FARF: A Fair and Adaptive Random Forests Classifier

Wenbin Zhang[1], Albert Bifet[2,3], Xiangliang Zhang[4], Jeremy C. Weiss[5], and Wolfgang Nejdl[6]

[1] University of Maryland, Baltimore County, MD 21250, USA
[2] University of Waikato, Hamilton 3216, New Zealand
[3] Télécom Paris, Institut Polytechnique de Paris, Palaiseau 91764, France
[4] King Abdullah University of Science and Technology, Thuwal 23955, Saudi Arabia
[5] Carnegie Mellon University, Pittsburgh, PA 15213, USA
[6] L3S Research Center & Leibniz University Hannover, Hannover 30167, Germany
[1]wenbinzhang@umbc.edu, [2,3]abifet@waikato.ac.nz,
[4]xiangliang.zhang@kaust.edu.sa, [5]jeremyweiss@cmu.edu, [6]nejdl@L3S.de

Abstract. As Artificial Intelligence (AI) is used in more applications, the need to consider and mitigate biases from the learned models has followed. Most works in developing fair learning algorithms focus on the offline setting. However, in many real-world applications data comes in an online fashion and needs to be processed on the fly. Moreover, in practical application, there is a trade-off between accuracy and fairness that needs to be accounted for, but current methods often have multiple hyper-parameters with non-trivial interaction to achieve fairness. In this paper, we propose a flexible ensemble algorithm for fair decision-making in the more challenging context of evolving online settings. This algorithm, called FARF (Fair and Adaptive Random Forests), is based on using online component classifiers and updating them according to the current distribution, that also accounts for fairness and a single hyper-parameters that alters fairness-accuracy balance. Experiments on real-world discriminated data streams demonstrate the utility of FARF.

1 Introduction

AI-based decision-making systems are routinely being used across a wide plethora of online (e.g., the targeting of products, the setting of insurance rates) as well as offline services (e.g., the issuing of mortgage approval, the allocation of health resource). As AI becomes integrated into more systems, various AI-based discriminatory incidents have also been observed and reported [3, 18, 24].

A large number of methods have been proposed to address this issue, ranging from discrimination discovery to discrimination elimination and interpretation in order to provide ethical and accurate decisions [28, 30]. These studies have typically adopted one or more of the three following strategies: i) *Pre-processing solutions* aim to eliminate discrimination at the data level, including the most popular ones massaging [21] and reweighting [9]. ii) *In-processing approaches*

mitigate bias by modifying the algorithm design [4, 22]. As a recent example, the Bayesian probabilistic modeling is leveraged to account for fairness [15]. iii) *Post-processing techniques* consist of a-posteriori adjusting the output of the model [18, 19]. For instance, the decision boundary for the protected group is shifted based on the theory of margins for boosting [14].

However, most of these methods tackle fairness as a static problem, i.e., that all the data is available at training time. This does not satisfy situations that may require online learning due to a continuously drifting data distribution, or can not computationally afford to process all of their data in memory [29]. There is very little work in the area of online learning that includes any definition of fairness as a goal of the method [20, 27]. Our work seeks to fill this void.

Current methods also lack a mechanism for easily adjusting the trade-off that exists between accuracy and fairness [23]. For instance, the "business necessity" clause [2] states that a certain degree of disparate impact discrimination can be allowed for the sake of meeting certain performance-related business constrains, on the condition that such decision-making causes the least disparate impact when fulfilling the current business needs. If an initial model fails to meet the discrimination or accuracy requirement for practical use, we would prefer there exist a single parameter with a direct and predictable impact on this trade-off. However, current studies solely focus on preserving prediction performance while minimizing discrimination, and do not allow for fine-grained control between fairness and accuracy [3, 30].

To overcome these issues we propose FARF, an online statistical parity aware Random Forest (RF) model. Like prior online RF algorithms, it is built from a sampling approach for the ensemble creation. In creating this fair variant of RF, we develop a number of contributions: i) We study a new research direction of fairness-aware learning considering concept and fairness drift. We then propose FARF, a fairness-aware and fairness-updated ensemble method to tackle online fairness. ii) We study another research direction of fairness-aware learning with customized control, and design a clear mechanism for fine-grained fairness control, providing more flexibility than state of the art. iii) We theoretically analyze the inadequacy of current sampling approaches in fairness studies and introduce a new effective sampling direction with experimental verification. iv) Extensive experimental evaluation on real-world datasets demonstrates the capability of the proposed model in online settings.

2 Problem Definition

An online stream D consists of a sequence of instances arriving over time, potentially infinite. One instance x_t at time step t in D is described in a feature space $A = \{A_1, ..., A_n\}$ within respective domains $dom(A_i)$ and its class label C_t. An online classifier is trained incrementally by taking instances up to time t to predict C_{t+1} for the unlabeled instance arriving at time step $t+1$. Once C_{t+1} is predicted, the actual class label of x_{t+1} becomes available and can be used for model update, known as prequential evaluation [16].

We assume one of the attributes A is a special attribute S, referred to as *sensitive attribute* (e.g., gender) with a special value $s \in dom(S)$ referred to as *sensitive value* (e.g., female), from which the discriminated group is defined. For simplicity, we consider binary classification tasks assuming $dom(C) \in \{+, -\}$ and S also is binary with $dom(S) \in \{p, u\}$ (i.e., protected and unprotected respectively). Four fairness related groups can therefore be distinguished combining S and C. These groups are p^+, p^- and u^+, u^- representing protected group (e.g., female) receiving positive and negative classification and unprotected group (e.g., male) receiving positive and negative classification, respectively.

Although more than twenty notions have been proposed to measure the discriminative behavior of AI models [26], formalizing fairness is a hard topic per se, and there is no consensus which measure is more versatile than others [3]. In addition, what constitutes "fair" or "discriminative" is dependent on many factors and context, as well as philosophical questions that have been researched long before the AI communities interest [7]. In this work, we adopt the *statistical parity* because American user studies have found that it is a measure compatible with many users' intuition of what constitutes a "fair" decision [25], expecting a wide spectrum of applications of our method. Briefly, statistical parity examines whether the probability of being granted for a positive benefit (e.g., the provision of health care) is the same for both protected and unprotected groups. While statistical parity is designed for offline fairness, the discriminative behavior of the AI model up to time t in the online setting, which we term as *accumulated statistical parity*, can be analogically defined as:

$$Disc(D_t) = \frac{u_t^+}{u_t^+ + u_t^-} - \frac{p_t^+}{p_t^+ + p_t^-} \tag{1}$$

where u_t^+, u_t^-, p_t^+ and p_t^+ are up to time t the number of individuals from respective groups.

People from the protected group can claim they are discriminated up to time t when more of them are rejected a benefit comparing to the people of the unprotected group. The aim of online fairness-aware learning is therefore to provide real time accurate but also fair predictions from the massive data streams, where D needs to be processed on the fly without the need for storage and reprocessing, and data distribution including $Disc(D_t)$ could also evolve over time.

3 The Fair and Adaptive Random Forests

Ensemble learning combines multiple base learners to generate more robust descriptions. Three common strategies are bagging, boosting and random forests. Specific to online learning, there are multiple versions of bagging and boosting that are part of the state of the art ensemble methods for evolving online learning [6, 11], while random forests for non-stationary data stream are currently represented by [1, 17], which also show random forests approaches have

a superior performance comparing to bagging and boosting methods. One possible reason is that training on sampled data and selected features for splitting generalize more than adding more random weights to instances by bagging and adding weights to incorrectly classified instances by boosting. In this paper, we follow the idea of online random forests [1, 17] as a powerful tool to increase the generalization and fairness when constructing an ensemble of classifiers.

Specifically, the proposed Fair and Adaptive Random Forests (FARF) is an adaptation of the classical random forest algorithm [8], and can also be viewed as an updated and fairness-aware version of the previous attempts to perform this adaptation [1, 17]. In comparison to these attempts, FARF proposes a theoretically sound and fairness-oriented sampling (Section 3.2), an updated adaptive strategy (Section 3.3) as well as employing a fairness-aware base learner also for ensemble diversity (Section 3.1) to cope with discriminatory evolving data streams collectively. The following subsections elaborate these three improvements one by one.

3.1 Diversified Fairness-aware Base Learner

Most of the existing online ensemble approaches [1, 17] induce their base learners based on the Hoeffding Tree (HT) algorithm [13], which exploits the fact that an optimal splitting attribute can be determined by a small sample and the learned model is asymptotically nearly identical to that of a conventional non-incremental learner. However, such induction is based on the *information gain (IG)* aiming to optimize for predictive performance and does not account for fairness. In our previous work [27], the *fair information gain (FIG)* is proposed as an alternative tree splitting criterion to address the discrimination issue of IG, formally put,

$$FIG(D, A) = \begin{cases} IG(D, A), & \text{if } FG(D, A) = 0 \\ IG(D, A) \times FG(D, A), & \text{otherwise} \end{cases} \tag{2}$$

where *fairness gain* (FG) measures the discrimination difference due to the splitting and is formulated as:

$$FG(D, A) = |Disc(D)| - \sum_{v \in dom(A)} |Disc(D_v)| \tag{3}$$

where D is the collection of instances and A represents the attribute that under evaluation, $D_v, v \in dom(A)$ are the partitions induced by A, and the resultant discrimination value is assessed according to Equation (1). In FIG, multiplication is favoured, when combining IG and FG as a conjunctive objective, over other operations for example addition as the values of these two metrics could be in different scales, and in order to promote fair splitting which results in a reduction in the discrimination after split, i.e., FG is a positive value.

In FARF, other than the discrimination reduction merit similar to the previous fairness-driven IG reformulation efforts [22, 27], such splitting criterion also detects local discrimination to increase diversity for the sake of maximizing the

accumulated fairness. Specifically, each partition induced by the attribute A contributes equally to the accumulated fairness of A regardless the number and size of branches. In the context of ensemble learning, diversity of the each individual classifier plays a key role. Increasing diversity by eyeing on local discrimination, i.e., identifying certain attribute values with a high discrimination rate but small in representation size, could therefore induce diversified base classifiers, reflecting different discrimination representation and improving the final ensemble capability. Such emphasis can also be regarded as selecting those attributes that otherwise would not be used for splitting thus adding more randomization for the construction of the tree.

This diversified fairness-aware learner therefore learns different attribute value level discrimination during the tree construction to maximize the accumulated fairness, and is used as the base learner of FARF. To align with such diversity-promoting strategy, different from the base learner of the previous ensemble approaches [1, 17], FARF also does not perform early tree pruning for its base learners, and a random subset of fair features are selected for new split attempts to further encourage diversity.

3.2 Fairness-aware Sampling

In batch random forests, each base classifier is trained on a bootstrap of the entire training set. However, such bootstrap replicates sampling strategy is infeasible in online setting as each training instance needs to be processed once "on arrival" without reprocessing. Oza et al. [6] simulate the construction of bootstrap replicates in online context by sending K copies of each training instance to update the base classifier accordingly, where K is a suitable Poisson random variable. Considering the arbitrary length of online stream, we follow [6] that found setting

$$\mathrm{K} = Poisson(6) \tag{4}$$

to have the best accuracy by increasing diversity of the base learners. Others have consistently found this approach effective in accuracy and computing requirements [17]. Then the latest arriving instances can be classified by voting of the base learners, the same way in online and batch random forests. We will propose two different methods of altering the sampling of K to encourage fair tree induction.

Sampling techniques have been studied in recent fairness-aware learning approaches to alleviate discrimination [4, 19]. In these studies, they exclusively concentrate on **over-sampling the protected positive group** through different heuristics. However, we argue that such interventions are insufficient especially in online setting for two reasons. First, the protected positive group is normally the under-represented minority. Solely focusing on sparse representation might not have significant bias mitigation effect. Such ineffectiveness is further exacerbated in online setting as instances from the protected positive group could

discontinue for a certain period of time. Second, over-sampling protected positive group in random forest can be regarded as **minority over-sampling with replacement**. Previous research has noted that it does not significantly improve minority class recognition [10]. We interpret the underlying effect in terms of spreading the decision regions of protected positive group to mitigate biases. Essentially, as protected positive group is over-sampled by increasing amounts, the effect is to learn qualitatively similar but more specific regions that overfit the protected positive group rather than spreading its decision boundary into the unprotected positive group region.

Therefore, instead of over-sampling protected positive group, our ensemble learning method **under-samples the unprotected positive group** to mitigate the discrimination. We design the update rule for instance weight for sampling as:

$$fairK(x_t) = \begin{cases} Disc(D_t) * K, \text{if } x_t \in u^+ \& Disc(D_t) > 0 \\ K, \qquad\qquad \text{otherwise} \end{cases} \tag{5}$$

where $Disc(D_t)$ measures the accumulated discrimination up to the current instance at time t in the stream and K is the Poisson weight defined in Equation (4). When the current accumulated discrimination is positive ($Disc(D_t) > 0$), i.e., protected group has been discriminated, and the current instance is a member of unprotected positive group, the sampling weight $fairK(x_t)$ is down-scaled for the current instance x_t, making it to be $Disc(D_t)$ proportional of Poisson weight K. When there is no membership discrimination against the protected group or the current instance belongs to unprotected group, $fairK(x_t)$ is equivalent to the Poisson weight K. This allows our models to learn a more effective decision surface for the unprotected group, while avoiding prior shortcomings to sampling based fairness.

Other than exclusively focusing on over-sampling the protected positive group, the previous fair sampling studies also require additional neighborhood information through KNN [4] and clustering [19]. On the contrary, sampling in our work is directly defined in terms of the targeting discrimination. While enjoying simplicity, this also opens the door to flexible control on the degree of fairness. Specifically, we present a second method of altering the sampling ratio K that allows the user to control a trade off between model accuracy and fairness by manually customizing the re-scaling ratio in $fairK$ to manage the trade-off. This is done with a fixed under-sampling weight α that is incorporated into an alternative equation $customK$ as:

$$customK(x_t) = \begin{cases} \alpha * K, & \text{if } x_t \in u^+ \\ K, & \text{otherwise} \end{cases} \tag{6}$$

where α is the tunable parameter adjusting the sampling ratio. Note that like $fairK$, the under-sampling only occurs for positive instances of the unprotected group. Such flexible control on the degree of fairness instantiates application-wise fairness-aware learning to accommodate scenarios such as the "business necessity" clause [2].

3.3 FARF Algorithm

Online fairness additionally requires learning algorithms process each instance upon arrival as well as dealing with non-stationary data distribution indicating concept drifts and fairness implications. That is to say, the relationship between sensitive attribute and class variable might also change over time. A stream classifier pays attention to the boundary evolution but ignores fairness drift. To this end, FARF encapsulates the capability of fairness drift detection and adaptation as well as standby trees and weighted voting to address online fairness comprehensively.

Ensemble learning has been used as a powerful tool by resetting underperforming base learners to adapt to change quickly. The conventional approach resets base learners the moment a drift is detected [6]. However, such reseting could be ineffective since the reseted learner cannot have a positive impact on the ensemble process as it has not been well trained. To this end, FARF employs a more permissive threshold to detect potential drifts and builds standby trees for ensemble members who detect such drifts. The standby trees are trained along the ensemble without intervening the ensemble prediction, and appear on the stage when they outperform their respective ensemble members.

The ensemble design of FARF also offers space for different change detectors being incorporated. One possible detector is ADWIN [5], which recomputes online whether two "large enough" subwindows of the most recent data exhibit "distinct enough" averages, and the older portion of the data is dropped when such distinction is detected. Different from the previous non-stationary studies [11, 17], FARF employs ADWIN to detect changes in accuracy but also fairness, reflecting both concept and fairness drifts. That is to say drift is detected when either of them evolves.

FARF also weights the prediction of each base learner in proportion to their prequential evaluation [16] fairness since its last reset, reflecting the tree performance on the current fairness distribution. Such weighting scheme enjoys the merit of free of predefined window or fading factor to estimate fairness as in other stream ensembles [1, 17] (their estimation focus is accuracy to reflect concept drift though). Note that FARF prioritizes fairness over accuracy by weighting and replacing ensemble members according to fairness. Algorithm 1 shows the sketch of FARF.

For each new instance (line 2), FARF first decides its weight according to fairness-aware sampling based on its fairness information and the accumulated discrimination up to the current instance (line 5-7). When customizable fairness is deployed, the weight is set according to customized sampling ratio (line 3-4). FARF then trains each ensemble member (line 9) with this weight (line 10). When a change is detected (line 11) in one ensemble member who does not have a standby tree (line 12), a respective standby tree is created (line 13), otherwise performances between the ensemble member and its respective standby tree are compared (line 15) to decide ensemble membership replacement if needed (line 16). All standby trees are also trained along the ensemble (line 21-22). The weighted vote can be performed at anytime to predict the class of an instance

Algorithm 1: FARF Leaning Algorithm

Input: a discriminated data stream D, the number of base models M, optional sampling ratio α

1 Init base models h_m for all m $\in$ {1, 2, ..., M}
2 **for** *each instance* x_t *in* D **do**
3 **if** α *specified* **then**
4 $w_t \leftarrow customK(x_t)$ according to Equation (6);
5 **else**
6 Calculate $Disc(D_t)$ according to Equation (1);
7 $w_t \leftarrow fairK(x_t)$ according to Equation (5);
8 **end**
9 **for** *m= 1, 2, ..., M* **do**
10 Update h_m with x_t with weight w_t;
11 **if** *ADWIN detects a change in fairness or accuracy in* h_m **then**
12 **if** *standby learner* $h'_m = \emptyset$ **then**
13 Build a new diversified fair standby learner h'_m;
14 **else**
15 **if** $|Disc(h_m)| > |Disc(h'_m)|$ **then**
16 Replace h_m with h'_m;
17 **end**
18 **end**
19 **end**
20 **end**
21 **for** ***all*** h'_m **do**
22 Update h'_m with x_t with weight w_t;
23 **end**
24 **end**
25 **anytime output:** $h(x_t) = argmax_{c \in C} \sum_{m=1}^{M} W(h_m(x_t) = \mu_m(c))$

(line 25). Note that the replacement and voting could also be performed from the accuracy perspective, i.e., replacing the ensemble member when its error is higher and weighted vote on accuracy instead. FARF does fairness replacement and voting in order to prioritize fairness at these steps.

4 Experimental Evaluation

In the case of static datasets and evaluation, accepted benchmarks for evaluating fairness mitigating approaches are limited in number [3]. With respect to the highly under-explored online fairness, this challenge is further magnified by the drift and the demanding requirement of the number of instances contained therein. We evaluate our approach on the datasets used in the recent works of this research direction [20, 27], the *Adult* and the *Census* datasets [12] both targeting the learning task of determining whether a person earns more than 50K dollars per annum. We follow the same options in our experiments for fair comparison including the selection of sensitive attribute "gender" with female being

the sensitive value and processing them in sequence. One difference is that instead of randomizing the order, we order the datasets by the "race" attribute for both datasets to better simulate concept drift and possibly increase the learning bias. The previous discussed prequential evaluation is employed for evaluation.

4.1 Benchmark Performance

This section first investigates the theoretically designed fairness-aware and fairness-updated capabilities of FARF. For comparison, we implemented two recently proposed fair online learners, FEI [20] and FAHT [27]. While the paper of FEI did not compare with any baselines, FAHT studied two. We compare with these two baselines therein as well, namely the Hoeffding Tree (HT) and KHT in which the fairness-aware splitting criterion proposed in [22] is embedded into HT. We also trained the state of the art concept-adapting ensemble learner ARF [17] as another baseline. Other competing fairness methods, including recent proposed fairness ensemble methods which require multiple full data scan, are not considered as none of them can be transferred to online settings. All methods are trained the same way for fair comparison. Relevant results on all datasets are shown in Table 1. Note that since accuracy can be misleading for imbalanced class distributions, we also report Kappa statistics [16].

Table 1. The predictive performance-vs-discrimination between FARF and baseline models. Best results in **bold**, second best in *italics*.

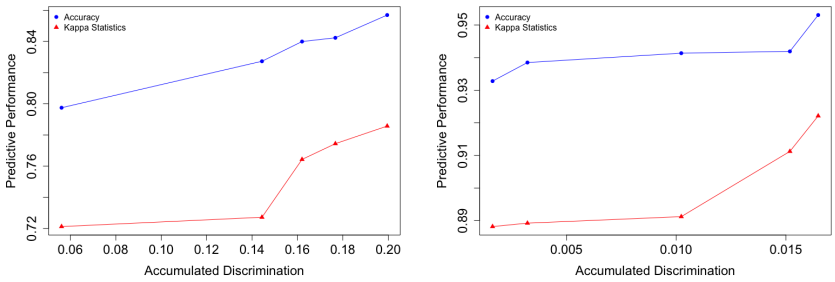
Metric / Methods	**Adult dataset**			**Census dataset**		
	Disc%	Acc%	Kappa%	Disc%	Acc%	Kappa%
HT	24.14	82.16	68.15	6.61	93.11	87.54
KHT	24.24	82.43	67.2	6.74	93.26	87.12
FAHT	*17.20*	81.62	70.48	*3.63*	93.06	88.14
ARF	24.17	**84.51**	**78.15**	6.64	*94.18*	**90.41**
FEI	23.06	74.27	54.27	6.64	80.06	84.27
FARF	**8.89**	*84.19*	*77.54*	**0.07**	**94.83**	*90.33*

As shown in Table 1 our new FARF method dominates all other baselines in terms of minimizing discrimination, and is best of second-best by both Accuracy and Kappa scores in all other cases. We note that when second best FARF is still highly competitive, being at most 0.78% within the top performer. This is a desirable trade-off since FARF reduced the discrimination score by a factor of 1.9× and 51.8× for Adult and Census dataset, respectively.

4.2 Accuracy-Fairness Control

The design of FARF provides a clear mechanism to manage the trade-off between fairness and accuracy. This can be necessary when an initial model does not meet one of these requirements, allowing the end-user to make adjustments. FARF

controls thus with the α parameter. As α is in proportion to accuracy, increasing its value leads to a higher accuracy at the expense of a higher discrimination. Such expected trend is clear from the results visualized in Figure 1. Clients can therefore accommodate their needs according to their respective constraints.



(a) Adult dataset (b) Census dataset

Fig. 1. The predictive performance and accumulated discrimination trade-off fined-grained by the tunable parameter α ranging from 0.3 to 1.5 with step size 0.3.

The x-axis of the above figure is with respect to the amount of discrimination that is present (larger values indicate more discrimination), and the y-axis is the predictive accuracy (larger is more accurate). With respect to both accuracy and Kappa scores we see a monotonic behavior with respect to the α parameter. This means it behaves as we desire: a simple and direct relationship controlling the trade-off between accuracy and statistical parity. This makes it easy to use, compared to most methods that have multiple parameters that all need to be adjusted to achieve a satisficing trade-off [23].

4.3 Justification of Sampling Component in FARF

Recent fairness-aware learning approaches employ sampling techniques to mitigate bias, which exclusively focus on over-sampling protected positive group through different heuristics. We theoretically discussed the drawbacks of these methods (c.f., Section 3.2). This section provides experimental justification and verifies our choice to instead under sample the protected positive group and that it is critical to our results. We perform two ablations to confirm this by replacing our sampling with: 1) over-sampling protected positive group, and 2) over-sampling protected positive group and under-sampling unprotected positive group. All other components of our approach remain the same so that we can isolate our sampling approach as the critical factor in results. These two types of ensemble are denoted as **FARFS**$^{-}$ and **FARFS**$^{-+}$ respectively in comparison with **RF**, which refers to random forests without sampling intervention, and our proposed FARF. The results are shown in Table 2.

Table 2. The predictive performance-vs-discrimination comparison between different sampling strategies. Best results in **bold** second best in *italics*.

Methods \ Metric	**Adult dataset**			**Census dataset**		
	Disc%	Acc%	Kappa%	Dis%	Acc%	Kappa%
RF	16.32	**84.31**	**78.05**	1.34	*94.13*	**90.37**
FARFS^{-}	19.36	83.26	73.47	1.10	94.17	90.24
FARFS^{-+}	*10.53*	81.64	72.49	*0.45*	93.95	89.15
FARF	**8.89**	*84.19*	*77.54*	**0.07**	**94.83**	*90.33*

As can be seen FARF is the only method that consistently obtains accuracy near that of an unconstrained Random Forest. At the same time, neither approach is able to reach discrimination rates as low as FARF. This shows that over-sampling approaches of prior fairness studies are not as effective as our under-sampling based approach.

5 Conclusions

Our work has proposed the first online version of Random Forests with fairness constraints. Our design includes a mechanism for altering the trade off between accuracy and fairness so that users can adjust it easily toward their specific applications. In doing so we have show positive results compared to alternative methods available, without compromising on the desirable properties of online Random Forests.

References

1. H. Abdulsalam, D. B. Skillicorn, and P. Martin. Classifying evolving data streams using dynamic streaming random forests. In *DEXA*, pages 643–651. Springer, 2008.
2. S. Barocas and A. D. Selbst. Big datas disparate impact. *California Law Review*, 104(3):671, 2016.
3. A. Beutel, J. Chen, T. Doshi, H. Qian, A. Woodruff, C. Luu, P. Kreitmann, J. Bischof, and E. H. Chi. Putting fairness principles into practice: Challenges, metrics, and improvements. *AIES*, 2019.
4. D. Bhaskaruni, H. Hu, and C. Lan. Improving prediction fairness via model ensemble. In *ICTAI*, pages 1810–1814, 2019.
5. A. Bifet and R. Gavalda. Learning from time-changing data with adaptive windowing. In *SDM*, pages 443–448, 2007.
6. A. Bifet, G. Holmes, and B. Pfahringer. Leveraging bagging for evolving data streams. In *ECML PKDD*, pages 135–150. Springer, 2010.
7. R. Binns. Fairness in machine learning: Lessons from political philosophy. In *Conference on Fairness, Accountability and Transparency*, pages 149–159, 2018.
8. L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.
9. T. Calders, F. Kamiran, and M. Pechenizkiy. Building classifiers with independency constraints. In *ICDMW*, pages 13–18, 2009.
10. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: synthetic minority over-sampling technique. *JAIR*, 16:321–357, 2002.

11. S.-T. Chen, H.-T. Lin, and C.-J. Lu. An online boosting algorithm with theoretical justifications. In *ICML*, pages 1873–1880, 2012.
12. D. Dheeru and E. Karra Taniskidou. UCI machine learning repository, 2017.
13. P. Domingos and G. Hulten. Mining high-speed data streams. In *KDD*, pages 71–80. ACM, 2000.
14. B. Fish, J. Kun, and Á. D. Lelkes. A confidence-based approach for balancing fairness and accuracy. In *SDM*, pages 144–152, 2016.
15. J. R. Foulds, R. Islam, K. N. Keya, and S. Pan. Bayesian modeling of intersectional fairness: The variance of bias. In *SDM*, pages 424–432, 2020.
16. J. Gama. *Knowledge discovery from data streams*. Chapman and Hall/CRC, 2010.
17. H. M. Gomes, A. Bifet, J. Read, J. P. Barddal, F. Enembreck, B. Pfharinger, G. Holmes, and T. Abdessalem. Adaptive random forests for evolving data stream classification. *Machine Learning*, 106(9-10):1469–1495, 2017.
18. M. Hardt, E. Price, N. Srebro, et al. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323, 2016.
19. V. Iosifidis, B. Fetahu, and E. Ntoutsi. Fae: A fairness-aware ensemble framework. In *IEEE International Conference on Big Data (Big Data)*, pages 1375–1380, 2019.
20. V. Iosifidis, T. N. H. Tran, and E. Ntoutsi. Fairness-enhancing interventions in stream classification. In *DEXA*, pages 261–276. Springer, 2019.
21. F. Kamiran and T. Calders. Classifying without discriminating. In *2nd International Conference on Computer, Control and Communication*, pages 1–6, 2009.
22. F. Kamiran, T. Calders, and M. Pechenizkiy. Discrimination aware decision tree learning. In *ICDM*, pages 869–874, 2010.
23. J. Kleinberg, S. Mullainathan, and M. Raghavan. Inherent Trade-Offs in the Fair Determination of Risk Scores. In *FAT ML Workshop*, 2016.
24. D. Meyer. Amazon reportedly killed an ai recruitment system because it couldnt stop the tool from discriminating against women. fortune, october 10, 2018.
25. M. Srivastava, H. Heidari, and A. Krause. Mathematical notions vs. human perception of fairness: A descriptive approach to fairness for machine learning. In *KDD*, pages 2459–2468, 2019.
26. S. Verma and J. Rubin. Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*, pages 1–7. IEEE, 2018.
27. W. Zhang and E. Ntoutsi. Faht: an adaptive fairness-aware decision tree classifier. In *IJCAI*, pages 1480–1486, 2019.
28. W. Zhang, X. Tang, and J. Wang. On fairness-aware learning for non-discriminative decision-making. In *ICDMW*, pages 1072–1079. IEEE, 2019.
29. W. Zhang and J. Wang. A hybrid learning framework for imbalanced stream classification. In *2017 IEEE International Congress on Big Data (BigData Congress)*, pages 480–487. IEEE, 2017.
30. I. Zliobaite. A survey on measuring indirect discrimination in machine learning. *arXiv preprint arXiv:1511.00148*, 2015.