Fairness-enhancing Ensemble Classification in Water Distribution Networks

Janine Strotherm[0009−0005−7013−5167] and Barbara Hammer[0000−0002−0935−5591]

Bielefeld University, 33501 Bielefeld, Germany
{js,bhammer}@techfak.uni-bielefeld.de

Abstract. As relevant negative examples such as the future criminal detection software show, fairness of AI-based and social domain affecting decision support tools constitutes an important area of research. In this contribution, we investigate the applications of AI to socioeconomically relevant infrastructures such as those of water distribution networks (WDNs), where fairness issues have yet to gain a foothold. To establish the notion of fairness in this domain, we propose an appropriate definition of protected groups and group fairness in WDNs as an extension of existing definitions. We demonstrate that typical methods for the detection of leakages in WDNs are unfair in this sense. Further, we therefore propose a remedy to increase the fairness which can be applied even to non-differentiable ensemble classification methods as used in this context.

Keywords: Fairness · Disparate Impact · Equal Opportunity · Leakage Detection in Water Distribution Networks.

1 Introduction

Due to the increasing usage of artificial intelligence (AI)-based decision making systems in socially relevant fields of application the question of *fair decision making* gained much importance in recent years (cf. [1], [5]). Fairness is hereby related to the several (protected) groups or individuals which are affected by the algorithmic decision making and characterized by *sensitive features* such as gender or ethnicity. Most algorithms on which these tools are based on rely on data which can be biased with respect to the questions of fairness without intention, resulting in skewed models. Also the algorithm itself can discriminate protected groups or individuals without explicitly aiming to do so due to an undesirable algorithmic bias (cf. [11]).

Therefore, several definitions of fairness as well as approaches to achieve these fairness standards have been theoretically discussed and tested in practise (cf. [3,4,6,10,11,15]). From a legal perspective, one distinguishes between *disparate treatment* and *disparate impact* (cf. [3]). While disparate treatment occurs whenever a group or an individual is intentionally treated differently because of their membership in a protected class, disparate impact is a consequence of indirect discrimination happening despite "seemingly neutral policy" (cf. [11]). From a

scientific viewpoint, the variety of fairness notions is much larger where many popular approaches focus mainly on (binary) classification tasks (cf. [4,10,11]).

Besides the definition of fairness, the problem arises how to enhance fairness in well known AI methods while maintaining a reasonable overall performance of the model. Approaches can hereby be grouped in three categories: Pre-processing such as the modification of the features or labels (cf. [11]), in-processing such as adding a regularization term or constraints to the training objective (cf. [15]), or post-processing such as individualized choices of thresholds (cf. [11]).

The question of fairness becomes especially relevant when the decisions of a machine learning (ML) model take impact on socioeconomic infrastructure, such as water distribution networks (WDNs). To the best of our knowledge, the question of fairness has not been approached within this domain. We address the important problem of leakage detection in WDNs and investigate in how far typical models treat different groups of consumers of the WDN (in)equally. We hereby focus on *group fairness*, which in contrast to *individual fairness* focuses on treating different groups among the WDN equally instead of treating similar individuals similarly (cf. [10]).

To come up with a first approach to improve fairness in such a domain of high social and ethical relevance, based on [15], we consider the empirical covariance between the sensitive features and the model's prediction as a proxy for the fairness measure. Moreover, we also present algorithms that can handle multiple non-binary sensitive features and that satisfy both the concept of disparate treatment *and* disparate impact simultaneously, which is an asset towards most fairness enhancing algorithms (cf. [15,11]). In addition, we extend the theory of [15] by (a) giving explicit generalized definitions of well-known fairness measures for multiple non-binary sensitive features, (b) modifying their idea to any possibly non-convex classification model instead of convex margin-based classifiers, which allows to apply this approach to an ensemble classifier instead of a single classifier, and (c) presenting a method to handle the problem of potential non-differentiability in order to enlarge the space of AI models to which our approach can be applied.

To be able to introduce the notion of fairness into the application domain of leakage detection in WDNs, the rest of the work is structured as followed: In section 2, we introduce two definitions of group fairness for multiple non-binary sensitive features. Afterwards, in section 3, we present a standard methodology to detect leakages in WDNs and investigate whether the resulting model makes fair decisions with respect to the previously defined notions of fairness. Then, in section 4, we propose and evaluate several adaptations to this methodology that enhance fairness. Finally, our findings are summarized section 5.

2 Fairness in Machine Learning

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a population space of interest and $\hat{Y} : \Omega \to \mathcal{Y}$ some binary classifier, i.e., $\mathcal{Y} = \{0, 1\}$, being trained to model some true labels $Y : \Omega \to \mathcal{Y}$.

Usually, $\hat{Y}$ can be written as some model $f : \mathcal{X} \to \{0, 1\}$, applied to the features $X : \Omega \to \mathcal{X}$, i.e., $\hat{Y} = f(X)$ holds. In recent years, the interest towards the question of such classifier $\hat{Y}$ being fair with respect to some additional, sensitive feature $S : \Omega \to \mathcal{S}$ has risen. Mostly, $\mathcal{S} = \{0, 1\}$ gives binary information about the membership or non-membership of a protected class, such as some certain gender or ethnicity (cf. [11]). While the majority of the literature focuses on a single sensitive feature S (cf. [4,10,11]), in this work, we generalize the understanding of fairness to multiple binary sensitive features $S_1, ..., S_K$ that model K different groups. This also allows to work with (multiple) non-binary sensitive features S, i.e., sensitive features for which $\mathcal{S} = \{s_1, ..., s_K\}$ holds, by encoding a single sensitive feature S to K binary sensitive features $S_1, ..., S_K$.

Within this work, we will focus on group fairness. Assuming that all of the following conditional probabilities exist, one well-known notion of group fairness based on the predictor $\hat{Y}$ and the binary sensitive feature S is called disparate impact (DI), requiring that

$$\frac{\mathbb{P}(\hat{Y} = 1 \mid S = 0)}{\mathbb{P}(\hat{Y} = 1 \mid S = 1)} \geq 1 - \epsilon$$

is satisfied for some given value $\epsilon \in [0, 1]$, assuming that $\{S = 0\}$ is the protected group and that the nominator is smaller than the denominator (cf. [11]). The disparate impact notion is also known as the p%-rule and received its importance because it is "designed to mathematically represent the legal notion of *disparate impact*" (cf. [11]). Here, p is given by $p = 100(1 - \epsilon)$ and $p = 80$ is a desirable choice (cf. [15]). Disparate impact assures that the relative amount of positive predictions within the protected group $\{S = 0\}$ deviates at most $(100 - p)\% = 100\epsilon\%$ from the relative amount of positive predictions within the non-protected group $\{S = 1\}$.

We generalize this definition to multiple binary sensitive features by

$$\text{DI} := \min_{k_1, k_2 \in \{1,...,K\}} \frac{\mathbb{P}(\hat{Y} = 1 \mid S_{k_1} = 1)}{\mathbb{P}(\hat{Y} = 1 \mid S_{k_2} = 1)} \geq 1 - \epsilon. \tag{1}$$

Criticism of the disparate impact score DI could be the missing dependence on the true label Y (cf. [3]). We thus introduce another notion of group fairness, called equal opportunity (EO). In standard definition, equal opportunity holds whenever

$$\left| \mathbb{P}(\hat{Y} = 1 \mid Y = 1, S = 1) - \mathbb{P}(\hat{Y} = 1 \mid Y = 1, S = 0) \right| \leq \epsilon$$

is satisfied for some given value $\epsilon \in [0, 1]$ (cf. [10,11]). Equal opportunity ensures the true positive rates (TPRs) among protected and non-protected groups to differ at most $100\epsilon\%$.

Similarly, we generalize this definition to multiple binary sensitive features:

$$\text{EO} := \max_{\substack{k_1, k_2 \\ \in \{1,...,K\}}} \left| \mathbb{P}(\hat{Y} = 1 \mid Y = 1, S_{k_1} = 1) - \mathbb{P}(\hat{Y} = 1 \mid Y = 1, S_{k_2} = 1) \right| \leq \epsilon. \tag{2}$$

Remark 1. Our generalized notions of fairness go hand in hand with the conventional ones: In the conventional definitions, a single random variable S gives information about the membership of a protected group $\{S = 0\}$ or the membership of a non-protected group $\{S = 1\}$. Our definition handles the existence of K different groups without defining which of the groups are protected in advance. By defining the protected group $\{S = 0\}$ as group 1 and the non-protected group $\{S = 1\}$ as group 2 and the random variables S_k giving information about the membership of group k for $k = 1, 2$, the conventional definitions and our definitions (cf. eq. (1) and (2)) coincide.

Remark 2. Note the difference between the disparate impact *score* DI and the phrase that disparate impact, i.e., that DI $\geq 1-\epsilon$, *holds*, for some given $\epsilon \in [0, 1]$. Usually it is clear from the context what of both is meant and mostly, the disparate impact score DI is measured. Afterwards, it is checked whether the score is sufficiently large. The same holds for the analogy between the equal opportunity *score* EO and the phrase that equal opportunity *holds*.

Remark 3. A more powerful notion of group fairness using the true labels Y, the predictor $\hat{Y}$ and the sensitive feature S is equalized odds. This notion of fairness simultaneously considers the (largest) absolute differences between the TPRs as well as the false positive rates (FPRs) among sensitive groups (cf. [11]). However, as we will see in the proof of lemma 1, the FPRs $\mathbb{P}(\hat{Y} = 1 \mid Y = 0, S_k = 1)$ do not exist in our domain of application.

3 Leakage Detection in Water Distribution Networks

A key challenge in the domain of WDNs is to detect leakages. In this task, Ω corresponds to possible states of a WDN, given by time-dependant demands of the end users of the D nodes in the network. We assume that among those, d nodes are provided with sensors (usually, $D \gg d$), which deliver pressure measurements $p(t) \in \mathbb{R}^d$ for different times $t \in \mathbb{R}$ and which can be used for the task at hand. As we usually measure pressure values within fixed time intervals $\delta \in \mathbb{R}_+$, we introduce the notation $t_i := t_0 + i\delta$, where t_0 is some fixed reference point with respect to time.

3.1 Methodology

There are several methodologies that make use of pressure measurements to approach the problem of leakage detection using ML, i.e., by training a classifier $\hat{Y} \in \{0, 1\} = \mathcal{Y}$ that predicts the true state of the WDN $Y \in \mathcal{Y}$ with respect to the question whether a leak is active (1) or not (0). The standard approaches come in two steps: In first instance, so called *virtual sensors* are trained, i.e., regression models being able to predict the pressure at a given node $j \in \{1, ..., d\}$ and some time $t_i \in \mathbb{R}$, based on measured pressure at the remaining nodes $\hat{j} \neq j$ and over some discrete time interval of size $T_r + 1 \in \mathbb{N}$. Subsequently, these virtual sensors are used to compute *pressure residuals* of measured and predicted pressure to train an ensemble classifier that is able to predict whether a leakage is present in the WDN at the time of the used residual (cf. [14]).

Virtual Sensors The virtual sensors $f_j^r : \mathbb{R}^{d_r} \to \mathbb{R}$ for each node $j \in \{1, ..., d\}$ and $d_r := d-1$ are linear regression models trained on leakage free training data $\mathcal{D}_j^r = \{(\overline{p}_{\neq j}(t_i), p_j(t_i)) \in \mathbb{R}^{d_r} \times \mathbb{R} \mid i = 0, ..., n_r\}$. More precisely, $y(t_i) = 0 \in \mathcal{Y}$ holds for all realisations $i = 0, ..., n_r$ of Y and the inputs are given by the rolling means $\overline{p}_{\neq j}(t_i) := (T_r + 1)^{-1} \sum_{\iota=0}^{T_r} p_{\neq j}(t_i - \iota\delta)$ at all nodes except the node j, which is the only preprocessing required for the training pipeline (cf. [2]).

Ensemble Leakage Detection Standard leakage detection methods rely on the residuals $r_j(t_i) := |p_j(t_i) - f_j^r(\overline{p}_{\neq j}(t_i))| \in \mathbb{R}_+$ we obtain from the true pressure measurements $p(t_i) \in \mathbb{R}^d$ and the virtual sensor predictions $f_j^r(\overline{p}_{\neq j}(t_i)) \in \mathbb{R}$ for each sensor node $j \in \{1, ...d\}$ and (possibly unseen) times $t_i \in \mathbb{R}$ (cf. [8,14]).

A simple detection method performing good on standard benchmarks is the *threshold-based ensemble classification* introduced by [2]: Without any further training, we can define a classifier $f_j^c : \mathbb{R}_+ \to \mathcal{Y}$ by

$$f_j^c(r_j(t_i)) = f_j^c(r_j(t_i), \theta_j) := \mathbb{1}_{\{r_j(t_i) \,>\, \theta_j\}}$$

for each sensor node $j \in \{1, ...d\}$ and a node-dependant hyperparameter $\theta_j \in \mathbb{R}_+$. We easily obtain an ensemble classifier $f^c : \mathcal{X} \to \mathcal{Y}$, called the H-method, with hyperparameter $\Theta := (\theta_j)_{j=1,...,d} \in \mathcal{X}$ for $\mathcal{X} := \mathbb{R}_+^{d_c}$ and $d_c := d$ that predicts whether there is a leakage present in the WDN at time $t_i \in \mathbb{R}$ or not, defined by

$$f^c(r(t_i)) = f^c(r(t_i), \Theta) := \mathbb{1}_{\left\{\sum_{j=1}^{d_c} f_j^c(r_j(t_i)) \,\geq\, 1\right\}}. \tag{3}$$

Evaluation We evaluate the H-method in terms of general performance, measured by accuracy (ACC), and in terms of fairness, measured by disparate impact as well as equal opportunity score DI and EO, respectively (cf. eq. (1) and (2)).

3.2 Application Domain and Data Set

One key contribution of this work is to introduce the notion of fairness in the application domain of WDNs. The WDN considered is *Hanoi* (cf. [12,14]) displayed in figure 1. It consists of 32 nodes and 34 links.

To evaluate the H-method presented in section 3.1 on Hanoi, we generate pressure measurements with a time window of $\delta = 10$min. using the atmn toolbox (cf. [13]). The pressure is simulated at the sensor nodes displayed in figure 1 and for different leakage scenarios, which differ in the leakage location and size. As the WDN is relatively small, we are able

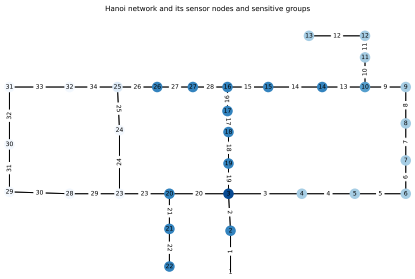

Fig. 1: The Hanoi WDN, its sensor nodes (IDs 3, 10 and 25) and the protected groups, each highlighted in another color. The sensor nodes are colored slightly darker than the node color of the protected group they belong to.

to simulate a leakage at each node in the network and for three different diameters $d \in \{5, 10, 15\}$cm. However, the methods do not require a simulation at every possible leakage location. For the preprocessing according to section 3.1, we choose $T_r = 2$ such as [2] do.

The question arises how the leakage detection is affected by the question of fairness. Knowing that each node of the network corresponds to a group of consumers, the natural question arises whether these local groups benefit from the WDN and its related services in equal degree. To ensure that the algorithms that will be presented in section 4.1 scale to larger WDNs, we do not consider single nodes but groups of nodes in the WDN as protected groups in terms of fairness. Thus, for all $k = 1, ..., K$, the sensitive feature $S_k \in \{0, 1\}$ gives answer to the question whether (1) or whether not (0) a leakage is active in the protected group k. In terms of equal service one would expect an equally good detection of leakages independent on the leakage location, i.e., the protected group. For Hanoi, we work with $K = 3$ different groups, also displayed in figure 1.

Given this definition of sensitive features S_k for $k = 1, ..., K$ in the WDN, we obtain the following important results with regard to the notions of fairness.

Lemma 1 (Disparate impact and equal opportunity are equivalent). *Let S_k be the sensitive feature describing whether a leakage is active in the protected group k of the WDN for each $k = 1, ..., K$. Moreover, let $\epsilon, \tilde{\epsilon} \in [0, 1]$ and define* $\max_k := \max_{k \in \{1,...,K\}} \mathbb{P}(\hat{Y} = 1 \mid S_k = 1)$.

1. *If disparate impact holds with ϵ, equal opportunity holds with $\tilde{\epsilon} = \epsilon \max_k$.*
2. *If equal opportunity holds with $\tilde{\epsilon}$, disparate impact holds with $\epsilon = \tilde{\epsilon}(\max_k)^{-1}$.*

Proof. First of all, note that for any $k \in \{1, ..., K\}$ and $\omega \in \Omega$, $S_k(\omega) = 1$ implies $Y(\omega) = 1$ by definition of the sensitive feature S_k. Therefore, $\{Y = 0, S_k = 1\}$ is empty. Subsequently, we obtain $\{Y = 1, S_k = 1\} = \{S_k = 1\}$ and thus, $\mathbb{P}(\hat{Y} = 1 \mid Y = 1, S_k = 1) = \mathbb{P}(\hat{Y} = 1 \mid S_k = 1)$.

Secondly, we also define $\min_k := \min_{k \in \{1,...,K\}} \mathbb{P}(\hat{Y} = 1 \mid S_k = 1)$. We then obtain $\mathrm{DI} = \frac{\min_k}{\max_k}$ and, together with the first observation, $\mathrm{EO} = \max_k - \min_k$.

Now the rest follows by simple equivalent transformations: In setting 1, we find that

$$\frac{\min_k}{\max_k} \geq 1 - \epsilon \iff \min_k \geq (1 - \epsilon) \max_k \iff \max_k - \min_k \leq \epsilon \max_k$$

holds. In setting 2, we obtain

$$\max_k - \min_k \leq \tilde{\epsilon} \iff 1 - \frac{\min_k}{\max_k} \leq \frac{\tilde{\epsilon}}{\max_k} \iff \frac{\min_k}{\max_k} \geq 1 - \frac{\tilde{\epsilon}}{\max_k}.$$

Corollary 1. *Given the setting of lemma 1,*

1. $\mathrm{EO} = (1 - \mathrm{DI}) \cdot \max_k$ *and*
2. $\mathrm{DI} = 1 - \frac{\mathrm{EO}}{\max_k}$ *holds.*

Proof. This is a direct consequence of lemma 1, where we choose $\epsilon := 1 - \mathrm{DI}$ in setting 1 and $\tilde{\epsilon} := \mathrm{EO}$ in setting 2, and where we can work with equalities instead of estimations.

3.3 Experimental Results and Analysis: Residual-Based Ensemble Leakage Detection Does Not Obey Disparate Impact or Equal Opportunity

In table 1, the results of the H-method presented in section 3.1 are shown. The choice of hyperparameter $\Theta \in \mathcal{X} = \mathbb{R}_+^{d_c}$ is chosen manually such that the test accuracy is increased. On the one hand, we see that the method in general performs better the larger the leakage size is (measured in ACC), because larger leakages are associated with larger pressure drops. Note that the method was originally tested on a larger WDN, where even small leakages can be detected with high accuracy due to a more complex network structure (cf. [2]).

On the other hand, and more importantly, we see that the method is unfair in terms of disparate impact score DI, where a value of 0.8 or larger is desirable (cf. [15]), and equal opportunity score EO. However, comparing the column of disparate impact score calculated according to equation (1) to the one according to corollary 1.2 ($\hat{\text{DI}}$), and the column of equal opportunity score calculated according to equation (2) to the one according to corollary 1.1 ($\hat{\text{EO}}$), we not only empirically prove this theoretical finding, but also justifies that in our setting, the usage of only one of the two measures is sufficient. Therefore, from now on, we work with the disparate impact score DI only.

Table 1: Results of the H-method with $\max_k = \max_k \mathbb{P}(\hat{Y} = 1 \mid S_k = 1)$ and $\min_k = \min_k \mathbb{P}(\hat{Y} = 1 \mid S_k = 1)$ according to (the proof of) lemma 1 and disparate impact score $\hat{\text{DI}} := 1 - \text{EO} \cdot (\max_k)^{-1}$ and equal opportunity score $\hat{\text{EO}} = (1 - \text{DI}) \cdot \max_k$ according to corollary 1.

d	ACC	$\max_k$	$\min_k$	DI	EO	$\hat{\text{DI}}$	$\hat{\text{EO}}$
5	0.6223	0.8468	0.4880	0.5763	0.3558	0.5763	0.3588
10	0.7998	0.9983	0.6372	0.6383	0.3611	0.6383	0.3611
15	0.8837	1.0000	0.6402	0.6402	0.3598	0.6402	0.3598

4 Fairness-Enhancing Leakage Detection in Water Distribution Networks

Motivated by the result that the standard leakage detection method presented in section 3.1 does not satisfy the notions of fairness, as another main contribution of this work, we modify this H-method to enhance fairness as introduced in section 2. The main idea is based on the fact that in the H-method the only models trained are the virtual sensors f_j^r for all $j = 1, ..., d$. However, given these virtual sensors and resulting residuals $r(t_i) \in \mathcal{X} = \mathbb{R}_+^{d_c}$ for times $t_i \in \mathbb{R}$, we can turn the choice of the hyperparameters $\Theta := (\theta_j)_{j=1,...,d} \in \mathcal{X}$ of the ensemble classifier f^c (cf. eq. (3)) into an optimization problem (OP). The corresponding function space is $\mathcal{H} := \{f^c : \mathcal{X} \to \mathcal{Y},\ r \mapsto f^c(r, \Theta) \mid \Theta \in \mathcal{X}\}$. In the following section, we present different, in contrast to the H-method optimization-based, baselines as well as fairness enhancing methods that aim at optimizing the parameter $\Theta \in \mathcal{X}$ in order to obtain an optimal classifier $f^c(\cdot, \Theta_{\text{opt.}}) \in \mathcal{H}$.

4.1 Methodology

The following methods define training algorithms based on labeled training data $\mathcal{D}^c = \{(r(t_i), y(t_i)) \in \mathcal{X} \times \mathcal{Y} \mid i = 1, ..., n_c\}$[1] for an $n_c > n_r$, which also holds data based on leaky states of the WDN. For simplicity, we omit the dependence of all (loss) functions on the training data.

Optimizing Loss with Fairness Constraints *Definition of the Learning Problem* The idea of an OP is to minimize some suitable loss function $L : \mathcal{X} \to \mathbb{R}$ with respect to the threshold $\Theta \in \mathcal{X}$, i.e.,

$$\Big\{ \min_{\Theta \in \mathcal{X}} \; L(\Theta). \tag{4}$$

The advantage of redefining the choice of hyperparameters Θ (H-method) as an OP is that we can now extend this OP by fairness constraints, which can be given by side constraints $C_k : \mathcal{X} \to \mathbb{R}$ of the underlying OP:

$$\begin{cases} \min_{\Theta \in \mathcal{X}} & L(\Theta), \\ \text{s.t.} & C_k(\Theta) \geq 0 \; \forall k = 1, ..., \hat{K}. \end{cases} \tag{5}$$

Choice of Loss Functions In view of the notions of fairness (cf. section 2), an intuitive and by means of linearity easily to differentiate loss function is given by the difference of the FPR and TPR, i.e., $L_1(\Theta) := -\operatorname{TPR}(\Theta) + \operatorname{FPR}(\Theta)$. Another classical evaluation score which we can use as a loss function is the accuracy $L_2(\Theta) := -\operatorname{ACC}(\Theta)$.

Choice of Fairness Constraints In terms of fairness constraints, [15] introduce the covariance between a single binary sensitive feature and the signed distance of a feature vector and the decision boundary of a convex margin-based classifier as a proxy to fairness measurements. We adapt this idea to our setting by considering the covariance of each sensitive feature and replacing the signed distance by the prediction of the ensemble classifier $\hat{Y} = f^c(X, \Theta)$. Using that $\hat{y}(t_i) = f^c(r(t_i), \Theta)$ holds for all realisations $i = 1, ..., n_c$, for all sensitive features S_k for $k = 1, ..., K$, the *empirical* covariance is given by

$$\operatorname{Cov}_{\text{emp.}}(S_k, \hat{Y}) = \frac{1}{n_c} \sum_{i=1}^{n_c} (s_k(t_i) - \overline{s_k}) \cdot f^c(r(t_i), \Theta). \tag{6}$$

Remark 4. The empirical covariance is based on the covariance of each sensitive feature S_k for $k = 1, ..., K$ and the prediction of the ensemble classifier $\hat{Y} = f^c(X, \Theta)$. By linearity, this covariance is given by

$$\operatorname{Cov}(S_k, \hat{Y}) = \mathbb{E}((S_k - \mathbb{E}(S_k)) \cdot (\hat{Y} - \mathbb{E}(\hat{Y})) = \mathbb{E}((S_k - \mathbb{E}(S_k)) \cdot \hat{Y})$$

However, as the probability measure $\mathbb{P}(S_k, \hat{Y})^{-1}$ on $\mathcal{Y} \times \mathcal{Y}$ is unknown, we replace it by its empirical approximation $\frac{1}{n_c} \sum_{i=1}^{n_c} \delta_{(s_k(t_i), \hat{y}(t_i))}$ and obtain the empirical covariance (6).

[1] In practise, we train and test the (ensemble) classifier(s) on unseen data for times $i \geq n_r + 1$. However, for the sake of readability, we choose the indices $i = 1, ..., n_c$ instead of $i = n_r + 1, ..., n_c$ here.

Assuming that a comparatively high (empirical) covariance (in either positive or negative direction) between a sensitive feature S_k for $k \in \{1, ..., K\}$ and the model's prediction $\hat{Y} = f^c(X, \Theta)$ implies a significant difference in the relative amount of positive predictions in contrast to the remaining sensitive features leads to the idea of constraining the absolute value of the (empirical) covariance as a side constraint in the above considered OP (cf. eq. (4)) (cf. [15]).

Table 2: Overview of the proposed methods.

Method	Loss	Constraints
T-F-PR	L_1	-
T-F-PR+F	L_1	emp. Cov.
ACC	L_2	-
ACC+F	L_2	emp. Cov.

Motivated by that, we require $\mathrm{Cov}_{\mathrm{emp.}}(S_k, \hat{Y}) \leq c$ and $\mathrm{Cov}_{\mathrm{emp.}}(S_k, \hat{Y}) \geq -c$ or, equivalently formulated in standard form, $C_k(\Theta) = c - \mathrm{Cov}_{\mathrm{emp.}}(S_k, \hat{Y}) \geq 0$ and $C_k(\Theta) = c + \mathrm{Cov}_{\mathrm{emp.}}(S_k, \hat{Y}) \geq 0$ to hold for all $k = 1, ..., K$ (i.e., $\hat{K} = 2K$ in equation (5)). Hereby, the hyperparameter $c \in [0, \infty)$ regulates how much the covariance's absolute value is bounded and therefore, the desired fairness.

Explicit Methods The resulting methods as a combination of used loss function with or without the fairness-enhancing side constraint (cf. OP (4) or (5)) deliver two baseline and two fairness-enhancing ensemble leakage detection algorithms, summarized in table 2.

Differentiable Approximation of the Learning Problems Loss function and side constraint (cf. eq. (6)) clearly depend on the model's prediction $\hat{Y} = f^c(X, \Theta)$ resp. $y(t_i) = f^c(r(t_i, \Theta))$ for all $i = 1, ..., n_c$. However, in view of the model's definition (cf. eq. (3)), f^c is not differentiable with respect to Θ. To make $\hat{Y} = f(X, \cdot)$ differentiable, we approximate the indicator function $\mathbb{1}_{\{v>0\}}$ by the sigmoid function $\mathrm{sgd}_b(v) = (1 + \exp^{-bv})^{-1}$. The larger we choose $b \in \mathbb{R}_+$, the better the indicator function is approximated, however, the more extreme gradients appear in the optimization scheme. All in all, we obtain a differentiable OP by replacing the ensemble classifier $f^c(r(t_i), \Theta)$ (cf. eq. (3)) by

$$\hat{f}^c(r(t_i), \Theta) := \mathrm{sgd}_b \left(\sum_{j=1}^{d} \mathrm{sgd}_b(r_j(t_i) - \theta_j) - T \right) \tag{7}$$

for all $i = 1, ..., n_c$, where we replace the threshold 1 of the exact ensemble classifier f^c by a hyperparameter $T \in [0, 1]$ to handle the insecurity of the sigmoid function around zero. Then, by expressing the losses $L_1 = -\mathrm{FPR} + \mathrm{TPR}$ and $L_2 = -\mathrm{ACC}$ by

$$L_1(\Theta) = -\frac{\sum_{i=1}^{n_c} y(t_i) \cdot f^c(r(t_i), \Theta)}{\sum_{i=1}^{n_c} y(t_i)} + \frac{\sum_{i=1}^{n_c} (1 - y(t_i)) \cdot f^c(r(t_i), \Theta)}{\sum_{i=1}^{n_c} (1 - y(t_i))},$$
$$L_2(\Theta) = \frac{\sum_{i=1}^{n_c} y(t_i) \cdot f^c(r(t_i), \Theta) + \sum_{i=1}^{n_c} (1 - y(t_i)) \cdot (1 - f^c(r(t_i), \Theta))}{n_c},$$

their approximated versions using $\hat{f}^c$ instead of f^c will be differentiable with respect to Θ as well, and so is the empirical covariance (cf. eq. (6)) when using $\hat{f}^c$ instead of f^c. The resulting approximated OPs can therefore be optimized with a gradient-based optimization technique.

Optimizing Fairness with Accuracy Constraints Instead of optimizing some loss function L under some fairness side constraints, [15] suggest to optimize a fairness proxy under loss constraints. They use the covariance as a proxy while constraining the training loss by some percentage of the optimal loss obtained when training without fairness considerations. As a variation, we use the disparate impact score DI directly as a loss function and the accuracy ACC for the constraint. The resulting DI+ACC-method is therefore given by

$$\begin{cases} \min_{\Theta \in \mathcal{X}} & -\,\mathrm{DI}(\Theta), \\ \text{s.t.} & \mathrm{ACC}(\Theta) \geq (1-\lambda)\,\mathrm{ACC}_{\text{opt.}}\,. \end{cases} \tag{8}$$

The hyperparameter $\lambda \in [0, 1]$ hereby regulates how much the obtained accuracy $\mathrm{ACC}(\Theta)$ is allowed to differ from the optimal accuracy $\mathrm{ACC}_{\text{opt.}}$ received in the ACC-method (cf. table 2).

In contrast to the methods proposed in section 4.1, we like to test the OP (8) as a non-differentiable OP, which therefore requires a non-gradient-based optimization technique.

Evaluation We evaluate all presented methods, i.e., the standard H-method (section 3.1), the optimization-based baselines T-F-PR- and ACC-method as well as the fairness-enhancing T-F-PR+F-, ACC+F- (cf. section 4.1 and table 2) and DI+ACC-method (cf. section 4.1), in terms of general performance, measured by accuracy (ACC), and in terms of fairness, measured by disparate impact as well as equal opportunity score DI and EO, respectively (cf. eq. (1) and (2)).

4.2 Experimental Results

Based on the pressure measurements in the Hanoi WDN as introduced in section 3.2 and the resulting residuals, we test all six methods introduced in section 3.1 (H-method) and section 4.1 (T-F-PR-, ACC-, T-F-PR+F, ACC+F, DI+ACC-method, also see Evaluation in section 4.2) in practise. The implementation can be found on GitHub[2].

Setup *H-method* We use the H-method presented in section 3.1 and tested in section 3.3 as a baseline. Subsequently, we use the hyperparameter found here as an initial parameter $\Theta_0 \in \mathcal{X}$ for the remaining optimization-based methods.

Gradient-Based Methods While the T-F-PR- and the ACC-method are used as another baseline, the remaining methods are fairness-enhancing methods. The the magnitude of fairness can be regulated by a hyperparameter: The T-F-PR+F- and ACC+F-method ensure fairness by bounding the empirical covariance of each sensitive feature and the models approximated prediction (cf. eq. (6) and (7)) by the hyperparameter $c \in [0, \infty)$. For $c = \infty$, the T-F-PR+F-method equals the T-F-PR-method and the same holds for the accuracy versions. In addition, for all these methods, we choose $b = 100$ [3] and $T = 0.8$.

[2] https://github.com/jstrotherm/FairnessInWDNS

[3] We choose b in such a way that $\hat{f}^c$ approximates f^c sufficiently well without choosing b too large such that gradients get too small in the optimization scheme.

Non-Gradient-Based Method In contrast, the DI+ACC-method regulates fairness by different choices of the hyperparameter $\lambda \in [0, 1]$ that controls how much loss in accuracy is allowed while increasing fairness.

Transforming Constraint OPs in Non-Constraint OPs For all OPs, we use the log-barrier method (cf. [9]) to transform the constrained OP into a non-constrained one and tune the regularization term per method.

Optimization Techniques For the differentiable OPs (T-F-PR-, ACC-, T-F-PR+F- and ACC+F-method), we use BFGS (cf. [9]) to find the optimal parameter $\Theta_{\text{opt.}} \in \mathcal{X}$. For the non-differentiable OP (DI+ACC-method), we use Downhill-Simplex-Search, also known as Nelder-Mead (cf. [7]). Each method is trained on 40% of the data and evaluated on the remaining data.

Results In figure 2, we see the performance of each ensemble classifier measured in accuracy and disparate impact score.[4] For the fairness-enhancing methods, we test different hyperparameters, causing error bars for these methods. We start with a hyperparameter c and λ that causes an accuracy of 0.5 and disparate impact score of 1.0 whenever possible and in- resp. decrease the hyperparameter by 0.01 until the disparate impact score of the fairness-enhancing model achieves the disparate impact score of the corresponding baseline (T-F-PR for T-F-PR+F and ACC for ACC+F and DI+ACC). The height of the bars with error bars correspond to the mean accuracy and disparate impact score, achieved by each method over all hyperparameters tested. The error bars themselves reach from the lowest to the largest score of the two scores considered.

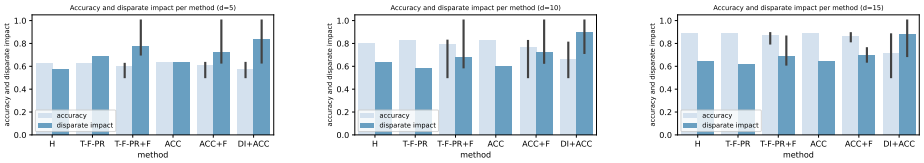

Fig. 2: Accuracy and disparate impact score per method and leakage diameter.

We see that for all fairness-enhancing methods and all leakage sizes, the fairness-enhancing methods on average increase fairness, measured by disparate impact score, while on average, mostly decreasing accuracy by only some small percentage compared to their corresponding baselines. For $d = 5$, all fairness-enhancing methods allow a large range of fairness improvement at cost of a small range in accuracy. For $d = 10$, the ranges of fairness and accuracy are similarly large. In contrast, for $d = 15$ both ranges for the gradient-based methods are small, while the non-gradient-based method offers large ranges again.

While figure 2 only hints at the relationship between fairness and overall performance, measured in disparate impact and accuracy score, respectively, a more detailed visualization of how fairness is related to the overall performance of the model can be found in figure 3. For each tested hyperparameter c and λ, respectively, depending on what fairness-enhancing method was used, the

[4] Note that by lemma 1, considering the equal opportunity score is redundant.

obtained disparate impact score is plotted together with the observed accuracy. For better readability, we split these observations by the leakage sizes tested.

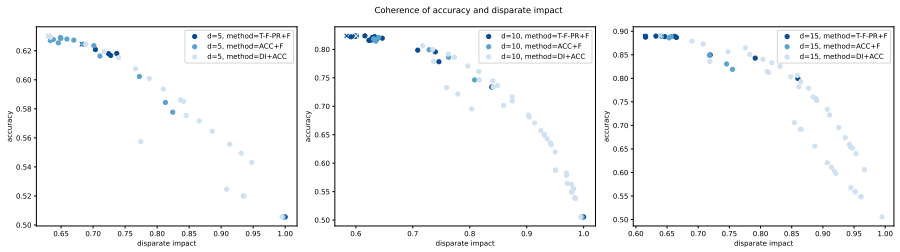

Fig. 3: Coherence of accuracy and disparate impact for the different fairness-enhancing methods and different leakage sizes. The cross data points visualize the disparate impact and accuracy of the non-fairness-enhancing baselines (T-F-PR, dark blue, for T-F-PR+F and ACC, light blue, for ACC+F and DI+ACC).

The characteristic curve that can be observed in all subimages is the so-called pareto front, showing that the improvement of the model fairness competes with the overall performance of the model and vice versa. More precisely, we see that the increase in fairness is accompanied by the reduction in accuracy score. A desired disparate impact score of about 0.8 can be achieve by a decrease of accuracy by approximately 0.03 - 0.05 points below the optimal accuracy obtained. The largest accuracies of the fairness-enhancing methods are approximately as good as the accuracy of their baseline methods while achieving equal or better fairness results. In opposite direction, perfect fairness of 1.0 can be achieved at a cost of the worst possible accuracy of 0.5. While for the covariance-based algorithms (T-F-PR+F- and ACC+F-method), the jump in disparate impact and accuracy score is rather abrupt when reaching the extreme of (1.0, 0.5), the method relying on the optimization of fairness while constraining on the accuracy (DI+ACC-method) allows more fine-grained variations in both scores.

In figure 4, we show how the hyperparameters are related to disparate impact and accuracy score. Each of the two scores are plotted against the used hyperparameter for all fairness-enhancing methods and leakage diameters tested.

For the T-F-PR+F- and the ACC+F-method, we see that the larger the covariance hyperparameter c, the smaller the disparate impact score and the larger the accuracy becomes. In contrast, for the DI+ACC-method, the disparate impact score increases with increasing hyperparameter λ while the accuracy decreases, except for some little jumps.

At this point we like to mention that due to space constraints, we can not present the coherence of equal opportunity and accuracy score and the hyperparameters, respectively. However, what our results confirm are the observations from lemma 1: For the coherence of equal opportunity and accuracy score, the results observed are equal to the ones in figure 3, but reflected along a vertical axis. For the coherence of equal opportunity score and the hyperparameters, the results equal the ones for disparate impact score in figure 4, but reflected along the horizontal axis through the point (0,0.5).

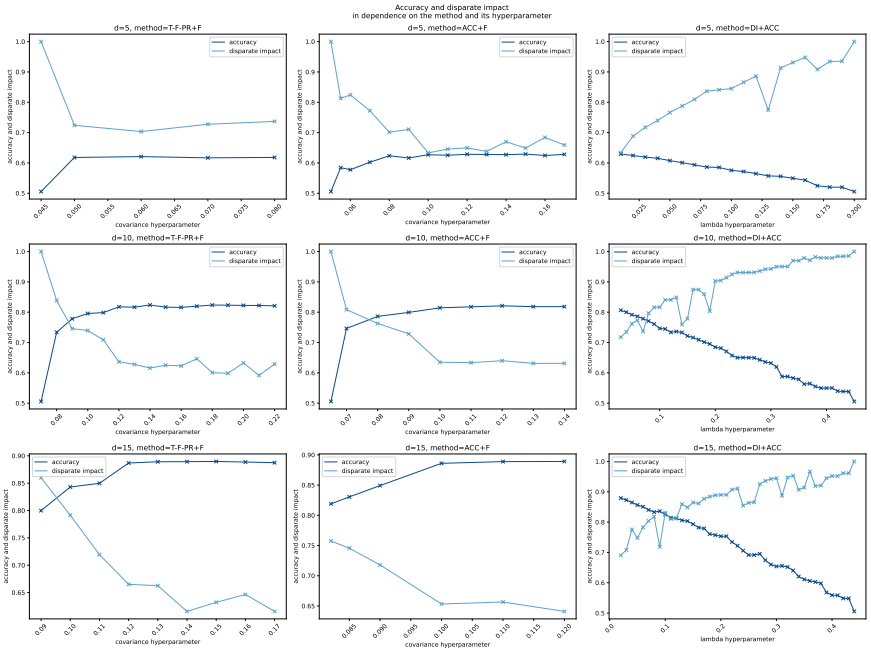

Fig. 4: Coherence of accuracy, disparate impact and the training hyperparameter.

4.3 Analysis

Increasing Fairness For all fairness-enhancing methods we could empirically prove that they lead to the desired improvement of fairness compared to their respective baselines. The relatively small range in accuracy compared to the large range of disparate impact score for $d = 5$cm is only due to the relatively poor accuracy of the leakage detector in this scenario in general. Looking at the other scenarios, one can therefore say that fairness and overall performance are mutually dependent to about the same extent.

The Coherence of Fairness and Overall Performance This equilateral dependence leads to the fact that accuracy and fairness measure define a pareto front. Both can be influenced by the fairness hyperparameters c and λ, respectively. Deciding which choice of fairness hyperparameter is optimal is a difficult task that depends on the extent of the decisions of the underlying ML model as well as legal requirements. Regarding legal requirements, by not using the sensitive features for the decision making of the algorithm, the methods presented can satisfy the legal definition of disparate treatment *and* disparate impact (depending in the hyperparameter chosen) simultaneously.

Hereby, the observation that the DI+ACC method allows more fine-grained combinations of disparate impact and accuracy score is due to the fact that the hyperparameter λ used here regulates the accuracy and not the fairness measure. The accuracy constraints are less sensitive to the log barrier method than the covariance constraints, since a too small choice of the hyperparameter c quickly sets all punishment terms to infinity and thus outputs the trivial solution.

The Influence of the Hyperparameters on Fairness and Overall Performance The observation that for the covariance-based fairness-enhancing methods the decrease of the hyperparameter c is accompanied by the improvement of the fairness measure as well as the decrease of the performance measure can be explained by the intuition described before: A high empirical covariance of a sensitive feature and the prediction of the approximated ensemble model means that the relative number of positive predictions within the related group differs significantly from the relative number of positive predictions within a group with small covariance. Thus, the more the covariance is constrained, the less such extreme differences in the relative number of positive predictions across groups occur, leading to a better fairness score. In the case of disparate impact, therefore, a higher score at the expense of a lower overall performance - compared to the overall performance that occurs in the unconstrained case or for a looser constraint, that is a larger bound c, - appears.

In contrast, the observation that for the DI+ACC method the increase of the hyperparameter λ is accompanied by the improvement of the fairness measure as well as the decrease of the performance measure is due to the fact that a higher hyperparameter λ allows a larger deviation of the optimal accuracy score. Thus, a worse accuracy is penalized less or not at all, so that the fairness measure can be optimized to a larger extend.

Last but not least, note that the non-optimal solutions and the local jumps recognized in figure 3 and 4, respectively, can be explained by the to be optimized objectives not being convex. Therefore, the found solutions strongly depend on the initialized parameter Θ_0 and might not correspond to the global optimum.

5 Conclusion

In this work, we introduced the notion of fairness in an application domain of high social and ethical relevance, namely in the field of water distribution networks (WDNs). This required the extension of fairness definitions for a single binary sensitive feature to multiple, even non-binary, sensitive features. We then investigated on the fairness issue in the area of leakage detection within WDNs. We showed that standard approaches are not fair in the context of different groups related to the locality within the network. As a remedy, we presented methods that increase fairness of the ensemble classification model with respect to the introduced fairness notion while satisfying the legal notions of disparate treatment and disparate impact simultaneously. We empirically demonstrated that fairness and overall performance of the model are interdependent and the use of hyperparameters provides the ability to trade off fairness and overall performance. However, this trade off lies in the responsibility of the policy maker.

To allow more fine-grained steps between improving fairness and decreasing overall performance in the presented covariance-based approaches, next steps would be to swap loss function and constraint to achieve similar results as in the approach with accuracy constraint. Moreover, the notion of fairness within the water domain is still in its beginning and extensions to more complex WDNs as well as more powerful ML algorithms is essential.

6 Acknowledgments

We gratefully acknowledge funding from the European Research Council (ERC) under the ERC Synergy Grant Water-Futures (Grant agreement No. 951424).

References

1. Angwin, J., Larson, J., Mattu, S., Lauren Kirchner, L.: Machine Bias - There's software used across the country to predict future criminals. And it's biased against blacks. ProPublica (2016)
2. Artelt, A., Vrachimis, S., Eliades, D., Polycarpou, M., Hammer, B.: One Explanation to Rule them All – Ensemble Consistent Explanations. arXiv preprint arXiv:2205.08974 (2022)
3. Barocas, S., Hardt, M., Narayanan, A.: Fairness and Machine Learning: Limitations and Opportunities. fairmlbook.org (2019), http://www.fairmlbook.org
4. Castelnovo, A., Crupi, R., Greco, G., Regoli, D., Penco, I.G., Cosentini, A.C.: A clarification of the nuances in the fairness metrics landscape. Scientific Reports **12**(1), 4209 (2022)
5. Commission, E., Directorate-General for Communications Networks, C., Technology: Ethics guidelines for trustworthy AI. Publications Office (2019). https://doi.org/doi/10.2759/346720
6. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.: Fairness Through Awareness. In: Proceedings of the 3rd innovations in theoretical computer science conference. pp. 214–226 (2012)
7. Gao, F., Han, L.: Implementing the Nelder-Mead simplex algorithm with adaptive parameters. Computational Optimization and Applications **51**(1), 259–277 (2012)
8. Isermann, R.: Fault-Diagnosis Systems. Springer Berlin Heidelberg, Berlin, Heidelberg (2006)
9. Jorge Nocedal, S.J.W.: Numerical Optimization, vol. 02. Springer New York (2006)
10. 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)
11. Pessach, D., Shmueli, E.: A Review on Fairness in Machine Learning. ACM Computing Surveys (CSUR) **55**(3), 1–44 (2022)
12. Santos-Ruiz, I., López-Estrada, F.R., Puig, V., Valencia-Palomo, G., Hernández, H.R.: Pressure Sensor Placement for Leak Localization in Water Distribution Networks Using Information Theory. Sensors **22**(2), 443 (2022)
13. Vaquet, J.: Automation Toolbox for Machine learning in water Networks. https://pypi.org/project/atmn/ (2023)
14. Vaquet, V., Ashraf, I., Hinder, F., Artelt, A., Heihoff, B., Lammers, K., Vaquet, J., Strotherm, J., Brinkrolf, J., Hammer, B.: ML Challenges in Water Distribution Networks - a Survey and a Graph Neural Network-Based Solution for Leakage Detection and Localization. Submitted to IJCNN 2023 (2023)
15. Zafar, M.B., Valera, I., Rogriguez, M.G., Gummadi, K.P.: Fairness Constraints: Mechanisms for Fair Classification. In: Artificial intelligence and statistics. pp. 962–970. PMLR (2017)