Reduce model unfairness with maximal correlation based fairness optimization

Wenxuan Huang

Master thesis

February 2022

TU Delft

Reduce model unfairness with maximal correlation based fairness optimization

by

Wenxuan Huang

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Friday February 25, 2022 at 11:00 AM.

Student number: 4925777
Project duration: November 26, 2020 – February 25, 2022

Thesis committee:		
	Prof.dr. Jan van Gemert,	TU Delft
	Prof.dr. Christoph Lofi,	TU Delft
	Prof.dr. Marco Loog,	TU Delft, Supervisor

TU Delft

Preface

You are reading the thesis “Reduce model unfairness with maximal-correlation-based fairness optimization”. This paper describes a framework that identifies the source of model unfairness and alleviates its influence. The research paper is a part of the master thesis to fulfil the requirement of the Computer Science master’s degree.

For me personally, this thesis is more than a description of the method. In the name of efficiency and predictability, institutions pleasingly replaced human workflow with machine learning algorithms. In my short experience of helping LGBTQ refugees in the Netherlands, I witnessed their confusion when they had to evidentiarily prove that they belong to the Queer community. As underprivileged individuals, there is a sense of powerlessness against a metal-cold adversary that degrades him/her as merely a set of features. Will there be a prospective algorithm that predicts whether someone is “sufficiently” gay?

During this project, Marco Loog, as daily supervisor, provides me with countless feedback for conducting proper research and method validation. I want to thank him for his guidance in more than 80 emails, many hours of meetings and written advice. He made me feel safe when there were difficulties in conducting experiments and when I was alone in the situation of a ruthless pandemic. I want to thank Jan van Gemert and Christoph Lofi for their guidance in the form of emails and research guidelines. At last, I want to thank my families for their unwavering support.

I hope you enjoy reading my thesis.

Wenxuan Huang
Feb 2022

Reduce model unfairness with maximal-correlation-based fairness optimization

Wenxuan Huang
w.huang-10@student.tudelft.nl
Technical University of Delft
Delft, The Netherlands

ABSTRACT

Supervised machine learning is a growing assistive framework for professional decision-making. Yet bias that causes unfair discrimination has already been presented in the datasets. This research proposes a method to reduce model unfairness during the machine learning training process without altering the sample value or the prediction value. Using an objective function that identifies the biased feature with maximal correlation estimation, the method selects samples to train the updated classifier model. The quality of the sample selection determines the extent of unfairness reduction. With an adequate sample size, we demonstrate that the method is valid in reducing model unfairness without severely sacrificing classification accuracy. We tested our method on multiple benchmark datasets with demographic parity and feature independence as the notions for a statistically fair classification model.

KEYWORDS

Demographic Parity, Independence, Maximal Correlation, Sensitive Feature, Objective Function

1 INTRODUCTION

Supervised machine learning, accompanied with feature-rich or sample-rich datasets, became a novel framework to generate statistical insight & prediction [1]. In growing cases, it influences and assists the professional decision-making process and arbitrates the outcome of critical applications including loan acceptance or chance of bail and parole [2, 3]. Yet supervised machine learning is susceptible to dataset biases that originate from data collection [4] or model goal [5] because the dataset biases could be reinforced into the model [6]. The reinforcement begins when the machine learning model generates functions that map dataset features and ground truth value during the training process, and then employs feature-label patterns to facilitate prediction. Label, in this paper, indicates ground truth given in the dataset. One instance of dataset biases is unintended associations between classifier predictions and the sensitive feature, since predictions using the information in this feature lead to unequal rates of the outcome proportion for different sensitive classes, suggested by cases in Zliobaite's survey on indirect data discrimination [7]. For instance, some African American criminal offenders have a higher risk score for the potential to re-offend though they committed lighter crimes, compared to Caucasian offenders [8]. The determination of sensitive features is contextual. Common choices include inherent attributes like race, gender, and sexual orientation.

This research aims to minimize chances where a classifier delivers unfair model predictions. The fairness notion we adopt is independence. Independence between sensitive feature and classifier prediction states that the prediction does not use the information from sensitive feature[9]. Demographic parity (i.e. DP), as a mathematical interpretation of independence, pursues an equal proportion of positive prediction outcomes for each sensitive group [9]. The fairness in DP is defined by the consistency of the prediction's positivity-negativity ratio per sensitive group. Classifier models with a perfect DP assume an equal rate of positive prediction over different sensitive groups [10]. In Barocas's review of fairness notion [9], for a model with demographic parity, classifier prediction and sensitive feature are assumed to be statistically independent and their probabilities of occurrence are not affected by each other). In recidivism prediction, a classifier satisfies this fairness criterion if both African-American and Caucasian groups have an equal probability of being assigned to the positive recidivist classification.

We propose a method to improve model fairness by reducing dependence between the sensitive feature and model predictions. Calders' research and several others on anti-discriminatory machine learning similarly approach fairness through modifying dependence between sensitive feature and model prediction [11–13]. These researches did not consider the correlation between the sensitive feature and other non-sensitive features in the same dataset. Non-sensitive features collected along the sensitive feature can inherit the bias of the sensitive feature through an associative mapping [14, 15]. Hoffmann's research on the intersectionality of data biases shows that features other than racial identity are similarly biased with race-related assumptions [16]. Our unfairness reduction method identifies biases from prominent non-sensitive feature and reduces the dependence between biased non-sensitive feature and the model prediction. Here the method uses maximal correlation powered by the ACE algorithm to estimate the dependence between the two above-mentioned variables. Maximal correlation is a population correlation that expresses the degree of association between two variables [17]. By estimating the maximal correlation between non-sensitive features and sensitive feature, our method targets non-sensitive feature that implicitly reinforces unfairness onto model predictions. On the other hand, the estimation of maximal correlation between biased feature and model prediction provides insight into the negative impact of this feature over the prediction value. These two maximal correlation measures are core to the fairness constraint objective function the method uses to reduce model unfairness.

If a non-sensitive feature has a higher maximal correlation with the sensitive feature, our method identifies it as the biased feature. Belitz et al.'s research use forward feature selection to procedurally remove biased features from the training process [18]. Selecting

only non-biased features to train models, however, could potentially weaken the classification accuracy of the model severely if the dataset features are intertwined with redundant encoding or inexplicit connections [19]. Therefore, Belitz et al.'s feature-based approach could risk over-sacrificing the classifier's performance. In the logistic regression training process, the feature weight of a biased feature determines the degree of impact the biased feature can have on the classifier's prediction. Since logistic regression is a general linear model, the model prediction is a result of a linear operation of feature weights and corresponding feature values. Kamiran's research suppresses biased feature's impact on model classifier by directly changing ground-truth value (i.e. dataset massaging). This triggers logistic regression to re-weight [20]. Kamiran's research provides a straightforward re-weighing technique to alleviate data biases, but modified sample values change the distribution of the dataset, which potentially causes incompatibility for downstream processes including the objective function and model prediction on testing dataset [21].

Our proposed method similarly uses reweighing of the logistic regression to adjust feature importance. Instead of triggering reweigh by modifying sample value, this method only selects a limited number of samples from the validation pool as the classifier trains. The selection requirement of samples is based on the value of an objective function that is fairness constrained. Previously Kamishima's research [15] reduce the weight of biased features externally from the internal working of the logistic regression. Our method similarly would not directly change the weight of the biased feature numerically, since feature weight assignment of logistic regression is not a discrete process. A non-intrusive reweigh also enables the adjustment of weight for other non-sensitive features, and manual weight adjustment could not reproduce this dynamic weight adjustment. After the samples are selected, this method expects the updated classifier, trained with selected samples, to output a fairer model prediction.

1.1 Main contribution

The main contribution of the paper includes:

- An novel objective function to facilitate feature weight reweigh by selecting samples to include in the training dataset for a fairer classifier.
- An in-processing method that reduce the linear and complex dependence between sensitive feature and model prediction supported by maximal correlation measure and ACE estimation.
- A framework to reduce model unfairness by imposing demographic parity as a fairness notion into the machine learning training process.

2 BACKGROUND INFORMATION

This section defines two concepts our research adopts for identifying model unfairness, measuring feature dependence, and fairness notion.

2.1 Demographic parity

The demographic parity notion will assume that the model is fair if the model prediction and the sensitive feature are stochastically independent of each other [9], thus their probability of occurrence is not affected by each other. Mathematically, the burden of proof for fulfilling demographic parity is to prove that:

$$P(\hat{Y} = 1|S = 1) = P(\hat{Y} = 1|S = 0) \tag{1}$$

This is the definition of demographic parity where $\hat{Y}$ is model prediction and S is the sensitive feature. The measure of demographic parity during the training process between two variables is the maximal correlation between the two variables.

2.2 Maximal correlation estimation

The maximal correlation estimation is a measure of association between two variables. The measure of demographic parity between model prediction $\hat{Y}$ and sensitive feature S requires an estimation that can capture independence between the two variables. And the maximal correlation estimation can capture linear, non-linear, and polynomial dependence and independence. Similarly, the maximal correlation between feature X and S provides a numerical measure of whether a feature has a close association with the sensitive feature. This enables the method to select the most biased feature among all non-sensitive features available. In the sample selection process, to examine a sample's effect on feature weight reduction, the maximal correlation of training samples' X feature and y ground truth will be compared before and after the addition of each sample.

Given two variables S and $\hat{Y}$, the maximal correlation between these two variables is:

$$mCor(S, \hat{Y}) = \max_{f,g} Corr(f(S), g(\hat{Y})) \tag{2}$$

where the maximum correlation value is obtained from all functions $f : S \to \mathbb{R}$ and $g : \hat{Y} \to \mathbb{R}$. Function selection aims to capture the most correlated function mappings of two variables. The selected functions extract the most correlated aspect of S and $\hat{Y}$.

If the dependence between two variables is non-linear (e.g. inversely proportional) or polynomial, the estimation of dependence cannot rely on an estimator that only works well linearly. Therefore, we propose to use maximal correlation as the measure of independence. Appropriate estimation of independence between two features or between features and prediction can provide a framework for estimating and reducing unfairness. The appropriateness of maximal correlation is based on its coverage of dependence estimation without restriction on monotonicity or linearity. There are three reasons for using maximal correlation as the measure of bivariate independence:

- Maximal correlation has the desired property to prove independence, while prominent correlation measures (e.g. Pearson's correlation) can only disprove independence. Maximal correlation between two variables is 0 if and only if they are independent [22].
- Maximal correlation accounts for non-linear and polynomial association, which can estimate a wider spectrum of dependence relationships. Maximal correlation is 1 if the two variables are deterministically associated by a function [23].

- Maximal correlation estimation has a $[0, 1]$ range and it is independent to the marginal probabilities of the two variables. This enables the method to use maximal correlation as independence measurement with absolute threshold [23].

The maximal correlation has the following properties, take example of the maximal correlation between $\hat{Y}$ and S [24]:

- $mCorr(S, \hat{Y}) = 0$ if and only if S and $\hat{Y}$ are independent.
- $mCorr(S, \hat{Y}) = 1$ if there exist a dependence function between S and $\hat{Y}$, so that $S = g(\hat{Y})$ or $\hat{Y} = f(S)$.

The maximal correlation is the maximized correlation value of maximal correlation functions $f(S)$ and $g(\hat{Y})$ (i.e. Equation 4). For maximal correlation functions $f(S)$ and $g(\hat{Y})$:

$$f(S) = (f(s_1), ..., f(s_n)), g(\hat{Y}) = (g(\hat{y}_1), ..., g(\hat{y}_n)) \tag{3}$$

where n is the number of samples in sensitive feature S and model prediction $\hat{Y}$ (i.e. n is the same in both case).

The maximal correlation functions have the following prerequisites to ensure output of the maximal correlation function is centered and scaled:

- Centering: $\bar{f}(S) = \bar{g}(\hat{Y}) = 0$
- Scaling: $f(S) \cdot f^T(S) = g(\hat{Y}) \cdot g^T(\hat{Y}) = \mathbf{I}$

Therefore, based on Equation 3 and the prerequisite above, the maximal correlation between S and $\hat{Y}$ is:

$$mCor(S, \hat{Y}) = \max_{\substack{\bar{f}(S)=\bar{g}(\hat{Y})=0 \\ \bar{f^2}(S)=\bar{g^2}(\hat{Y})=1}} f(S)g(\hat{Y}) \tag{4}$$

To estimate the maximum of $f(S)g(\hat{Y}))$, maximal correlation converts $(f(S) - g(\hat{Y}))^2$ to a linear component composing $f(S)g(\hat{Y})$. The conversion suggests that the objective of maximizing $f(S)g(\hat{Y})$ is equivalent to the objective of minimizing $(f(S) - g(\hat{Y}))^2$ (see in Equation 7).

$$\begin{aligned}(f(S) - g(\hat{Y}))^2 &= f^2(S) + g^2(\hat{Y}) - 2(f(S)g(\hat{Y})) \\ &= 2 - 2(f(S)g(\hat{Y}))\end{aligned} \tag{5}$$

Hence, the maximal correlation function $f()$ and $g()$ are the optimal function for maximal correlation estimation in Equation 5 if and only if they are optimal in the following optimization problem:

$$\min_{\substack{\bar{f}(S)=\bar{g}(\hat{Y})=0 \\ \bar{f^2}(S)=\bar{g^2}(\hat{Y})=1}} (f(S) - g(\hat{Y}))^2 \tag{6}$$

In case of a fixed $f(S)$, the optimization problem is to minimizing over function $g(\hat{Y})$:

$$\begin{aligned}&\min_{g(\hat{Y})}[(f(S) - g(\hat{y}))^2 | \hat{Y} = \hat{y}] \\ &\Rightarrow g(\hat{y}) = f(S) | \hat{Y} = \hat{y} \\ &\Rightarrow g(\hat{Y}) = f(S) | \hat{Y}\end{aligned} \tag{7}$$

If the prerequisite of zero mean and norm being one is considered, the maximal correlation function $g(\hat{Y})$ is:

$$g(\hat{Y}) = \frac{f(S)|\hat{Y}}{\sqrt{(f(S)|\hat{Y})^2}} \tag{8}$$

In case of a fixed $g(\hat{Y})$, the optimization problem is to minimizing over function $f(S)$:

$$\begin{aligned}&\min_{f(S)}[(f(s) - g(\hat{Y}))^2 | f(S) = s] \\ &\Rightarrow f(s) = g(\hat{Y}) | S = s \\ &\Rightarrow f(S) = g(\hat{Y}) | S\end{aligned} \tag{9}$$

If the prerequisite of zero mean and norm being one is considered, the maximal correlation function $f(S)$ is:

$$f(S) = \frac{g(\hat{Y})|S}{\sqrt{(g(\hat{Y})|S)^2}} \tag{10}$$

This method uses ACE (alternating conditional expectation) algorithm to estimate maximal correlation (shown in Algorithm 1).

Algorithm 1 Alternating conditional expectation algorithm

1: **procedure** ACE
2: $f_0(S) \leftarrow \frac{S-\bar{S}}{\sqrt{(S-\bar{S})^2}}$
3: **for** $k = 1, 2, ...,$ till $f_k(S)g_k(\hat{Y}) = f_{k-1}(S)g_{k-1}(\hat{Y})$ **do**
4: $g_k(\hat{Y}) = \frac{f_{k-1}(S)|\hat{Y}}{\sqrt{(f_{k-1}(S)|\hat{Y})^2}}$
5: $f_k(S) = \frac{g_k(\hat{Y})|S}{\sqrt{(g_k(\hat{Y})|S)^2}}$
6: $mCorr(S, \hat{Y}) = f_k(S)g_k(\hat{Y})$

3 METHOD

Our research aims to reduce unfairness of the model prediction and expects the updated classifier to produce predictions that have a minimum maximal correlation with the sensitive feature. Anahideh's research on fairness [25] proved that the covariance between sensitive feature S and model prediction $\hat{Y}$ attributes to both the feature (non-sensitive)-feature (sensitive) covariance and the feature weight of non-sensitive features. This research similarly attributes maximal correlation between features and model prediction to (Specified in Appendix A):

- The maximal correlation between feature i in X and sensitive feature: $mCorr(X_i, S)$.
- The feature weight of feature X_i: θ_i.

Maximal correlation between feature X_i and sensitive feature S (i.e. $mCorr(X_i, S)$) estimates the dependence between them. A higher maximal correlation value suggests a larger deviation from the concept of independence, thus also further away from the notion of demographic parity. Therefore, this method focuses on feature i with the highest value of $mCorr(X_i, S)$ among all feature in X, since this feature has, among all, the strongest association with the sensitive feature (i.e. strong biases) and this association is a factor determining biases of the model prediction.

Once we identify feature X_i that has a higher association with the sensitive feature, the method evaluates the value of the feature weight θ_i in the current classifier. If the feature weight θ_i is higher compared to other features, it will be the unfairness reduction target. This is because the pattern of X_i is highly predictive of ground truth

Y in logistic regression. The predictability of this feature is high since the model prediction of the logistic classifier is based linearly on feature weights, thus higher weight θ_i increases the proportion of X_i in the constitution of the prediction $\hat{Y}$, shown in Equation 11.

$$\hat{y}_i = log_b \frac{p_i}{1-p_i} = \theta_0 + \theta_1 X_1 + \cdots + \theta_d X_d = \theta^T X \tag{11}$$

By reducing the higher weight on a feature that has the top $mCorr(X, S)$ among all features, the method attempts to lower the linear ratio of this biased feature on the constitution of model prediction $\hat{Y}$.

3.1 Feature weight reduction

Previously we stated that adding limited samples to the training dataset could help reduce model unfairness. This section details this process inspired by An's approach to the study of noise addition in machine learning training [26]. Equation 12 is the likelihood function of logistic regression where y is the ground-truth of training data and θ is the feature weights of feature X.

$$L(\theta) = \prod_{i=1}^{n} P(Y = y_i | X = \mathbf{x}_i) = P(y_1|\mathbf{x}_1) \cdot P(y_2|\mathbf{x}_2) \cdot \cdots \cdot P(y_m|\mathbf{x}_m) \tag{12}$$

The likelihood function of logistic regression uses the likelihood of individual training samples to determine feature weight. It is the conditional probability, per sample, of ground-truth value given feature values [27]. Therefore, our research reduces the weight of the biased feature by changing the likelihood of the training samples. This can be achieved by adding samples that reduce the likelihood of the training data. By adding validation samples that dilute the association between biased feature and model prediction, the updated classifier model relies comparatively more on other features, and thus assigns a lower feature weight to the biased feature.

The weight reduction process adopts a pre-processing approach - by transforming training sample data before the training process of the logistic regression. The ACE algorithm is applied to transform data, acting as a black box function that performs the maximal correlation estimation.

Algorithm 2 Expected maximal correlation selection

```
procedure EXPECTED SELECTION
    for i = 1, 2, ..., for all features do
        for j = 1, 2, ..., for all samples in validation pool v do
            current mCorr ← mCorr(X_i, Y)
            x_i^(j) ← Value of feature i of sample j in v
            X_i' ← union(X_i, X_i^(j))
            Y' ← union(Ŷ, Y^(j))
            expected mCorr ← mCorr(X_i', Y')
            δ ← current mCorr - expected mCorr
            obj_i ← append(θ_i · mCorr(S, X_i) · δ)
        obj ← append(obj_i)
    index ← argmax(obj, i) for i with highest mCorr(S, X_i)
```

For training samples, consider a feature X_i that is highly dependent on S and this feature comes with a high feature weight θ_i. Due to the likelihood estimation of the logistic regression, feature X_i that is predictive of the ground-truth Y will be assigned a higher feature weight. To bring down the value of this feature weight, the lower maximal correlation $mCorr(X_i, Y)$ value of the training dataset could indicate a weakening association between the biased feature and the predictions [28].

One way to lower $mCorr(X_i, Y)$ is to add specific samples to the training dataset from validation pool. This method evaluate the effect of the addition of every single validation sample to the training dataset separately, and select a subset of samples that lower the $mCorr(X_i, Y)$ more significantly. In this case, consider a point in the validation pool $P_v = (X^{(v)}, Y^{(v)}, S^{(v)})$. If the maximal correlation between feature i and sensitive feature is high, this method add P_v into the training dataset, and record the change of $mCorr(X_i, Y)$.

$$\Delta mCorr(X_i, Y, X_i^{(v)}) = mCorr(X_i, Y) - mCorr(X_i \cup X_i^{(v)}, Y \cup Y^{(v)}) \tag{13}$$

It will be iterated for every sample in the validation set (e.g. line 3-10 in algorithm 2). In case where feature i has the top $mCorr(X_i, S)$ value across all features, this method proposes the multiplication of $mCorr(X_i, S)$ and $\Delta mCorr(X_i, , X_i^{(v)})$ as the objective function.

3.2 Objective function

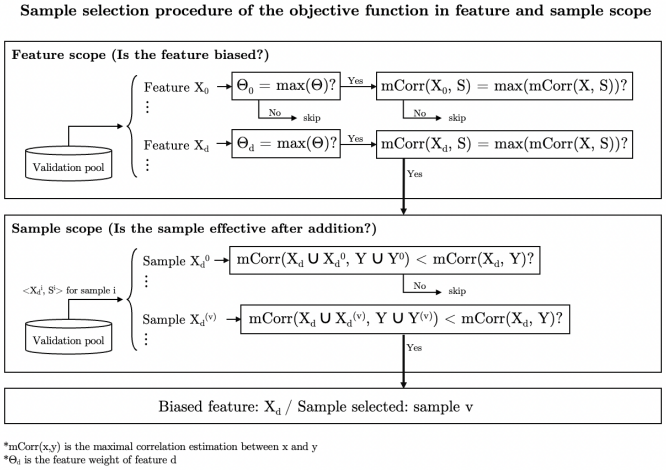

Figure 1: Procedure of objective function selecting sample from validation set

The objective is to add the most effective set of samples to the training data in terms of lowering $mCorr(X_i, Y)$ after their addition. The objective function provides numerical metrics to rank each sample in the validation pool based on sample effectiveness. Specifically, the objective function aims to aggregate a value on the individual sample's effect on changing the association between biased feature and ground truth. The output value of the objective function per sample is used for the decision on sample selection, without changing logistic regression weight or parameters directly. Given that a feature i is the most biased feature, the value of the objective function needs to indicate the sample's potential to lower the feature weight of feature i. There are three factors to consider for a sample to be selected into the training dataset:

- The current feature weight of the biased feature (i.e. θ_i)
- The maximal correlation between the feature and sensitive feature (i.e. $mCorr(X_i, S)$)
- The addition of the samples' effect on the feature weight (i.e. $\Delta mCorr(X_i, Y, X_i^{(v)})$)

This paper calls the first factor **feature weight**, second factor **maximal sensitive correlation**, and the third factor **delta difference**. These components form an optimization problem

$$\max(\Phi(\theta_i, mCorr(X_i, S), \Delta mCorr(X_i, Y, X_i^{(v)}))) \tag{14}$$

The objective function attempt to maximize the multiplication product of these three measures. Samples that have the highest value of the objective function are to be added to the training data in a batch. The number of samples to be added serves as a parameter to control the precision of sample effectiveness. The procedure of selecting samples is illustrated in figure 1. The objective function combines the value indicating these three factors through multiplication to increase the variance of the objective value per sample. Multiplication also prevents these three requirements from canceling out each other.

$$\Phi_{max} = \sum_{i=1}^{d} |\theta_i \cdot mCorr(X_i, S)| \cdot \Delta mCorr(X_i, Y, X_i^{(v)}) \tag{15}$$

The research from H.Anahideh(2021) also uses a similar objective function, since both methods lower feature weight by adding samples to the training data. The difference is that, in their research, the data from the validation pool is not labeled (i.e. lacking ground truth), while this method directly uses the ground truth of validation data instead. However, both use feature weight, and the association relationship as multiplication factors to construct the objective function.

3.3 Evaluation of method effectiveness

The objective function selects samples to be added to the training dataset. With the addition of samples from the validation pool, model predictions from training data are expected to be fairer (i.e. in the notion of demographic parity), since the classifier depends on the biased feature less (i.e. less on the sensitive feature as well) with the lowering of maximal correlation between the biased feature and ground truth.

To evaluate the improvement of demographic parity fairness, our method measures mutual information between the model prediction and sensitive feature, before and after the addition of samples. Mutual information measures the information of one variable given another variable [29], or specifically the reduction of uncertainty about one variable that results from observing the other variable [30].

Since demographic parity pursues stochastic independence between model prediction and sensitive feature, mutual information score of model prediction and the sensitive feature can measure the improvement of fairness from the objective function [31]. A decrease of mutual information score after applying the objective function and sample addition reflects an improvement in model fairness, in the scope of demographic parity notion of fairness.

$$MI(S, \hat{y}) = KL(p(S, \hat{y}) || p(S)p(\hat{y})) \tag{16}$$

The mutual information between sensitive feature set S and model prediction $\hat{y}$ is determined by how different it is between the joint distribution of S and $\hat{y}$ and the product of the marginal probability of the two variables [32].

$$MI(S, \hat{y}) = \sum_{i=1}^{|S|} \sum_{j=1}^{|\hat{Y}|} p_{(S,\hat{Y})}(s, \hat{y}) log(\frac{p_{(S,\hat{Y})}(s, \hat{y})}{p_S(s) p_{\hat{Y}}(\hat{y})}) \tag{17}$$

where $|S|$ is the size of a collection of sensitive feature values s and $|\hat{Y}|$ is the size of a collection of model prediction values $\hat{y}$.

4 EXPERIMENT

The experiment uses both synthetic dataset and realistic dataset to answer the question of whether maximal-correlation-based objective function reduces model unfairness effectively, without significantly reducing classification accuracy.

For all experiments, each dataset is divided into validation pool and testing test by K-fold cross-validation. Several random samples (e.g. 6 or 10, depending on overall sample size) will be initialized as training samples, and more samples will be added from the validation set to the training set according to the objective of reducing model unfairness. The data preparation step is detailed in Appendix B.

The experiment will use mutual information (i.e. MI) between model prediction and sensitive feature as the measure of unfairness result. Lower MI indicates the classifier relies its model less on the value of the sensitive feature, thus closer to the stochastic independence pursued by the notion of demographic parity. The classification accuracy is represented by Area Under the Curve (i.e. AUC) values to indicate the classifier's ability to distinguish classes [33].

4.1 Experiment I: Applying synthetic dataset

The synthetic dataset contains 500 samples and 30 new samples will be added into the training dataset that contains initially 10 samples.

In the synthetic dataset simplified from the COMPAS dataset, an artificially biased feature has the overlapped feature value with the sensitive feature, but in an adjustable 90%, 70% and 50% overlaps. For example, 70% overlap entails that 70% of the sample from the biased feature and sensitive feature has identical binary value. This is achieved by flipping 10%, 30%, and 50% of the binary values in the artificial feature (originally 100% identical to sensitive feature value). An artificial feature overlapped with the sensitive feature creates a scenario where this feature is known to highly correlates the sensitive feature, making it biased with a known and adjustable extent. Ground truth in this synthetic dataset is identical to the value of the sensitive feature. This guarantees a high feature weight for the artificial biased feature. This is necessary since a biased feature with a relatively smaller influence over the model prediction does not trigger the objective function to select samples that reduce its feature weight values.

From Equation 15, the objective function selects samples that reduce the feature(biased)-prediction dependence. It targets the biased feature that has high feature weight and a high maximal correlation with the sensitive feature.

The section illustrates the method mechanism of a manipulated training round and presents the result from the synthetic experiment. Section 3.2 stated that the objective function selects samples with three factors. The value of these factors by applying the synthetic dataset helps show the objective function's alignment with the goal of unfairness reduction.

Three factors within the objective function the synthetic dataset aims to manipulate are:

- Feature weight: Since the ground truth (identical to sensitive feature value) and biased feature is 90%, 70% or 50% overlapped, the feature weight of the synthetic biased feature is expected to be higher than other features if the algorithm is effective.
- Maximal sensitive correlation: Since the sensitive feature value and biased feature value is 90%, 70% or 50% overlapped, the maximal correlation between sensitive feature and biased feature is expected to be considerably higher compared to other features.
- Delta difference: The maximal correlation difference of biased feature and ground truth before and after sample addition (i.e. $\Delta mCorr(X_i, Y, X_i^{(v)})$) is expected to be higher than other features if the right samples are selected and trained.

4.2 Experiment II: Applying COMPAS datasets

4.2.1 Data description. The maximal correlation method and the objective function are tested against a variant of the COMPAS dataset. It has 5875 individual criminal records of juvenile felonies, published by ProPublica [8]. The experiment sets 'race' as the sensitive feature and 'two- year-recid' as ground truth. 'two-year-recid' for each sample entry is a binary value differentiating whether sample individuals re-offend in a future time-frame of two years. The 9 training features include marriage status, age, prior convictions, degree charged, and more. This dataset is normalized with zero mean and unit variance on the requirement of maximal correlation estimation.

4.2.2 Method comparison. The effectiveness of this method on reducing model unfairness is compared against two other machine learning methods. The first method uses randomized sampling for sample addition. Compared to the proposed method, the samples added to the training dataset are randomly selected samples through a Python randomizer, without considering the effect of sample addition to unfairness. This research needs to ensure that the unfairness reduction results from an effective objective function and sample selection, instead of the effect of selecting a smaller training sample size from the validation pool. Thus the unfairness measure between random sampling and maximal correlation sampling is expected to differ (similar to the comparison of RS and MC method in the synthetic experiment).

Another method is the FBC (Fairness by covariance) method, used in Anahideh et al.'s research on fair active learning. Although FBC is proposed as the primitive method in their research [25], it shares a similar objective function and sampling algorithm. FBC uses covariance value between the biased feature and sensitive feature while this research uses maximal correlation. Comparing FBC's result against the proposed maximal correlation sampling is, essentially, to compare the ability to reduce model unfairness from samples selected by maximal correlation-based objective function and covariance-based objective function.

4.2.3 Sample selection size. These three methods select and add the same amount of samples to the training dataset per experiment (i.e. 4300, 1000, and 200 validation samples). Adding 4300 samples to the training dataset mimics the conventional machine learning training process, in which almost all non-testing samples are added as a part of the training dataset. For the proposed maximal correlation method, unfairness reduction is achieved by adding 1000 or 200 samples of 4387 samples in the validation pool that best reduces dependence between model prediction and the biased feature. In the 4300 sample addition case, there is essentially no sample selection for all methods, thus it is expected to not affect unfairness reduction and has similar classification and fairness measures before and after sample addition. The purpose of the experiment that adds 4300 samples is to serve as a baseline (with sample selection having no effect of reducing unfairness). The baseline result represents the accuracy and unfairness value of a model generated by using the only logistic regression learning method. Results from the baseline (i.e. 4300 added samples) will be used as a default result to see how much each method reduce unfairness, with effective sample selection, through percentage decrease, shown in table 6 and 7 in the result section.

4.2.4 result acquisition. The AUC and MI values are measured after adding the specified amount of samples (i.e. 1000 and 200 samples) into the training dataset, for 500 instances of training each. The mean and standard deviation of both AUC and MI are compared across three methods. Higher AUC and lower MI values are desirable, as they indicate improved model fairness while retaining decent classification accuracy.

4.3 Supplementary experiment: Applying Adult dataset

4.3.1 Dataset description. This variation of the Adult dataset has 500 individual income data from the US Census Bureau, with 97 feature attributes, excluding the sensitive feature. The research sets 'gender' as a sensitive feature and 'income' as the ground truth. 'income' for each sample entry is a binary value differentiating sample individuals with income lower and higher than $50000. The features include *personal capital*, *level of education*, *marital status*, *occupation*, *native country* and more. The dataset is normalized with zero mean and unit variance, based on the requirement of maximal correlation estimation. All results in the Adult experiment are significant at 95% significance of the K-fold cross-validated paired t-test by Mlxtend.

4.3.2 Experiment description. This experiment shows the method's relevance to the non-racially sensitive dataset. The use of Adult data is to validate the effectiveness of the proposed method in datasets that are low in sample count and high in feature count, opposite to the COMPAS experiment. Except for the dataset and the sample selection size, the setup and goal for the Adult data experiment follow the COMPAS setup. The 370 sample addition case serves as the baseline. Since all samples from the validation pool is selected for training, there is no sample selection, similar

	90% overlap	70% overlap	50% overlap
Feature weight	([[0.20496376], [0.44800426], [0.51815331], [0.23798448], [0.37344904], [0.78492843], [**1.04620673**]])	([[0.32657519], [0.32841899], [0.34667147], [0.64679488], [0.09596769], [0.03695392], [**0.52859995**]])	([[0.28604877], [0.37431886], [0.], [0.10742105], [0.0984099], [0.11193722], [**0.46866759**]])
Maximal sensitive correlation	([[0.3245813], [0.556544], [0.0209041], [0.14016506], [0.01565823], [0.13341595], [**0.89607144**]])	([[0.34770085], [0.56025497], [0.03292416], [0.14576171], [0.01548655], [0.13303456], [**0.89585086**]])	([[0.34534887], [0.57055435], [0.03439379], [0.13084751], [0.01582718], [0.13152772], [**0.89551598**]])
Delta difference	([0.5595, 0.5721, 0.5675, -0.579 , 0.5110, 0.5891, **0.68887**])	([0.005215, 0.0002795, 0.0001138, 0.0007796, -0.10162, -0.11043, **1.54438**])	([0.08399, -0.10426, 0.01673, 0.01748, 0.01700, 0.0147, **0.08148**])

Table 1: Values of three factors of the objective function

Synthetic-30-90%	RS	MC
Accuracy (AUC)	0.792 → 0.904	0.792 → 0.736
Unfairness (MI)	0.2125 → 0.3813	0.2125 → 0.1200

Table 2: AUC (area under the curve) and MI (mutual information) of Synthetic dataset (90% overlap) with 30 samples addition for RS (random sampling method) and MC (proposed maximal correlation method)

to using only logistic regression. For the second experiment, 30 samples are selected from 370 samples in the validation dataset.

5 RESULTS

5.1 Result of the synthetic experiment

The value of the objective function in a one-time run of the algorithm using the synthetic dataset is shown in table 1. Bold values are from the known biased feature. Since the three factors multiply into the result of the objective function (see Equation 15), the higher values each factor has for biased feature are desirable.

From the one-time run, the feature weight and maximal correlation value of the biased feature (shown in table 1) are comparatively higher than other features.

The AUC (Area under the curve) value and MI (Mutual information) value before and after supplying 30 samples to the training dataset are shown in table 2, 3, and 4 for different percentages of overlap between the biased feature and the sensitive feature. The

Synthetic-30-70%	RS	MC
Accuracy (AUC)	0.592 → 0.704	0.592 → 0.640
Unfairness (MI)	0.01217 → 0.04727	0.01217 → 0.03614

Table 3: AUC and MI of Synthetic dataset (70% overlap) with 30 samples addition for RS and MC method

Synthetic-30-50%	RS	MC
Accuracy (AUC)	0.568 → 0.584	0.568 → 0.608
Unfairness (MI)	0.01096 → 0.02287	0.01096 → 0.02814

Table 4: AUC and MI of Synthetic dataset (50% overlap) with 30 samples addition for RS and MC method

AUC and MI value before sample addition is identical for both random sampling and maximal correlation sampling, since the 10 initial training samples are the same for both methods. With 30 more samples, each added to different sampling methods, the accuracy and unfairness measures start to differ. In general, the classification accuracy of the classifier with the random sampling method increases more significantly than that with the maximal correlation method. Conversely, the unfairness value decreases more after training with the maximal correlation method. With a decreasing overlap, the fairness improvement of the maximal correlation method diminishes.

5.2 Result of the COMPAS experiment

Table 5, 6 and 7 are AUC and MI results across three methods, with three different sample selection sizes. All experiment values are significant at 95% significance level in the K-fold cross-validated paired t-test by Mlxtend (based on p-value and critical value) [34], in which the other two methods significantly differs from the random sampling method.

Table 5 presents baseline results, applying almost all validation samples to training datasets. The experiment ran 500 training instances, each with a machine restart and Python reset. It contains the average mean of 500 AUC value and MI value from RS (random sampling), MC (maximal correlation), and FBC (covariance) methods. The AUC and MI for each method are relatively similar, which fits the expectation as a baseline result. This result shows the classification accuracy of the logistic regression model without factoring fairness issue.

Table 6 and 7 are similar to the experiment setup of table 5, with the only difference in sample selection size. For the proposed unfairness reduction method, the top 1000 out of 4387 samples are trained, compared to a more rigorous selection of 200 effective samples in table 7 (COMPAS dataset with 200 samples added to training dataset).

The bold value shows the highest result per row among three methods, and "win counts" show how many times the method "wins" over other methods. For table 6 and 7, the percentage decrease of AUC and MI compared to the baseline (table 5) is shown. This indicates the degree of unfairness reduction, as well as how classification accuracy changes with unfairness reduction.

COMPAS-4300 (full validation set)	RS_AUC	MC_AUC	FBC_AUC	RS_MI	MC_MI	FBC_MI
Mean	0.681	0.682	0.679	0.02670	0.02691	0.02712

Table 5: Baseline AUC (Classification accuracy, the higher the better) and MI (Unfairness measure, the lower the better) of COMPAS dataset with 4300 samples addition for RS (random sampling), MC (proposed maximal correlation) and FBC (covariance) method

COMPAS-1000	RS_AUC	MC_AUC	FBC_AUC	RS_MI	MC_MI	FBC_MI
Mean	**0.670**	0.649	0.654	0.02242	**0.01511**	0.02132
Standard deviation	0.0081	0.0171	0.0221	0.0027	0.0069	0.0070
Win counts	**337/500**	61/500	102/500	40/500	**349/500**	111/500
Percentage decrease (%)	**1.62**	4.84	3.68	16.03	**43.85**	21.39
Critical value	NIL	1.9633	1.9637	NIL	1.9637	1.9637
P-value	NIL	1.04E-98	2.66E-46	NIL	3.88E-80	0.00112

Table 6: Averaged AUC and MI of COMPAS dataset with 1000 samples addition for RS, MC and FBC method

COMPAS-200	RS_AUC	MC_AUC	FBC_AUC	RS_MI	MC_MI	FBC_MI
Mean	**0.635**	0.566	0.585	0.01722	**0.00773**	0.01177
Standard deviation	0.0196	0.0432	0.0444	0.02242	0.01511	0.02132
Win counts	**398/500**	30/500	72/500	39/500	**283/500**	178/500
Percentage decrease (%)	**6.76**	17.01	13.84	35.51	**71.28**	56.60
Critical value	NIL	1.9634	1.9634	NIL	1.9623	1.9628
P-value	NIL	4.35E-143	2.98E-87	NIL	2.32E-124	4.02E-31

Table 7: Averaged AUC and MI of COMPAS dataset with 200 samples addition for RS, MC and FBC method

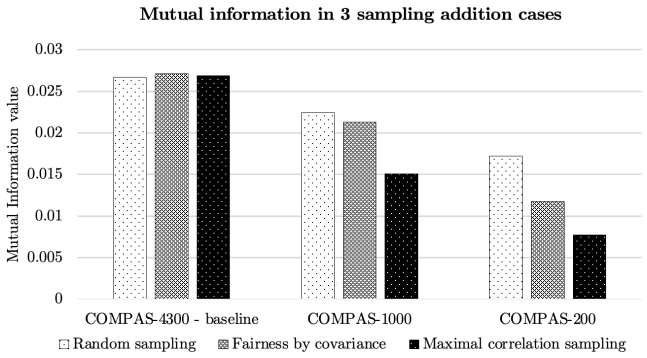

Figure 2: Bar plot of MI across 4300, 1000 and 200 sample intake from validation set to the training set (lower the better)

Regardless of the difference in sample selection size, maximal correlation sampling obtain a lower unfairness value with the highest percentage decrease in both experiments (43.85% and 71.28% decrease in unfairness). Models with random sampling have the lowest percentage decrease in unfairness measure while being able to retain higher classification accuracy. As shown in Figure 2, the FBC covariance method's MI value is closer to random sampling in the experiment with 1000 sample addition, while approaching the unfairness reduction performance of maximal correlation in the experiment with 200 sample addition.

5.3 Result of the Adult experiment

Results in table 8 and 9 use the Adult dataset with 370 and 30 samples added to the training dataset. The p-value and critical value indicate significant results at 95% significance level.

The maximal correlation method has a 59.47% decrease in unfairness after adding 30 top samples from the validation pool, in terms of each sample's effectiveness over reducing dependence between biased feature and model prediction. The classification accuracy decreases 2.05% compared to the baseline. It has a considerable advantage compared to both random sampling and covariance methods.

6 DISCUSSION AND CONCLUSION

6.1 Method analysis

From the one-time training instance of the synthetic experiment, the feature weight and maximal correlation value of the biased feature are considerably higher than other features (i.e. bold values in table 1), suggesting that the algorithm correctly identifies the problematic feature and its high influence over the model prediction. With only 50% of the value that correlates to the sensitive feature, the biased feature still has the highest maximal correlation value with the sensitive feature among all features. The average delta difference is higher on the biased feature compared to other features, suggesting that the algorithm can effectively select samples that can lower the $mCorr(X_i, Y)$ the most, given that X_i is the biased feature. A lowered $mCorr(X_i, Y)$ distances model prediction from

Adult-370 (full validation set)	RS_AUC	MC_AUC	FBC_AUC	RS_MI	MC_MI	FBC_MI
Mean	0.835	0.829	0.830	0.03702	0.04091	0.04200

Table 8: Baseline AUC and MI of Adult dataset with 370 samples addition for RS (Random sampling), MC (proposed maximal correlation) and FBC (covariance) method

Adult-30	RS_AUC	MC_AUC	FBC_AUC	RS_MI	MC_MI	FBC_MI
Mean	**0.841**	0.812	0.765	0.02432	**0.01658**	0.04115
Standard deviation	0.0179	0.0401	0.0419	0.0175	0.0142	0.0444
Win counts	**322/500**	166/500	12/500	117/500	**234/500**	149/500
Percentage decrease (%)	**-0.72**	2.05	7.83	34.31	**59.47**	2.02
Critical value	NIL	1.9634	1.9635	NIL	1.9625	1.9636
P-value	NIL	3.99E-42	1.53E-166	NIL	3.89E-14	1.42E-14

Table 9: Averaged AUC and MI of Adult dataset with 30 samples addition for RS, MC and FBC method

the biases of the biased feature, thus helping the model to approach the notion of demographic parity.

6.1.1 Limitation of the objective function. Comparing the unfairness measure from high overlap (i.e. table 2) to lower overlap in table 4, the advantage of the maximal correlation method is diminishing. From table 4, unfairness reduction on both methods achieve similar results. Since sensitive feature and model prediction is identical in the synthetic dataset, the lower overlap between the biased feature and sensitive feature results in a lower overlap between the biased feature and model prediction. For the experiment with low 50% overlap (e.g. table 4), half of the samples are potential candidates to be selected since the biased feature value of these samples is different from model prediction. With a flood of candidates, it is no longer sufficient to select samples based on different values between biased feature and model prediction. The selection of samples thus has to consider other features' weaker influence on reducing feature (biased)-model dependence. The effectiveness of these influence is not as significant (e.g. lower delta difference) and the precision of selecting effective samples are diluted. Real-world datasets often have even smaller percentage overlap, thus this limitation is easily reproducible.

Secondly, the decrease of the feature weight due to a 50% data difference can also contribute to ineffective sample selection, as other non-biased or less biased feature might be wrongly targeted by the objective function (e.g. when feature weight between other features and biased feature are marginal). This limitation entails the risk of using feature weight as a factor in the objective function - the wrongly selected sample could negatively affect the unfairness reduction by lowering the significance of the non-biased target feature.

Last but not least, the objective function selects samples by multiplication of the three factors (section 3.2). It is adopted since it highlights influential biased feature exponentially. However, it is prone to error if one of the factor's values is not prominent. For example, if samples from the validation pool are unable to reduce feature-model dependence, a low delta difference value could result in an overall low objective function result, even if the feature is the most biased.

6.1.2 Limitation on the training sample selection. Our objective function limited samples to be added to the training data for a fairer model. It does not interfere with the training process of the logistic regression, therefore, feature weights of both the biased feature and the other features are changed according to the need for a fairer classifier. However, this method targets datasets where all samples in the validation set are labeled (i.e. has ground truth by default). Compared to other unfairness reduction methods [18, 35], only using a slice of the validation sample to train results in the incomplete representation of accuracy and fairness value over the dataset, since the experiment did not account for the influence of samples that are not selected, and the influence could potentially be significant.

6.2 Real world result analysis

6.2.1 Accuracy & Fairness trade-off. In this research, methods with an aim to reduce model unfairness inevitably encounter a decrease in classification accuracy. Inversely, models became less fair with a larger sample selection size, while the classification accuracy increases. This is because the information and pattern given to the model by smaller training samples are more limited than larger sized samples [36, 37]. The result from selecting 200 and 1000 samples (from table 6 and 7) corroborate this claim. In the scope of each method, results from 1000 sample selection size, per method, have a higher classification accuracy and higher unfairness measure. Several other pieces of research also claimed the validity of the accuracy & fairness trade-off [38–40]. To determine whether the trade-off of classification accuracy for a method is more worthy for its fairness improvement, this research pursues scenarios where:

- The improvement of fairness is significant given the sacrifice for classification accuracy is relatively marginal
- Under similar level of sacrifice in classification accuracy, the improvement of fairness for one method is significantly larger than other methods

From table 5, the baseline result indicates that the classification accuracy of the COMPAS dataset, under the logistic regression training, can reach approximately 68% classification accuracy rate. For 1000 sample additions (table 6), compared to the percentage

decrease of classification accuracy (below 5% for all methods), the maximal correlation method can achieve a 43.85% decrease of unfairness measure, compared to 16% for random sampling and 21% for covariance method. The proposed method, in the case of 1000 sample addition, satisfy the pursuit of the first scenario.

In table 7, the maximal correlation method and covariance method have similar classification accuracy rates (i.e. 56.6% and 58.5%). The unfairness reduction for the maximal correlation model is approximately 15% higher than that of the covariance method. However, if the model is expected to function effectively, both 56.6% and 58.5% classification accuracy rate is not desirable as a reliable classifier. This could be attributed to the incompleteness of the data representation. 200 samples out of over 5000 samples in the original dataset are not representative of a classifier model, and the training process of logistic regression stopped immaturely as the training sample ran out. Although unfairness is drastically lower than the baseline, aside from the contribution of the objective function, lack of data representation could also result in low unfairness, which does not fit into the goal for a balanced accuracy/fairness trade-off scenario. For a COMPAS dataset, the experiment of 1000 sample additions is more applicable with its smaller impact on classification accuracy.

6.3 Conclusion

Overall, this maximal-correlation-based unfairness reduction method is empirically valid in reducing unfairness in both synthetic and real-world experiments, especially if the biased feature is sufficiently evident, and the parameters such as the size of the sample addition balance the accuracy/fairness trade-off. The deterioration of classification accuracy is inevitable with a limited training sample count, yet the percentage decrease is controllable and tolerable. Our research illustrates the maximal correlation estimation's versatility in its functionality and its effectiveness, and a novel method that uses training sample addition to reduce model unfairness. Under the guidance of the independence notion of demographic parity, the mutual information between biased features and the prediction outcome can be reduced effectively, achieving improved statistical model fairness.

ACKNOWLEDGMENTS

To my daily supervisor Marco Loog, for the invaluable feedback and inspirations.

REFERENCES

[1] Simon Caton and Christian Haas. Fairness in machine learning: A survey. *arXiv preprint arXiv:2010.04053*, 2020.
[2] Philippe Besse, Eustasio del Barrio, Paula Gordaliza, Jean-Michel Loubes, and Laurent Risser. A survey of bias in machine learning through the prism of statistical parity for the adult data set. *arXiv preprint arXiv:2003.14263*, 2020.
[3] Andrea Romei and Salvatore Ruggieri. A multidisciplinary survey on discrimination analysis. *The Knowledge Engineering Review*, 29(5):582–638, 2014.
[4] Hyojin Bahng, Sanghyuk Chun, Sangdoo Yun, Jaegul Choo, and Seong Joon Oh. Learning de-biased representations with biased representations. In *International Conference on Machine Learning*, pages 528–539. PMLR, 2020.
[5] Guilherme Alves, Maxime Amblard, Fabien Bernier, Miguel Couceiro, and Amedeo Napoli. Reducing unintended bias of ml models on tabular and textual data. In *2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–10. IEEE, 2021.
[6] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29, 2016.
[7] Indre Zliobaite. A survey on measuring indirect discrimination in machine learning. *arXiv preprint arXiv:1511.00148*, 2015.
[8] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias, 2016.
[9] Solon Barocas, Moritz Hardt, and Arvind Narayanan. Fairness in machine learning. *Nips tutorial*, 1:2017, 2017.
[10] Sahil Verma and Julia Rubin. Fairness definitions explained. In *2018 ieee/acm international workshop on software fairness (fairware)*, pages 1–7. IEEE, 2018.
[11] Toon Calders and Sicco Verwer. Three naive bayes approaches for discrimination-free classification. *Data mining and knowledge discovery*, 21(2):277–292, 2010.
[12] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29:3315–3323, 2016.
[13] 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.
[14] Ricardo Salazar, Felix Neutatz, and Ziawasch Abedjan. Automated feature engineering for algorithmic fairness. *Proceedings of the VLDB Endowment*, 14(9):1694–1702, 2021.
[15] 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.
[16] Anna Lauren Hoffmann. Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse. *Information, Communication & Society*, 22(7):900–915, 2019.
[17] Shao-Lun Huang and Xiangxiang Xu. On the sample complexity of hgr maximal correlation functions for large datasets. *IEEE Transactions on Information Theory*, 67(3):1951–1980, 2020.
[18] Clara Belitz, Lan Jiang, and Nigel Bosch. Automating procedurally fair feature selection in machine learning. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 379–389, 2021.
[19] Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 560–568, 2008.
[20] Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33, 2012.
[21] Andrija Petrović, Mladen Nikolić, Sandro Radovanović, Boris Delibašić, and Miloš Jovanović. Fair: Fair adversarial instance re-weighting. *arXiv preprint arXiv:2011.07495*, 2020.
[22] Alfréd Rényi. On measures of dependence. *Acta Mathematica Academiae Scientiarum Hungarica*, 10(3-4):441–451, 1959.
[23] Jérémie Mary, Clément Calauzenes, and Noureddine El Karoui. Fairness-aware learning for continuous attributes and treatments. In *International Conference on Machine Learning*, pages 4382–4391. PMLR, 2019.
[24] Joshua Lee, Yuheng Bu, Prasanna Sattigeri, Rameswar Panda, Gregory Wornell, Leonid Karlinsky, and Rogerio Feris. A maximal correlation approach to imposing fairness in machine learning. *arXiv preprint arXiv:2012.15259*, 2020.
[25] Hadis Anahideh, Abolfazl Asudeh, and Saravanan Thirumuruganathan. Fair active learning. *arXiv preprint arXiv:2001.01796*, 2020.
[26] Guozhong An. The effects of adding noise during backpropagation training on a generalization performance. *Neural computation*, 8(3):643–674, 1996.
[27] David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. *Applied logistic regression*, volume 398. John Wiley & Sons, 2013.
[28] Heinrich Jiang and Ofir Nachum. Identifying and correcting label bias in machine learning. In *International Conference on Artificial Intelligence and Statistics*, pages 702–712. PMLR, 2020.
[29] Ian H Witten, Eibe Frank, Mark A Hall, CJ Pal, and MINING DATA. Practical machine learning tools and techniques. In *DATA MINING*, volume 2, page 4, 2005.
[30] David JC MacKay and David JC Mac Kay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.
[31] Jaewoong Cho, Gyeongjo Hwang, and Changho Suh. A fair classifier using mutual information. In *2020 IEEE International Symposium on Information Theory (ISIT)*, pages 2521–2526. IEEE, 2020.
[32] Christopher M Biship. Pattern recognition and machine learning (information science and statistics), 2007.
[33] Tom Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006.
[34] Sebastian Raschka. Mlxtend: Providing machine learning and data science utilities and extensions to python's scientific computing stack. *Journal of open source software*, 3(24):638, 2018.
[35] Pieter Delobelle, Paul Temple, Gilles Perrouin, Benoît Frénay, Patrick Heymans, and Bettina Berendt. Ethical adversaries: Towards mitigating unfairness with adversarial machine learning. *ACM SIGKDD Explorations Newsletter*, 23(1):32–41, 2021.
[36] Margarita Sordo and Qing Zeng. On sample size and classification accuracy: A performance comparison. In *International Symposium on Biological and Medical Data Analysis*, pages 193–201. Springer, 2005.

[37] Carlton Chu, Ai-Ling Hsu, Kun-Hsien Chou, Peter Bandettini, ChingPo Lin, Alzheimer's Disease Neuroimaging Initiative, et al. Does feature selection improve classification accuracy? impact of sample size and feature selection on classification using anatomical magnetic resonance images. *Neuroimage*, 60(1):59–70, 2012.

[38] Benjamin Fish, Jeremy Kun, and Ádám D Lelkes. A confidence-based approach for balancing fairness and accuracy. In *Proceedings of the 2016 SIAM international conference on data mining*, pages 144–152. SIAM, 2016.

[39] Christian Haas. The price of fairness-a framework to explore trade-offs in algorithmic fairness. In *40th International Conference on Information Systems, ICIS 2019*. Association for Information Systems, 2019.

[40] Patrick Janssen and Bert M Sadowski. Bias in algorithms: On the trade-off between accuracy and fairness. 2021.

APPENDIX

A FEATURE-PREDICTION BIASES ATTRIBUTION

To prove that the feature weight θ and maximal correlation of sensitive feature and features $mCorr(X, S)$ are the factor of the value of maximal correlation of sensitive feature and model prediction $mCorr(\hat{Y}, S)$:

$$\begin{aligned} mCorr(S, \hat{Y}) &= [\mathbf{f}^*(S)\mathbf{g}^*(\hat{Y})] \\ &= [\mathbf{f}^*(S) \sum_{i=1}^{n} \theta_i \cdot \mathbf{g}^*(x_i)] \\ &= [\mathbf{f}^*(S) \cdot \theta_1 \cdot \mathbf{g}^*(x_1) + \cdots + \mathbf{f}^*(S) \cdot \theta_n \cdot \mathbf{g}^*(x_n)] \\ &= [\mathbf{f}^*(S) \cdot \theta_1 \cdot \mathbf{g}^*(x_1)] + \cdots + [\mathbf{f}^*(S) \cdot \theta_n \cdot \mathbf{g}^*(x_n)] \\ &= \theta_1 \cdot [\mathbf{f}^*(S) \cdot \mathbf{g}^*(x_1)] + \cdots + \theta_n \cdot [\mathbf{f}^*(S) \cdot \mathbf{g}^*(x_n)] \\ &= \sum_{i=1}^{n} \theta_i \cdot mCorr(S, \mathbf{X}_i) \\ &= \theta^T \cdot mCorr(S, X) \end{aligned} \tag{18}$$

B DATA PREPARATION

Specific to this research, the initial training sample count is small since we want the model training to be largely dependent on the samples selected later from the validation set. Here we illustrate how samples are categorized.

$$\begin{aligned} \#D_{total} &= n \\ D_{train} &= \{\} \\ D_{test} &= split(n, fold = 4) \\ D_{validation} &= D_{total} - D_{test} \rightarrow \textbf{Validation pool } V \\ init &= random(5, D_{validation}, y = 1) + random(5, D_{validation}, y = 0) \\ D_{train}&.add(init) \\ D_{validation} &= D_{validation} - D_{train} \end{aligned} \tag{19}$$