

Fairness in the Flow: An Overview and Drift-Aware Evaluation of Fair Classifiers for Drifting Data Streams

KATHRIN LAMMERS, Bielefeld University, Germany

MARYBETH DEFRANCE, Ghent University, Belgium

BARBARA HAMMER, Bielefeld University, Germany

VALERIE VAQUET, Bielefeld University, Germany

Machine learning systems increasingly inform high-stakes decisions, making fairness, transparency, and reliability essential. While algorithmic fairness has been widely studied in the traditional batch setting, its extension to stream learning, where data arrives continuously and distributions may evolve, remains underexplored. This paper surveys and systematizes the emerging field of fair stream learning, highlighting challenges such as fairness–performance trade-offs, temporal dynamics, and class imbalance. We first give an overview over the foundations of both fairness and stream learning as separate fields. Building on this foundation, we provide a comprehensive overview of existing fair stream learning methods and propose a drift-aware evaluation framework. Our empirical analyses, for the first time comprehensively utilizing both drift-aware as well as cumulative score aggregation, compare representative approaches, offering insights into their behavior under evolving data conditions and identifying directions for future research.

Keywords: Fairness, Stream Learning, Concept Drift, Evaluation

Reference Format:

Kathrin Lammers, MaryBeth Defrance, Barbara Hammer, and Valerie Vaquet. 2026. Fairness in the Flow: An Overview and Drift-Aware Evaluation of Fair Classifiers for Drifting Data Streams. In *Proceedings of Fifth European Conference on Algorithmic Fairness (ECAF’26)*. Proceedings of Machine Learning Research, 22 pages.

1 Introduction

Machine learning algorithms are becoming part of many decision-making processes in everyday life, increasingly affecting everything from which products or content people consume, whether their loan or job application is successful, to the medical treatments they receive. It is therefore ethically and legally imperative that the decisions these algorithms make are correct, transparent and fair [3]. While algorithmic fairness has become a fairly well-established topic in the typical batch classification set-up for machine learning, research into fair stream learning is still a growing but somewhat under-explored research field [22]. Fair stream classification models are challenging to develop and also to evaluate, as issues of fairness–performance trade-offs, temporal dynamics and classic machine

Authors’ Contact Information: Kathrin Lammers, Bielefeld University, Bielefeld, Germany, klammers@techfak.uni-bielefeld.de; MaryBeth Defrance, Ghent University, Ghent, Belgium, marybeth.defrance@ugent.be; Barbara Hammer, Bielefeld University, Bielefeld, Germany, bhammer@techfak.uni-bielefeld.de; Valerie Vaquet, Bielefeld University, Bielefeld, Germany, vvaquets@techfak.uni-bielefeld.de.

This paper is published under the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International (CC-BY-NC-ND 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.

ECAF’26, September 02–September 04, 2026, Ghent, BE

© 2026 Copyright held by the owner/author(s).

learning problems such as class imbalance not only all have to be addressed individually, but also have to be set in relation in a joint and cohesive manner.

Algorithmic fairness, itself an interdisciplinary and diverse field, has recently gained additional importance with the publication of the EU AI Act [5]. However, while it is generally agreed upon that fairness is important, there are several different and incompatible notions of fairness prominently found in the literature, and balancing requirements of non-discrimination with sufficient predictive performance is not trivial either [16]. Furthermore, it has been shown that a lack of fairness at one point in time can induce vicious cycles, increasing discrimination in the future [21].

In the domain of stream learning, temporal dynamics are even more important, as data is processed continuously, and changes in the distribution in the form of concept drift may occur at every step even when no vicious feedback loops are present. As stream learning is important for many technical and industrial applications [9], fairness research in this application domain becomes increasingly necessary [26].

In addition to providing a comprehensive overview of the area of fair stream learning for researchers (and practitioners) familiar either only with stream learning or fairness research, or perhaps even unfamiliar with both, this paper also presents a thorough evaluation of the currently available algorithmic approaches to fair stream learning, particularly with regards to a drift-aware evaluation [17], thereby capturing and comparing temporal dynamics affecting the fairness and performance of these classifiers.

We will therefore begin by laying out the foundations of both algorithmic fairness (2.1) and stream learning (2.2), before giving a thorough overview over the current landscape of fair stream learners in section 3, followed by a review on how to evaluate fairness in stream learning (4). Then we will apply these evaluation approaches to a selection of the fair stream learning classifiers (5), thoroughly analyzing the obtained results in section 6. Finally, we will conclude by summarizing our findings, as well as identifying promising future work in section 7.

2 Foundations of Fairness and Stream Learning

Here, we will give a brief introduction to fairness and its associated metrics as well as stream learning and its main challenge, concept drift.

2.1 Fairness

As algorithmic decision making has become more prevalent, there has been an increased attention on algorithmic fairness. Machine Learning algorithms are not inherently neutral, but can exhibit algorithmic bias both due to being trained on biased data as well as due to modelling decisions, such as regularization [22].

Data biases range from issues with data collection and modelling, such as *measurement bias* or *omitted variable bias* [25, 28], where the real world is distorted in a biased manner, to problems with equity and justice in the real world, where the human decisions collected for training algorithms are themselves discriminatory [20].

Apart from biases already exhibited by the data, almost every decision made in choosing and training a machine learning model, from model type, to regularization terms and loss function, priors and optimizers, induces model bias in one way or another. Trade-offs between generalization ability, model complexity and performance on minority samples are inevitable. Building a machine learning model is, in a way, nothing but *choosing a specific bias* [6], which is then applied to make predictions about the real world to base decision upon.

However, biases can also be – and often are – of the harmful variety that lead to discrimination against minority groups. The goal of research of algorithmic fairness is therefore to evaluate not just model performance but also whether algorithms make fair decisions, identify and mitigate harmful biases and prevent discrimination through algorithmic decision making.

Fairness is a complex construct, and many different definitions of what constitutes a "fair" model have been conceived, and it has been proven that many of these definitions can not be satisfied at the same time in most cases. Definitions of fairness, typically made quantifiable in the form of *fairness scores*, can be generally categorized as either *group-based*, *individual* or *causal* in nature. As the last category, causal fairness, is very domain-specific and requires expert knowledge of causal structures as well as ethical discussions on permissible causal graphs, we will not further consider it in this paper, as our goal is a general evaluation of the fairness and performance of fair stream learners and therefore removed from specific application contexts [27].

2.1.1 Group Fairness Definitions

Group fairness measures generally differentiate between equality of outcome, with scores such as *Demographic Parity* (also called *Statistical Parity*) and *Disparate Impact* comparing the difference of rates of positive outcomes between a privileged and a non-privileged group, and equality of opportunity, with scores such as *Equal Opportunity* or *Equal FPR* comparing the acceptance rates for individuals from both groups under consideration of the ground truth outcome as a proxy for their actual qualification. It has been shown that these definitions are somewhat incompatible for most datasets, as ground truth acceptance rates are often unequal.

For a sensitive attribute S , with $S = 1$ used to denote privileged individuals and $\hat{y} = 1$ signifying a positive model prediction, *Demographic Parity* [4] can be expressed as $|P(\hat{y} = 1|S = 1) - P(\hat{y} = 1|S = 0)|$. In contrast, *Disparate Impact* [8] compares relative rather than absolute acceptance rates between the groups: $1 - \min\left(\frac{P(\hat{y}=1|S=1)}{P(\hat{y}=1|S=0)}, \frac{P(\hat{y}=1|S=0)}{P(\hat{y}=1|S=1)}\right)$. As it does not rely on ground truth labels, this approach to fairness is especially suitable when the training data is suspected to contain discriminatory biases already.

Meanwhile, *Equal Opportunity* calculates the difference between the true positive rates between the groups, with $y = 1$ signifying a positive label: $|P(\hat{y} = 1|y = 1, S = 1) - P(\hat{y} = 1|y = 1, S = 0)|$. *Equal FPR*, meanwhile, measures the same for false positive rates with the formula $|P(\hat{y} = 1|y = 0, S = 1) - P(\hat{y} = 1|y = 0, S = 0)|$. Because these fairness definitions rely on ground truth labels, they should only be employed in contexts where this data is deemed trustworthy and possible base rate differences are justified [10].

2.1.2 Individual Fairness

Unlike group-based fairness measures, which work on the aggregate level, individual fairness measures seek to ensure that similar individuals are treated similarly. A simple, but often insufficient, way of ensuring this is to test whether a direct "flip" of the sensitive attribute would affect the prediction of a classifier. We can formalize this as $|P(\hat{y}_t = y_t|S, A) - P(\hat{y}_t = y_t|S^{-1}, A)|$. This fairness notion is also known as *Fairness through Unawareness*, as it ensures that a classifier makes a decision without taking the sensitive attribute into account. In contrast, *fairness through awareness* represents a more realistic but also more complex approach to individual fairness by defining an explicit similarity metric and comparing the treatment of samples it considers similar [27]. However, this method is

highly dependent on the definition of similarity, and the metric itself often requires expert knowledge to design. For this reason, our experiments will only consider *Fairness through Unawareness* as a proxy for individual fairness.

2.2 Stream Learning

Machine learning is usually practised in the batch set-up, where an entire dataset $D = \{(x_i, y_i)\}_{i=1}^N \in (X \times Y)^N$ which is drawn iid from the data generating distribution $\mathcal{D}$, is available. In contrast, in the *stream learning* set-up, data is arriving over time in form of a *data stream* of (potentially infinite) length T is a sequence $S = (x_t, y_t)_{t=1}^T$, $x_t \in X$, $y_t \in Y$, where each (x_t, y_t) is drawn from $\mathcal{D}_t$. Here, $(\mathcal{D}_t)_{t \in \mathcal{T}}$ constitutes a family of time-indexed data distributions on X and Y . The data stream might contain *concept drift*, i.e. a change in distribution over time. More formally, it contains drift iff there are time points $t_0, t_1 \in \mathcal{T}$ such that $\mathcal{D}_{t_0} \neq \mathcal{D}_{t_1}$.

The goal of stream learning is to accurately model the current concept, which, unlike in Continual Learning, requires strategic forgetting. To deal with drift, training follows the test-then-train scheme, where a model is first tested on x_t and afterwards trained with (x_t, y_t) .

Concept drift can be either *real* or *virtual*, depending on how the distribution changes. In the case of *real* drift, the relationship between the label y_t and the feature vector x_t changes over time, which means that the true decision boundary shifts as well. *Virtual* drift, on the other hand, affects on the distribution of the features (and thereby potentially the frequency of the labels), while the decision boundary stays the same. However, even virtual drift may have a strong negative impact on classifier performance [12].

We can also distinguish different types of drift based on their temporal characteristics, separating drift that occurs in an *abrupt* manner, in contrast to drift that occurs due to a *gradual* increase of samples from the new distribution, as well as *incremental* drift, where one distribution slowly morphs into another. Concept drift may also be *re-occurring*, for example when the same drift occurs every year due to seasonal changes [9].

3 Fair Stream Learning Classifiers

Recently, a few initial algorithms have been proposed to jointly address the challenges of both fairness and stream learning. As is typical in fairness, we can broadly group these algorithms into three categories based on when in the training process they address the issue of fairness: During pre-processing, in-processing or the post-processing phase. A brief overview can be found in the appendix in Table 5.

3.1 Pre-Processing

The earliest approaches to fair stream learning were *Massaging* and *Reweighting* [14]. These were both applied in a chunk-based manner, meaning that a fixed amount of samples was collected before the method was applied, somewhat diverging from a pure stream learning set-up. In the case of massaging, labels were swapped before training in order to create more training samples with positive outcomes for the disadvantaged and negative outcomes for the diasadvantaged group. In the case of reweighting, the weight of positive and negative predictions for each group were adjusted in order to create a more balanced training set.

CFSMOTe [18], a newer approach that can be applied in streaming fashion for continuously arriving samples, achieves a similar balance by synthetically upsampling underrepresented class-group combinations. Using Adaptive Windowing, it ensures the distribution it samples from is consistent with the current concept. It also aims to enhance

individual fairness by testing new samples and their counterfactuals (with flipped sensitive attribute) on a proxy classifier and only using samples that pass for training the actual classifier. As a variant of CSMOTE, CFSMOTE also deals with issues of class imbalance.

As pre-processing methods, *Massaging*, *Reweighting* and *CFSMOTE* are all model-agnostic, but as they are not designed to explicitly optimize specific fairness metrics, their effects are often somewhat limited.

3.2 In-Processing

In-processing methods, conversely, are often highly model-dependent, as they modify model-specific parameters during training. We therefore sub-categorize these methods further, dividing them by the underlying algorithmic approaches that they use.

3.2.1 Tree-based methods

Tree-based methods are fairness-modified versions of the Hoeffding (Adaptive) Tree algorithm, where the split-criterion was adjusted in order to account for fairness, with demographic parity used as an explicit metric to optimize. These tree-based methods are not suitable for highly imbalanced datasets, as the split-criterion does not take class balance into account. For the tree-based methods, we have a clear geneology, with increasing smoothness of the split criterion and varying drift-adaptability.

FAHT [31], based on the non-adaptive Hoeffding Tree, models fairness gain by the difference between the model's demographic parity before and after a split, calculating a weighed sum over the leaves in the new partition in order to achieve this. *FEAT* [29], in contrast, uses an unweighed sum for the fairness gain and employs the drift-aware Hoeffding Adaptive Tree as a base model.

While both *FAHT* and *FEAT* merely multiply fairness and information gain to achieve a final split criterion and use pure information gain if fairness does not change, *2FAHT* [32] uses a smooth exponential function with the fairness gain as an exponent in order to combine the two measures, also allowing for modification of the fairness-accuracy-trade-off through a hyperparameter multiplied with the fairness gain. *2FAHT* uses *CVFDT*, a window-based Hoeffding Tree variant, as its base model.

FARF [33], an adaptive random forest variant, uses the same fairness gain as *FEAT*, and additionally undersamples the protected, positive group in it's ensembling approach if the current model shows discriminatory tendencies, thereby combining two approaches to improve fairness.

3.2.2 Multi-Objective methods

As fairness can also be understood as a multi-objective optimization problem, there are also several approaches to stream fairness that explicitly use MOO approaches, rather than making an implicit fairness and performance trade-off. However, evolutionary and swarm-based approaches can be computationally costly and the MOO step has so far not been realized in a fully streaming manner, relying on small batches instead.

FAS Stream [23] utilizes Particle Swarm Optimization (PSO) to optimize the feature weights of a SAMDSOS classifier. The PSO algorithm selecting fairness-enhancing feature weights is employed whenever discrimination increases, while the classifier itself is concept-adapting through self-adjusting memory. Meanwhile, *FOMOS* [24] uses SMPPO in order to optimize the feature-wise trade-off between information and fairness gain in a Hoeffding Tree with feature weights. This method also aims to improve explainability by optimizing for a small and therefore

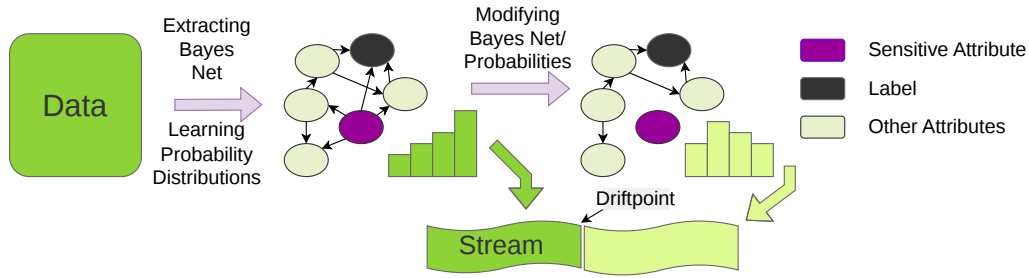


Fig. 1. An illustration of the generation process for the semi-synthetic benchmark data streams.

easily comprehensible tree. *EMO-SAM* [1], in a similar fashion, uses swarm-based evolutionary multi-objective optimization to optimize both fairness and accuracy by adjusting feature weights of a SMA-kNN classifier whenever an HP filter determines that the current discrimination level is unacceptably high. A genetic MOO algorithm is employed by *FairCMNB* [2] to determine how many samples to re-assign to different class-group clusters of a class-imbalance-aware Continuous Mixed Naive Bayes classifier, thereby introducing fairness somewhere between in- and post-processing steps. As *FairCMNB* makes an explicit effort to balance classes, it is also suitable for imbalanced datasets. It is suitable to optimize multiple different fairness metrics (though not simultaneously), including Demographic Parity, Equal Opportunity and the causal metric FACE.

3.3 Post-Processing

The only fully post-processing approach to stream learning fairness, called *FABBOO* [13], is also able to optimize for several different group fairness metrics, as well as explicitly imbalance-aware. Based on OzaBoost, an ensemble boosting stream learner, *FABBOO* optimizes a given fairness metric by dynamically adjusting the decision boundary for the underprivileged group so that a pre-defined fairness threshold is satisfied, allowing for strong fairness guarantees but little regard for potential trade-offs.

4 Evaluating Fair Stream Learners

In this section we will discuss how to evaluate fairness in a stream learning context, focussing on benchmark data streams as well as adapting static fairness definitions for drifting streams.

4.1 Benchmark Datasets

The evaluation of fairness-aware stream learners commonly employs stationary fairness benchmark datasets, such as *Adult*, *KDD Census*, *German Credit* or *COMPAS*. These datasets are randomly permuted and then simply treated as sequential data in order to gain fairness and performance insights [13, 29, 31]. Additionally, some evaluations also make use of stream-learning benchmarks, although these may often be poor benchmarks for fairness as the dataset itself is "unbiased" enough that even fairness-unaware classifiers do not show significant levels of discrimination. Additionally, it is unclear where drift occurs and if and how this impacts the fairness properties of the dataset. A purely synthetic dataset generation method has also been proposed [15]. While this method, based on Gaussian distributions for each attribute, can simulate changing demographic parity and concept drift, its

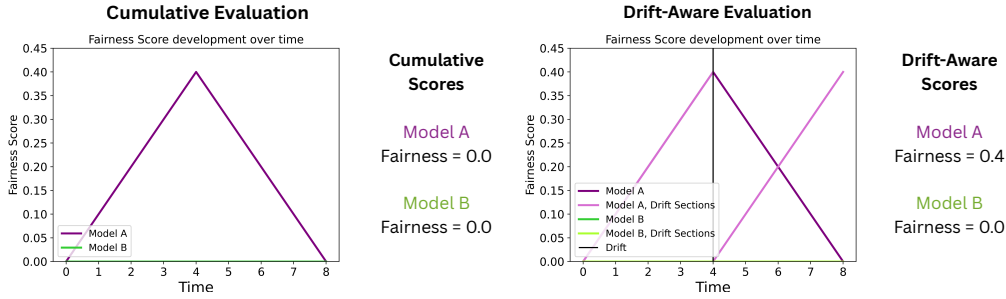


Fig. 2. An illustration highlighting the difference between cumulative and drift aware score aggregation. The cumulative approach takes the final, cumulative score at the end of the stream, while the drift-aware approach averages the scores at the end of each drift segment. In this example, there is one drift where the direction of discrimination changes, but the same pattern repeats with the other group for Model A.

properties are devoid of semantic meaning. But as there is a lack of real-world drifting datasets where nature and characteristic of the drift is known, particularly streams with fairness-relevant features, we are nevertheless reliant on (semi-)synthetic streams. We therefore chose to apply a method for generating semi-synthetic datasets based on Bayesian networks [19], which enables us to introduce controlled drifts by explicitly modifying the probability distributions of individual features as well as isolating features through marginalization. As we control the process, we maintain interpretability, while also being relatively close to original benchmarks (see fig. 1). We use synthetic datasets based on the *Adult* and *Student Performance* datasets.

4.2 Drift-Aware Scoring

For the evaluation, we compare the standard approach of cumulative score aggregation over the entire with a drift-aware score aggregation approach. This method uses either detected or ground-truth drift points to split the stream. It then aggregates unfairness over each drift-free segment and then accumulates the scores, weighed by the length of each segment. Due to the formalization of many fairness metrics, a fully cumulative approach is insensitive to both changes in the direction of the discrimination a classifier exhibits. Cumulative evaluation would, for example, not give different statistical parity scores to a classifier which is fully fair over the entire stream than to a classifier which first heavily favors men and then later heavily favors women, as long as the total number of accepted individuals from each group is equal. Drift-aware evaluation, meanwhile, detects drift points, particularly those relevant for fairness (i.e. drift affecting the sensitive attribute and/or label), and splits the stream to evaluate each individual segment with its different distribution (see illustration in fig. 2). This allows us to ensure fairness at every time step, rather than overlooking discriminatory patterns in some parts of the stream [17].

5 Experimental Set-Up

For our experiments, we ran selected fair stream learning classifiers (see section 3), chosen as representatives of their respective category, on ten slightly permuted versions (with the same drift points) of each of the examined data streams. We analyzed the performance with both cumulative and drift-aware score aggregation on select streams, comparing several different fairness and performance metrics. For our drift-aware scoring, we did not employ a

drift-point detector but rather utilized our ground truth knowledge about the drift points in order to get the most accurate results possible. Additionally, we used *autorank* [11] to compare the classifier's mean performance over all different streams derived from *Adult* and *Student* in order to analyze overall classifier fairness and performance on five different fairness metrics (statistical parity, disparate impact, equal opportunity, equal FPR and individual fairness) as well as balanced accuracy. Autorank enables to compare classifiers over multiple streams, assigning them a ranking based on their overall performance.

We used the single-drift datasets derived on the basis of *Adult* (four streams) and *Student Performance* (five streams) from Lammers et. al. [19], with abrupt drifts at clearly defined drift points allowing for a clean drift-aware score aggregation. Some of these datastreams contain fairness-relevant (real and virtual) drift, while for others the drift did not affect the sensitive groups differently. Additionally, we also studied several data stream with multiple driftpoints derived from these datasets by introducing more "cuts", so that these drifts can be characterized as abrupt and recurrent, as we abruptly change which of two distributions we sample from. One of these distributions is always the unmodified distribution of the base dataset, while the other is the one indicated by the dataset name. Streams based on *Adult* all contain approximately 48.800 samples, while streams based on *Student Performance* contain 5.000 samples each. An overview over the resulting streams and their characteristics can be found in Table 1.

For the *Adult* scenario, generally more extreme drifts were simulated, such as the "sensitive attribute switch" scenario, where the coding of the sensitive attribute switches with the drift. This is intended in order to examine the effects of very severe drift, although such a drift would not occur in the real world (except through changes in data entry). The "Fair World" scenario, likewise, introduces drift by switching to data derived from a Bayes Net which differs from the original one through complete marginalization and therefore mitigation of all effects of the sensitive attribute on other attributes. In contrast, the other scenarios represent more realistic scenarios, such as all monetary values increasing due to inflation, an increased likelihood of women to take jobs in STEM, or a decreased likelihood for all genders to get married.

The streams modelled on the *Student* dataset are all largely realistic, and also represent actionable interventions, such as gender-specific school support programs, grade inflation, or widespread internet access. The "Grade Category Adjustment" corresponds to a re-evaluation of what counts as a "high" grade and therefore a markedly increased threshold on the Portuguese 20-point scale from 10 to 14 points, similar to a switch from a C to a B as a "high" grade.

Our evaluated classifiers are Massaging and CFSMOTE for pre-processing, FAHT, FEAT, and 2FAHT as tree-based variants, EMOSAM for MOO-optimization algorithms and FABBOO (two variants, one each optimizing for statistical parity and equal opportunity) for post-processing. These classifiers were chosen both on the basis of representativeness (at least one per category), as well as for code availability. As there was no published code available for several algorithms (FEAT, 2FAHT and EMOSAM), we used our own *Python* implementation of these algorithms, validated for their performance on the streams used in the original papers [1, 29, 32]. However, EMOSAM turned out to be significantly more computationally expensive than all other algorithms, exceeding their runtime by several magnitudes, so we are only able to report on its results for the shorter *Student Performance* streams.

Table 1. Overview of the studied data streams.

	Stream	#Drifts	Driftpoints	Significant Fairness Drift	% pos.	$S = 1$	Dem.Par.	Disp.Imp.
Adult	Sensitive Attribute Switches	3	10,000, 25,000, 40,000	✓	24.01 ± 0.10	F	1.03 ± 0.26	4.37 ± 1.11
	Normal & Fair World Switches	3	5,000, 20,000, 35,000	✓	22.95 ± 0.13	M	7.18 ± 0.31	28.31 ± 1.13
	Normal & Fair World Switch	1	25,000	✓	23.14 ± 0.13	M	8.74 ± 0.26	33.53 ± 0.91
	Inflation Scenario	1	25,000	✗	31.40 ± 0.09	M	20.56 ± 0.31	53.78 ± 0.64
	Women in STEM Scenario	1	25,000	?	24.25 ± 0.11	M	16.54 ± 0.19	55.58 ± 0.60
	Less Marriage Scenario	1	25,000	✗	24.16 ± 0.14	M	17.07 ± 0.21	57.18 ± 0.57
Student	Grade Category Adjustments	4	1,000, 2,000, 3,000, 4,000	✓	63.24 ± 0.41	F	7.28 ± 0.59	11.00 ± 0.86
	Girls Support Mul.	3	500, 3,500, 4,000	✓	86.51 ± 0.23	F	8.99 ± 0.42	9.98 ± 0.43
	Boys Support Mul.	2	1,500, 2,500	✓	85.45 ± 0.32	F	4.73 ± 0.40	5.42 ± 0.44
	Boys Support Single	1	2,000	✓	85.81 ± 0.40	F	3.16 ± 0.67	3.62 ± 0.77
	Girls Support Single	1	2,000	✓	86.03 ± 0.32	F	7.83 ± 0.72	8.79 ± 0.79
	Grade Category Adjustment	1	2,000	✓	52.23 ± 0.38	F	9.01 ± 0.92	16.25 ± 1.52
	Grade Inflation	1	2,000	✗	87.49 ± 0.30	F	5.08 ± 0.61	5.67 ± 0.67
	Internet Era	1	2,000	✗	85.32 ± 0.31	F	5.36 ± 0.61	6.13 ± 0.68

6 Results

In the following section, we will discuss the results of our experiments. Comprehensive tables reporting on the classifiers fairness and performance, as measured by the drift-aware scoring method, as well as differences between cumulative and drift-aware evaluation for fairness the group metrics can be found in appendix A. In the following section, however, we will focus only on the most salient results as well as general trends in terms of classifier performance as identified by the autorank analysis.

Our hypotheses are as follows:

- *H1*: Drift-aware score aggregation will lead to increased reports of unfairness in comparison to cumulative scoring, particularly on datasets with large fairness relevant drifts.
- *H2*: There will be significant trade-offs between fairness definitions, likely most strongly notable for the post-processing classifiers.
- *H3*: Classifiers that are not designed for imbalanced data will overpredict the majority class on imbalanced streams, influencing fairness outcomes.

We assume that hypothesis *H1* holds because the classifiers optimize specifically for cumulative fairness, thereby potentially and, in some cases, deliberately ignoring temporal differences in treatment of the groups. Hypothesis *H2* assumes that well-known fairness trade-offs [16] will also apply in stream learning, while *H3* reasons that overprediction of a class will also affect fairness scores, as the underlying acceptance and error rates will necessarily be affected.

6.1 Effects of Drift and Drift-Aware Evaluation

To analyze the effects of drift and drift-aware evaluation, we will mainly focus on a few datasets with strong fairness drift, as well as one dataset with less fairness relevant drift. As can be seen in appendix A, for streams with no or low low fairness relevant drift, there is very little difference between drift-aware and fully cumulative evaluation, as neither data bias nor classifier performance changes particularly significantly due to the drift. We do not highlight

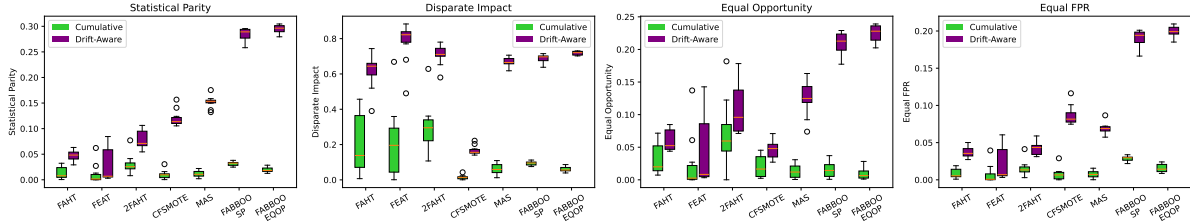


Fig. 3. Paired boxplots showing the differences between cumulatively and drift-awarely aggregated scores on the *sensitive attribute switches* stream with three drifts.

the difference for less drift-sensitive scores, as performance and individual fairness, due to their definition, do not change significantly based on score aggregation method, with only minor differences (explainable due to rounding errors) observed.

Classifier	Statistical Parity	Disparate Impact	Equal Opportunity	Equal FPR	Individual Fairness	Accuracy	Balanced Accuracy	Recall
FAHT	4.85 ± 1.14	61.4 ± 9.87	6.24 ± 1.75	3.79 ± 0.83	9.15 ± 1.56	75.1 ± 0.12	53 ± 0.73	10.4 ± 2.08
FEAT	2.88 ± 3.51	79.5 ± 10.2	4.4 ± 5.71	2.16 ± 2.42	3.12 ± 3.68	75.7 ± 0.13	51.5 ± 2.34	4.77 ± 6.61
2FAHT	7.89 ± 1.73	71.2 ± 5.35	11 ± 3.87	4.3 ± 0.94	0.29 ± 0.09	77.6 ± 0.84	58 ± 1.99	20.2 ± 4.42
CFSMOTE	12.1 ± 1.41	16.8 ± 2.25	4.69 ± 1.35	8.78 ± 1.12	2.28 ± 1.19	49.6 ± 1.7	62.6 ± 0.97	87.5 ± 0.87
MAS	15.2 ± 1.09	66.8 ± 2.61	12.3 ± 2.5	6.85 ± 0.73	0.9 ± 0.27	83.3 ± 0.17	72.7 ± 0.65	52.1 ± 1.72
FABBOO SP	28.5 ± 1.21	69 ± 2.17	20.9 ± 1.64	19.1 ± 1.12	5.09 ± 0.8	80.4 ± 0.25	78.8 ± 0.21	75.6 ± 0.66
FABBOO EQOP	29.4 ± 0.81	71.7 ± 1.11	22.3 ± 1.36	19.9 ± 0.75	3.34 ± 0.5	80.8 ± 0.15	78.9 ± 0.15	75.3 ± 0.6

Table 2. Fairness and performance comparison across metrics on the *sensitive attribute switches* stream with drift-aware scoring.

On the *sensitive attribute switches* stream, where the drift directly affects the correlation between sensitive attribute and distribution in a very significant manner, the differences between the two evaluation approaches is extremely large, with some unfairness indicators rising tenfold for several classifiers in the drift-aware evaluation, compared to the cumulative score (see fig. 3).

While this dataset presents an extreme and somewhat artificial situation, this shows that current fair stream learners seem to struggle in contexts where concept drift affects the "direction" of discrimination, i.e. the discriminated group. In this case, the fully aggregated measures of discrimination do not adequately reflect the discrimination occurring at any distinct time step, and as can be seen in our results, classifiers may struggle to suitably adapt in a drift-aware manner.

For example, a disparate impact of up to 20% (or the underprivileged group receiving at least 80% of the positive outcomes of the privileged group) is often deemed legally acceptable [8], though smaller effects may still be ethically problematic. While several fair stream learners adhere to this restriction when evaluated cumulatively on the *sensitive attribute switch dataset*, all but one exceed this relatively high limit by a significant margin in the drift-aware evaluation.

It is also evident that the difference is particularly large for our more imbalance-aware methods, *CFSMOTE*, *Massaging* and *FABBOO*. The latter, in particular, struggles greatly with the shift in privilege and with adapting to

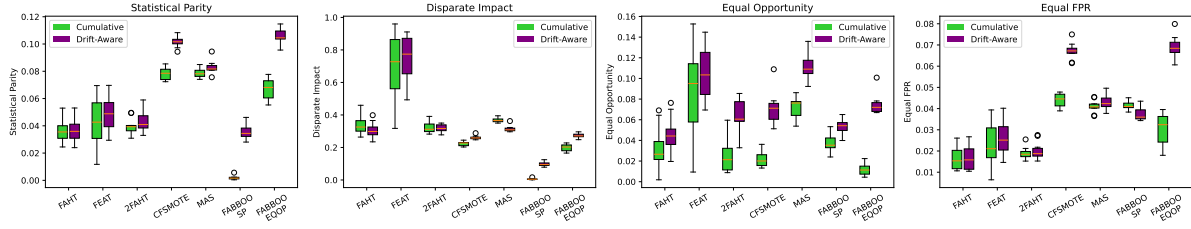


Fig. 4. Paired boxplots showing the differences between cumulatively and drift-awarely aggregated scores on the *normal and fair world switches* stream with multiple drifts.

fairness-relevant drift, unable to satisfy any post-processing guarantees and generally exceeding dataset-inherent unfairness in terms of statistical parity and disparate impact under the drift-aware aggregation.

While for the tree-based learners the effects were more moderate on the metrics focussing on absolute differences, the increase in the relative disparate impact score reveals that these unbalanced models are also unable to adequately adapt to this extreme drift, with effects simply masked in absolute scores due to extreme base rates.

We can see somewhat less pronounced differences between the drift-aware and the cumulative scores on the *normal and fair world switches* stream with three separate drifts, where the data bias is extremely reduced for some segments by sampling from a marginalized distribution (see fig. 4). However, the relative effect that the drift-aware evaluation has on FABBOO is also notable in this case, particularly on the specific metric that was minimized through dpost-processing. This might indicate that this classifier continues to compensate for discrimination even when the data bias has already lifted, biasing it against the previously privileged group.

If we compare these results with those for the *normal and fair world switch* stream with only one drift, we can see that this effect seems to be highly dependant on the number of drifts for FABBOO, but not for the other classifiers (see fig. 5). The number of drifts also appears to affect the fairness metrics measuring equality of outcome much more than those measuring equality of opportunity, suggesting a higher impact on overall acceptance rates than on error rates. As both the fair and the normal distribution contribute roughly the same number of samples to each stream with both one and three drifts, this finding suggests that temporal dynamics are highly relevant for bias mitigation, and more frequent drifts harder to adapt to.

The *more women in stem dataset* was chosen as a representative of streams with little fairness relevant drift, because its statistical parity in the data only lowers only 0.9% after the changepoint. Its results show that small drifts in the underlying data bias only lead to marginal differences between drift-aware evaluation and a fully cumulative evaluation for most classifiers (see fig. 6).

Similar tendencies as on the *Adult*-derived streams can also be observed for the *Student Performance* stream derivatives, such as the *grade category adjusted* stream with four separate drifts. However, while the differences are notable, the smaller "magnitude" of the drift w.r.t. to data fairness, lower level of disparity between genders in the data itself, as well as the higher class imbalance on this dataset all seem to contribute to smaller overall differences between cumulative and drift-aware scoring (see fig. 7).

In terms of performance, we can see the same tendencies as before (see tab. 3), with CFSMOTE, MAS and FABBOO showing relatively high balanced accuracy, while the tree-based classifiers struggle with low recall. We

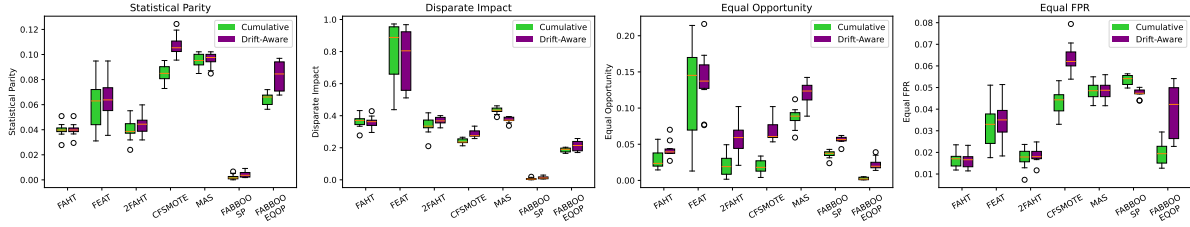


Fig. 5. Paired boxplots showing the differences between cumulatively and drift-awarely aggregated scores on the *normal and fair world switches* stream with only one drift.

Classifier	Statistical Parity	Disparate Impact	Equal Opportunity	Equal FPR	Individual Fairness	Accuracy	Balanced Accuracy	Recall
FAHT	3.7 ± 0.91	30.1 ± 4.83	4.59 ± 1.63	1.69 ± 0.59	3.99 ± 1.23	80.5 ± 0.16	62.2 ± 0.56	28.3 ± 1.44
FEAT	4.78 ± 1.24	74 ± 14.1	10.4 ± 2.48	2.58 ± 0.73	4.63 ± 1.42	78.2 ± 0.81	55.1 ± 2.7	12.7 ± 6.24
2FAHT	4.23 ± 0.78	31.1 ± 2.25	6.2 ± 1.56	1.96 ± 0.4	0.57 ± 0.14	80.8 ± 0.32	63.7 ± 0.83	32.2 ± 1.98
CFSMOTe	10 ± 0.4	25.7 ± 1.1	6.98 ± 1.53	6.64 ± 0.37	7.81 ± 0.63	78.8 ± 0.59	77.3 ± 0.33	74.5 ± 0.69
MAS	8.12 ± 0.46	30.8 ± 1.84	10.9 ± 1.12	4.2 ± 0.37	1.79 ± 0.28	82.6 ± 0.31	72.2 ± 0.37	53 ± 0.96
FABBOO SP	3.6 ± 0.53	9.91 ± 1.35	5.41 ± 0.7	3.79 ± 0.27	13.8 ± 0.29	78.4 ± 0.17	77.5 ± 0.21	76 ± 0.49
FABBOO EQOP	10.5 ± 0.58	27.1 ± 1.42	7.46 ± 0.96	6.87 ± 0.49	6.85 ± 0.6	80 ± 0.2	78.2 ± 0.25	74.9 ± 0.52

Table 3. Fairness and performance comparison across metrics on the *fair and normal world switches* stream with three drifts, evaluated with drift-aware scoring.

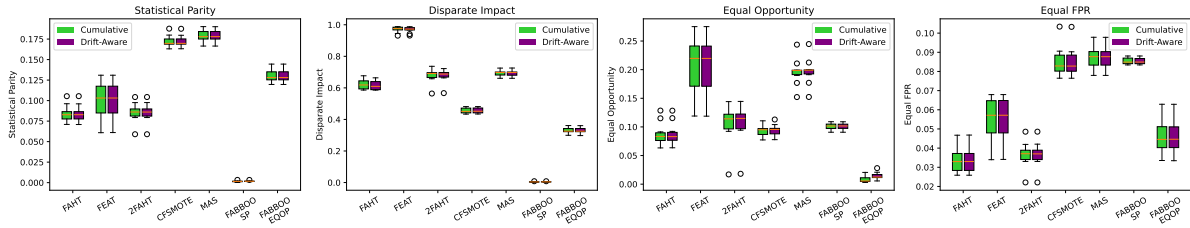


Fig. 6. Paired boxplots showing the differences between cumulatively and drift-awarely aggregated scores on the *women in stem scenario* stream with only one drift.

also observe the usual fairness-performance trade-offs, particularly with regards to the absolute scores, where classifiers that maintain balance between classes typically receive worse fairness outcomes in the drift-aware evaluation, with the exception of the relative disparate impact score, with high balanced accuracy coinciding especially with more individual discrimination.

6.2 Autorank Analysis

In addition to analysing fairness and performance and the effects of drift and drift-aware evaluation in detail for selected datasets, we also analyzed the overall performance of the stream learning classifiers on all data stream derived from both the *Adult* and the *Student Performance* datasets, based on the drift-aware evaluation.

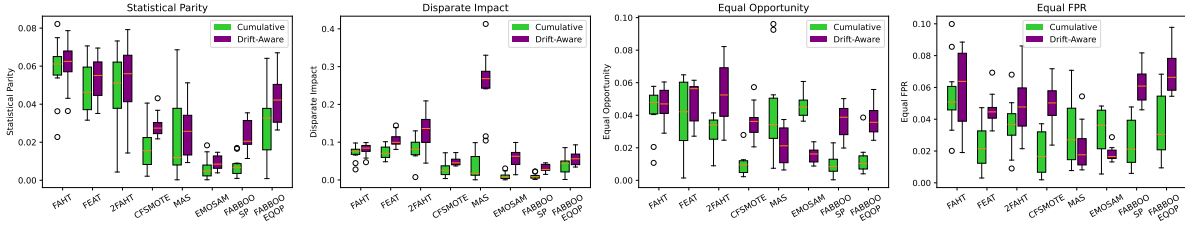


Fig. 7. Paired boxplots showing the differences between cumulatively and drift-awarely aggregated scores on the *grade category adjustments* stream based on the *Student* dataset with four drifts.

Classifier	Statistical Parity	Disparate Impact	Equal Opportunity	Equal FPR	Individual Fairness	Accuracy	Balanced Accuracy	Recall
FAHT	6.12 ± 1.29	8.02 ± 1.58	4.72 ± 0.95	5.9 ± 2.38	3.85 ± 0.93	65 ± 0.92	56.6 ± 1.4	83.7 ± 1.63
FEAT	5.36 ± 1.15	10.6 ± 2.05	4.87 ± 1.21	4.62 ± 0.97	3.12 ± 0.6	64.4 ± 1.31	56.7 ± 0.71	67.4 ± 2.1
2FAHT	5.36 ± 1.85	12.9 ± 4.78	5.34 ± 1.91	4.89 ± 1.95	3.08 ± 1.39	67.8 ± 2.47	54.8 ± 1.68	59.4 ± 8.4
CFSMOTE	2.88 ± 0.63	4.92 ± 1.16	3.61 ± 1.06	5.01 ± 1.39	6.86 ± 1.61	58.9 ± 1.2	55.1 ± 0.93	61.8 ± 2.51
MAS	2.54 ± 1.31	25.6 ± 8.67	2.16 ± 1.13	2.2 ± 1.44	1.48 ± 0.9	42.5 ± 4.38	53.2 ± 1.91	65.3 ± 7.54
EMOSAM	0.92 ± 0.37	5.74 ± 2.75	1.54 ± 0.47	1.81 ± 0.45	0 ± 0	75.8 ± 1.03	49.8 ± 0.46	60.8 ± 1.37
FABBOO SP	2.4 ± 0.76	3.17 ± 0.97	3.61 ± 1	6.19 ± 1.15	5.29 ± 0.75	67.6 ± 0.7	63.9 ± 0.6	83.6 ± 1.26
FABBOO EQOP	4.24 ± 1.26	5.66 ± 1.76	3.72 ± 0.97	7.04 ± 1.46	4.47 ± 1.04	67.8 ± 0.78	64.2 ± 0.68	82.9 ± 1.41

Table 4. Fairness and performance comparison across metrics on the *grade category adjusted* stream with drift-aware scoring.

On the *Adult* data streams (see Fig. 8), we can see that both FABBOO and CFSMOTE deliver both good performance (as measured by balanced accuracy), as well as good fairness according to both disparate impact and equal opportunity fairness definitions, with FABBOO optimized specifically for this metric also scoring well on statistical Parity. As the tree-based classifiers over-predict the negative class on all *Adult*-based datasets, the low FPR scores for them can be explained by overall very low false-positive-rates, as equal FPR relies on absolute differences. Therefore, in the case of streams with characteristics similar to *Adult*, particularly those that are not subject to significant changes of underlying data bias, FABBOO and, to a lesser degree also CFSMOTE depending on desired fairness and performance characteristics, can be recommended.

However, the results on the *Student Performance* data streams (see Fig. 9) show a significant trade-off between fairness and performance, as FABBOO and CFSMOTE, which have a higher degree of balanced accuracy, receive consistently poor fairness scores compared to the tree-based classifiers and EMOSAM, which in turn show low balanced accuracy. But upon closer examination, the low unfairness of EMOSAM and the tree-based classifiers is most likely only achieved through grossly under-predicting the negative minority class, leading to very high acceptance rate, FPR and TPR for both privileged and unprivileged group and therefore also very small absolute and relative differences between them. Massaging, meanwhile, achieves middling results for both fairness and performance, while also over-predicting the majority group. Whether the scores FABBOO and CFSMOTE achieve on the fairness scores show sufficient non-discrimination is therefore hard to answer only in comparison to these effectively dummy classifiers, though under drift-aware evaluation, absolute rates of statistical parity and equal opportunity exceed the promised threshold of the post-processing approach FABBOO considerably.

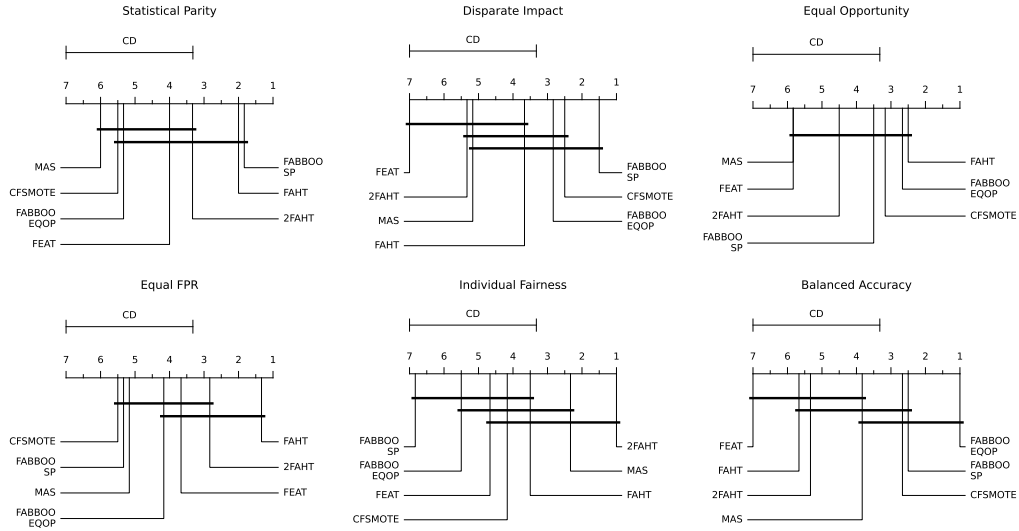


Fig. 8. Autorank [11] plot over all examined stream variants derived from the *Adult* dataset. It shows a relative ranking, with the thick horizontal lines showing which classifiers must be "grouped" together, as their score differences have not been found to be significant enough.

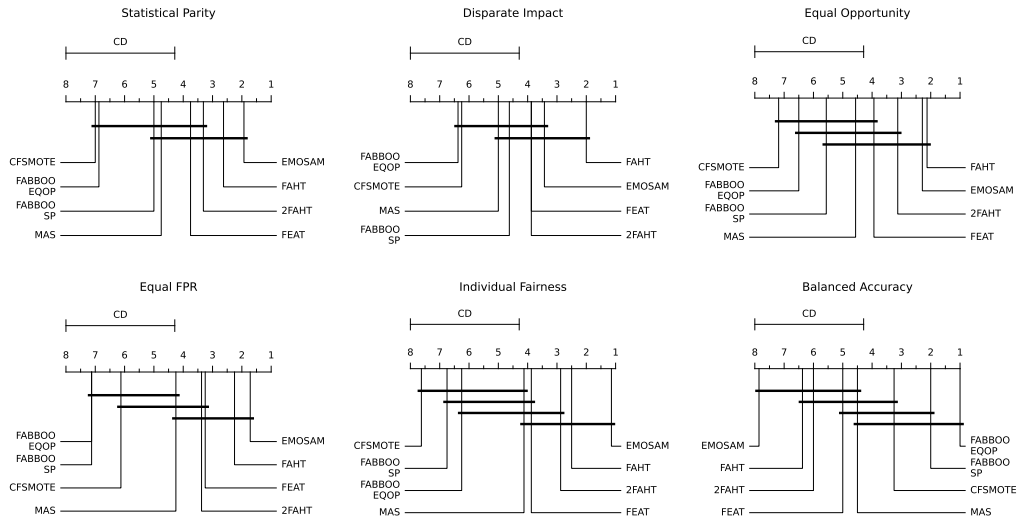


Fig. 9. Autorank plot over all examined stream variants derived from the *Student Performance* dataset.

In general, these results show that drift-aware fairness in stream learning, i.e. classifiers acting fairly for every distribution in a given stream, might still be under-explored and more adaptive algorithms are required to ensure fairness at all times.

6.3 Results Summary

Our results showed that there is a significant difference between cumulative and drift-aware scoring, particularly on data streams with strong fairness-relevant drift, confirming hypothesis *H1*. This effect is especially strong for FABBOO, the only true post-processing classifier currently available. However, while FABBOO generally achieves good fairness outcomes on the especially optimized metric, particularly when considering its high degree of performance, it underperforms considerably on streams where drift changes the direction or presence of discrimination considerably. Smaller differences between drift-aware and cumulative evaluation were also found for other classifiers on these streams. However, if drift does not significantly affect data biases, both score aggregation approaches lead to markedly similar results. We also found that on our largely imbalanced data streams, the not specifically imbalance-aware classifiers tended to perform badly, in according with *H3*, but generally seemed to achieve good model fairness. However, this is likely partially or even mostly explained by them over-predicting the majority class to such a degree that absolute fairness scores become meaningless, as they tended to still show discrimination when assessed with the disparate impact score. This shows that for imbalanced streams, classifiers should be evaluated with relative metrics and, perhaps equally curcially, explicitly optimize for relative metrics. We could also refer to this finding as a fairness-performance trade-off, however, as the high-performing FABBOO and CFSMOTE still scored well on some fairness metrics, we can not conclusively state that we truly observed such a trade-off, particularly as we only performed fair stream learners and did not compare to non-fairness-optimizing base models. We did, however, observe trade-offs between some fairness metrics, particularly equality of outcome with equality of opportunity and equality of outcome with individual fairness, partially confirming *H2*.

7 Conclusion

In this paper, we first gave an overview over the field of fair stream learning research, before we proceeded to systematically evaluate a select number of fair stream learning algorithms in a drift-aware manner. This comprehensive use of a novel evaluation protocol revealed that current fair stream classifiers work relatively well under small, non-fairness-relevant drift, but are not robust to more extreme changes in fairness. We also showed that class imbalance is an extreme challenge for several classifiers, and that for imbalanced streams, relative fairness metrics may be of greater use than absolute ones. We further examined inter-fairness trade-offs, finding them particularly significant for post-processing approaches. However, more research into suitable real-world benchmarks may be beneficial for a more realistic evaluation approach. Additionally, more suitable base datasets for the semi-synthetic data stream generation might improve the evaluation, e.g. streams based on *Folktables* [7] instead of *Adult*. Research into drift-aware post-processing for fair stream learning classification might also be promising for a more robust adaptation to more extreme fairness-relevant drifts.

Acknowledgments

This research was funded by the BMFTR through grant no. 16IS24057A (KI Akademie OWL).

References

- [1] Pivithuru Amarasinghe, Diem Pham, Binh Tran, Su Nguyen, Yuan Sun, and Daminda Alahakoon. 2024. Evolutionary Multi-Objective Optimisation for Fairness-Aware Self Adjusting Memory Classifiers in Data Streams. In *Proceedings of the Genetic and Evolutionary*

- Computation Conference* (Melbourne, VIC, Australia) (*GECCO '24*). Association for Computing Machinery, New York, NY, USA, 258–266. doi:10.1145/3638529.3654038
- [2] Maryam Badar and Marco Fisichella. 2024. Fair-CMNB: Advancing Fairness-Aware Stream Learning with Naïve Bayes and Multi-Objective Optimization. *Big Data and Cognitive Computing* 8 (01 2024), 16. doi:10.3390/bdcc8020016
 - [3] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2018. *Fairness and Machine Learning Limitations and Opportunities*. The MIT Press. <https://api.semanticscholar.org/CorpusID:113402716>
 - [4] Toon Calders and Sicco Verwer. 2010. Three naive Bayes approaches for discrimination-free classification. *Data Min. Knowl. Discov.* 21 (09 2010), 277–292. doi:10.1007/s10618-010-0190-x
 - [5] Council of European Union. 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance). <http://data.europa.eu/eli/reg/2024/1689/oj/eng> Legislative Body: CONSIL, EP.
 - [6] AlgorithmsThomas G. Dietterich and Eun Bae Kong. 1995. Machine Learning Bias, Statistical Bias, and Statistical Variance of Decision Tree. <https://api.semanticscholar.org/CorpusID:16497709>
 - [7] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. 2021. Retiring Adult: New Datasets for Fair Machine Learning. *Advances in Neural Information Processing Systems* 34 (2021).
 - [8] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Eduardo Scheidegger, and Suresh Venkatasubramanian. 2014. Certifying and Removing Disparate Impact. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2014). <https://api.semanticscholar.org/CorpusID:2077168>
 - [9] Joao Gama, Indre Zliobaite, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. 2014. A survey on concept drift adaptation. 46, 4, Article 44 (March 2014), 37 pages. doi:10.1145/2523813
 - [10] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of opportunity in supervised learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems* (Barcelona, Spain) (*NIPS'16*). Curran Associates Inc., Red Hook, NY, USA, 3323–3331.
 - [11] Steffen Herbold. 2020. Autorank: A Python package for automated ranking of classifiers. *Journal of Open Source Software* 5, 48 (2020), 2173. doi:10.21105/joss.02173
 - [12] Fabian Hinder, Johannes Brinkrolf, Kathrin Lammers, and Barbara Hammer. 2026. Real vs. Virtual Drift: Creating Realistic Stream Learning Benchmarks. 133–138. doi:10.14428/esann/2026.ES2026-200
 - [13] Vasileios Iosifidis and Eirini Ntoutsi. 2020. FABBOO - Online Fairness-Aware Learning Under Class Imbalance. In *IFIP Working Conference on Database Semantics*. <https://api.semanticscholar.org/CorpusID:221742304>
 - [14] Vasileios Iosifidis, Thi Ngoc Han Tran, and Eirini Ntoutsi. 2019. Fairness-Enhancing Interventions in Stream Classification. In *Database and Expert Systems Applications*, Sven Hartmann, Josef Küng, Sharma Chakravarthy, Gabriele Anderst-Kotsis, A Min Tjoa, and Ismail Khalil (Eds.). Vol. 11706. Springer International Publishing, Cham, 261–276. doi:10.1007/978-3-030-27615-7_20
 - [15] Vasileios Iosifidis, Thi Ngoc Han Tran, and Eirini Ntoutsi. 2019. Fairness-Enhancing Interventions in Stream Classification. In *Database and Expert Systems Applications*, Sven Hartmann, Josef Küng, Sharma Chakravarthy, Gabriele Anderst-Kotsis, A Min Tjoa, and Ismail Khalil (Eds.). Springer International Publishing, Cham, 261–276.
 - [16] Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent Trade-Offs in the Fair Determination of Risk Scores. In *Information Technology Convergence and Services*. <https://api.semanticscholar.org/CorpusID:12845273>
 - [17] Kathrin Lammers, Fabian Hinder, Barbara Hammer, and Valerie Vaquet. 2026. Drift-Aware Evaluation of Fair Stream Learning. 127–132. doi:10.14428/esann/2026.ES2026-152
 - [18] Kathrin Lammers, Valerie Vaquet, and Barbara Hammer. 2025. Continuous Fair SMOTE – Fairness-Aware Stream Learning from Imbalanced Data. In *Artificial Neural Networks and Machine Learning – ICANN 2025: 34th International Conference on Artificial Neural Networks, Kaunas, Lithuania, September 9–12, 2025, Proceedings, Part I* (Kaunas, Lithuania). Springer-Verlag, Berlin, Heidelberg, 336–348. doi:10.1007/978-3-032-04558-4_27
 - [19] K. Lammers et al. 2025. Realistic Benchmarks for Fair Stream Learning. *Neural Information Processing* (2025). <https://www.springerprofessional.de/en/realistic-benchmarks-for-fair-stream-learning/51716664>

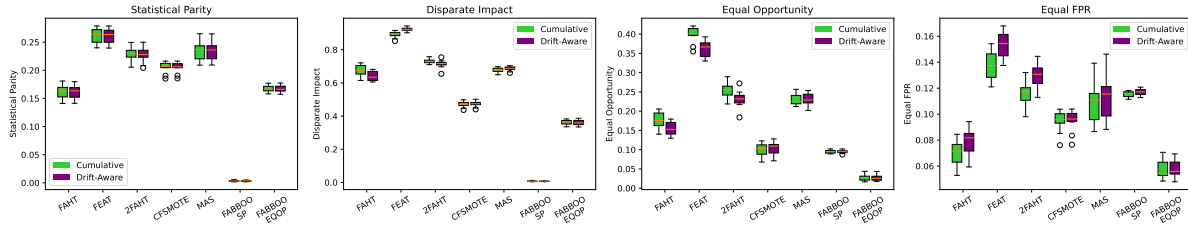


Fig. 10. Paired boxplots showing the differences between cumulatively and drift-awarely aggregated scores on the *inflation scenario* stream with one drift.

- [20] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. *ACM Comput. Surv.* 54, 6, Article 115 (jul 2021), 35 pages. doi:10.1145/3457607
- [21] Benjamin Paaßen, Astrid Bunge, Carolin Hainke, Leon Sindelar, and Matthias Vogelsang. 2019. Dynamic fairness - Breaking vicious cycles in automatic decision making. <https://www.esann.org/sites/default/files/proceedings/legacy/es2019-134.pdf>
- [22] Dana Pessach and Erez Shmueli. 2022. A Review on Fairness in Machine Learning. *ACM Comput. Surv.* 55, 3, Article 51 (feb 2022), 44 pages. doi:10.1145/3494672
- [23] Diem Pham, Binh Tran, Su Nguyen, and Daminda Alahakoon. 2022. Fairness Aware Swarm-based Machine Learning for Data Streams. In *AI 2022: Advances in Artificial Intelligence: 35th Australasian Joint Conference, AI 2022, Perth, WA, Australia, December 5–8, 2022, Proceedings* (Perth, WA, Australia). Springer-Verlag, Berlin, Heidelberg, 205–219. doi:10.1007/978-3-031-22695-3_15
- [24] Diem Pham, Binh Tran, Su Nguyen, Daminda Alahakoon, and Mengjie Zhang. 2024. Fairness optimisation with multi-objective swarms for explainable classifiers on data streams. *Complex & Intelligent Systems* (2024). <https://api.semanticscholar.org/CorpusID:268897732>
- [25] Harini Suresh and John Gutttag. 2021. A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. In *Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '21)*. ACM. doi:10.1145/3465416.3483305
- [26] Valerie Vaquet, Ulrike Kuhl, Saša Brdnic, and Benjamin Paaßen. 2026. Beyond Performance: Comprehensive Evaluation Strategies for Impactful Machine Learning. doi:10.14428/esann/2026.ES2026-3
- [27] Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness* (Gothenburg, Sweden) (*FairWare '18*). Association for Computing Machinery, New York, NY, USA, 1–7. doi:10.1145/3194770.3194776
- [28] R. Wilms, E. Mäthner, L. Winnen, and R. Lanwehr. 2021. Omitted variable bias: A threat to estimating causal relationships. *Methods in Psychology* 5 (2021), 100075. doi:10.1016/j.metip.2021.100075
- [29] Wenbin Zhang and Albert Bifet. 2020. FEAT: A Fairness-Enhancing and Concept-Adapting Decision Tree Classifier. In *Discovery Science: 23rd International Conference, DS 2020, Thessaloniki, Greece, October 19–21, 2020, Proceedings* (Thessaloniki, Greece). Springer-Verlag, Berlin, Heidelberg, 175–189. doi:10.1007/978-3-030-61527-7_12
- [30] Wenbin Zhang, Albert Bifet, Xiangliang Zhang, Jeremy C. Weiss, and Wolfgang Nejdl. 2021. FARF: A Fair and Adaptive Random Forests Classifier. In *Advances in Knowledge Discovery and Data Mining*, Kamal Karlapalem, Hong Cheng, Naren Ramakrishnan, R. K. Agrawal, P. Krishna Reddy, Jaideep Srivastava, and Tanmoy Chakraborty (Eds.). Vol. 12713. Springer International Publishing, Cham, 245–256. doi:10.1007/978-3-030-75765-6_20
- [31] Wenbin Zhang and Eirini Ntoutsi. 2019. FAHT: An Adaptive Fairness-aware Decision Tree Classifier. In *International Joint Conference on Artificial Intelligence*. <https://api.semanticscholar.org/CorpusID:197431421>
- [32] Wenbin Zhang, Mingli Zhang, Ji Zhang, Zhen Liu, Zhiyuan Chen, Jianwu Wang, Edward Raff, and Enza Messina. 2020. Flexible and Adaptive Fairness-aware Learning in Non-stationary Data Streams. *2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI)* (2020), 399–406. <https://api.semanticscholar.org/CorpusID:229703513>
- [33] Wenbin Zhang and Liang Zhao. 2020. Online Decision Trees with Fairness. doi:10.48550/ARXIV.2010.08146

Table 5. Overview of fair stream learning algorithms

ML Algorithm			Fairness pre-/post/in		Drift	Imbalance	Code published
		target metric					
pre	Massaging [14]	model agnostic	DP	pre	✗	✗	✓
	Reweighting [14]	model agnostic	DP	pre	✗	✗	✓
	CFSMOTE [18]	model agnostic	individual	pre	✓	✓	✓
tree	FAHT [31]	Hoeffding Tree	DP	in	✗	✗	✓
	FEAT [29]	HAT	DP	in	✓	✗	✗
	2FAHT [32]	CVFDT	DP	in	✓	✗	✗
	FARF [30]	ARF	DP	in	✓	✗	✗
evolutionary	FOMOS [24]	Hoeffding Tree	DP	pre/in	✗	✗	✗
	FAS Stream [23]	SAM, DSOS	DP	pre/in	✓	✗	✗
	EMO-SAM [1]	SAM-kNN	DP	pre/in	✓	✗	✗
	Fair-CMNB [2]	Mixed Naive Bayes		in/post	✓	✓	✗
post	FABBOO [13]	OzaBoost (model agnostic)	DP/EqOp/ EqFPR	post	✓	✓	✓

Classifier	Statistical Parity	Disparate Impact	Equal Opportunity	Equal FPR	Individual Fairness	Accuracy	Balanced Accuracy	Recall
FAHT	3.94 ± 0.52	35.5 ± 3.58	4.27 ± 1.14	1.61 ± 0.34	3.64 ± 0.9	80.4 ± 0.19	62.3 ± 0.4	28.5 ± 0.95
FEAT	6.26 ± 1.65	74.7 ± 18.8	13.7 ± 4.01	3.4 ± 0.87	6.45 ± 1.47	78 ± 0.31	55.5 ± 1.04	13.7 ± 2.4
2FAHT	4.34 ± 0.75	36.1 ± 2.2	5.69 ± 2.07	1.84 ± 0.35	0.57 ± 0.05	80.7 ± 0.31	63.4 ± 1.23	31.2 ± 2.98
CFSMOTE	10.5 ± 0.88	28.2 ± 2.56	6.79 ± 1.55	6.25 ± 0.73	6.81 ± 0.74	78.5 ± 0.28	76.9 ± 0.15	73.9 ± 0.78
MAS	9.37 ± 0.54	36.5 ± 1.7	11.7 ± 1.52	4.74 ± 0.45	3.41 ± 0.4	82.8 ± 0.38	72.5 ± 0.24	53.3 ± 0.59
FABBOO SP	0.44 ± 0.24	1.39 ± 0.8	5.51 ± 0.5	4.64 ± 0.2	16.6 ± 0.4	77.7 ± 0.24	77.2 ± 0.15	76.4 ± 0.62
FABBOO EQOP	8.11 ± 1.14	20.8 ± 3.13	2.27 ± 0.8	3.88 ± 1.18	10.6 ± 0.79	79.5 ± 0.26	78.1 ± 0.23	75.6 ± 0.68

Table 6. Fairness and performance comparison across metrics on the *normal and fair world switch* stream (1 drift) with drift-aware scoring.

Classifier	Statistical Parity	Disparate Impact	Equal Opportunity	Equal FPR	Individual Fairness	Accuracy	Balanced Accuracy	Recall
FAHT	16.4 ± 1.15	64.1 ± 2.83	15.7 ± 1.76	8.02 ± 1.01	8.98 ± 1.96	76.7 ± 0.13	65.7 ± 0.22	39.7 ± 0.51
FEAT	26.6 ± 1.27	92.1 ± 1.21	36.6 ± 1.99	15.6 ± 1	17.1 ± 2.21	76.4 ± 0.56	64.9 ± 1.07	40 ± 2.55
2FAHT	23.1 ± 1.28	71.5 ± 2.41	23.4 ± 2.16	13.1 ± 0.92	1.35 ± 0.34	78.2 ± 0.17	68.8 ± 0.92	48.6 ± 2.43
CFSMOTE	20.5 ± 0.94	47.3 ± 1.77	10.4 ± 1.8	9.46 ± 0.82	6.56 ± 0.82	78.7 ± 0.35	77.2 ± 0.23	74.1 ± 0.78
MAS	23.5 ± 1.67	68.6 ± 1.35	23.1 ± 1.54	11.4 ± 1.77	4.96 ± 0.66	80.4 ± 0.3	74.5 ± 0.33	60.1 ± 2.23
FABBOO SP	0.35 ± 0.1	0.74 ± 0.18	9.47 ± 0.4	11.8 ± 0.25	30 ± 0.56	74.3 ± 0.32	76.2 ± 0.16	82.1 ± 0.41
FABBOO EQOP	16.8 ± 0.6	36 ± 1.58	2.67 ± 0.76	5.76 ± 0.67	16.8 ± 1.15	78.3 ± 0.34	78.6 ± 0.16	79.9 ± 0.49

Table 7. Fairness and performance comparison across metrics on the *inflation scenario* stream (1 drift) with drift-aware scoring.

A Additional Tables and Results

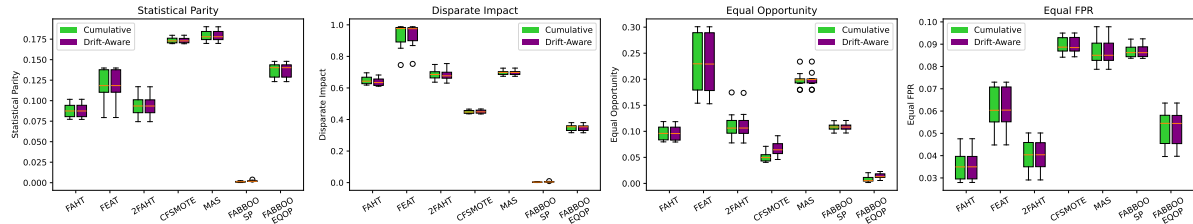
B Online Resources

The code can be found on GitHub.

Classifier	Statistical Parity	Disparate Impact	Equal Opportunity	Equal FPR	Individual Fairness	Accuracy	Balanced Accuracy	Recall
FAHT	8.44 ± 1	61.7 ± 2.94	8.88 ± 1.91	3.41 ± 0.66	4.57 ± 1.1	79.7 ± 0.12	62.8 ± 0.55	30.1 ± 1.43
FEAT	10.2 ± 2.23	96.9 ± 1.81	21.1 ± 4.91	5.51 ± 1.12	8.06 ± 1.69	77.3 ± 0.53	57 ± 1.72	17.8 ± 4.09
2FAHT	8.56 ± 1.15	67.7 ± 4.09	10.6 ± 3.37	3.67 ± 0.65	0.53 ± 0.16	79.4 ± 0.28	61.8 ± 0.63	27.5 ± 1.59
CFSMO HAT	16.6 ± 2.19	36.7 ± 3.48	6.17 ± 2.67	9.42 ± 2.39	2.11 ± 0.42	73.5 ± 1.21	74.3 ± 0.65	75.9 ± 1
CFSMO ARF	17.2 ± 0.69	45.5 ± 1.69	9.37 ± 0.93	8.46 ± 0.76	7.89 ± 1.45	78.9 ± 0.34	76.7 ± 0.24	72.5 ± 0.6
MAS	17.8 ± 0.74	69.2 ± 1.92	19.7 ± 2.21	8.77 ± 0.61	4.74 ± 1	82.4 ± 0.3	73 ± 0.22	54.6 ± 0.63
FABBOO SP	0.18 ± 0.06	0.46 ± 0.14	10.1 ± 0.55	8.54 ± 0.16	29.5 ± 0.64	74.8 ± 0.29	76.2 ± 0.11	79 ± 0.42
FABBOO EQOP	13 ± 0.71	32.9 ± 1.73	1.42 ± 0.6	4.57 ± 0.8	18.3 ± 1.06	78.1 ± 0.3	77.7 ± 0.14	77 ± 0.48

Table 8. Fairness and performance comparison across metrics on the *women in stem scenario* stream with drift-aware scoring.

Classifier	Statistical Parity	Disparate Impact	Equal Opportunity	Equal FPR	Individual Fairness	Accuracy	Balanced Accuracy	Recall
FAHT	8.87 ± 0.86	64 ± 2.33	9.78 ± 1.45	3.58 ± 0.64	4.67 ± 1.15	79.9 ± 0.21	63 ± 0.54	30.4 ± 1.43
FEAT	11.9 ± 1.96	93.5 ± 7.43	23.4 ± 5.66	6.18 ± 0.92	8.71 ± 2.21	78.2 ± 0.86	59.5 ± 2.31	23.3 ± 5.16
2FAHT	9.42 ± 1.16	68.2 ± 3.56	11.2 ± 2.49	4.02 ± 0.67	0.72 ± 0.24	79.8 ± 0.36	62.9 ± 1.02	30.1 ± 2.43
CFSMOTTE	17.3 ± 0.34	44.9 ± 1.06	6.57 ± 1.29	8.93 ± 0.36	7.07 ± 0.96	78.8 ± 0.38	77.3 ± 0.26	74.3 ± 0.31
MAS	17.9 ± 0.65	69.7 ± 1.54	19.9 ± 1.45	8.68 ± 0.63	4.73 ± 0.99	82.5 ± 0.33	73 ± 0.22	54.6 ± 0.52
FABBOO SP	0.21 ± 0.08	0.53 ± 0.19	10.8 ± 0.73	8.69 ± 0.28	30.2 ± 1.02	74.8 ± 0.26	76.3 ± 0.11	79.1 ± 0.52
FABBOO EQOP	13.7 ± 0.86	34.9 ± 2	1.38 ± 0.56	5.2 ± 0.81	18.2 ± 0.8	78.3 ± 0.24	77.9 ± 0.24	77 ± 0.55

Table 9. Fairness and performance comparison across metrics on the *less marriage scenario* stream with drift-aware scoring.Fig. 11. Paired boxplots showing the differences between cumulatively and drift-awarely aggregated scores on the *less marriage scenario* stream with one drift.

Classifier	Statistical Parity	Disparate Impact	Equal Opportunity	Equal FPR	Individual Fairness	Accuracy	Balanced Accuracy	Recall
FAHT	0.26 ± 0.11	0.26 ± 0.11	0.27 ± 0.1	0.3 ± 0.31	0 ± 0.01	86.3 ± 0.22	50.1 ± 0.23	99.7 ± 0.13
FEAT	1.26 ± 0.59	1.28 ± 0.6	0.81 ± 0.43	2.57 ± 1.24	0.1 ± 0.08	86.4 ± 0.28	52.5 ± 1.05	99 ± 0.36
2FAHT	0.51 ± 0.25	0.51 ± 0.25	0.44 ± 0.14	1.21 ± 0.65	0.02 ± 0.01	86.2 ± 0.35	50.5 ± 0.43	99.3 ± 0.28
CFSMOTTE	6.45 ± 1.78	7.57 ± 2.03	5.56 ± 1.22	6.63 ± 3.15	6.39 ± 2.95	77.1 ± 2.14	57.2 ± 2.68	84.5 ± 2.48
MAS	2.55 ± 1.4	2.6 ± 1.43	1.58 ± 1.02	4.84 ± 2.45	0.34 ± 0.3	86.4 ± 0.35	53.7 ± 1.4	98.5 ± 0.66
EMOSAM	0.34 ± 0.23	0.34 ± 0.24	0.36 ± 0.28	0.48 ± 0.45	0 ± 0	86 ± 0.4	50 ± 0.27	99.3 ± 0.36
FABBOO SP	2.95 ± 0.92	3.22 ± 0.99	2.02 ± 0.74	9.82 ± 1.51	7.74 ± 0.46	84.8 ± 0.42	62.4 ± 0.91	93 ± 0.62
FABBOO EQOP	5.19 ± 1.41	5.67 ± 1.55	3.68 ± 0.99	10.4 ± 2.83	6.48 ± 0.58	84.6 ± 0.45	63.4 ± 1.12	92.4 ± 0.76

Table 10. Fairness and performance comparison across metrics on the *girls support* stream (three drifts) with drift-aware scoring.

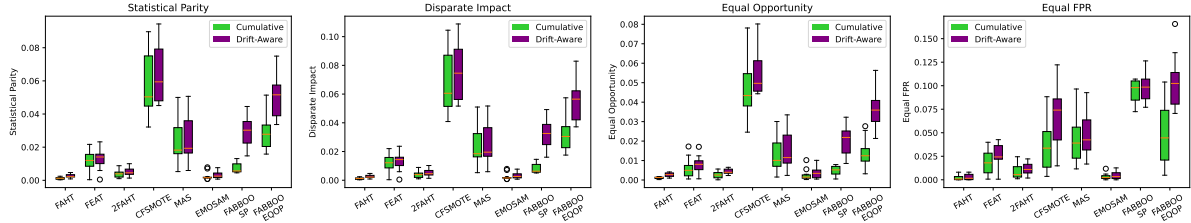


Fig. 12. Paired boxplots showing the differences between cumulatively and drift-awarely aggregated scores on the *girls support* stream with three drifts.

Classifier	Statistical Parity	Disparate Impact	Equal Opportunity	Equal FPR	Individual Fairness	Accuracy	Balanced Accuracy	Recall
FAHT	0.12 ± 0.07	0.12 ± 0.07	0.13 ± 0.09	0.2 ± 0.15	0.01 ± 0.02	85.9 ± 0.32	50 ± 0.1	99.8 ± 0.07
FEAT	1.83 ± 0.53	1.87 ± 0.55	1.11 ± 0.34	2.51 ± 1.63	0.1 ± 0.1	85.9 ± 0.42	53 ± 1.04	98.8 ± 0.34
2FAHT	0.69 ± 0.39	0.7 ± 0.4	0.54 ± 0.33	1.61 ± 0.65	0.05 ± 0.04	85.7 ± 0.34	51.3 ± 1.04	99.1 ± 0.43
CFSMOTE	6.53 ± 1.37	7.65 ± 1.73	5.54 ± 1.13	6.75 ± 2.97	7.89 ± 4.1	76.7 ± 2.05	58.3 ± 1.69	83.8 ± 2.57
MAS	2.55 ± 1.09	2.63 ± 1.14	1.55 ± 0.81	3.98 ± 2.23	0.52 ± 0.37	85.9 ± 0.56	54.8 ± 1.72	98 ± 1.09
EMOSAM	0.33 ± 0.22	0.34 ± 0.22	0.3 ± 0.24	0.64 ± 0.39	0 ± 0	85.3 ± 0.62	49.9 ± 0.23	99 ± 0.52
FABBOO SP	2.45 ± 0.74	2.75 ± 0.81	1.35 ± 0.67	5.8 ± 2.29	4.96 ± 0.88	83.7 ± 0.68	63.8 ± 0.61	91.3 ± 1.04
FABBOO EQOP	4.29 ± 1.16	4.8 ± 1.29	2.66 ± 0.85	4.47 ± 1.07	4.1 ± 1.04	83.5 ± 0.69	64.6 ± 0.53	90.8 ± 1.09

Table 11. Fairness and performance comparison across metrics on the *girls support* stream (one drift) with drift-aware scoring.

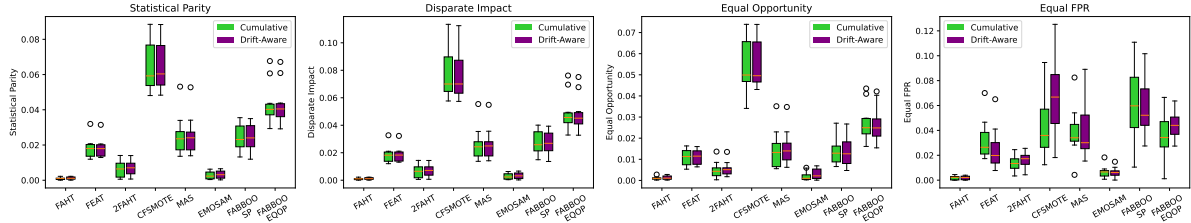
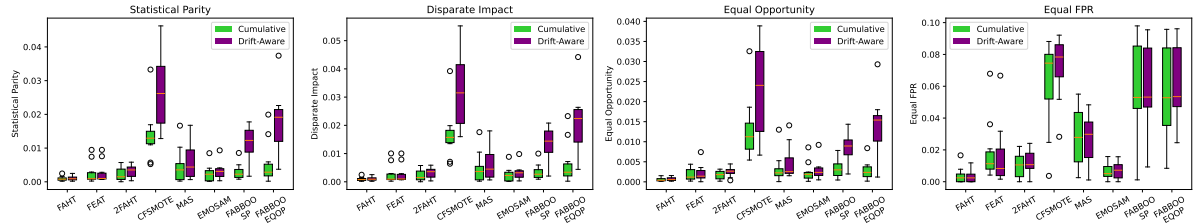


Fig. 13. Paired boxplots showing the differences between cumulatively and drift-awarely aggregated scores on the *girls support* stream with one drift.

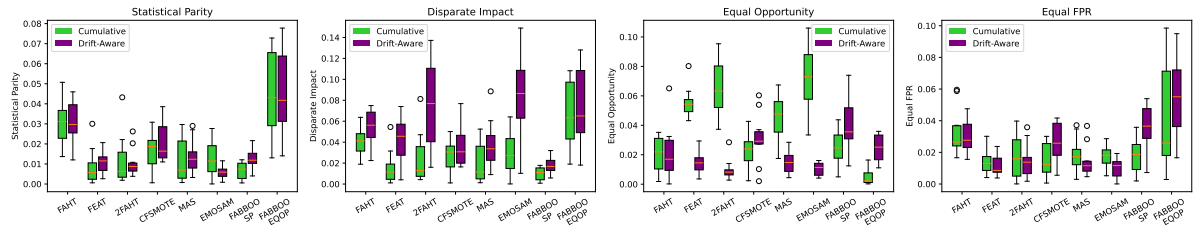
Classifier	Statistical Parity	Disparate Impact	Equal Opportunity	Equal FPR	Individual Fairness	Accuracy	Balanced Accuracy	Recall
FAHT	0.16 ± 0.15	0.16 ± 0.15	0.16 ± 0.14	0.24 ± 0.22	0.01 ± 0.02	85.3 ± 0.34	50 ± 0.14	99.7 ± 0.11
FEAT	0.74 ± 0.31	0.77 ± 0.33	0.52 ± 0.22	1.41 ± 0.95	0.14 ± 0.08	85.4 ± 0.36	53.1 ± 0.85	98.6 ± 0.37
2FAHT	0.42 ± 0.15	0.43 ± 0.16	0.34 ± 0.13	2.03 ± 0.98	0.07 ± 0.06	85.1 ± 0.38	51 ± 0.49	99.1 ± 0.39
CFSMOTE	2.6 ± 0.73	3.15 ± 0.9	2.28 ± 0.6	5.26 ± 2.29	5.9 ± 2.85	76.7 ± 1.58	59.3 ± 2.48	83.8 ± 1.79
MAS	1.09 ± 0.71	1.14 ± 0.77	0.76 ± 0.66	2.87 ± 1.18	0.44 ± 0.39	85.3 ± 0.33	55.3 ± 2.51	97.6 ± 1.23
FABBOO SP	2.07 ± 0.6	2.43 ± 0.7	1.64 ± 0.52	8.51 ± 1.96	2.65 ± 0.53	82.4 ± 0.6	66.5 ± 0.8	88.9 ± 0.75
FABBOO EQOP	2.74 ± 0.6	3.22 ± 0.71	2.06 ± 0.56	8.5 ± 2.83	2.6 ± 0.58	82.3 ± 0.55	66.7 ± 0.64	88.7 ± 0.7

Table 12. Fairness and performance comparison across metrics on the *boys support* stream (two drifts) with drift-aware scoring.

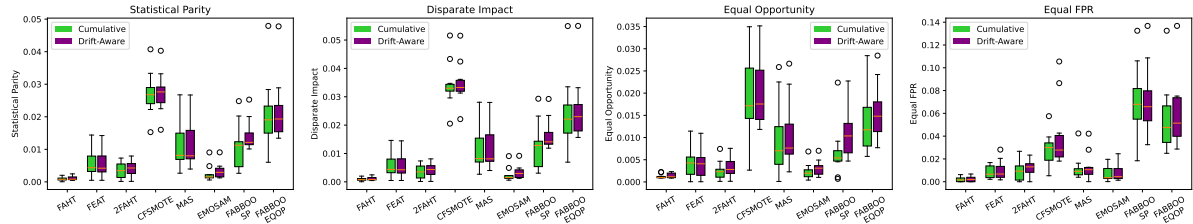
Classifier	Statistical Parity	Disparate Impact	Equal Opportunity	Equal FPR	Individual Fairness	Accuracy	Balanced Accuracy	Recall
FAHT	0.11 ± 0.06	0.11 ± 0.06	0.07 ± 0.04	0.34 ± 0.38	0.03 ± 0.07	85.7 ± 0.44	50.1 ± 0.47	99.7 ± 0.17
FEAT	0.27 ± 0.3	0.28 ± 0.32	0.22 ± 0.2	1.67 ± 1.9	0.12 ± 0.15	85.7 ± 0.45	53 ± 2.14	98.7 ± 0.96
2FAHT	0.32 ± 0.17	0.32 ± 0.18	0.25 ± 0.12	1.27 ± 0.75	0.06 ± 0.04	85.6 ± 0.4	51 ± 0.58	99.3 ± 0.31
CFSMOTE	2.73 ± 1.15	3.27 ± 1.36	2.32 ± 1.1	7.24 ± 1.87	6.48 ± 3.16	77.2 ± 1.1	59.3 ± 2.39	84.3 ± 1.05
MAS	0.64 ± 0.56	0.67 ± 0.6	0.48 ± 0.41	2.69 ± 1.57	0.52 ± 0.35	85.6 ± 0.45	54.5 ± 2.44	97.9 ± 1.18
EMOSAM	0.32 ± 0.24	0.33 ± 0.25	0.29 ± 0.23	0.73 ± 0.52	0 ± 0	85 ± 1.13	49.9 ± 0.34	98.9 ± 1.35
FABBOO SP	1.12 ± 0.53	1.32 ± 0.62	0.88 ± 0.36	5.9 ± 2.7	2.26 ± 0.54	82.5 ± 0.73	66.5 ± 0.67	88.9 ± 0.93
FABBOO EQOP	1.8 ± 0.88	2.13 ± 1.04	1.37 ± 0.74	6.05 ± 2.42	2.37 ± 0.64	82.5 ± 0.73	66.6 ± 0.58	88.8 ± 0.95

Table 13. Fairness and performance comparison across metrics on the *boys support* stream (one drift) with drift-aware scoring.Fig. 14. Paired boxplots showing the differences between cumulatively and drift-awarely aggregated scores on the *boys support* stream with one drift.

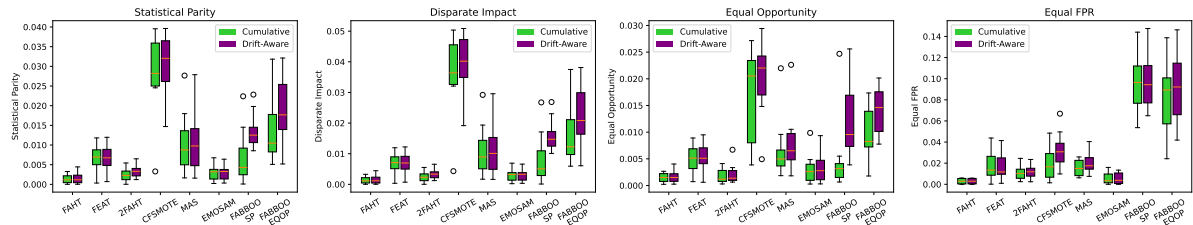
Classifier	Statistical Parity	Disparate Impact	Equal Opportunity	Equal FPR	Individual Fairness	Accuracy	Balanced Accuracy	Recall
FAHT	3.14 ± 1.05	5.46 ± 1.64	2.13 ± 1.78	3.03 ± 1.02	2.24 ± 0.82	64.5 ± 1	51.9 ± 0.26	75.7 ± 2.88
FEAT	1.06 ± 0.52	4.17 ± 2.14	1.5 ± 0.74	1.19 ± 0.65	0.54 ± 0.24	71.1 ± 0.95	50.8 ± 0.63	54.3 ± 2.72
2FAHT	1.06 ± 0.68	7.65 ± 4.09	0.98 ± 0.69	1.5 ± 1.04	0.46 ± 0.52	73.5 ± 1.34	51 ± 0.78	47.1 ± 4.59
CFSMOTE	2.05 ± 0.93	3.62 ± 1.9	3.08 ± 1.65	2.71 ± 1.18	7.83 ± 2.24	62.7 ± 1.21	54.8 ± 1.47	63.7 ± 1.85
MAS	1.38 ± 0.82	3.78 ± 2.24	1.5 ± 0.77	1.44 ± 0.97	0.36 ± 0.38	69 ± 0.74	51.8 ± 1.61	59.4 ± 1.4
EMOSAM	0.58 ± 0.29	8.49 ± 3.68	1.05 ± 0.43	1.02 ± 0.59	0 ± 0	74.1 ± 0.97	50 ± 0.31	42 ± 1.17
FABBOO SP	1.27 ± 0.51	1.8 ± 0.81	3.93 ± 1.82	3.53 ± 1.53	5.93 ± 0.65	64.7 ± 0.5	60.8 ± 0.4	76 ± 1.16
FABBOO EQOP	4.68 ± 2.07	7.54 ± 3.59	2.43 ± 0.89	5.49 ± 2.36	6.29 ± 1.26	65.2 ± 0.63	61 ± 0.4	74.6 ± 1.25

Table 14. Fairness and performance comparison across metrics on the *grade category adjustment* stream (one drift) with drift-aware scoring.Fig. 15. Paired boxplots showing the differences between cumulatively and drift-awarely aggregated scores on the *grade category adjustment* stream with one drift.

Classifier	Statistical Parity	Disparate Impact	Equal Opportunity	Equal FPR	Individual Fairness	Accuracy	Balanced Accuracy	Recall
FAHT	0.11 ± 0.07	0.11 ± 0.07	0.14 ± 0.06	0.22 ± 0.24	0 ± 0.01	87.3 ± 0.28	50 ± 0.07	99.8 ± 0.08
FEAT	0.6 ± 0.46	0.61 ± 0.47	0.46 ± 0.37	0.97 ± 0.83	0.06 ± 0.07	87.2 ± 0.32	51.7 ± 0.94	99.1 ± 0.38
2FAHT	0.41 ± 0.22	0.42 ± 0.23	0.33 ± 0.2	1.23 ± 0.61	0.04 ± 0.03	87.2 ± 0.38	51 ± 0.57	99.2 ± 0.35
CFSMOTE	2.74 ± 0.61	3.49 ± 0.73	2.01 ± 0.75	4.08 ± 2.88	5.48 ± 2.56	74.9 ± 3.2	59.7 ± 1.78	80.2 ± 4.14
MAS	1.24 ± 0.79	1.29 ± 0.84	1.1 ± 0.74	1.34 ± 1.19	0.54 ± 0.33	86.9 ± 0.5	53.8 ± 2.23	98 ± 1.23
EMOSAM	0.34 ± 0.23	0.34 ± 0.23	0.32 ± 0.17	0.77 ± 0.71	0 ± 0	87 ± 0.4	50 ± 0.2	99.3 ± 0.39
FABBOO SP	1.42 ± 0.46	1.64 ± 0.54	1.09 ± 0.54	7.25 ± 2.92	3.93 ± 0.68	83.7 ± 0.85	64.7 ± 0.75	89.9 ± 1.08
FABBOO EQOP	2.21 ± 0.97	2.57 ± 1.11	1.59 ± 0.61	6.04 ± 3	3.55 ± 0.93	83.5 ± 0.85	65.1 ± 0.6	89.6 ± 1.08

Table 15. Fairness and performance comparison across metrics on the *grade inflation* stream (one drift) with drift-aware scoring.Fig. 16. Paired boxplots showing the differences between cumulatively and drift-awarely aggregated scores on the *grade inflation* stream with one drift.

Classifier	Statistical Parity	Disparate Impact	Equal Opportunity	Equal FPR	Individual Fairness	Accuracy	Balanced Accuracy	Recall
FAHT	0.15 ± 0.14	0.16 ± 0.14	0.17 ± 0.12	0.29 ± 0.24	0.01 ± 0.02	85.1 ± 0.29	50 ± 0.09	99.7 ± 0.1
FEAT	0.7 ± 0.33	0.72 ± 0.34	0.54 ± 0.25	1.69 ± 1.18	0.16 ± 0.1	85.4 ± 0.4	53.3 ± 0.87	98.7 ± 0.36
2FAHT	0.34 ± 0.16	0.34 ± 0.16	0.21 ± 0.18	1.2 ± 0.63	0.05 ± 0.04	85.1 ± 0.35	51.1 ± 0.68	99.3 ± 0.32
CFSMOTE	3.08 ± 0.75	3.95 ± 0.92	2.03 ± 0.66	3.26 ± 1.57	5.07 ± 1.94	75.9 ± 2.53	64.3 ± 2.48	80.8 ± 2.63
MAS	1.08 ± 0.75	1.14 ± 0.8	0.79 ± 0.56	2.15 ± 1.05	0.59 ± 0.4	85.3 ± 0.5	55.7 ± 2.54	97.6 ± 1.26
EMOSAM	0.3 ± 0.18	0.31 ± 0.19	0.33 ± 0.27	0.56 ± 0.52	0 ± 0	84.5 ± 0.95	49.9 ± 0.3	98.9 ± 0.92
FABBOO SP	1.37 ± 0.42	1.61 ± 0.5	1.23 ± 0.73	9.76 ± 2.43	3.13 ± 0.64	82.8 ± 0.69	66.9 ± 0.37	89.5 ± 0.85
FABBOO EQOP	1.93 ± 0.83	2.28 ± 0.98	1.4 ± 0.42	9.07 ± 3.13	2.96 ± 0.85	82.8 ± 0.66	67.1 ± 0.34	89.3 ± 0.83

Table 16. Fairness and performance comparison across metrics on the *internet era* stream (one drift) with drift-aware scoring.Fig. 17. Paired boxplots showing the differences between cumulatively and drift-awarely aggregated scores on the *internet era* stream with one drift.