

We first investigate the effectiveness of the RegressorChains. We train one chain for each permutation of the bias scores. Figure 7 (left) demonstrates that the order matters only moderately – for Group Fairness, two permutations result in the strongest RegressorChains before there is a bump in RMSE. We observe a similar effect for Disparate Parity. However, when predicting all targets, no single chain seems superior.

Since the order matters, we train an ensemble of RegressorChains to combine all chains into one final prediction by averaging. Similarly, we train CLUS [30, 31], which predicts all targets simultaneously rather than chaining them. Figure 7 (right) shows an overall ranking of these multi-target approaches compared to the single-target regressors. We observe that while the best chain per bias score ("BestChain") is statistically equivalent with the single-target models, it does not improve upon their performance. The same holds for the ensemble approach and CLUS. Although there is a correlation among bias scores, this behavior indicates that the meta-features are sufficient for the chains to not consider the other scores.

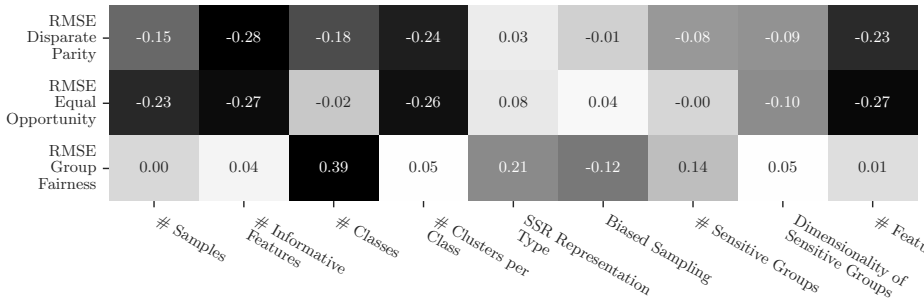


Fig. 8: Pearson correlation of dataset and bias generation characteristics to RMSE when predicting the dataset's bias scores.

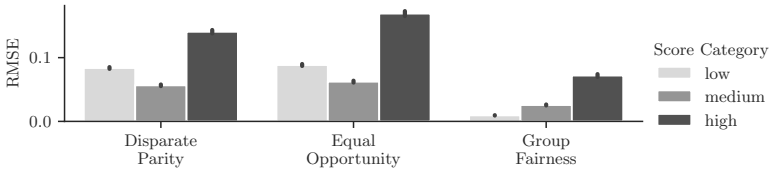


Fig. 9: RMSE on datasets with low/medium/high bias, separated by different bias scores

4.4 Which dataset characteristics or meta-features are particularly relevant to quantify bias?

In order to investigate the impact of certain characteristics of synthetic datasets and biases, we calculate Pearson correlation coefficients between these characteristics and the RMSE obtained when predicting bias scores for these datasets with LightGBM. Figure 8 illustrates the results. We observe a negative correlation between dataset size and error, which might stem from meta-features being statistically more robust on larger datasets. A higher dimensionality reduces the error. This is due to bias in only a small subset of the dimensions often being less pronounced, falling into the low bias category where our performance is generally better. A similar argument can be made for the number of classes or clusters per class. One exception is Group Fairness, for which the number of classes correlates strongly positively with the prediction error. Since Group Fairness expects to encode classes so they reflect how beneficial each class is, as the number of classes increases, the difference of benefit across entries in the dataset also increases. This results in a harder prediction task for our meta-model.

Next, we investigate how the strength of the induced bias affects the predictability of said bias. To achieve this, we divide the meta-data, for each score separately, into three quantile-based bins of approximately equal size based on the bias scores, i.e., into data with low/medium/high bias. For each bin, we assess the RMSE of a tuned LightGBM model trained on the entire dataset separately. Figure 9 shows that

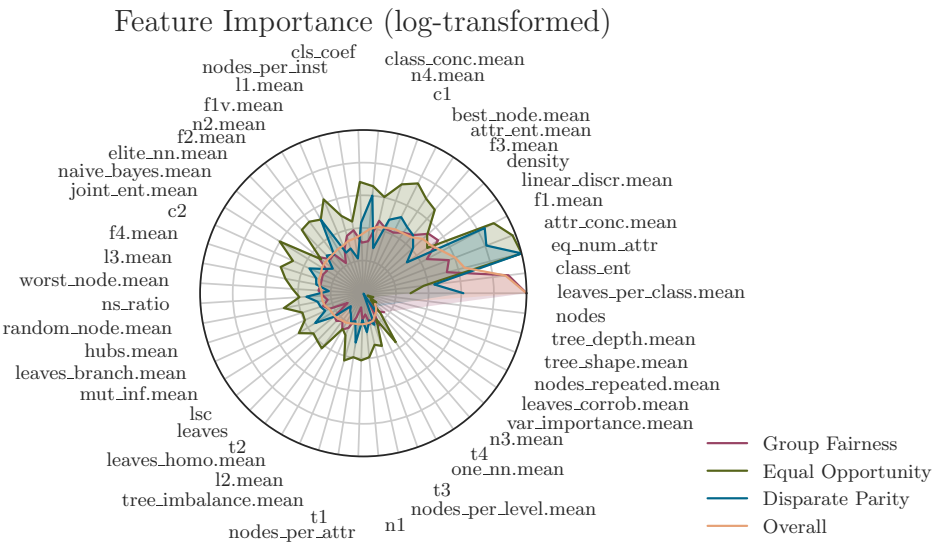


Fig. 10: Log-transformed feature importances produced by CLUS, sorted by overall feature importances. Outer rings imply higher importances. Importances are normalized per target such that the most important feature touches the outermost ring.

our meta-learner performs substantially better on quantifying low to medium bias than high bias. As shown in Figure 3, this reflects the bias distribution in our meta-database. This emphasizes the need to dedicate further research into the detection and quantification of strong bias in future work. Particularly, further strongly-biased datasets should be included in the meta-database or weighting approaches during training should place more emphasis on higher scores.

Finally, we study the importance of the meta-features. Although it is not the strongest model, we use CLUS as it can obtain both feature importance scores for individual bias scores but also for their simultaneous prediction with Genie3 [33]. Note that the single-target features CLUS identifies are echoed by LightGBM-based feature importance. We provide those results in our supplementary materials, alongside the legend of all feature names. Figure 10 shows the CLUS-based feature importances.

We observe that while Disparate Parity and Equal Opportunity use a large number of features, Group Fairness relies heavily on "leaves_per_class", the average number of leaves dedicated to a class, and "class_ent", the features' average Shannon entropy. Group Fairness seems to be strongly impacted by the number of classes (see Figure 8), and the number of leaves per class may be an adequate proxy. Since the importance of these two features is comparatively high for predicting Group Fairness, they also dominate the overall feature importance. The next three overall most important features are those essential to predicting Disparate Parity and Equal Opportunity: "eq_num_attr", the number of equivalent attributes for a predictive task [34], "attr_conc", the average concentration coefficient between pairs of features [35], and "f1", the maximum Fisher discriminant ratio measuring the overlap of features in different classes [25].

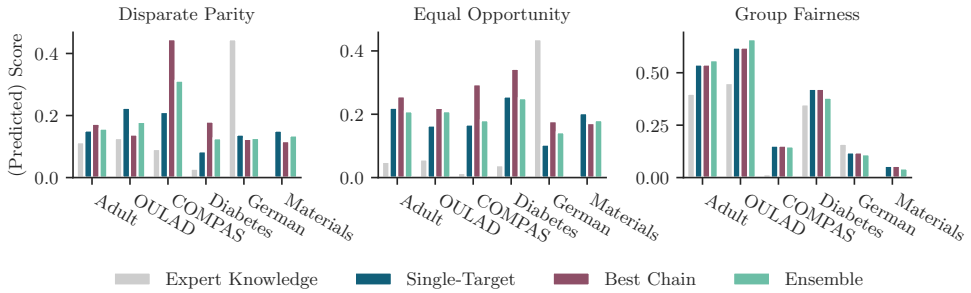


Fig. 11: Predictions of meta-models on real world datasets' bias metrics compared to those calculated from expert knowledge

Other features standing out for the prediction of Disparate Parity are "class_conc", the concentration coefficient between the features and the target [35], and "n2", the average ratio of intra/inter class nearest neighbor distance [25]. Prediction of Equal Opportunity considers a blend of most features. Overall, the least important features are the ones generally describing the shape of the landmarker trees.

4.5 Case Study: Can our meta-features confirm expert-identified bias in well-researched datasets?

To understand the effectiveness of our proposed framework with respect to real-world applications, we apply our meta-learning models on the well-studied datasets specified above. The goal of our case study is to assess whether bias metrics predicted by our meta-model match the known bias. We therefore express the following bias types in our SBC framework: (i) *Racial Bias* against non-white individuals, (ii) *Gender Bias* against females, (iii) *Age Bias* against individuals over 65 years, (iv) *Disability Bias* against disabled individuals (for OULAD), (v) *Marital Status Bias*, where a married female person is less favorable than a single female, and a married male is more favorable than a single male [36] (for German), and (vi) *Foreigner Status Bias* against foreigners (for German). The corresponding bias scores yield the "expert knowledge" and are considered the ground truth in this analysis. Our meta-models shall predict bias scores that are as close as possible to this ground truth.

Figure 11 shows that all our meta-models, single-target LightGBM, the best RegressorChain, the average of all RegressorChains and the ensemble of RegressorChains, predict scores that are relatively close to the expert knowledge, with a few exceptions. All meta-models fail to predict Disparate Parity and Equal Opportunity for the German dataset accurately. This is likely because the German bias scores are unusually high compared to our training dataset (see Figure 8). For the other datasets, all meta-models are overestimating Equal Opportunity, which could point to additional bias that has not yet been explored by the literature, especially on the Diabetes dataset, where numeric attributes are included. Overall, the meta-learning approach seems to generalize reasonably well to real-world datasets.

In addition to the well-researched datasets from the area of fair ML, we applied our meta-model to "Materials", a real-world dataset containing mechanical properties of different metals. While we cannot confirm or deny the produced results, our meta-model detects high Disparate Parity and Equal Opportunity scores in this dataset, implying that there are some regions in the feature space with a high risk of overly positive predictions and others being neglected – a warning for practitioners to assess their models more closely when trained on this dataset.

4.6 Computational Requirements

Extracting meta-features is computationally intensive, with both runtime and memory usage growing exponentially with dataset size. On average, extraction takes 283.48 seconds for large datasets and 36.00 seconds for small ones. Similarly, generating bias complexes becomes more demanding with increasing feature dimensionality; patch-like biases require more time than continuous ones. Bias generation averages 9.77 seconds for large datasets and 4.27 seconds for small ones. Overall, generating the complete meta-dataset took approximately 261 hours. Training meta-models is relatively efficient: linear models take under 1 second, while XGBoost, LightGBM, Random Forest, and MLP require 11.71, 2.23, 0.92, and 10.45 seconds respectively per model. However, hyperparameter tuning using 5-fold grid search across all combinations increases total training time substantially—up to 73 minutes for XGBoost. Inference is fast, but extracting meta-features for new datasets remains the dominant computational bottleneck. All experiments were conducted on the same device, equipped with an Intel(R) Core(TM) i5-10500H CPU @ 2.50 GHz, NVIDIA GeForce GTX 1650 GPU, integrated Intel(R) UHD Graphics, and 16 GB of RAM.

5 Conclusions

Dataset bias will likely be propagated into machine learning models and cause them to make unfair, discriminatory, or simply wrong predictions. Current approaches to ensure fairness in machine learning focus on categorical features in personal data such as age groups, gender, or race and demand preliminary suspicion of a concrete bias. However, these requirements are not always met as datasets may mask sensitive attributes for privacy reasons (e.g., drug response prediction tasks may focus on cancer cell mutations [12] rather than the patient demographics, which have been omitted due to privacy concerns [37]), or they may not contain human-centered data at all such as in our materials science case study.

In this paper, we introduced a novel method to detect and quantify previously unseen discriminatory bias in datasets with numeric, non-discrete, continuous features in an automated manner using meta-learning. In the process, we contributed several novel artifacts: First, we extended current definitions of bias to continuous datasets so that we can investigate more tangential and composite discriminatory bias. Second, we propose several dataset-level bias metrics independently of a specific choice of model to quantify potential risks of bias for practitioners. Third, based on the new bias definitions for real-valued features, we introduce a novel bias generator, which we use

for the generation of a well-fed meta-database. The bias generator could also support benchmarking the future bias detection or mitigation approaches we hope to motivate.

Our proposed meta-learning approach shows promise – our comprehensive set of experiments confirms that bias can indeed be detected, even if not perfectly quantified, highlighting that meta-features do carry information on unseen bias. However, we have yet to fully understand how these meta-features convey information related to bias. In addition, as our generated meta-dataset mostly consists of low-bias-level datasets, our meta-learning models tend to underestimate bias levels in our experiments.

In future research, we will incorporate a more diverse range of bias types and further explore how dataset complexity affects the detectability of bias. Additionally, we will develop techniques to inform practitioners further about the nature of the identified bias.

Statements and Declarations

Funding. Dost is supported by the European Union's Horizon Europe research and innovation programme under the Marie Skłodowska-Curie Postdoctoral Fellowship Programme, SMASH co-funded under the grant agreement No. 101081355. The SMASH project is co-funded by the Republic of Slovenia and the European Union from the European Regional Development Fund. Džeroski is supported by the Slovenian Research and Innovation Agency (under grant P2-0103).

Competing Interests. Kim is employed by Callaghan Innovation, a Crown Research Institute in New Zealand.

Ethics Approval and Consent to Participate. Not applicable.

Consent for Publication. Not applicable.

Data, Materials & Code Availability. All data, materials, code, results, scripts, and supplementary materials are available in our repository: github.com/Hydracerynitis/Assessing-Risk-of-Bias-in-Datasets.

Author Contribution. All authors contributed to the study conception and design. Implementation and experiments were performed by Dai. The first draft of the manuscript was written by Dai, Dost, and Kim. All authors commented on previous versions of the manuscript, and they read and approved the final manuscript.

References

[1] Mukerjee, A., Biswas, R., Deb, K., Mathur, A.P.: Multi–objective evolutionary algorithms for the risk–return trade–off in bank loan management. International Transactions in Operational Research **9**, 583–597 (2002)

[2] Faliagka, E., Ramantas, K., Tsakalidis, A., Tzimas, G.: Application of machine learning algorithms to an online recruitment system. In: Proc. International Conference on Internet and Web Applications and Services, pp. 215–220 (2012)

[3] Yeh, I.-C., Lien, C.-h.: The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Systems with Applications **36**, 2473–2480 (2009)

[4] Julia Angwin, S.M. Jeff Larson, Lauren Kirchner, P.: Machine Bias (2016). www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing Accessed 2024-07-24

[5] Lambrecht, A., Tucker, C.: Algorithmic bias? an empirical study of apparent gender-based discrimination in the display of STEM career ads. Management Science **65**, 2966–2981 (2019)

[6] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ACM Computing Surveys **54** (2021)

[7] Bellamy, R.K., Dey, K., Hind, M., Hoffman, S.C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilović, A., *et al.*: AI fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. IBM Journal of Research and Development **63**, 4–1 (2019)

[8] Saleiro, P., Kuester, B., Hinkson, L., London, J., Stevens, A., Anisfeld, A., Rodolfa, K.T., Ghani, R.: Aequitas: A bias and fairness audit toolkit. arXiv preprint arXiv:1811.05577 (2018)

[9] Chen, F., Wang, L., Hong, J., Jiang, J., Zhou, L.: Unmasking bias in artificial intelligence: a systematic review of bias detection and mitigation strategies in electronic health record-based models. Journal of the American Medical Informatics Association **31**, 1172–1183 (2024)

[10] Dost, K., Taskova, K., Riddle, P., Wicker, J.: Your best guess when you know nothing: Identification and mitigation of selection bias. In: Proceedings of the 2020 IEEE International Conference on Data Mining, pp. 996–1001 (2020)

[11] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ACM computing surveys (CSUR) **54**(6), 1–35 (2021)

[12] Adam, G., Rampášek, L., Safikhani, Z., Smirnov, P., Haibe-Kains, B., Goldenberg, A.: Machine learning approaches to drug response prediction: challenges and recent progress. NPJ precision oncology **4**(1), 19 (2020)

[13] Hort, M., Chen, Z., Zhang, J.M., Harman, M., Sarro, F.: Bias mitigation for machine learning classifiers: A comprehensive survey. ACM Journal on Responsible Computing **1**, 1–52 (2024)

[14] Feldman, M., Friedler, S.A., Moeller, J., Scheidegger, C., Venkatasubramanian, S.: Certifying and removing disparate impact. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 259–268 (2015)

[15] Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning.

Advances in Neural Information Processing Systems **29** (2016)

[16] Caliskan, A., Bryson, J.J., Narayanan, A.: Semantics derived automatically from language corpora contain human-like biases. Science **356**(6334), 183–186 (2017)

[17] Alelyani, S.: Detection and evaluation of machine learning bias. Applied Sciences **11**, 6271 (2021)

[18] Moreno-Torres, J.G., Raeder, T., Alaiz-Rodríguez, R., Chawla, N.V., Herrera, F.: A unifying view on dataset shift in classification. Pattern Recognition **45**, 521–530 (2012)

[19] Azzalini, F., Criscuolo, C., Tanca, L.: E-fair-db: Functional dependencies to discover data bias and enhance data equity. Journal of Data and Information Quality **14** (2022)

[20] Jiang, L., Belitz, C., Bosch, N.: Synthetic dataset generation for fairer unfairness research. In: Proceedings of the 14th Learning Analytics and Knowledge Conference, pp. 200–209 (2024)

[21] Simonetta, A., Trenta, A., Paoletti, M.C., Vetrò, A., *et al.*: Metrics for identifying bias in datasets. In: International Conference of Yearly Reports on Informatics Mathematics and Engineering, pp. 10–17 (2021)

[22] Le Quy, T., Roy, A., Iosifidis, V., Zhang, W., Ntoutsi, E.: A survey on datasets for fairness-aware machine learning. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **12**, 1452 (2022)

[23] Jesus, S., Pombal, J., Alves, D., Cruz, A., Saleiro, P., Ribeiro, R., Gama, J., Bizarro, P.: Turning the tables: Biased, imbalanced, dynamic tabular datasets for ml evaluation. Advances in Neural Information Processing Systems **35**, 33563–33575 (2022)

[24] Barbierato, E., Vedova, M.L.D., Tessera, D., Toti, D., Vanoli, N.: A methodology for controlling bias and fairness in synthetic data generation. Applied Sciences **12**, 4619 (2022)

[25] Lorena, A.C., Garcia, L.P., Lehmann, J., Souto, M.C., Ho, T.K.: How complex is your classification problem? a survey on measuring classification complexity. ACM Computing Surveys **52**, 1–34 (2019)

[26] Vilalta, R., Giraud-Carrier, C.G., Brazdil, P., Soares, C.: Using meta-learning to support data mining. International Journal of Computer Science & Applications **1**, 31–45 (2004)

[27] Balte, A., Pise, N., Kulkarni, P.: Meta-learning with landmarking: A survey. International Journal of Computer Applications **105** (2014)

[28] Speicher, T., Heidari, H., Grgic-Hlaca, N., Gummadi, K.P., Singla, A., Weller, A., Zafar, M.B.: A unified approach to quantifying algorithmic unfairness: Measuring individual & group unfairness via inequality indices. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 2239–2248 (2018)

[29] Read, J., Pfahringer, B., Holmes, G., Frank, E.: Classifier chains for multi-label classification. Machine Learning **85**, 333–359 (2011)

[30] Struyf, J., Džeroski, S.: Constraint based induction of multi-objective regression trees. In: Knowledge Discovery in Inductive Databases, pp. 222–233 (2006)

[31] Petković, M., Levatić, J., Kocev, D., Breskvar, M., Džeroski, S.: CLUSplus: A decision tree-based framework for predicting structured outputs. SoftwareX **24**, 101526 (2023)

[32] Demšar, J.: Statistical comparisons of classifiers over multiple data sets. The Journal of Machine Learning Research **7**, 1–30 (2006)

[33] Petković, M., Kocev, D., Džeroski, S.: Feature ranking for multi-target regression. Machine Learning **109**(6), 1179–1204 (2020)

[34] Michie, D., Spiegelhalter, D.J., Taylor, C.C., Campbell, J.: Machine Learning, Neural and Statistical Classification. Ellis Horwood, USA (1995)

[35] Kalousis, A., Hilario, M.: Model selection via meta-learning: a comparative study. In: Proceedings 12th IEEE Internationals Conference on Tools with Artificial Intelligence, pp. 406–413 (2000)

[36] Jordan, A.H., Zitek, E.M.: Marital status bias in perceptions of employees. Basic and Applied Social Psychology **34**, 474–481 (2012)

[37] Smirnov, P., Kofia, V., Maru, A., Freeman, M., Ho, C., El-Hachem, N., Adam, G.-A., Ba-Alawi, W., Safikhani, Z., Haibe-Kains, B.: Pharmacodb: an integrative database for mining in vitro anticancer drug screening studies. Nucleic acids research **46**(D1), 994–1002 (2018)