Information Sciences 451–452 (2018) 348–370

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

A large-scale comparison of concept drift detectors

Roberto Souto Maior Barros*, Silas Garrido T. Carvalho Santos

Centro de Informática, Universidade Federal de Pernambuco, Cidade Universitária, Recife, PE, 50740-560, Brazil

ARTICLE INFO

Article history:
Received 16 November 2017
Revised 3 February 2018
Accepted 3 April 2018
Available online 3 April 2018

Keywords:
Concept drift
Drift detection
Large-scale comparison
Data stream
Online learning

ABSTRACT

Online learning involves extracting information from large quantities of data (streams) usually affected by changes in the distribution (concept drift). A drift detector is a small program that estimates the positions of these changes to replace the base learner and ultimately improve overall accuracy. This article reports on a large-scale comparison of 14 concept drift detector configurations for mining fully labeled data streams with concept drift, using a large number of artificial datasets and two different base classifiers (Naive Bayes and Hoeffding Tree). The goal is to adequately measure how good the existent concept drift detectors really are and also to verify and challenge a common belief in the area, namely that the best drift detection methods are necessarily those that detect all the existing drifts closer to their correct positions, and only them, irrespective of the fact that different objectives usually require alternative solutions. Finally, to some extent, this article may also be seen as an extensive literature survey of concept drift detectors.

© 2018 Elsevier Inc. All rights reserved.

1. Introduction

In data stream environments, large amounts of data (possibly infinite) flow rapidly and continuously. Learning from these streams therefore requires online methods and is often restricted in terms of memory and run-time usage. Also, a common constraint is that each data instance can be read only once. Finally, concept drifts [20,21], usually seen as changes in data distribution, may occur.

A very common categorization of concept drift is based on the speed of change. Sudden and/or rapid changes between concepts are called *abrupt* and transitions from one concept to another over a larger number of instances are called *gradual* [22–24,38].

In real-world applications concept drift may occur for several reasons, such as equipment failure, intrusion, seasonal changes, etc. and thus it is important to detect these drifts in many applications. Examples of real-world applications with concept drift include the detection of changes in the weather or water temperature, detecting movement in data collected from sensors [31], filtering spam in e-mail messages [29], predicting equipment failure, and many others [20,51].

Different directions have been investigated to learn from data streams containing concept drift. One that is very common involves concept drift detectors [24], lightweight software that focuses on identifying changes in data distribution by monitoring the prediction results of a base classifier.

Several ensembles employing a base learner have also been proposed to deal with concept drift, sometimes using different strategies and/or weighting functions to compute the resulting classification, including Dynamic Weighted Majority

* Corresponding author.
E-mail addresses: roberto@cin.ufpe.br (R.S.M. Barros), sgtcs@cin.ufpe.br (S.G.T.C. Santos).

https://doi.org/10.1016/j.ins.2018.04.014
0020-0255/© 2018 Elsevier Inc. All rights reserved.

(DWM) [30], Diversity for Dealing with Drifts (DDD) [35], Adaptable Diversity-based Online Boosting (ADOB) [48], Fast Adaptive Stacking of Ensembles (FASE) [19], and Boosting-like Online Learning Ensemble (BOLE) [7]. Some methods reuse previous classifiers on recurring concepts, e.g. Recurring Concept Drifts (RCD) [23]. Recurring concept drift has been also addressed by fuzzy methods [43]. In addition, it is worth pointing out that some of these ensembles also rely on an auxiliary drift detection method [7,23,35,48].

Ensembles of concept drift detection methods sharing the same base classifier is another approach that has been comparatively less explored [16,33]. Sequential Extreme Learning Machine (ELM) based computation has also been proposed to deal with concept drift and imbalanced data [36].

A number of concept drift detectors have been proposed over the years and the most well-known are Drift Detection Method (DDM) [22], Early Drift Detection Method (EDDM) [3], Adaptive Windowing (ASWIN) [9], Statistical Test of Equal Proportions (STEPD) [38], Page-Hinkley Test (PHT) [39], Paired Learners (PL) [2], and EWMA for Concept Drift Detection (ECDD) [45]. Of these, DDM and STEPD are among the most simple methods and present a good all-round performance [24].

Other more recent concept drift detection methods have also been proposed, including Sequential Drift (SEQDRIFT) [40], SEED Drift Detector (SEED) [26], Drift Detection Methods based on Hoeffding's Bounds (HDDM) [18] with versions $HDDM_A$ and $HDDM_W$, Fast Hoeffding Drift Detection Method (FHDDM) [42], Equal Means Z-Test Drift Detector (EMZD) [12], Fisher Test Drift Detector (FTDD) [12,13], Reactive Drift Detection Method (RDDM) [4,5], and Wilcoxon Rank Sum Test Drift Detector (WSTD) [4,6].

One of the objectives of this work is to verify and challenge the common wisdom in the area that the best drift detection methods are necessarily those that detect all the existing concept drifts closer to their correct points, ideally detecting them alone. However, the real world often does not behave according to the expectations or predictions of currently accepted theoretical models. In addition, in most given problems, different objectives usually require alternative solutions.

For example, cars are normally expected to remain in the main driveway at all times, irrespective of the reason why they are being driven. This is certainly part of the best practices to avoid accidents in normal everyday driving. Nevertheless, in car racing, where the objective is to go a certain distance as quickly as possible, this is not always the best strategy. In this scenario, a certain amount of sliding of the cars out of the main path of the track is often beneficial, especially in go-karts and rally racing, and even in formula one racing.

Similarly, it might be that detecting the existing concept drifts very close to their positions, and detecting only these drifts, is not the best strategy to maximize the accuracy of the classifiers. In other words, minimizing the distance of the true positive detections as well as the false negative and false positive ones might not be the best strategy to maximize accuracy in some scenarios. A small number of false negatives and/or false positives might indeed be beneficial, helping to improve the global accuracy of the classifiers in some problems.

This article presents a large-scale comparison of concept drift detection methods, including detailed information on all relevant aspects of the experiments and analyzing their results. More explicitly, 14 different configurations of drift detectors are compared in terms of accuracy and their detections in a fully labeled setting (supervised learning). The results of this research give indications of the best methods available. It summarizes and extends the experiments reported in [4]. To the best of our knowledge, these are the largest and most comprehensive comparisons of drift detectors ever reported in the area of data stream mining.

Another comparative study of concept drift detectors was published a few years ago [24]. However, that earlier work had a much more limited scope, with only the most well-known drift detectors, considerably fewer datasets, only one base learner, and limited analysis of the drift detections. In addition, it aimed to identify which parameters had more influence on the predictive accuracy of the methods using a 2^k factorial design. On the other hand, the current work has its focus on providing a deep analysis of the performance of both traditional and newer drift detection methods in different scenarios.

The experiments reported here were performed using several artificial dataset generators, configured with both abrupt and gradual drift versions of several sizes, using two different base classifiers — Naive Bayes (NB) [28] and Hoeffding Tree (HT) [27], and run in the Massive Online Analysis (MOA) framework [10], release 2014.11.

More specifically, these experiments were designed to answer research questions **RQ1** to **RQ5** below:

- **RQ1:** What are the best drift detectors in terms of accuracy in abrupt and gradual concept drift datasets?
- **RQ2:** What are the best concept drift detectors in terms of detections, measured by *Precision* and *Recall* [41] and the Matthews Correlation Coefficient (MCC) metric [34], in the abrupt datasets?
- **RQ3:** Do the answers to **RQ1** and **RQ2** vary with the different dataset generators used in the experiments? If so, to what extent?
- **RQ4:** Do the answers of **RQ1** and **RQ2** depend on the size of the concepts included in the datasets? If so, to what extent?
- **RQ5:** In the same datasets, are the best methods of **RQ1** and **RQ2** the same? If so, to what extent?

The rest of the document is organized in six sections: Section 2 reviews the published literature on concept drift detection methods; Section 3 shows the configuration of the experiments, also including brief descriptions of the datasets used in the tests; Section 4 analyses the accuracy results of the tested concept drift detection methods configurations and evaluates them statistically to answer **RQ1**; Section 5 inspects the results of the confusion matrix regarding the detections of the methods, i.e. false negatives (FN), false positives (FP), true negatives (TN), and true positives (TP), to answer **RQ2**; Section 6 provides additional perspectives on the results of the experiments, answering **RQ3**, **RQ4**, and **RQ5**; and, finally, Section 7 presents some conclusions and proposes future work.

2. Related work: Drift detection methods

It is fairly common to use a concept drift detection method together with a base classifier to learn from data streams. In general, the drift detector analyses the prediction results of the base learner and applies a particular decision model to attempt to detect changes in the data distribution. The most well-known methods that follow this approach are DDM [22], EDDM [3], and STEPD [38].

Provided with a sequence of examples in the form of pairs $(\vec{x}_i, y_i)$, where $\vec{x}_i$ is a vector of attributes and y_i is its corresponding class, for each example, the base learner makes a prediction $(\hat{y}_i)$, which is then compared to the actual result (y_i) to decide whether the prediction was correct $(\hat{y}_i = y_i)$ or not $(\hat{y}_i \neq y_i)$.

Distinct drift detection methods use different strategies and/or statistics to monitor the performance of the base classifier and to decide when concept drifts have occurred. A lower confidence level to indicate warnings is also common and these warnings mean that concept drifts may have occurred. At such points, the methods prescribe that a new instance of the base classifier is created and starts to be trained in parallel. Eventually, when a concept drift is confirmed, the new learner will replace the original learner. On the other hand, when the warning is a false alarm, the new instance will be discarded.

Nevertheless, it is important to clarify that, in the MOA framework [10], the drift detection methods merely signal the *warning* and *drift* positions. The interface with the base learner is actually handled by other classes of MOA.

DDM detects concept drifts in streams by analyzing the error rate and its corresponding standard deviation. For each position i, DDM defines the error rate p_i as the probability of making an incorrect prediction and its standard deviation as $s_i = \sqrt{p_i \times (1 - p_i) \,/\, i}$. Based on the Probably Approximately Correct (PAC) learning model [37], the authors of DDM argue that, provided that the distribution of the examples remains stationary, the error rate p_i should decrease as the number of examples i increases. If the error rate increases, DDM assumes there was a change in the data distribution and the current base learner is outdated.

EDDM is similar to DDM but it monitors the distance between two consecutive errors, rather than the error rate. Accordingly, when the concepts are stationary, the distance between errors tends to increase and, when it decreases, warnings and drifts are triggered. Its authors claim that EDDM is more suitable than DDM for detecting gradual concept drifts while DDM is better suited for abrupt concept drifts.

Both methods use parametrized threshold values for the detection of the warning and drift levels and their default values represent 95% and 99% confidence intervals, respectively.

The parameters of DDM are thresholds for warnings and drifts, α_w and α_d, respectively, and the minimum number of instances n before the detection of concept drifts is permitted. The default values chosen by Gama et al. [22] for these parameters are 2.0, 3.0, and 30, respectively.

In EDDM, the parameters and respective default values are the thresholds, $w = 0.95$ and $d = 0.9$, and the minimum number of errors before drift detections are permitted, $e = 30$.

STEPD uses a statistical test of equal proportions, with continuity correction, calculated over two windows of the processed data, named *recent* and *older*. The accuracies of the base learner over these two windows are expected to be the same within each concept. Warnings and drifts are signaled when a significant difference is detected in the accuracy of the *recent* window. The parameters of STEPD with their respective default values are the *recent* window size ($w = 30$) and the significance levels for detecting drifts ($\alpha_d = 0.003$) and warnings ($\alpha_w = 0.05$).

ADWIN [9] uses a sliding window of instances (W) with a variable size. When drifts are detected, the size of W is reduced and the longer the concept the larger the size of W. Two dynamically adjusted sub-windows are stored, representing older and recent data. Drifts are detected when the difference of the means of these sub-windows is higher than a given threshold. The parameters of ADWIN are a confidence level to reduce the window size – $\delta \in (0, 1)$ – and the minimum frequency of instances needed for the window size to be reduced – f. The default values of ADWIN in its implementation in the MOA framework are $\delta = 0.002$ and $f = 32$.

PHT [39] is a sequential analysis technique which is also used for concept drift detection. It computes the observed values (the actual accuracy of the base learner) and their mean up to the processed instance. When a drift occurs, the classifier starts to fail to correctly classify new instances, making the current and the mean accuracy decrease. The cumulative (U_T) and the minimum (m_T) differences between these two values are computed. Higher U_T values mean the observed values differ considerably from their previous ones. When $U_T - m_T$ is above a specified threshold (λ), the permitted magnitude of changes, a drift is detected. Higher values of λ result in fewer false positives but more false negatives and may delay some detections. An extended version of the test was also recently proposed [49].

PL [2] uses two learners. The *stable* learner (S), which uses all known instances for training, and the *reactive* learner (R), that only trains on the last W instances, which is a parameter. The number of instances incorrectly classified by S but correctly classified by R is kept updated and, if its proportion of W is greater then a parametrized percentage threshold θ, a drift is detected. After concept drifts are triggered, S is replaced by R and R is reset. The parameters of PL and their defaults in MOA are $W = 12$ and $\theta = 0.2$.

ECDD [45] was adapted from Exponentially Weighted Moving Average (EWMA) [44] to be used in data streams subject to concept drifts. EWMA detects significant changes in the mean of a sequence of random variables provided that the mean and standard deviation of the data are known in advance. However, in ECDD, the mean and standard deviation are not needed. The authors of ECDD defined three parameters but its MOA implementation has only two: the weights used to

differentiate recent from old instances (λ) and the minimum number of instances before the detection of drifts is permitted (n). The default parameter values of ECDD in the MOA framework is one of the configurations used by its authors: $\lambda = 0.2$ and $n = 30$.

The authors of SEQDRIFT [40] stated that it was proposed to improve on some deficiencies of ADWIN. It also uses two sub-windows to represent old and new data. In its newer version, SEQDRIFT2, an extended version of SEQDRIFT1 [46], the old data is managed by the use of reservoir sampling, a one-pass method to obtain a random sample of fixed size from a data pool whose size is not known in advance. This technique presents computational efficiency in maintaining and sampling the reservoir. SEQDRIFT2 also uses the Bernstein bound [8] to compare the sample means of both sub-windows and, according to its authors, it presents good results compared to other published bounds, especially in distributions with low variance. The proposed parameters and respective default values are the size of the pool ($b = 200$) and the drift level ($\delta = 0.01$).

SEED [26] also draws on ADWIN and compares two sub-windows within a window W. Whenever these two sub-windows of W exhibit distinct averages higher than a chosen threshold δ, the older portion of the window (W_L) is dropped. SEED uses the Hoeffding Inequality with Bonferroni correction, proposed in ADWIN, to calculate its test statistic and it also performs block compression to eliminate unnecessary cut points and merge blocks that are homogeneous in nature. The authors of SEED claim it is faster and more memory-efficient than ADWIN. The parameters of SEED and their respective default values in MOA are a block size ($b = 32$), a compress term ($c = 75$), the threshold ($\delta = 0.05$), the growth parameter that controls the magnitude of the increment in a linear fashion ($\alpha = 0.8$), and the base value for the linear function ($\epsilon = 0.01$).

The HDDM authors [18] propose to monitor the performance of the base learner by applying "some probability inequalities that assume only independent, uni-variate and bounded random variables to obtain theoretical guarantees for the detection of such distributional changes". This is different than DDM, EDDM, and ECDD, for example, which assume that measured values are given according to a Bernoulli distribution. HDDM also provides bounds on both false positive and false negative rates, whereas ECDD only focuses on the false positive rate. Two main approaches have been proposed. The authors claim [18] that the first (A_Test, HDDM$_A$) "involves moving averages and is more suitable to detect abrupt changes", and the second (W_Test, HDDM$_W$) "follows a widespread intuitive idea to deal with gradual changes using weighted moving averages". They have three common parameters: the confidence values for drifts (α_D =0.001) and warnings ($\alpha_W = 0.005$), and the direction of the error, which can be one-sided ($t = 0$, only increments), default for HDDM$_W$, or two-sided ($t = 1$, error increments and decrements), and default for HDDM$_A$. Finally, HDDM$_W$ has an extra parameter ($\lambda = 0.05$) that is used to control how much weight is given to more recent data in comparison to older data.

The algorithm of FHDDM [42] uses a sliding window and Hoeffding's inequality [25] to calculate and compare the maximum (best) probability of the correct predictions observed so far with the most recent probability of correct predictions for the purpose of drift detection. The authors of FHDDM claim that their algorithm results in less detection delay, fewer false positives and fewer false negatives when compared to state-of-the-art detectors. The parameters of FHDDM and their provided default values are the size of the sliding window ($n = 25$) and the probability of error allowed ($\delta = 0.000001$).

FTDD [12,13] is one of three concept drift detection methods based on an efficient implementation of Fisher's Exact test [17]. It draws on STEPD [38] and on the deficiency of its statistical test of equal proportions in situations where the data samples are small or imbalanced. This particular method detects drifts using Fisher's Exact test instead of the test of equal proportions in all situations. Its two sibling methods adopt hybrid applications of Fisher's Exact test. Both Fisher Proportions Drift Detector (FPDD) and Fisher Squared Drift Detector (FSDD) use Fisher's Exact test only in situations in which either the number of errors or correct predictions in either of the two windows of STEPD, also adopted in all these three methods, is small. Otherwise, FPDD uses the test of equal proportions, just like STEPD, and FSDD adopts the chi-squared statistical test for homogeneity of proportions [11]. The three methods have the same three parameters of STEPD: recent window size $w = 30$ and significance levels $\alpha_d = 0.003$ and $\alpha_w = 0.05$.

RDDM [4,5] was proposed to overcome/alleviate a performance loss problem of DDM when concepts are very large, caused by decreased sensitivity, requiring too many instances to detect the changes. RDDM adds an explicit mechanism to discard older instances of very long concepts, periodically recalculating the RDDM statistics responsible for detecting the warning and drift levels. In addition, it forces concept drifts when the number of instances of the warning period reaches a parametrized threshold. The authors claim RDDM delivers higher global accuracy than DDM in most situations, especially in gradual concept drift datasets, by detecting more drifts and detecting them earlier, despite a small increase in false positives and memory consumption.

WSTD [4,6] also detects concept drifts in data streams using two windows of data, similarly to STEPD. More specifically, WSTD detects drifts based on an efficient implementation of the Wilcoxon rank sum statistical test [50], instead of the test of equal proportions used in STEPD. The authors claim that WSTD detects fewer false positives than STEPD and is particularly efficient in detecting abrupt drifts. In addition to the three parameters of STEPD with the same default values, WSTD limits the size of the *older* window using an extra parameter ($w_2 = 4{,}000$).

To conclude this section, Table 1 shows the number of parameters of each method.

3. Experimental setting

This section provides all the relevant information on the experiments reported in this article. All the concept drift detection methods have been tested with both Naive Bayes (NB) and Hoeffding Tree (HT) as base learners because they are

Table 1
Number of parameters of the methods.

DDM	EDDM	ADWIN	PHT	PL	ECDD	STEPD	SeqDr2	SEED	$HDDM_A$	$HDDM_W$	FHDDM	FTDD	RDDM	WSTD
3	3	2	4	2	2	3	2	5	3	4	2	3	6	4

the most frequently used classifiers in experiments in the area and their implementations are available within the MOA framework.

The tested detection methods are DDM, EDDM, ADWIN, ECDD, STEPD, SEQDRIFT2, SEED, $HDDM_A$, $HDDM_W$, FHDDM, FTDD, RDDM, and WSTD, all previously described in Section 2. In the case of RDDM, two different sets of values were used on its first three parameters. The first version, referred to as $RDDM_{30}$, uses the configuration tested in [5], following the default parameters of DDM ($n = 30,\ \alpha_w = 2,$ and $\alpha_d = 3$), whereas the second version, named RDDM, uses its proposed default parameter values: $n = 129,\ \alpha_w = 1.773,$ and $\alpha_d = 2.258$.

It is worth pointing out that, in the experiments reported in [4], three versions of both DDM and RDDM were tested to allow a fairer comparison of these two methods. In addition to these two sets of parameter values, both methods have also been tested with the configuration proposed as default for BOLE [7], i.e. $n = 7,\ \alpha_w = 1.2,$ and $\alpha_d = 1.95$. The reason why they were not included here was to make room for the inclusion of SEED and FHDDM.

Six artificial dataset generators were chosen to build abrupt and gradual concept drift datasets of seven different sizes, i.e., there are seven sizes of each generator, with 10K, 20K, 50K, 100K, 500K, 1 million, and 2 million instances, respectively. In the experiments reported in Chapter 6 of [4], the 1 million and 2 million instances datasets were tested using only NB.

In all generated datasets, four concept drifts were distributed at regular intervals and, thus, the size of the concepts in each dataset version of the same generator is different, covering different scenarios. For instance, in the 10K instances datasets, the four concept drifts are in positions 2K, 4K, 6K, and 8K.

Concept drifts were simulated by joining different concepts. In all the gradual concept drift datasets the changes lasted for 500 instances and were generated by a probability function, available in the MOA framework. This means that, in the 250 instances before and the 250 instances after the specified points, the class labels are progressively less likely to be based on the older concept and more likely to be based on the newer concept. At the change point, the probability should be 50% for each concept.

The experiments using the datasets with up to 100K instances were executed 30 times to calculate the accuracies of the methods and the mean results were computed with 95% confidence intervals. In the larger datasets, there were only 10 repetitions.

Finally, the accuracy evaluation adopted the Prequential methodology [14] with a sliding window of size 1,000 as its forgetting mechanism, the default in MOA. In this methodology, each incoming instance is used initially for testing and subsequently for training, and the accuracy is based on the cumulative sum of the sequential errors over time, i.e., of the loss function between the predictions and the observed values. In its sliding windows variation, the portion of the data considered in the calculation in each instance is the previous 1,000 instances.

3.1. Datasets

The dataset generators that were chosen for the experiments reported in Section 4 have all been previously used in the area and are all publicly available, most of them in the MOA framework, from the MOA website, or at https://sites.google.com/site/moaextensions.

The specific dataset generators selected for these experiments are Agrawal, LED, Mixed, Random RBF, Sine, and Waveform. In the case of Agrawal, it was used twice: $Agrawal_1$ uses its first five functions (F1 to F5) and $Agrawal_2$ uses its remaining functions (F6 to F10), as these provide very different datasets.

The Agrawal generator [1,33] stores information from people willing to receive a loan of a given amount. From this data, they are classified as belonging to group A or group B. The attributes are: salary, commission, age, education level, make of their car, zip code, value of the house, number of years house is owned, and the total amount of the loan. To perform the classification, the authors proposed ten functions, each with different forms of evaluation. In addition, it is possible to add noise.

The LED generator [18,23,48] represents the problem of predicting the digit shown by a seven-segment LED display. It has 24 categorical attributes, 17 of which are irrelevant, and a categorical class, with ten possible values. Also, each attribute has a 10% probability of being inverted (noise). Concept drifts are simulated by changing the position of the relevant attributes.

The Mixed generator [3,22,23] has two boolean attributes (v and w) and two numeric ones (x and y). Each instance can be classified as positive or negative. They are positive if at least two of the three following conditions are met: $v, w,$ and $y < 0.5 + 0.3\sin(3\pi x)$. To simulate a concept drift, the labels of the aforementioned conditions are reversed.

The RandomRBF generator [33,40] uses n centroids with their centers, labels, and weights randomly defined, and a Gaussian distribution to determine the values of m attributes. The chosen centroid also determines the class label of the example. This effectively creates a normally distributed hyper-sphere of examples surrounding each central point with varying densities and this represents a problem very hard to learn. A concept drift is simulated by changing the positions of the centroids. This dataset generator was used with six classes, 40 attributes, and 50 centroids.

The Sine generator [3,22,45] has two numeric attributes (x and y) and two contexts (Sine1 and Sine2). In the former, a given instance will be classified as positive if the point (x, y) is below the curve $y = sin(x)$. In the latter, the condition $y < 0.5 + 0.3\sin(3\pi x)$ must be satisfied. Concept drifts can be simulated either by alternating between Sine1 and Sine2 or by reversing the aforementioned conditions, i.e. points below the curves become negative.

The Waveform generator [40,47] has three classes and 40 numerical attributes, with the last 19 used to produce noise. The goal of the problem is to detect the waveform generated by combining two of three base waves. To perform changes, the positions of the attributes representing a certain context are reversed.

4. Accuracy results and analysis

This section introduces the accuracy results of all the configurations of the concept drift detection methods tested and examines them, including several statistical evaluations, to thoroughly answer **RQ1**.

Tables 2 and 3 present the accuracy results of the methods (split into two parts), as well as their ranks in the *abrupt* datasets using NB as base learner, also including the ranks considering all datasets (for completeness). In each dataset and in the ranks, the best result is shown in **bold**.

Similarly, Tables 4 and 5 present the corresponding accuracy results of the methods (also split into two parts) and their ranks in the datasets configured with *gradual* concept drifts using NB as base learner, again including the ranks considering all datasets (for completeness). In each dataset and in the ranks, the best result is also shown in **bold**.

Tables 6 and 7 are similar to Tables 2 and 3, respectively, but refer to the results in the *abrupt* datasets using HT as base learner, instead of NB.

Likewise, Tables 8 and 9 are very similar to Tables 4 and 5, respectively, but refer to the results in the *gradual* datasets using HT as base learner, rather than NB.

Complementing the analysis of the results, different views of the accuracy results in Tables 2 to 9 were compared using the F_F statistic [15]. Note that the null hypothesis states that all methods are statistically equal but, when rejected, it is necessary to use a post-hoc test to discover in what method(s) there is a statistical difference. We used the Nemenyi post-hoc test to compare all the methods against all the others. The results are presented using graphics in which the critical difference (CD) is represented by bars and methods connected by a bar are *not* statistically different.

Fig. 1 presents the evaluation of the concept drift detection methods based on the results of the experiments in the *abrupt* datasets using NB, i.e., those presented in Tables 2 and 3. According to the ranks, RDDM, HDDM_A, WSTD, and FTDD are the best configurations in this subset of the tests, with no statistical differences between them or the next three methods (FHDDM, HDDM_W, and SEQDRIFT_2), despite the comparatively worse ranks of the latter four. Also, notice that, in spite of this, only RDDM is statistically better than ADWIN and only RDDM, HDDM_A, and WSTD are different from the following two configurations, RDDM_{30} and SEED.

Similarly, Fig. 2 presents the corresponding evaluation based on the results of the *gradual* datasets using NB, i.e., those presented in Tables 4 and 5. In these datasets, the best results were those of RDDM, HDDM_A, and FHDDM, with no statistical differences between them or the next two methods, SEQDRIFT_2 and RDDM_{30}. However, in this scenario, only RDDM is statistically superior to the following three methods: HDDM_W, WSTD, and FTDD. In addition, only RDDM and HDDM_A are better than ADWIN and SEED.

Fig. 3 evaluates the accuracy results of the methods aggregating *all* the tests executed using NB as base learner. With this broader view of the data, the best methods are RDDM and HDDM_A, though WSTD, FHDDM, and FTDD were also statistically similar to them. In this case, the statistical differences between the best two methods and the others are the same, with both being statistically superior to the remaining nine configurations. However, the other three aforementioned methods are *not* statistically different to HDDM_W, SEQDRIFT_2, and RDDM_{30}.

Figs. 4–6 represent the evaluations based on views similar to those of Figs. 1–3, respectively, but based on the tests using HT as base classifier. Fig. 4 refers to the results of the experiments in the *abrupt* datasets, i.e., those presented in

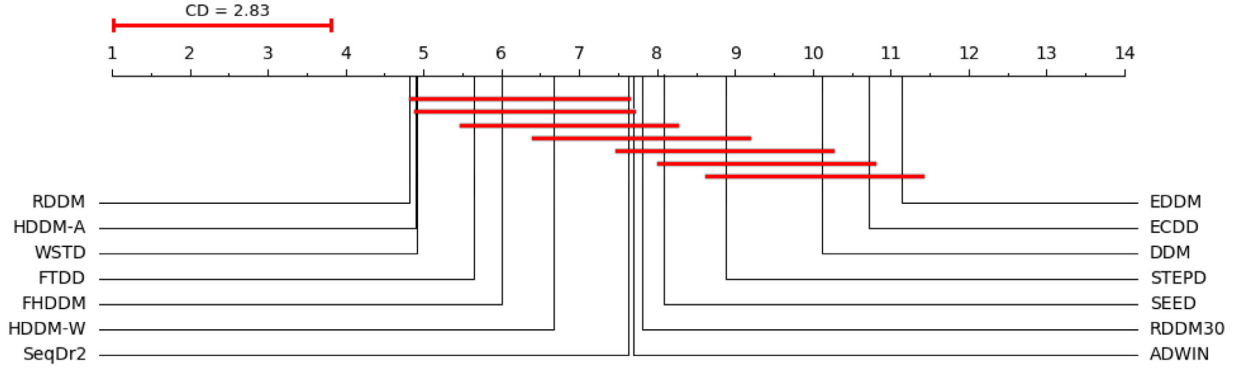

Fig. 1. Comparison results using the Nemenyi test of Detectors with NB in the abrupt datasets with a 95% confidence interval.

Table 2
Mean accuracies of Drift Detectors in percentage (%) in the abrupt datasets, with a 95% confidence interval, using NB (Part 1).

DS Type and Size	DATASET	DDM $HDDM_A$	EDDM $HDDM_W$	ADWIN FHDDM	ECDD FTDD	STEPD $RDDM_{30}$	SeqDrift2 RDDM	SEED WSTD
	AGRAW$_1$	61.56 (±0.52)	60.49 (±0.38)	61.52 (±0.30)	61.81 (±0.31)	62.72 (±0.35)	61.21 (±0.37)	62.25 (±0.25)
		63.17 (±0.32)	63.14 (±0.26)	63.33 (±0.29)	60.85 (±0.29)	62.54 (±0.28)	**63.56 (±0.26)**	62.07 (±0.36)
	AGRAW$_2$	72.68 (±1.59)	70.83 (±1.94)	78.36 (±0.31)	80.99 (±1.15)	81.43 (±0.47)	77.61 (±0.36)	80.08 (±0.42)
		80.11 (±0.66)	**81.99 (±0.34)**	81.48 (±0.50)	79.15 (±0.67)	73.73 (±1.18)	79.63 (±0.96)	80.69 (±0.52)
	LED	69.57 (±0.30)	67.52 (±0.40)	62.40 (±0.46)	67.48 (±0.41)	61.03 (±1.92)	58.87 (±0.97)	55.40 (±0.62)
		69.72 (±0.29)	69.28 (±0.43)	69.32 (±0.37)	67.20 (±0.75)	69.54 (±0.29)	**69.80 (±0.29)**	67.60 (±0.80)
ABRUPT	**MIXED**	89.74 (±0.29)	88.78 (±0.33)	88.82 (±0.26)	89.06 (±0.29)	90.40 (±0.28)	83.31 (±0.20)	89.14 (±0.23)
10K		90.39 (±0.21)	90.39 (±0.22)	90.20 (±0.21)	90.39 (±0.22)	89.87 (±0.23)	90.22 (±0.23)	**90.41 (±0.22)**
	RandRBF	30.87 (±0.59)	30.40 (±0.45)	30.40 (±0.50)	30.87 (±0.64)	30.01 (±0.51)	30.91 (±0.53)	30.23 (±0.53)
		30.56 (±0.43)	29.51 (±0.43)	29.98 (±0.47)	**31.08 (±0.53)**	30.77 (±0.49)	30.53 (±0.43)	30.70 (±0.56)
	SINE	85.10 (±0.69)	85.35 (±0.42)	85.59 (±0.23)	86.23 (±0.22)	**86.78 (±0.21)**	80.92 (±0.21)	85.73 (±0.19)
		86.62 (±0.21)	86.77 (±0.22)	86.67 (±0.22)	86.75 (±0.23)	86.03 (±0.24)	86.58 (±0.24)	86.76 (±0.22)
	WAVEF.	78.49 (±0.45)	78.53 (±0.43)	78.37 (±0.39)	78.33 (±0.39)	**79.25 (±0.42)**	78.41 (±0.41)	78.66 (±0.43)
		78.73 (±0.48)	79.18 (±0.44)	79.02 (±0.46)	78.06 (±0.61)	78.56 (±0.42)	79.12 (±0.47)	78.79 (±0.51)
	AGRAW$_1$	63.08 (±0.59)	61.73 (±0.32)	64.09 (±0.17)	62.37 (±0.15)	64.38 (±0.18)	63.70 (±0.34)	64.17 (±0.18)
		64.82 (±0.17)	64.16 (±0.19)	64.41 (±0.17)	62.02 (±0.35)	64.32 (±0.18)	**64.89 (±0.15)**	64.48 (±0.27)
	AGRAW$_2$	79.00 (±0.87)	71.11 (±1.52)	81.23 (±0.35)	83.30 (±0.18)	83.82 (±0.37)	80.88 (±0.34)	83.31 (±0.25)
		83.00 (±0.50)	**84.13 (±0.19)**	84.05 (±0.22)	81.90 (±0.41)	79.50 (±0.82)	83.18 (±0.56)	83.41 (±0.40)
	LED	71.32 (±0.25)	69.17 (±0.27)	63.42 (±0.57)	68.15 (±0.42)	65.79 (±1.63)	60.95 (±1.04)	56.24 (±0.73)
		71.52 (±0.18)	70.79 (±0.40)	71.28 (±0.31)	70.55 (±0.47)	71.39 (±0.18)	**71.74 (±0.16)**	70.60 (±0.44)
ABRUPT	**MIXED**	90.26 (±0.67)	89.79 (±0.19)	90.46 (±0.12)	89.41 (±0.20)	90.95 (±0.19)	87.62 (±0.14)	90.27 (±0.16)
20K		91.10 (±0.12)	**91.19 (±0.13)**	91.11 (±0.13)	91.18 (±0.13)	90.78 (±0.14)	91.03 (±0.15)	**91.19 (±0.13)**
	RandRBF	30.79 (±0.47)	30.92 (±0.50)	30.46 (±0.35)	**31.25 (±0.62)**	29.65 (±0.39)	31.14 (±0.48)	29.94 (±0.46)
		30.69 (±0.41)	29.32 (±0.32)	29.76 (±0.37)	31.15 (±0.46)	30.76 (±0.42)	30.50 (±0.41)	30.70 (±0.57)
	SINE	83.67 (±1.77)	85.60 (±0.60)	86.66 (±0.17)	86.42 (±0.16)	87.18 (±0.16)	84.26 (±0.17)	86.56 (±0.17)
		87.08 (±0.18)	**87.22 (±0.18)**	87.17 (±0.18)	87.21 (±0.19)	86.51 (±0.23)	87.02 (±0.19)	87.21 (±0.18)
	WAVEF.	78.98 (±0.29)	78.86 (±0.28)	79.28 (±0.29)	78.67 (±0.25)	79.74 (±0.22)	79.44 (±0.27)	79.69 (±0.26)
		79.60 (±0.26)	**79.85 (±0.24)**	79.69 (±0.28)	79.12 (±0.44)	79.32 (±0.30)	79.78 (±0.29)	79.71 (±0.28)
	AGRAW$_1$	63.64 (±0.63)	62.81 (±0.24)	65.51 (±0.13)	62.80 (±0.13)	65.12 (±0.15)	65.55 (±0.10)	65.35 (±0.13)
		65.67 (±0.16)	64.61 (±0.11)	65.03 (±0.14)	63.55 (±0.51)	65.36 (±0.17)	**65.73 (±0.11)**	65.57 (±0.14)
	AGRAW$_2$	82.40 (±1.16)	70.56 (±0.86)	84.97 (±0.20)	84.27 (±0.09)	85.33 (±0.19)	83.32 (±0.39)	85.54 (±0.13)
		85.34 (±0.22)	85.60 (±0.10)	**85.63 (±0.15)**	83.86 (±0.50)	84.24 (±0.43)	85.38 (±0.20)	85.30 (±0.25)
	LED	71.66 (±0.71)	70.06 (±0.17)	64.76 (±0.53)	68.73 (±0.31)	68.73 (±0.79)	65.00 (±1.41)	56.69 (±0.52)
		72.81 (±0.16)	71.63 (±0.27)	72.12 (±0.29)	72.23 (±0.21)	72.67 (±0.15)	**72.89 (±0.15)**	72.10 (±0.33)
ABRUPT	**MIXED**	90.85 (±0.96)	90.07 (±0.59)	91.43 (±0.11)	89.82 (±0.14)	91.43 (±0.14)	90.29 (±0.10)	91.38 (±0.12)
50K		91.63 (±0.11)	91.72 (±0.10)	91.68 (±0.10)	91.72 (±0.10)	91.41 (±0.11)	91.57 (±0.10)	**91.73 (±0.10)**
	RandRBF	31.06 (±0.50)	31.14 (±0.42)	30.58 (±0.34)	**31.32 (±0.65)**	29.28 (±0.30)	31.24 (±0.51)	29.87 (±0.27)
		30.91 (±0.40)	29.19 (±0.31)	29.58 (±0.34)	31.03 (±0.49)	30.95 (±0.39)	30.73 (±0.36)	30.39 (±0.54)
	SINE	84.21 (±1.32)	85.46 (±0.66)	87.14 (±0.12)	86.44 (±0.11)	87.27 (±0.12)	86.19 (±0.12)	87.09 (±0.11)
		87.26 (±0.10)	**87.40 (±0.12)**	87.39 (±0.12)	**87.40 (±0.12)**	86.79 (±0.21)	87.22 (±0.13)	**87.40 (±0.11)**
	WAVEF.	79.60 (±0.18)	79.21 (±0.16)	80.12 (±0.13)	79.02 (±0.16)	80.06 (±0.13)	80.08 (±0.13)	80.14 (±0.13)
		80.13 (±0.15)	80.15 (±0.14)	80.18 (±0.13)	79.92 (±0.25)	79.93 (±0.17)	80.16 (±0.14)	**80.21 (±0.13)**
	AGRAW$_1$	64.17 (±0.68)	63.31 (±0.21)	66.00 (±0.08)	62.89 (±0.08)	65.40 (±0.08)	66.06 (±0.08)	65.91 (±0.11)
		66.06 (±0.08)	64.81 (±0.09)	65.27 (±0.10)	65.04 (±0.47)	65.73 (±0.17)	**66.08 (±0.08)**	65.96 (±0.11)
	AGRAW$_2$	84.29 (±0.62)	70.20 (±0.37)	86.05 (±0.05)	84.49 (±0.07)	85.84 (±0.09)	84.49 (±0.45)	**86.23 (±0.07)**
		86.14 (±0.09)	86.09 (±0.07)	86.14 (±0.06)	84.60 (±0.47)	85.26 (±0.34)	86.13 (±0.04)	85.84 (±0.31)
	LED	72.54 (±0.40)	70.45 (±0.17)	65.21 (±0.53)	69.02 (±0.22)	69.46 (±0.60)	67.82 (±1.18)	57.51 (±0.74)
		73.37 (±0.11)	71.99 (±0.23)	72.53 (±0.23)	72.94 (±0.19)	73.23 (±0.12)	**73.39 (±0.12)**	72.85 (±0.20)
ABRUPT	**MIXED**	90.70 (±1.17)	90.02 (±1.02)	91.75 (±0.06)	89.81 (±0.09)	91.54 (±0.08)	91.19 (±0.06)	91.69 (±0.07)
100K		91.81 (±0.07)	**91.90 (±0.06)**	91.88 (±0.06)	**91.90 (±0.06)**	91.67 (±0.06)	91.78 (±0.06)	**91.90 (±0.06)**
	RandRBF	31.38 (±0.42)	31.49 (±0.38)	30.59 (±0.30)	31.51 (±0.58)	29.12 (±0.18)	31.41 (±0.42)	30.02 (±0.28)
		31.13 (±0.34)	29.30 (±0.21)	29.53 (±0.25)	**31.65 (±0.45)**	31.24 (±0.35)	31.16 (±0.28)	30.69 (±0.44)
	SINE	83.77 (±1.40)	85.75 (±0.52)	87.28 (±0.08)	86.45 (±0.10)	87.30 (±0.08)	86.82 (±0.08)	87.24 (±0.08)
		87.27 (±0.10)	**87.43 (±0.09)**	87.42 (±0.09)	**87.43 (±0.09)**	86.85 (±0.20)	87.31 (±0.10)	**87.43 (±0.09)**
	WAVEF.	79.67 (±0.22)	79.36 (±0.21)	80.27 (±0.10)	79.13 (±0.13)	80.21 (±0.10)	80.27 (±0.10)	80.26 (±0.11)
		80.27 (±0.11)	80.29 (±0.12)	80.31 (±0.11)	80.23 (±0.18)	80.05 (±0.14)	80.25 (±0.11)	**80.33 (±0.10)**

Tables 6 and 7. In this subset of the tests, seven different configurations are statistically similar: $HDDM_A$, FTDD, FHDDM, RDDM, $HDDM_W$, WSTD, and $RDDM_{30}$. Despite this, only $HDDM_A$ is statistically better than the remaining seven methods. For instance, the methods with the following three best ranks, FTDD, FHDDM, and RDDM, were *not* superior to DDM in this subset of the tests.

Accordingly, Fig. 5 corresponds to the evaluation of the results of the experiments in the *gradual* datasets using HT, i.e., those presented in Tables 8 and 9. In these datasets, nine methods presented statistically similar results: $HDDM_A$, RDDM, FHDDM, $RDDM_{30}$, $HDDM_W$, DDM, FTDD, WSTD, and $SEQDRIFT_2$. However, analogously to the other previously discussed

Table 3
Mean accuracies of Drift Detectors in percentage (%) in the abrupt datasets, with a 95% confidence interval, using NB (Part 2).

DS Type and Size	DATASET	DDM HDDM$_A$	EDDM HDDM$_W$	ADWIN FHDDM	ECDD FTDD	STEPD RDDM$_{30}$	SeqDrift2 RDDM	SEED WSTD
	AGRAW$_1$	64.72 (±0.74)	63.73 (±0.16)	66.39 (±0.05)	62.96 (±0.06)	65.58 (±0.08)	**66.44 (±0.04)**	66.28 (±0.04)
		66.40 (±0.05)	64.88 (±0.08)	65.35 (±0.06)	66.32 (±0.07)	66.11 (±0.22)	66.39 (±0.06)	66.23 (±0.06)
	AGRAW$_2$	85.89 (±0.90)	70.43 (±0.04)	86.84 (±0.03)	84.75 (±0.05)	86.28 (±0.09)	86.54 (±0.47)	**86.85 (±0.03)**
		86.83 (±0.06)	86.56 (±0.04)	86.64 (±0.03)	86.17 (±0.64)	86.59 (±0.18)	86.78 (±0.06)	86.74 (±0.11)
	LED	72.63 (±0.60)	70.79 (±0.16)	67.61 (±1.10)	69.18 (±0.14)	70.03 (±0.29)	72.79 (±0.38)	59.21 (±1.54)
		73.77 (±0.11)	72.39 (±0.26)	72.93 (±0.13)	73.49 (±0.28)	73.59 (±0.12)	73.75 (±0.08)	73.45 (±0.10)
ABRUPT	**MIXED**	91.21 (±1.20)	90.68 (±0.10)	92.04 (±0.03)	89.94 (±0.07)	91.64 (±0.05)	91.93 (±0.04)	92.00 (±0.03)
500		92.02 (±0.05)	**92.07 (±0.03)**	92.06 (±0.03)	**92.07 (±0.03)**	91.97 (±0.05)	92.01 (±0.03)	**92.07 (±0.03)**
	RandRBF	**33.42 (±0.35)**	33.36 (±0.36)	30.78 (±0.28)	33.26 (±0.66)	29.07 (±0.12)	33.15 (±0.44)	30.38 (±0.31)
		32.54 (±0.29)	29.40 (±0.09)	29.93 (±0.18)	33.12 (±0.31)	32.49 (±0.25)	32.13 (±0.26)	31.00 (±0.29)
	SINE	79.19 (±4.28)	83.63 (±2.66)	87.36 (±0.06)	86.45 (±0.05)	87.35 (±0.05)	87.28 (±0.05)	87.36 (±0.06)
		87.33 (±0.07)	87.39 (±0.06)	87.40 (±0.06)	**87.41 (±0.06)**	87.21 (±0.10)	87.40 (±0.06)	87.40 (±0.06)
	WAVEF.	79.80 (±0.30)	79.23 (±0.33)	**80.39 (±0.12)**	79.19 (±0.10)	80.26 (±0.11)	**80.39 (±0.12)**	80.38 (±0.11)
		80.38 (±0.12)	80.34 (±0.11)	**80.39 (±0.11)**	**80.39 (±0.11)**	80.07 (±0.20)	80.37 (±0.11)	80.38 (±0.11)
	AGRAW$_1$	64.28 (±1.19)	63.62 (±0.17)	66.49 (±0.04)	62.98 (±0.05)	65.67 (±0.05)	**66.51 (±0.04)**	66.41 (±0.04)
		66.46 (±0.05)	64.95 (±0.06)	65.39 (±0.05)	66.45 (±0.07)	66.35 (±0.05)	66.49 (±0.05)	66.30 (±0.05)
	AGRAW$_2$	85.99 (±1.15)	70.44 (±0.03)	**86.95 (±0.02)**	84.77 (±0.04)	86.34 (±0.05)	86.89 (±0.06)	**86.95 (±0.03)**
		86.91 (±0.03)	86.61 (±0.02)	86.70 (±0.02)	86.70 (±0.27)	86.64 (±0.27)	86.86 (±0.02)	86.89 (±0.05)
	LED	72.95 (±0.35)	70.85 (±0.16)	68.49 (±1.32)	69.25 (±0.12)	70.16 (±0.30)	73.34 (±0.20)	61.36 (±1.76)
		73.84 (±0.06)	72.47 (±0.14)	72.99 (±0.11)	73.49 (±0.27)	73.70 (±0.05)	73.82 (±0.06)	73.52 (±0.10)
ABRUPT	**MIXED**	90.11 (±3.29)	89.60 (±1.98)	92.09 (±0.03)	89.97 (±0.05)	91.67 (±0.03)	92.03 (±0.03)	92.06 (±0.03)
1M		92.08 (±0.03)	**92.10 (±0.03)**	**92.10 (±0.03)**	**92.10 (±0.03)**	92.03 (±0.04)	92.04 (±0.04)	**92.10 (±0.03)**
	RandRBF	33.40 (±0.23)	**33.51 (±0.21)**	30.74 (±0.13)	33.16 (±0.44)	29.03 (±0.08)	33.49 (±0.28)	30.73 (±0.35)
		32.93 (±0.21)	29.32 (±0.07)	29.87 (±0.18)	33.27 (±0.21)	32.55 (±0.19)	32.16 (±0.13)	31.07 (±0.23)
	SINE	81.76 (±4.29)	83.47 (±2.59)	87.42 (±0.05)	86.48 (±0.03)	87.36 (±0.04)	87.38 (±0.05)	87.40 (±0.05)
		87.38 (±0.07)	87.43 (±0.05)	87.44 (±0.05)	**87.45 (±0.05)**	87.32 (±0.07)	87.44 (±0.04)	87.44 (±0.05)
	WAVEF.	79.85 (±0.36)	79.23 (±0.31)	80.43 (±0.07)	79.20 (±0.08)	80.27 (±0.07)	**80.44 (±0.07)**	80.42 (±0.07)
		80.41 (±0.09)	80.38 (±0.08)	80.42 (±0.08)	80.40 (±0.10)	80.35 (±0.08)	80.41 (±0.07)	80.40 (±0.06)
	AGRAW$_1$	64.12 (±0.99)	63.37 (±0.59)	66.55 (±0.04)	62.98 (±0.03)	65.66 (±0.03)	**66.56 (±0.04)**	66.46 (±0.04)
		66.49 (±0.05)	64.95 (±0.04)	65.40 (±0.02)	66.53 (±0.04)	66.44 (±0.04)	66.52 (±0.03)	66.31 (±0.02)
	AGRAW$_2$	84.94 (±1.40)	70.45 (±0.03)	**87.01 (±0.02)**	84.78 (±0.03)	86.34 (±0.03)	86.98 (±0.03)	86.99 (±0.02)
		86.97 (±0.02)	86.63 (±0.01)	86.72 (±0.02)	86.97 (±0.04)	86.79 (±0.09)	86.90 (±0.02)	86.94 (±0.03)
	LED	72.78 (±0.72)	70.95 (±0.12)	70.54 (±0.98)	69.31 (±0.08)	70.24 (±0.23)	73.60 (±0.20)	62.94 (±1.69)
		73.89 (±0.05)	72.52 (±0.07)	73.07 (±0.09)	73.77 (±0.19)	73.78 (±0.09)	73.87 (±0.04)	73.64 (±0.06)
ABRUPT	**MIXED**	89.91 (±2.38)	89.96 (±1.62)	92.06 (±0.03)	89.95 (±0.04)	91.64 (±0.03)	92.03 (±0.02)	92.04 (±0.03)
2M		92.03 (±0.02)	**92.07 (±0.02)**	**92.07 (±0.02)**	**92.07 (±0.02)**	92.00 (±0.03)	92.01 (±0.03)	**92.07 (±0.02)**
	RandRBF	33.58 (±0.28)	**33.86 (±0.15)**	30.98 (±0.25)	33.07 (±0.19)	29.00 (±0.09)	33.27 (±0.18)	30.76 (±0.28)
		33.02 (±0.20)	29.31 (±0.07)	29.84 (±0.10)	33.23 (±0.14)	32.67 (±0.18)	32.13 (±0.13)	31.16 (±0.16)
	SINE	77.48 (±6.00)	84.62 (±2.50)	87.44 (±0.03)	86.47 (±0.02)	87.36 (±0.03)	87.41 (±0.03)	87.42 (±0.02)
		87.41 (±0.02)	87.44 (±0.02)	87.44 (±0.03)	87.44 (±0.03)	87.36 (±0.02)	**87.47 (±0.03)**	87.44 (±0.02)
	WAVEF.	79.55 (±0.29)	79.23 (±0.29)	80.46 (±0.03)	79.21 (±0.04)	80.31 (±0.04)	**80.48 (±0.03)**	80.46 (±0.03)
		80.47 (±0.04)	80.43 (±0.04)	80.45 (±0.04)	80.47 (±0.04)	80.39 (±0.04)	80.46 (±0.03)	80.46 (±0.04)
NB	**RANK**	10.11220	11.13270	7.65306	10.72450	8.87755	7.63265	8.09184
ABRUPT		4.89796	6.67347	6.01020	5.64286	7.81633	**4.81633**	4.91837
NB	**RANK**	9.65306	10.60710	7.95918	10.67860	9.34184	7.06122	8.18878
ALL		4.78061	6.80102	5.95408	6.36735	7.28061	**4.40306**	5.92347

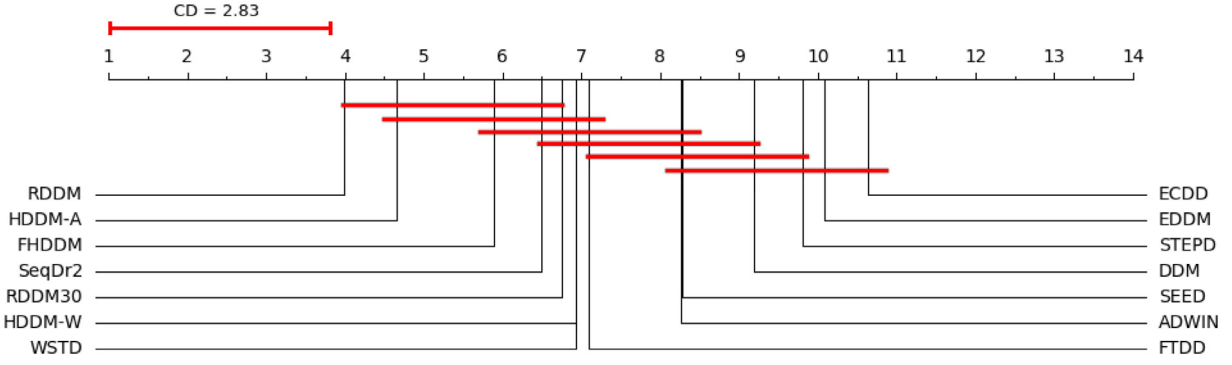

Fig. 2. Comparison results using the Nemenyi test of Detectors with NB in the gradual datasets with a 95% confidence interval.

Table 4
Mean accuracies of Drift Detectors in percentage (%) in the gradual datasets, with a 95% confidence interval, using NB (Part 1).

DS Type and Size	DATASET	DDM HDDM$_A$	EDDM HDDM$_W$	ADWIN FHDDM	ECDD FTDD	STEPD RDDM$_{30}$	SeqDrift2 RDDM	SEED WSTD
GRAD. 10K	**AGRAW$_1$**	60.56 (±0.38)	59.97 (±0.32)	60.74 (±0.18)	60.84 (±0.26)	61.28 (±0.30)	60.23 (±0.27)	61.04 (±0.23)
		61.25 (±0.34)	61.83 (±0.25)	61.53 (±0.30)	59.27 (±0.52)	60.20 (±0.35)	**62.05 (±0.27)**	60.80 (±0.30)
	AGRAW$_2$	69.38 (±1.23)	69.93 (±1.53)	74.45 (±0.48)	77.03 (±1.08)	76.90 (±0.89)	73.57 (±0.65)	75.74 (±0.47)
		76.12 (±0.85)	**77.89 (±0.78)**	77.33 (±0.83)	71.82 (±1.30)	69.29 (±1.32)	74.29 (±1.44)	75.61 (±1.08)
	LED	67.78 (±0.40)	66.69 (±0.37)	60.92 (±0.49)	65.16 (±0.40)	59.51 (±1.55)	58.41 (±0.71)	55.99 (±0.81)
		67.65 (±0.30)	66.72 (±0.36)	66.91 (±0.35)	63.11 (±0.90)	67.83 (±0.34)	**67.85 (±0.29)**	64.40 (±0.72)
	MIXED	83.65 (±0.28)	**84.16 (±0.25)**	83.04 (±0.25)	83.02 (±0.31)	83.28 (±0.32)	83.50 (±0.25)	82.72 (±0.22)
		83.61 (±0.27)	83.74 (±0.29)	84.13 (±0.27)	83.74 (±0.24)	83.88 (±0.27)	83.89 (±0.29)	83.42 (±0.27)
	RandRBF	30.81 (±0.61)	30.46 (±0.45)	30.50 (±0.43)	30.91 (±0.65)	29.78 (±0.51)	**30.92 (±0.52)**	30.12 (±0.52)
		30.55 (±0.47)	29.41 (±0.39)	29.95 (±0.45)	30.90 (±0.56)	30.89 (±0.52)	30.39 (±0.44)	30.73 (±0.61)
	SINE	81.32 (±0.27)	81.71 (±0.20)	80.90 (±0.20)	81.20 (±0.19)	80.73 (±0.24)	81.10 (±0.22)	80.36 (±0.23)
		81.51 (±0.20)	81.63 (±0.22)	**82.01 (±0.18)**	81.26 (±0.20)	81.78 (±0.22)	81.85 (±0.18)	81.32 (±0.21)
	WAVEF.	77.99 (±0.43)	78.25 (±0.37)	77.86 (±0.39)	78.06 (±0.37)	78.33 (±0.38)	77.73 (±0.41)	78.21 (±0.39)
		77.82 (±0.49)	78.24 (±0.39)	77.78 (±0.40)	76.65 (±0.46)	77.87 (±0.41)	**78.46 (±0.37)**	77.54 (±0.54)
GRAD. 20K	**AGRAW$_1$**	62.62 (±0.51)	61.90 (±0.34)	63.07 (±0.18)	61.85 (±0.13)	63.31 (±0.23)	62.25 (±0.38)	63.44 (±0.20)
		63.92 (±0.14)	63.26 (±0.16)	63.40 (±0.14)	61.14 (±0.35)	63.62 (±0.16)	**63.98 (±0.13)**	63.15 (±0.41)
	AGRAW$_2$	75.89 (±1.02)	70.81 (±1.58)	79.06 (±0.31)	81.02 (±0.92)	81.64 (±0.38)	79.33 (±0.25)	80.64 (±0.36)
		80.47 (±0.51)	**82.27 (±0.19)**	81.07 (±0.97)	78.79 (±0.60)	76.58 (±1.07)	80.79 (±0.91)	79.76 (±1.35)
	LED	70.54 (±0.19)	69.29 (±0.24)	62.53 (±0.52)	67.21 (±0.43)	64.11 (±1.41)	60.56 (±0.92)	55.85 (±0.59)
		70.43 (±0.18)	69.36 (±0.38)	69.81 (±0.34)	67.66 (±0.87)	70.61 (±0.18)	**70.66 (±0.18)**	68.68 (±0.52)
	MIXED	87.85 (±0.17)	87.98 (±0.18)	87.12 (±0.15)	86.84 (±0.19)	87.50 (±0.16)	87.83 (±0.14)	87.11 (±0.16)
		87.80 (±0.18)	87.99 (±0.18)	**88.19 (±0.18)**	87.63 (±0.16)	88.01 (±0.16)	88.01 (±0.18)	87.71 (±0.16)
	RandRBF	30.89 (±0.51)	30.77 (±0.45)	30.37 (±0.41)	**31.27 (±0.61)**	29.70 (±0.41)	31.06 (±0.47)	29.88 (±0.41)
		30.68 (±0.44)	29.31 (±0.33)	29.56 (±0.40)	31.26 (±0.45)	30.95 (±0.42)	30.53 (±0.47)	30.78 (±0.57)
	SINE	84.64 (±0.20)	84.73 (±0.17)	84.03 (±0.15)	84.17 (±0.15)	84.36 (±0.19)	84.38 (±0.19)	83.93 (±0.19)
		84.97 (±0.15)	84.92 (±0.16)	**85.07 (±0.16)**	84.74 (±0.17)	84.92 (±0.19)	84.98 (±0.15)	84.60 (±0.16)
	WAVEF.	78.46 (±0.29)	78.74 (±0.25)	78.62 (±0.22)	78.53 (±0.23)	**79.31 (±0.26)**	78.68 (±0.26)	79.10 (±0.31)
		78.90 (±0.30)	79.19 (±0.26)	78.78 (±0.24)	78.15 (±0.41)	78.76 (±0.25)	79.22 (±0.31)	78.73 (±0.28)
GRAD. 50K	**AGRAW$_1$**	63.92 (±0.57)	62.80 (±0.26)	65.21 (±0.13)	62.53 (±0.11)	64.77 (±0.14)	65.28 (±0.13)	65.07 (±0.13)
		65.43 (±0.11)	64.30 (±0.12)	64.68 (±0.15)	62.87 (±0.42)	65.13 (±0.17)	65.38 (±0.11)	65.17 (±0.14)
	AGRAW$_2$	82.41 (±0.96)	70.40 (±0.80)	83.99 (±0.30)	83.53 (±0.09)	84.31 (±0.21)	83.07 (±0.42)	84.65 (±0.22)
		84.57 (±0.31)	**84.82 (±0.10)**	84.81 (±0.12)	82.75 (±0.51)	83.75 (±0.46)	84.77 (±0.22)	83.90 (±0.92)
	LED	72.30 (±0.26)	70.25 (±0.18)	64.49 (±0.50)	68.35 (±0.33)	68.17 (±0.80)	64.62 (±1.26)	56.92 (±0.63)
		72.48 (±0.15)	71.17 (±0.27)	71.68 (±0.26)	71.62 (±0.19)	72.50 (±0.14)	**72.63 (±0.15)**	71.48 (±0.28)
	MIXED	90.42 (±0.11)	90.17 (±0.11)	90.05 (±0.10)	88.78 (±0.15)	90.12 (±0.11)	90.37 (±0.09)	90.14 (±0.12)
		90.45 (±0.11)	90.47 (±0.11)	**90.52 (±0.10)**	90.42 (±0.11)	90.48 (±0.11)	90.50 (±0.09)	90.40 (±0.10)
	RandRBF	30.94 (±0.49)	31.10 (±0.43)	30.59 (±0.35)	**31.31 (±0.66)**	29.32 (±0.30)	31.08 (±0.48)	29.97 (±0.31)
		30.92 (±0.37)	29.28 (±0.31)	29.61 (±0.35)	31.00 (±0.49)	30.94 (±0.41)	30.81 (±0.35)	30.43 (±0.53)
	SINE	86.31 (±0.26)	85.98 (±0.17)	86.06 (±0.09)	85.62 (±0.12)	86.24 (±0.12)	86.29 (±0.12)	86.11 (±0.11)
		86.76 (±0.10)	86.76 (±0.10)	86.76 (±0.11)	86.58 (±0.11)	86.48 (±0.21)	**86.78 (±0.11)**	86.63 (±0.11)
	WAVEF.	79.51 (±0.21)	79.26 (±0.20)	79.88 (±0.18)	78.96 (±0.16)	79.90 (±0.12)	79.85 (±0.20)	**80.04 (±0.13)**
		79.89 (±0.18)	79.91 (±0.14)	79.75 (±0.18)	79.42 (±0.20)	79.78 (±0.16)	79.95 (±0.13)	79.80 (±0.17)
GRAD. 100K	**AGRAW$_1$**	64.06 (±0.63)	63.34 (±0.22)	65.84 (±0.09)	62.77 (±0.08)	65.16 (±0.09)	65.88 (±0.08)	65.75 (±0.11)
		65.93 (±0.09)	64.65 (±0.08)	65.07 (±0.09)	64.33 (±0.49)	65.71 (±0.12)	65.92 (±0.08)	65.69 (±0.11)
	AGRAW$_2$	83.83 (±0.81)	70.29 (±0.32)	85.65 (±0.10)	84.08 (±0.11)	85.42 (±0.09)	84.16 (±0.45)	**85.82 (±0.06)**
		85.70 (±0.13)	85.62 (±0.08)	85.70 (±0.07)	84.03 (±0.43)	85.07 (±0.36)	85.79 (±0.06)	85.51 (±0.26)
	LED	72.34 (±0.51)	70.51 (±0.15)	65.42 (±0.47)	68.83 (±0.23)	69.17 (±0.62)	68.51 (±0.77)	57.28 (±0.45)
		73.22 (±0.12)	71.76 (±0.23)	72.31 (±0.23)	72.47 (±0.17)	73.18 (±0.12)	**73.30 (±0.12)**	72.46 (±0.18)
	MIXED	91.22 (±0.07)	90.61 (±0.09)	91.03 (±0.07)	89.29 (±0.09)	90.88 (±0.08)	91.21 (±0.06)	91.07 (±0.07)
		91.25 (±0.07)	91.27 (±0.07)	**91.30 (±0.06)**	91.23 (±0.07)	91.27 (±0.07)	91.29 (±0.06)	91.23 (±0.07)
	RandRBF	31.41 (±0.40)	31.56 (±0.34)	30.62 (±0.27)	31.53 (±0.58)	29.14 (±0.19)	31.48 (±0.43)	30.07 (±0.28)
		31.16 (±0.33)	29.25 (±0.25)	29.52 (±0.25)	**31.64 (±0.45)**	31.20 (±0.35)	31.23 (±0.30)	30.79 (±0.40)
	SINE	86.58 (±0.29)	86.16 (±0.17)	86.74 (±0.08)	86.01 (±0.09)	86.73 (±0.07)	86.86 (±0.09)	86.71 (±0.07)
		87.14 (±0.09)	**87.18 (±0.09)**	87.17 (±0.08)	87.04 (±0.09)	86.84 (±0.20)	87.16 (±0.09)	87.05 (±0.09)
	WAVEF.	79.52 (±0.18)	79.25 (±0.20)	80.20 (±0.12)	79.10 (±0.13)	80.13 (±0.10)	**80.25 (±0.10)**	80.21 (±0.11)
		80.20 (±0.11)	80.16 (±0.12)	80.18 (±0.12)	79.95 (±0.19)	80.02 (±0.13)	80.13 (±0.11)	80.21 (±0.10)

scenarios, only the first three are statistically superior to all the remaining five configurations. The latter six configurations are *not* superior to EDDM, whereas WSTD and SEQDRIFT$_2$ are also statistically indistinguishable from STEPD.

Fig. 6 captures the evaluation of the accuracy results of the methods aggregating *all* the tests executed using HT as base learner, which is similar to the aggregation carried out for NB and represented in Fig. 3. With this subset of data, the best configurations are HDDM$_A$, RDDM, FHDDM, FTDD, HDDM$_W$, and RDDM$_{30}$, with no statistical differences between these six methods. Once again, HDDM$_A$ was the only one significantly superior to all the other eight tested methods. In this scenario, the others were *not* superior to WSTD or DDM, whereas RDDM$_{30}$ was *not* superior to SEQDRIFT$_2$, either.

Table 5
Mean accuracies of Drift Detectors in percentage (%) in the gradual datasets, with a 95% confidence interval, using NB (Part 2).

DS Type and Size	DATASET	DDM $HDDM_A$	EDDM $HDDM_W$	ADWIN FHDDM	ECDD FTDD	STEPD $RDDM_{30}$	SeqDrift2 RDDM	SEED WSTD
	AGRAW$_1$	63.93 (±0.91)	63.61 (±0.17)	66.38 (±0.04)	62.94 (±0.06)	65.55 (±0.07)	**66.41 (±0.04)**	66.23 (±0.05)
		66.38 (±0.05)	64.85 (±0.08)	65.32 (±0.06)	66.27 (±0.06)	66.22 (±0.13)	66.36 (±0.04)	66.20 (±0.05)
	AGRAW$_2$	86.16 (±0.34)	70.43 (±0.04)	86.74 (±0.04)	84.68 (±0.06)	86.19 (±0.07)	86.25 (±0.63)	**86.79 (±0.04)**
		86.74 (±0.05)	86.42 (±0.04)	86.52 (±0.05)	86.05 (±0.62)	86.49 (±0.18)	86.67 (±0.07)	86.61 (±0.11)
	LED	72.69 (±0.89)	70.80 (±0.16)	67.26 (±0.93)	69.15 (±0.15)	69.98 (±0.27)	72.66 (±0.34)	59.20 (±1.04)
		73.74 (±0.11)	72.31 (±0.26)	72.86 (±0.13)	73.35 (±0.26)	73.55 (±0.13)	73.72 (±0.09)	73.35 (±0.06)
GRAD.	**MIXED**	91.76 (±0.22)	90.70 (±0.10)	91.88 (±0.04)	89.83 (±0.07)	91.51 (±0.05)	91.93 (±0.04)	91.89 (±0.03)
500K		**91.94 (±0.04)**	91.93 (±0.03)	**91.94 (±0.03)**	91.92 (±0.03)	91.91 (±0.03)	91.91 (±0.03)	91.93 (±0.03)
	RandRBF	**33.31 (±0.43)**	33.27 (±0.34)	30.80 (±0.31)	33.26 (±0.66)	29.08 (±0.13)	33.18 (±0.36)	30.41 (±0.48)
		32.49 (±0.40)	29.40 (±0.08)	29.94 (±0.16)	33.14 (±0.32)	32.46 (±0.36)	32.14 (±0.26)	30.99 (±0.26)
	SINE	84.02 (±3.07)	84.05 (±2.27)	87.26 (±0.05)	86.36 (±0.05)	87.24 (±0.05)	87.29 (±0.05)	87.25 (±0.06)
		87.31 (±0.06)	87.35 (±0.07)	87.36 (±0.05)	87.33 (±0.04)	87.23 (±0.13)	**87.39 (±0.05)**	87.33 (±0.04)
	WAVEF.	79.81 (±0.29)	79.19 (±0.22)	80.37 (±0.12)	79.18 (±0.11)	80.24 (±0.11)	**80.38 (±0.12)**	80.36 (±0.12)
		80.35 (±0.12)	80.32 (±0.11)	80.36 (±0.11)	80.35 (±0.11)	80.19 (±0.16)	80.33 (±0.14)	80.33 (±0.15)
	AGRAW$_1$	64.57 (±0.85)	63.78 (±0.30)	66.47 (±0.04)	62.96 (±0.05)	65.64 (±0.05)	**66.49 (±0.04)**	66.38 (±0.04)
		66.45 (±0.05)	64.92 (±0.06)	65.36 (±0.05)	66.43 (±0.07)	66.33 (±0.10)	66.44 (±0.06)	66.29 (±0.05)
	AGRAW$_2$	86.55 (±0.29)	70.44 (±0.03)	**86.90 (±0.02)**	84.73 (±0.04)	86.28 (±0.04)	86.34 (±0.61)	86.89 (±0.03)
		86.83 (±0.05)	86.47 (±0.11)	86.63 (±0.02)	86.56 (±0.29)	86.65 (±0.27)	86.81 (±0.03)	86.76 (±0.10)
	LED	72.21 (±1.15)	70.82 (±0.12)	68.94 (±0.66)	69.23 (±0.12)	70.12 (±0.30)	73.38 (±0.19)	60.82 (±0.91)
		73.84 (±0.06)	72.44 (±0.14)	72.96 (±0.11)	73.61 (±0.24)	73.72 (±0.04)	73.79 (±0.05)	73.46 (±0.12)
GRAD.	**MIXED**	91.89 (±0.17)	90.70 (±0.15)	92.00 (±0.03)	89.92 (±0.05)	91.60 (±0.03)	92.03 (±0.03)	92.00 (±0.03)
1M		92.03 (±0.03)	92.03 (±0.03)	**92.04 (±0.03)**	92.03 (±0.03)	91.98 (±0.04)	91.99 (±0.03)	92.03 (±0.03)
	RandRBF	33.37 (±0.28)	**33.49 (±0.15)**	31.02 (±0.24)	33.16 (±0.44)	29.03 (±0.08)	**33.49 (±0.26)**	30.68 (±0.28)
		32.95 (±0.25)	29.33 (±0.09)	29.88 (±0.16)	33.25 (±0.21)	32.65 (±0.24)	32.10 (±0.15)	31.07 (±0.24)
	SINE	79.93 (±5.03)	85.98 (±0.09)	87.37 (±0.05)	86.43 (±0.03)	87.30 (±0.04)	87.38 (±0.05)	87.36 (±0.05)
		87.38 (±0.07)	87.43 (±0.05)	87.43 (±0.05)	87.41 (±0.05)	87.32 (±0.05)	**87.45 (±0.04)**	87.40 (±0.05)
	WAVEF.	79.89 (±0.32)	79.13 (±0.27)	80.40 (±0.08)	79.19 (±0.08)	80.27 (±0.06)	**80.43 (±0.07)**	80.42 (±0.07)
		80.40 (±0.09)	80.37 (±0.08)	80.40 (±0.09)	80.39 (±0.10)	80.25 (±0.20)	80.40 (±0.06)	80.38 (±0.07)
	AGRAW$_1$	64.06 (±1.10)	63.91 (±0.23)	66.54 (±0.04)	62.98 (±0.03)	65.65 (±0.04)	**66.55 (±0.04)**	66.46 (±0.03)
		66.49 (±0.05)	64.94 (±0.04)	65.39 (±0.03)	66.52 (±0.04)	66.40 (±0.05)	66.51 (±0.03)	66.30 (±0.03)
	AGRAW$_2$	85.85 (±0.87)	70.45 (±0.03)	**86.98 (±0.02)**	84.76 (±0.04)	86.32 (±0.03)	86.97 (±0.03)	86.97 (±0.01)
		86.95 (±0.02)	86.58 (±0.01)	86.68 (±0.02)	86.87 (±0.10)	86.77 (±0.08)	86.88 (±0.02)	86.87 (±0.03)
	LED	71.68 (±1.42)	71.00 (±0.13)	70.62 (±0.62)	69.31 (±0.08)	70.23 (±0.23)	73.58 (±0.15)	61.60 (±1.18)
		73.89 (±0.05)	72.50 (±0.07)	73.06 (±0.09)	73.74 (±0.18)	73.78 (±0.10)	73.86 (±0.04)	73.60 (±0.06)
GRAD.	**MIXED**	88.51 (±3.84)	89.50 (±2.32)	92.02 (±0.03)	89.92 (±0.04)	91.60 (±0.03)	**92.03 (±0.02)**	92.01 (±0.03)
2M		92.02 (±0.02)	**92.03 (±0.03)**	**92.03 (±0.02)**	**92.03 (±0.03)**	91.99 (±0.03)	91.98 (±0.03)	**92.03 (±0.03)**
	RandRBF	33.57 (±0.23)	**33.83 (±0.15)**	30.90 (±0.15)	33.07 (±0.19)	29.00 (±0.09)	33.26 (±0.20)	30.74 (±0.20)
		33.07 (±0.19)	29.31 (±0.07)	29.85 (±0.11)	33.21 (±0.14)	32.46 (±0.20)	32.13 (±0.10)	31.15 (±0.15)
	SINE	80.81 (±5.41)	82.86 (±4.43)	87.40 (±0.03)	86.45 (±0.02)	87.33 (±0.03)	87.41 (±0.02)	87.41 (±0.02)
		87.09 (±0.67)	87.45 (±0.02)	87.45 (±0.03)	87.44 (±0.03)	87.35 (±0.02)	**87.46 (±0.03)**	87.43 (±0.02)
	WAVEF.	79.53 (±0.30)	79.21 (±0.29)	80.45 (±0.03)	79.21 (±0.04)	80.31 (±0.04)	**80.48 (±0.04)**	80.46 (±0.04)
		80.46 (±0.04)	80.42 (±0.05)	80.44 (±0.04)	80.47 (±0.04)	80.40 (±0.05)	80.45 (±0.04)	80.45 (±0.04)
NB	**RANK**	9.19388	10.08160	8.26531	10.63270	9.80612	6.48980	8.28571
GRAD.		4.66327	6.92857	5.89796	7.09184	6.7449	**3.98980**	6.92857
NB	**RANK**	9.65306	10.60710	7.95918	10.67860	9.34184	7.06122	8.18878
ALL		4.78061	6.80102	5.95408	6.36735	7.28061	**4.40306**	5.92347

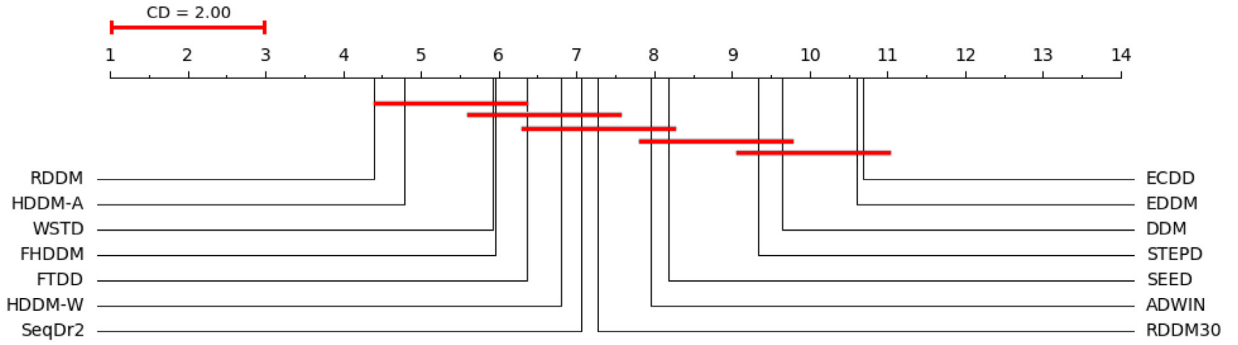

Fig. 3. Comparison results using the Nemenyi test of Detectors with NB in all artificial datasets with a 95% confidence interval.

Table 6
Mean accuracies of Drift Detectors in percentage (%) in the abrupt datasets, with a 95% confidence interval, using HT (Part 1).

DS Type and Size	DATASET	DDM $HDDM_A$	EDDM $HDDM_W$	ADWIN FHDDM	ECDD FTDD	STEPD $RDDM_{30}$	SeqDrift2 RDDM	SEED WSTD
	AGRAW$_1$	63.13 (±0.56)	62.08 (±0.34)	62.44 (±0.25)	63.26 (±0.33)	63.78 (±0.38)	62.87 (±0.33)	63.24 (±0.31)
		64.47 (±0.34)	64.31 (±0.37)	64.48 (±0.39)	62.64 (±0.38)	63.16 (±0.40)	**64.69 (±0.30)**	63.44 (±0.43)
	AGRAW$_2$	75.22 (±1.80)	75.89 (±1.76)	79.16 (±0.35)	81.95 (±0.27)	81.97 (±0.43)	77.68 (±0.35)	80.74 (±0.35)
		81.56 (±0.44)	**82.59 (±0.28)**	82.28 (±0.29)	79.41 (±0.66)	76.13 (±1.87)	81.58 (±0.85)	81.07 (±0.51)
	LED	69.56 (±0.30)	67.46 (±0.40)	61.98 (±0.66)	67.35 (±0.43)	60.78 (±1.68)	58.49 (±0.95)	55.19 (±0.74)
		69.68 (±0.30)	69.24 (±0.43)	69.29 (±0.37)	67.01 (±0.74)	69.52 (±0.29)	**69.78 (±0.29)**	67.08 (±1.00)
ABRUPT	**MIXED**	89.70 (±0.29)	88.84 (±0.29)	88.74 (±0.27)	89.01 (±0.28)	90.30 (±0.29)	83.23 (±0.20)	89.02 (±0.24)
10K		90.32 (±0.23)	90.35 (±0.22)	90.14 (±0.21)	90.33 (±0.23)	89.82 (±0.24)	90.17 (±0.24)	**90.36 (±0.22)**
	RandRBF	31.92 (±0.44)	31.89 (±0.41)	31.87 (±0.39)	30.83 (±0.64)	31.23 (±0.50)	**32.32 (±0.47)**	31.59 (±0.45)
		32.06 (±0.37)	31.00 (±0.41)	31.34 (±0.37)	32.26 (±0.50)	32.22 (±0.43)	32.01 (±0.39)	30.93 (±0.60)
	SINE	87.01 (±0.72)	85.86 (±0.28)	86.86 (±0.17)	86.57 (±0.27)	88.06 (±0.19)	82.37 (±0.14)	86.88 (±0.23)
		88.39 (±0.17)	**88.40 (±0.14)**	88.29 (±0.15)	88.37 (±0.17)	87.82 (±0.16)	87.98 (±0.20)	88.38 (±0.15)
	WAVEF.	78.45 (±0.46)	78.55 (±0.45)	78.33 (±0.39)	78.30 (±0.39)	**79.21 (±0.42)**	78.38 (±0.41)	78.66 (±0.44)
		78.69 (±0.48)	79.15 (±0.44)	78.99 (±0.45)	78.07 (±0.58)	78.54 (±0.42)	79.09 (±0.47)	78.77 (±0.51)
	AGRAW$_1$	64.93 (±1.28)	64.79 (±0.61)	64.27 (±0.23)	63.83 (±0.44)	64.96 (±0.31)	64.15 (±0.48)	64.45 (±0.19)
		68.12 (±0.48)	66.60 (±0.71)	67.26 (±0.59)	64.04 (±0.76)	67.31 (±0.49)	**68.19 (±0.45)**	65.33 (±0.48)
	AGRAW$_2$	81.19 (±1.60)	78.28 (±1.76)	83.27 (±0.24)	83.49 (±0.17)	84.56 (±0.23)	82.81 (±0.20)	83.76 (±0.17)
		84.44 (±0.26)	**84.88 (±0.14)**	84.68 (±0.17)	84.13 (±0.24)	82.06 (±1.24)	83.11 (±1.31)	84.52 (±0.23)
	LED	71.31 (±0.25)	69.14 (±0.26)	63.22 (±0.58)	68.09 (±0.42)	65.16 (±1.59)	60.97 (±1.17)	55.44 (±0.61)
		71.52 (±0.18)	70.76 (±0.41)	71.26 (±0.32)	70.51 (±0.43)	71.38 (±0.18)	**71.73 (±0.17)**	70.25 (±0.60)
ABRUPT	**MIXED**	88.96 (±0.54)	89.30 (±0.39)	90.13 (±0.13)	89.37 (±0.20)	90.65 (±0.17)	87.80 (±0.21)	90.02 (±0.14)
20K		90.29 (±0.15)	**90.67 (±0.14)**	90.58 (±0.14)	90.64 (±0.15)	90.47 (±0.16)	90.66 (±0.14)	90.64 (±0.15)
	RandRBF	31.82 (±0.43)	32.28 (±0.44)	32.10 (±0.37)	31.23 (±0.62)	31.43 (±0.40)	**32.70 (±0.41)**	31.86 (±0.42)
		32.40 (±0.34)	31.15 (±0.28)	31.39 (±0.32)	32.60 (±0.45)	32.34 (±0.36)	32.30 (±0.37)	31.12 (±0.54)
	SINE	89.31 (±0.14)	87.21 (±0.19)	88.67 (±0.14)	86.90 (±0.19)	89.22 (±0.20)	86.80 (±0.10)	88.26 (±0.19)
		89.89 (±0.13)	89.92 (±0.12)	89.87 (±0.12)	89.89 (±0.13)	89.48 (±0.14)	89.46 (±0.14)	**89.93 (±0.12)**
	WAVEF.	78.89 (±0.22)	78.86 (±0.27)	79.10 (±0.26)	78.65 (±0.25)	79.69 (±0.22)	79.25 (±0.24)	79.49 (±0.25)
		79.41 (±0.25)	**79.74 (±0.25)**	79.52 (±0.26)	79.05 (±0.36)	79.27 (±0.27)	79.64 (±0.26)	79.46 (±0.29)
	AGRAW$_1$	68.03 (±1.98)	67.45 (±0.82)	65.73 (±0.15)	64.76 (±0.64)	66.18 (±0.29)	66.84 (±0.35)	65.24 (±0.15)
		72.57 (±0.33)	70.68 (±0.58)	71.50 (±0.28)	67.23 (±0.89)	71.43 (±0.80)	72.43 (±0.31)	69.16 (±0.72)
	AGRAW$_2$	83.60 (±1.14)	74.08 (±2.06)	84.98 (±0.17)	84.40 (±0.08)	85.95 (±0.16)	84.19 (±0.43)	85.23 (±0.14)
		85.76 (±0.32)	**86.57 (±0.09)**	86.54 (±0.13)	84.46 (±0.44)	84.79 (±0.49)	86.09 (±0.19)	85.86 (±0.42)
	LED	71.93 (±0.48)	69.95 (±0.24)	64.15 (±0.55)	68.69 (±0.32)	67.85 (±1.10)	64.85 (±1.35)	55.99 (±0.54)
		72.81 (±0.16)	71.60 (±0.27)	72.10 (±0.29)	72.20 (±0.21)	72.66 (±0.15)	**72.88 (±0.15)**	71.99 (±0.31)
ABRUPT	**MIXED**	91.28 (±0.37)	90.30 (±0.17)	91.46 (±0.12)	89.78 (±0.14)	91.14 (±0.10)	90.56 (±0.12)	90.88 (±0.14)
50K		92.11 (±0.07)	**92.14 (±0.08)**	92.13 (±0.08)	92.05 (±0.09)	91.78 (±0.11)	91.60 (±0.14)	92.03 (±0.11)
	RandRBF	32.54 (±0.37)	32.45 (±0.33)	32.23 (±0.27)	**33.17 (±0.32)**	31.10 (±0.23)	32.62 (±0.36)	31.89 (±0.29)
		32.57 (±0.30)	30.97 (±0.21)	31.35 (±0.25)	32.70 (±0.42)	32.52 (±0.28)	32.40 (±0.28)	31.81 (±0.38)
	SINE	91.06 (±0.15)	88.97 (±0.24)	89.88 (±0.10)	87.12 (±0.13)	90.37 (±0.21)	90.23 (±0.11)	89.24 (±0.14)
		91.52 (±0.14)	91.54 (±0.13)	91.52 (±0.13)	**91.55 (±0.15)**	91.25 (±0.13)	91.19 (±0.14)	91.52 (±0.13)
	WAVEF.	79.28 (±0.20)	79.06 (±0.15)	79.53 (±0.16)	79.00 (±0.17)	**80.00 (±0.15)**	79.46 (±0.21)	79.73 (±0.16)
		79.58 (±0.16)	79.98 (±0.15)	79.81 (±0.14)	79.37 (±0.21)	79.47 (±0.19)	79.94 (±0.16)	79.63 (±0.18)
	AGRAW$_1$	71.01 (±2.08)	69.42 (±1.05)	66.48 (±0.12)	66.25 (±0.71)	66.89 (±0.27)	68.38 (±0.32)	65.59 (±0.12)
		74.74 (±0.34)	72.46 (±0.39)	72.91 (±0.25)	70.38 (±1.01)	74.19 (±0.98)	**74.85 (±0.28)**	71.62 (±0.74)
	AGRAW$_2$	84.81 (±0.83)	72.67 (±1.67)	85.98 (±0.11)	84.65 (±0.07)	86.70 (±0.10)	85.76 (±0.55)	85.83 (±0.09)
		87.44 (±0.14)	87.67 (±0.05)	**87.69 (±0.05)**	85.86 (±0.51)	85.84 (±0.66)	87.17 (±0.19)	87.40 (±0.15)
	LED	72.65 (±0.30)	70.32 (±0.18)	64.79 (±0.46)	68.97 (±0.23)	69.04 (±0.73)	67.90 (±1.11)	56.59 (±0.91)
		73.37 (±0.11)	71.97 (±0.23)	72.51 (±0.23)	72.93 (±0.18)	73.23 (±0.12)	**73.39 (±0.12)**	72.81 (±0.20)
ABRUPT	**MIXED**	92.79 (±0.12)	91.42 (±0.11)	91.77 (±0.10)	89.75 (±0.10)	91.23 (±0.09)	92.39 (±0.07)	91.26 (±0.13)
100K		93.12 (±0.07)	**93.15 (±0.06)**	**93.15 (±0.06)**	93.13 (±0.06)	92.89 (±0.09)	92.46 (±0.21)	93.09 (±0.06)
	RandRBF	33.64 (±0.26)	33.36 (±0.32)	32.42 (±0.24)	**34.86 (±0.23)**	31.17 (±0.16)	33.51 (±0.18)	31.91 (±0.22)
		32.87 (±0.27)	31.08 (±0.12)	31.30 (±0.13)	33.32 (±0.32)	33.08 (±0.27)	32.80 (±0.20)	32.45 (±0.30)
	SINE	92.31 (±0.09)	90.49 (±0.21)	90.33 (±0.08)	87.15 (±0.10)	90.96 (±0.16)	91.91 (±0.10)	89.56 (±0.11)
		92.57 (±0.11)	**92.61 (±0.10)**	92.59 (±0.10)	**92.61 (±0.10)**	92.39 (±0.10)	92.41 (±0.11)	92.59 (±0.10)
	WAVEF.	79.35 (±0.24)	79.40 (±0.17)	79.55 (±0.11)	79.12 (±0.14)	**80.16 (±0.11)**	79.53 (±0.17)	79.77 (±0.11)
		79.52 (±0.17)	80.03 (±0.10)	79.85 (±0.13)	79.59 (±0.19)	79.60 (±0.16)	79.97 (±0.14)	79.83 (±0.15)

4.1. Discussion and Answer to RQ1

One telling fact that can easily be identified in these reported evaluations is that the most well-known and cited concept drift detection methods, namely DDM, EDDM, ADWIN, ECDD, and STEPD, are consistently ranked among the worst configurations in *all* of them. The same can be said of SEED.

It is also worth observing that WSTD and FTDD delivered stronger performances in the abrupt datasets than in the gradual ones. On the other hand, the two configurations of RDDM and, to a lesser extent $SEQDRIFT_2$, were generally better in the gradual datasets.

Table 7
Mean accuracies of Drift Detectors in percentage (%) in the abrupt datasets, with a 95% confidence interval, using HT (Part 2).

DS Type and Size	DATASET	DDM HDDM$_A$	EDDM HDDM$_W$	ADWIN FHDDM	ECDD FTDD	STEPD RDDM$_{30}$	SeqDrift2 RDDM	SEED WSTD
	AGRAW$_1$	76.88 (±1.17)	73.47 (±3.58)	66.86 (±0.09)	66.81 (±0.50)	68.36 (±0.43)	71.46 (±0.76)	65.85 (±0.13)
		77.99 (±0.79)	75.04 (±0.19)	75.90 (±0.57)	76.88 (±1.59)	78.23 (±1.10)	**78.61 (±0.52)**	76.81 (±0.65)
	AGRAW$_2$	88.22 (±0.86)	75.48 (±2.81)	86.86 (±0.08)	84.85 (±0.05)	87.35 (±0.14)	89.24 (±0.12)	86.38 (±0.09)
		89.29 (±0.08)	89.17 (±0.05)	89.21 (±0.04)	88.72 (±0.52)	88.69 (±0.21)	88.79 (±0.14)	89.15 (±0.09)
	LED	70.95 (±0.92)	71.29 (±0.23)	65.09 (±0.30)	69.15 (±0.15)	69.51 (±0.38)	70.38 (±0.69)	57.57 (±1.21)
		73.57 (±0.08)	72.37 (±0.26)	72.94 (±0.15)	**73.60 (±0.10)**	73.32 (±0.13)	73.58 (±0.09)	73.25 (±0.09)
ABRUPT	**MIXED**	94.80 (±0.06)	93.72 (±0.08)	92.15 (±0.07)	89.88 (±0.07)	91.59 (±0.12)	94.72 (±0.05)	91.96 (±0.11)
500K		**94.89 (±0.05)**	**94.89 (±0.05)**	**94.89 (±0.05)**	94.87 (±0.07)	94.40 (±0.19)	94.22 (±0.14)	94.86 (±0.06)
	RandRBF	35.88 (±0.56)	37.06 (±0.33)	32.56 (±0.19)	**38.19 (±0.09)**	31.18 (±0.11)	35.80 (±0.61)	32.24 (±0.16)
		34.11 (±0.27)	31.00 (±0.05)	31.42 (±0.09)	35.39 (±0.39)	34.78 (±0.34)	33.73 (±0.24)	32.47 (±0.15)
	SINE	95.63 (±0.13)	94.75 (±0.26)	90.64 (±0.08)	87.20 (±0.07)	91.84 (±0.39)	95.57 (±0.15)	89.73 (±0.07)
		95.75 (±0.18)	95.80 (±0.19)	95.79 (±0.19)	**95.82 (±0.20)**	95.31 (±0.23)	94.89 (±0.30)	**95.82 (±0.18)**
	WAVEF.	81.69 (±0.19)	81.62 (±0.08)	79.78 (±0.16)	79.18 (±0.11)	80.20 (±0.10)	81.70 (±0.14)	79.84 (±0.15)
		81.09 (±0.20)	80.07 (±0.11)	80.51 (±0.32)	**81.74 (±0.14)**	80.62 (±0.27)	80.17 (±0.14)	80.75 (±0.24)
	AGRAW$_1$	76.26 (±3.86)	73.85 (±2.34)	66.98 (±0.08)	66.95 (±0.30)	68.30 (±0.36)	71.56 (±0.86)	65.94 (±0.12)
		79.50 (±0.70)	75.82 (±0.15)	76.91 (±0.87)	79.24 (±0.95)	78.99 (±0.84)	79.25 (±1.09)	78.23 (±0.34)
	AGRAW$_2$	88.45 (±0.78)	79.87 (±3.52)	87.00 (±0.07)	84.87 (±0.03)	87.58 (±0.14)	**89.67 (±0.09)**	86.54 (±0.07)
		89.66 (±0.05)	89.48 (±0.04)	89.53 (±0.04)	89.32 (±0.19)	88.85 (±0.32)	88.83 (±0.26)	89.41 (±0.12)
	LED	71.31 (±0.99)	71.41 (±0.16)	65.45 (±0.48)	69.23 (±0.13)	69.85 (±0.36)	71.04 (±0.57)	58.30 (±0.55)
		73.60 (±0.05)	72.46 (±0.14)	72.99 (±0.11)	**73.64 (±0.06)**	73.41 (±0.08)	73.62 (±0.06)	73.37 (±0.09)
ABRUPT	**MIXED**	95.91 (±0.09)	95.04 (±0.09)	92.23 (±0.05)	89.92 (±0.06)	91.53 (±0.09)	95.92 (±0.10)	91.98 (±0.17)
1M		95.98 (±0.13)	**96.00 (±0.13)**	95.99 (±0.13)	95.97 (±0.12)	94.89 (±0.34)	94.62 (±0.23)	95.93 (±0.12)
	RandRBF	37.17 (±0.59)	**38.38 (±0.16)**	32.64 (±0.14)	38.18 (±0.09)	31.14 (±0.07)	35.56 (±0.47)	32.29 (±0.17)
		35.01 (±0.23)	31.00 (±0.05)	31.39 (±0.06)	36.17 (±0.46)	35.04 (±0.26)	34.27 (±0.28)	32.45 (±0.14)
	SINE	96.92 (±0.09)	96.35 (±0.10)	90.68 (±0.03)	87.23 (±0.04)	92.11 (±0.42)	96.88 (±0.07)	89.90 (±0.05)
		97.02 (±0.15)	97.01 (±0.14)	97.00 (±0.14)	**97.03 (±0.16)**	95.95 (±0.30)	95.44 (±0.19)	**97.03 (±0.15)**
	WAVEF.	82.40 (±0.13)	82.33 (±0.09)	79.84 (±0.11)	79.19 (±0.07)	80.23 (±0.06)	82.51 (±0.07)	79.93 (±0.09)
		81.96 (±0.14)	80.12 (±0.08)	81.22 (±0.50)	**82.64 (±0.05)**	80.84 (±0.15)	80.25 (±0.12)	81.20 (±0.37)
	AGRAW$_1$	80.24 (±1.85)	77.61 (±3.05)	67.02 (±0.04)	67.01 (±0.25)	68.52 (±0.21)	72.83 (±0.86)	66.05 (±0.05)
		80.98 (±1.37)	76.26 (±0.10)	78.16 (±1.07)	80.65 (±0.86)	79.48 (±0.79)	79.34 (±0.45)	80.58 (±1.17)
	AGRAW$_2$	89.08 (±0.88)	82.81 (±2.00)	87.03 (±0.08)	84.88 (±0.02)	87.61 (±0.08)	**89.97 (±0.09)**	86.59 (±0.06)
		89.96 (±0.05)	89.76 (±0.02)	89.80 (±0.03)	89.79 (±0.17)	88.88 (±0.25)	89.08 (±0.05)	89.72 (±0.13)
	LED	70.00 (±0.68)	71.50 (±0.13)	65.60 (±0.35)	69.30 (±0.07)	69.99 (±0.25)	70.88 (±0.36)	58.52 (±0.61)
		73.65 (±0.03)	72.50 (±0.07)	73.07 (±0.09)	**73.72 (±0.04)**	73.47 (±0.06)	73.66 (±0.04)	73.46 (±0.05)
ABRUPT	**MIXED**	96.85 (±0.10)	96.14 (±0.06)	92.24 (±0.05)	89.91 (±0.05)	91.47 (±0.05)	96.88 (±0.11)	92.07 (±0.13)
2M		**96.94 (±0.10)**	96.93 (±0.09)	96.92 (±0.10)	96.91 (±0.09)	94.97 (±0.18)	94.54 (±0.14)	96.91 (±0.10)
	RandRBF	37.43 (±0.13)	**39.10 (±0.06)**	32.63 (±0.07)	38.63 (±0.08)	31.14 (±0.07)	36.04 (±0.19)	32.26 (±0.08)
		35.62 (±0.29)	30.99 (±0.05)	31.36 (±0.04)	36.48 (±0.25)	35.13 (±0.28)	34.26 (±0.18)	32.42 (±0.09)
	SINE	97.98 (±0.04)	97.62 (±0.06)	90.71 (±0.03)	87.23 (±0.02)	92.19 (±0.28)	97.98 (±0.05)	89.88 (±0.06)
		98.04 (±0.08)	98.04 (±0.09)	98.04 (±0.09)	**98.06 (±0.10)**	96.20 (±0.15)	95.68 (±0.10)	98.05 (±0.10)
	WAVEF.	83.16 (±0.07)	82.88 (±0.03)	79.93 (±0.07)	79.20 (±0.04)	80.27 (±0.04)	83.13 (±0.08)	79.97 (±0.08)
		82.56 (±0.14)	80.15 (±0.04)	82.06 (±0.37)	**83.23 (±0.07)**	81.04 (±0.17)	80.37 (±0.08)	81.63 (±0.33)
HT	**RANK**	7.52041	9.30612	10.83670	11.30610	9.12245	8.27551	11.16330
ABRUPT		**3.88776**	5.59184	5.30612	4.92857	6.55102	5.43878	5.76531
HT	**RANK**	6.93878	8.75510	10.88780	11.21430	9.40816	7.80612	11.27550
ALL		**4.22449**	5.79592	5.23980	5.72959	6.11735	5.22959	6.37755

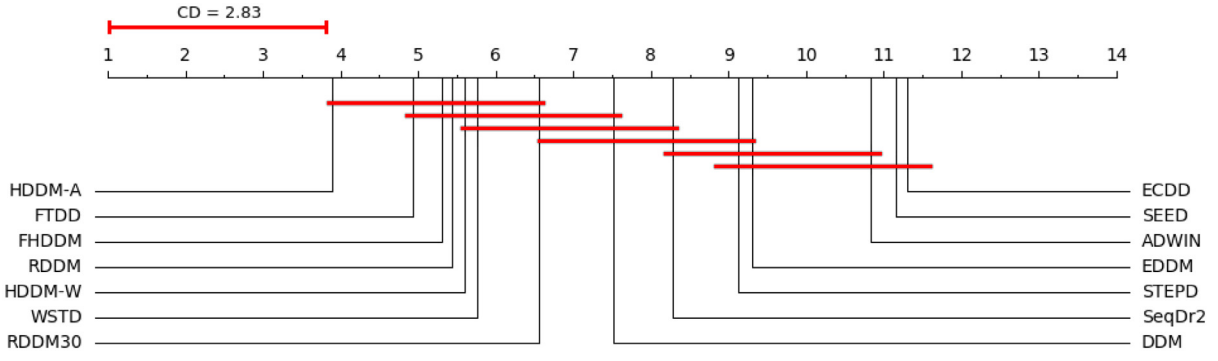

Fig. 4. Comparison results using the Nemenyi test of Detectors with HT in the abrupt datasets with a 95% confidence interval.

Table 8
Mean accuracies of Drift Detectors in percentage (%) in the gradual datasets, with a 95% confidence interval, using HT (Part 1).

DS Type and Size	DATASET	DDM $HDDM_A$	EDDM $HDDM_W$	ADWIN FHDDM	ECDD FTDD	STEPD $RDDM_{30}$	SeqDrift2 RDDM	SEED WSTD
GRAD. 10K	**AGRAW$_1$**	61.57 (±0.47)	61.48 (±0.27)	61.62 (±0.19)	61.98 (±0.20)	62.15 (±0.25)	61.57 (±0.26)	61.83 (±0.24)
		62.27 (±0.36)	62.71 (±0.22)	62.56 (±0.24)	61.33 (±0.29)	61.18 (±0.36)	**62.92 (±0.27)**	61.77 (±0.38)
	AGRAW$_2$	73.62 (±1.59)	73.68 (±1.57)	76.36 (±0.31)	79.16 (±0.25)	78.59 (±0.50)	76.15 (±0.33)	78.05 (±0.35)
		78.27 (±0.54)	**79.66 (±0.22)**	78.81 (±0.37)	74.56 (±0.76)	74.00 (±1.58)	78.65 (±0.98)	77.35 (±0.66)
	LED	67.76 (±0.42)	66.62 (±0.39)	60.20 (±0.52)	65.10 (±0.40)	59.39 (±1.39)	58.10 (±0.67)	55.40 (±0.78)
		67.58 (±0.31)	66.68 (±0.36)	66.86 (±0.35)	62.88 (±0.89)	67.80 (±0.34)	**67.81 (±0.29)**	63.99 (±0.81)
	MIXED	83.49 (±0.28)	**84.00 (±0.26)**	82.85 (±0.27)	82.84 (±0.31)	83.04 (±0.31)	83.27 (±0.24)	82.47 (±0.23)
		83.39 (±0.27)	83.55 (±0.28)	83.98 (±0.28)	83.50 (±0.23)	83.70 (±0.27)	83.70 (±0.31)	83.26 (±0.27)
	RandRBF	32.00 (±0.46)	31.74 (±0.36)	31.73 (±0.38)	30.87 (±0.65)	30.99 (±0.51)	**32.22 (±0.48)**	31.49 (±0.40)
		32.02 (±0.39)	30.94 (±0.35)	31.21 (±0.43)	32.10 (±0.52)	32.09 (±0.44)	31.93 (±0.38)	31.12 (±0.64)
	SINE	82.43 (±0.28)	82.26 (±0.22)	81.43 (±0.23)	81.50 (±0.22)	81.48 (±0.24)	81.97 (±0.24)	81.11 (±0.26)
		82.41 (±0.27)	82.57 (±0.24)	**82.95 (±0.21)**	82.28 (±0.24)	82.65 (±0.23)	82.66 (±0.19)	82.14 (±0.22)
	WAVEF.	77.97 (±0.43)	78.21 (±0.40)	77.81 (±0.38)	78.05 (±0.37)	78.29 (±0.39)	77.71 (±0.40)	78.21 (±0.37)
		77.82 (±0.47)	78.21 (±0.39)	77.77 (±0.39)	76.68 (±0.43)	77.86 (±0.41)	**78.42 (±0.37)**	77.57 (±0.51)
GRAD. 20K	**AGRAW$_1$**	64.10 (±1.17)	64.00 (±0.74)	63.38 (±0.24)	62.94 (±0.32)	64.05 (±0.21)	62.72 (±0.65)	63.66 (±0.18)
		66.30 (±0.43)	65.02 (±0.58)	65.73 (±0.50)	61.47 (±0.81)	65.95 (±0.39)	**66.84 (±0.40)**	64.61 (±0.37)
	AGRAW$_2$	79.00 (±1.91)	77.64 (±1.69)	81.93 (±0.20)	82.27 (±0.14)	83.08 (±0.24)	82.41 (±0.24)	82.48 (±0.18)
		82.98 (±0.28)	**83.42 (±0.12)**	82.98 (±0.17)	82.21 (±0.35)	79.82 (±1.65)	82.64 (±0.96)	82.94 (±0.22)
	LED	70.54 (±0.19)	69.25 (±0.23)	62.43 (±0.66)	67.16 (±0.43)	64.25 (±1.39)	60.15 (±0.82)	55.11 (±0.49)
		70.42 (±0.19)	69.32 (±0.38)	69.79 (±0.34)	67.67 (±0.84)	70.60 (±0.18)	**70.66 (±0.19)**	68.57 (±0.50)
	MIXED	87.29 (±0.19)	**87.59 (±0.16)**	86.88 (±0.14)	86.69 (±0.18)	87.30 (±0.15)	87.53 (±0.14)	87.00 (±0.13)
		87.23 (±0.18)	87.32 (±0.17)	87.52 (±0.17)	87.11 (±0.16)	87.44 (±0.17)	87.53 (±0.18)	87.17 (±0.17)
	RandRBF	31.80 (±0.54)	32.33 (±0.40)	32.02 (±0.34)	31.25 (±0.62)	31.21 (±0.36)	**32.76 (±0.37)**	31.73 (±0.42)
		32.32 (±0.37)	31.06 (±0.34)	31.26 (±0.35)	32.69 (±0.44)	32.44 (±0.34)	32.19 (±0.35)	31.06 (±0.53)
	SINE	86.68 (±0.14)	86.53 (±0.13)	85.43 (±0.11)	85.04 (±0.18)	85.89 (±0.12)	86.55 (±0.14)	85.34 (±0.15)
		86.79 (±0.13)	86.77 (±0.17)	**87.03 (±0.13)**	86.76 (±0.11)	86.88 (±0.12)	86.83 (±0.13)	86.67 (±0.10)
	WAVEF.	78.52 (±0.25)	78.73 (±0.25)	78.70 (±0.21)	78.50 (±0.23)	**79.29 (±0.23)**	78.63 (±0.24)	79.04 (±0.27)
		78.78 (±0.28)	79.20 (±0.24)	78.84 (±0.26)	78.36 (±0.30)	78.74 (±0.24)	79.10 (±0.28)	78.78 (±0.24)
GRAD. 50K	**AGRAW$_1$**	68.46 (±1.74)	67.30 (±0.82)	65.32 (±0.16)	64.95 (±0.74)	65.77 (±0.25)	66.55 (±0.28)	64.91 (±0.13)
		71.39 (±0.26)	70.06 (±0.52)	70.35 (±0.31)	65.95 (±0.82)	70.84 (±0.40)	**71.43 (±0.31)**	67.93 (±0.65)
	AGRAW$_2$	83.17 (±1.16)	73.97 (±2.05)	84.28 (±0.25)	83.81 (±0.10)	85.45 (±0.15)	84.08 (±0.43)	84.72 (±0.15)
		85.19 (±0.32)	**85.95 (±0.11)**	85.87 (±0.13)	83.88 (±0.45)	84.38 (±0.52)	85.69 (±0.22)	85.29 (±0.42)
	LED	72.33 (±0.23)	70.22 (±0.19)	63.84 (±0.62)	68.32 (±0.33)	67.97 (±0.93)	64.66 (±1.29)	55.91 (±0.48)
		72.47 (±0.14)	71.15 (±0.28)	71.66 (±0.26)	71.61 (±0.17)	72.50 (±0.14)	**72.62 (±0.15)**	71.36 (±0.32)
	MIXED	90.84 (±0.10)	90.33 (±0.12)	89.97 (±0.11)	88.69 (±0.14)	89.81 (±0.10)	90.62 (±0.10)	89.67 (±0.14)
		90.78 (±0.09)	90.87 (±0.09)	**90.93 (±0.09)**	90.75 (±0.08)	90.85 (±0.09)	90.74 (±0.09)	90.68 (±0.09)
	RandRBF	32.53 (±0.34)	32.56 (±0.37)	32.17 (±0.28)	**33.17 (±0.32)**	31.08 (±0.28)	32.71 (±0.35)	31.88 (±0.28)
		32.58 (±0.29)	31.01 (±0.22)	31.34 (±0.23)	32.69 (±0.43)	32.60 (±0.31)	32.38 (±0.28)	31.91 (±0.38)
	SINE	90.27 (±0.10)	89.00 (±0.25)	88.51 (±0.10)	86.33 (±0.14)	89.16 (±0.20)	90.11 (±0.10)	88.13 (±0.11)
		90.33 (±0.11)	90.35 (±0.10)	**90.42 (±0.10)**	90.27 (±0.11)	90.35 (±0.11)	90.34 (±0.09)	90.24 (±0.12)
	WAVEF.	79.18 (±0.23)	79.19 (±0.16)	79.39 (±0.18)	78.95 (±0.16)	**79.85 (±0.13)**	79.13 (±0.23)	79.56 (±0.18)
		79.44 (±0.18)	79.76 (±0.15)	79.49 (±0.16)	79.01 (±0.21)	79.53 (±0.20)	79.71 (±0.14)	79.44 (±0.19)
GRAD. 100K	**AGRAW$_1$**	71.72 (±1.76)	69.25 (±1.18)	66.23 (±0.12)	65.79 (±0.64)	66.63 (±0.27)	68.25 (±0.24)	65.37 (±0.10)
		74.25 (±0.29)	71.98 (±0.36)	72.61 (±0.31)	69.51 (±1.08)	73.37 (±0.95)	**74.43 (±0.33)**	70.90 (±0.80)
	AGRAW$_2$	84.47 (±0.82)	72.65 (±1.65)	85.73 (±0.09)	84.40 (±0.06)	86.53 (±0.12)	85.61 (±0.56)	85.55 (±0.10)
		87.14 (±0.15)	**87.36 (±0.06)**	87.31 (±0.05)	85.59 (±0.51)	85.68 (±0.66)	86.97 (±0.18)	86.98 (±0.33)
	LED	72.54 (±0.34)	70.40 (±0.17)	64.60 (±0.50)	68.79 (±0.23)	68.75 (±0.71)	68.48 (±0.75)	56.20 (±0.48)
		73.21 (±0.12)	71.74 (±0.23)	72.30 (±0.23)	72.53 (±0.15)	73.18 (±0.12)	**73.30 (±0.12)**	72.40 (±0.18)
	MIXED	92.42 (±0.08)	91.49 (±0.11)	91.01 (±0.10)	89.20 (±0.10)	90.64 (±0.09)	92.32 (±0.07)	90.65 (±0.14)
		92.43 (±0.08)	92.49 (±0.07)	**92.52 (±0.07)**	92.43 (±0.06)	92.48 (±0.07)	92.37 (±0.08)	92.38 (±0.06)
	RandRBF	33.67 (±0.24)	33.34 (±0.31)	32.51 (±0.21)	**34.85 (±0.20)**	31.16 (±0.15)	33.59 (±0.20)	31.94 (±0.23)
		32.92 (±0.26)	31.04 (±0.12)	31.30 (±0.13)	33.27 (±0.32)	32.86 (±0.21)	32.84 (±0.19)	32.26 (±0.25)
	SINE	92.00 (±0.09)	90.62 (±0.20)	89.60 (±0.07)	86.80 (±0.12)	90.38 (±0.15)	91.89 (±0.09)	88.88 (±0.08)
		91.98 (±0.09)	92.00 (±0.10)	**92.04 (±0.08)**	91.92 (±0.09)	92.02 (±0.09)	91.99 (±0.09)	91.93 (±0.10)
	WAVEF.	79.37 (±0.25)	79.41 (±0.17)	79.49 (±0.12)	79.09 (±0.13)	**80.10 (±0.10)**	79.46 (±0.17)	79.70 (±0.10)
		79.47 (±0.12)	79.95 (±0.10)	79.67 (±0.14)	79.28 (±0.21)	79.57 (±0.17)	79.81 (±0.13)	79.47 (±0.16)

The description of **RQ1** was: *What are the best drift detectors in terms of accuracy in abrupt and gradual concept drift datasets?*

Based on the reported experiments, the answer to **RQ1** is: even though there were variations in the results using the two base learners (NB and HT), as well as in the datasets configured with abrupt and gradual concept drifts, the overall best two concept drift detectors in terms of accuracy were clearly *$HDDM_A$* and *RDDM*, followed by FHDDM, FTDD, and WSTD. Fig. 7 corroborates this answer; it captures the evaluation of the accuracy results of the methods aggregating all the tests performed using both base classifiers.

Table 9
Mean accuracies of Drift Detectors in percentage (%) in the gradual datasets, with a 95% confidence interval, using HT (Part 2).

DS Type and Size	DATASET	DDM HDDM$_A$	EDDM HDDM$_W$	ADWIN FHDDM	ECDD FTDD	STEPD RDDM$_{30}$	SeqDrift2 RDDM	SEED WSTD
	AGRAW$_1$	77.15 (±1.11)	73.27 (±2.14)	66.91 (±0.12)	66.74 (±0.52)	68.48 (±0.32)	71.40 (±0.72)	65.90 (±0.09)
		78.14 (±0.89)	74.82 (±0.23)	75.45 (±0.07)	77.36 (±1.46)	**79.24 (±0.97)**	78.12 (±0.84)	77.66 (±1.36)
	AGRAW$_2$	88.19 (±0.87)	75.48 (±2.82)	86.83 (±0.09)	84.79 (±0.04)	87.38 (±0.14)	**89.22 (±0.11)**	86.34 (±0.10)
		89.20 (±0.08)	89.08 (±0.04)	89.12 (±0.06)	88.63 (±0.50)	88.63 (±0.20)	88.74 (±0.17)	89.14 (±0.07)
	LED	71.62 (±0.86)	71.35 (±0.22)	65.48 (±0.45)	69.12 (±0.15)	69.43 (±0.39)	70.52 (±0.45)	57.92 (±1.00)
		73.53 (±0.08)	72.30 (±0.26)	72.88 (±0.15)	73.53 (±0.10)	73.33 (±0.13)	**73.57 (±0.09)**	73.19 (±0.04)
GRAD.	**MIXED**	94.76 (±0.05)	93.66 (±0.11)	92.06 (±0.06)	89.76 (±0.08)	91.46 (±0.13)	94.70 (±0.04)	91.79 (±0.19)
500K		94.76 (±0.04)	94.76 (±0.04)	**94.78 (±0.05)**	94.69 (±0.04)	94.35 (±0.15)	94.01 (±0.14)	94.72 (±0.05)
	RandRBF	36.05 (±0.63)	36.94 (±0.30)	32.61 (±0.22)	**38.19 (±0.09)**	31.18 (±0.11)	35.59 (±0.53)	32.19 (±0.22)
		34.26 (±0.29)	31.00 (±0.05)	31.42 (±0.08)	35.42 (±0.36)	34.52 (±0.37)	33.98 (±0.28)	32.44 (±0.14)
	SINE	95.57 (±0.15)	94.62 (±0.17)	90.55 (±0.06)	87.14 (±0.08)	91.92 (±0.53)	95.56 (±0.14)	89.78 (±0.09)
		95.52 (±0.19)	95.58 (±0.17)	**95.61 (±0.17)**	95.55 (±0.16)	95.31 (±0.22)	94.89 (±0.32)	95.56 (±0.17)
	WAVEF.	81.58 (±0.19)	81.62 (±0.08)	79.86 (±0.13)	79.17 (±0.11)	80.19 (±0.10)	**81.76 (±0.21)**	79.82 (±0.14)
		81.09 (±0.24)	80.05 (±0.11)	80.46 (±0.40)	81.55 (±0.19)	80.63 (±0.21)	80.18 (±0.20)	80.62 (±0.34)
	AGRAW$_1$	76.14 (±3.93)	71.82 (±3.74)	66.95 (±0.09)	66.92 (±0.30)	68.33 (±0.29)	72.03 (±0.75)	65.96 (±0.12)
		79.46 (±0.93)	75.78 (±0.16)	76.34 (±0.19)	78.78 (±1.34)	79.05 (±0.60)	78.95 (±0.60)	79.34 (±1.06)
	AGRAW$_2$	88.45 (±0.79)	79.85 (±3.52)	86.90 (±0.09)	84.84 (±0.03)	87.49 (±0.14)	89.55 (±0.18)	86.54 (±0.05)
		89.62 (±0.06)	89.44 (±0.03)	89.49 (±0.03)	89.33 (±0.19)	88.80 (±0.30)	88.81 (±0.19)	89.47 (±0.05)
	LED	70.80 (±0.98)	71.27 (±0.42)	65.73 (±0.30)	69.21 (±0.14)	69.84 (±0.36)	70.69 (±0.28)	58.06 (±0.57)
		73.58 (±0.06)	72.43 (±0.14)	72.96 (±0.11)	73.60 (±0.06)	73.34 (±0.06)	**73.61 (±0.06)**	73.30 (±0.08)
GRAD.	**MIXED**	**95.93 (±0.12)**	95.07 (±0.12)	92.16 (±0.05)	89.86 (±0.05)	91.45 (±0.09)	95.87 (±0.08)	91.92 (±0.09)
1M		**95.93 (±0.10)**	95.90 (±0.11)	95.92 (±0.13)	95.89 (±0.11)	94.92 (±0.32)	94.38 (±0.24)	95.86 (±0.09)
	RandRBF	36.73 (±0.45)	**38.38 (±0.11)**	32.63 (±0.14)	38.19 (±0.09)	31.13 (±0.07)	35.65 (±0.48)	32.25 (±0.16)
		34.86 (±0.21)	31.00 (±0.04)	31.38 (±0.06)	36.15 (±0.45)	35.01 (±0.25)	34.22 (±0.31)	32.46 (±0.12)
	SINE	96.87 (±0.11)	96.35 (±0.10)	90.60 (±0.03)	87.19 (±0.04)	92.12 (±0.46)	96.90 (±0.09)	89.85 (±0.07)
		96.87 (±0.08)	96.90 (±0.12)	96.86 (±0.12)	**96.92 (±0.11)**	95.88 (±0.31)	95.46 (±0.29)	96.84 (±0.12)
	WAVEF.	82.48 (±0.10)	82.33 (±0.09)	79.84 (±0.09)	79.18 (±0.08)	80.23 (±0.06)	**82.52 (±0.12)**	79.87 (±0.14)
		81.86 (±0.18)	80.13 (±0.08)	81.21 (±0.55)	82.49 (±0.15)	81.00 (±0.27)	80.31 (±0.14)	81.28 (±0.35)
	AGRAW$_1$	80.72 (±1.80)	75.14 (±3.75)	67.01 (±0.05)	66.99 (±0.27)	68.59 (±0.20)	73.52 (±0.80)	65.99 (±0.05)
		80.46 (±1.11)	76.28 (±0.07)	78.08 (±1.01)	80.78 (±1.23)	79.55 (±0.62)	79.53 (±0.59)	**81.02 (±0.87)**
	AGRAW$_2$	89.06 (±0.87)	82.78 (±2.00)	87.01 (±0.09)	84.87 (±0.02)	87.65 (±0.09)	**89.95 (±0.10)**	86.60 (±0.04)
		89.92 (±0.04)	89.71 (±0.04)	89.79 (±0.02)	89.70 (±0.19)	88.88 (±0.27)	89.10 (±0.06)	89.69 (±0.13)
	LED	70.31 (±0.95)	71.47 (±0.08)	65.81 (±0.19)	69.29 (±0.07)	69.98 (±0.25)	70.93 (±0.13)	58.70 (±0.42)
		73.64 (±0.03)	72.49 (±0.08)	73.05 (±0.09)	**73.69 (±0.03)**	73.45 (±0.06)	73.65 (±0.04)	73.43 (±0.06)
GRAD.	**MIXED**	**96.91 (±0.09)**	96.10 (±0.07)	92.20 (±0.05)	89.87 (±0.04)	91.43 (±0.04)	96.83 (±0.08)	92.00 (±0.11)
2M		96.86 (±0.08)	96.89 (±0.09)	96.89 (±0.09)	96.86 (±0.09)	94.94 (±0.14)	94.55 (±0.16)	96.85 (±0.09)
	RandRBF	37.44 (±0.13)	**39.10 (±0.07)**	32.69 (±0.09)	38.60 (±0.07)	31.13 (±0.07)	36.26 (±0.33)	32.21 (±0.08)
		35.67 (±0.27)	30.99 (±0.05)	31.36 (±0.04)	36.48 (±0.27)	35.41 (±0.44)	34.34 (±0.15)	32.46 (±0.11)
	SINE	97.97 (±0.06)	97.59 (±0.06)	90.66 (±0.04)	87.21 (±0.02)	92.12 (±0.28)	97.98 (±0.07)	89.84 (±0.05)
		97.93 (±0.07)	97.99 (±0.06)	97.98 (±0.05)	97.99 (±0.06)	96.24 (±0.19)	95.71 (±0.11)	**98.00 (±0.06)**
	WAVEF.	83.17 (±0.08)	82.88 (±0.03)	79.89 (±0.05)	79.20 (±0.04)	80.27 (±0.04)	83.08 (±0.08)	79.96 (±0.06)
		82.46 (±0.07)	80.15 (±0.04)	82.16 (±0.42)	**83.30 (±0.07)**	81.12 (±0.16)	80.32 (±0.11)	81.45 (±0.36)
HT	**RANK**	6.35714	8.20408	10.93880	11.12240	9.69388	7.33673	11.38780
GRAD.		**4.56122**	6.00000	5.17347	6.53061	5.68367	5.02041	6.98980
HT	**RANK**	6.93878	8.75510	10.88780	11.21430	9.40816	7.80612	11.27550
ALL		**4.22449**	5.79592	5.23980	5.72959	6.11735	5.22959	6.37755

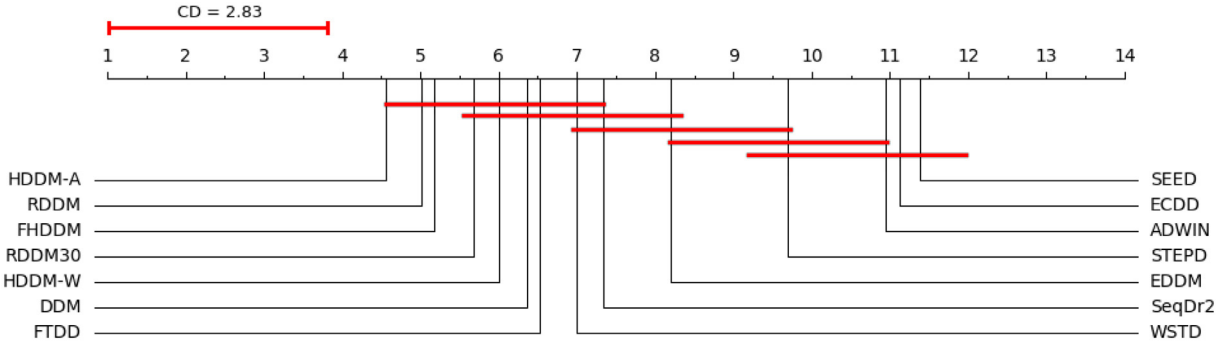

Fig. 5. Comparison results using the Nemenyi test of Detectors with HT in the gradual datasets with a 95% confidence interval.

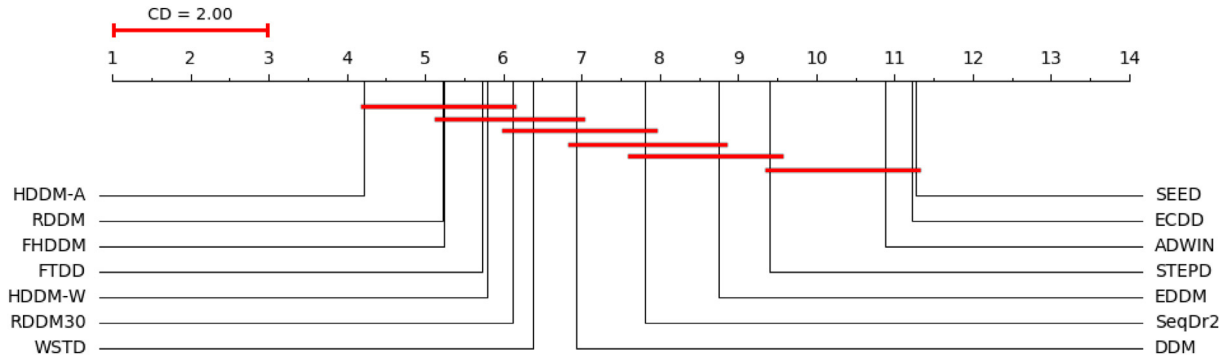

Fig. 6. Comparison results using the Nemenyi test of Detectors with HT in all artificial datasets with a 95% confidence interval.

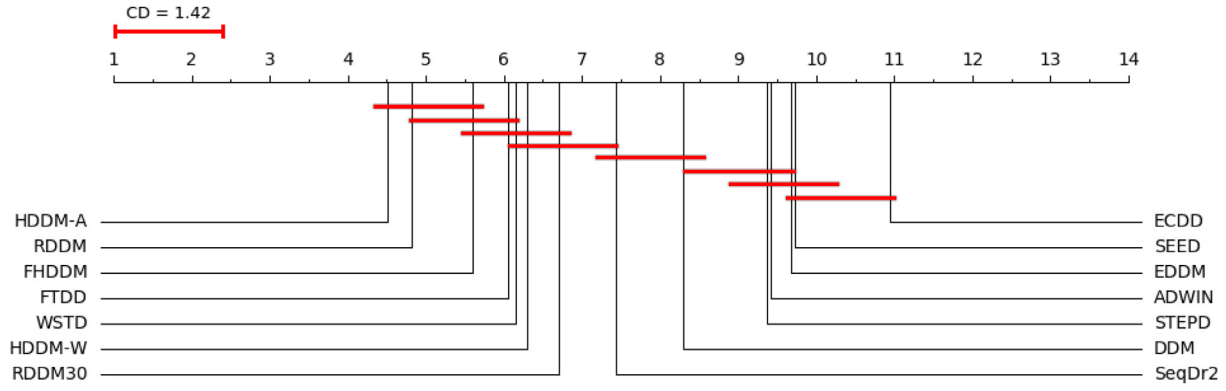

Fig. 7. Comparison of accuracy results using the Nemenyi test of Detectors including all tested datasets with a 95% confidence interval.

Another point to notice regards the performance dependency of some of the concept drift detectors on the chosen classifier (NB or HT). SEED and ADWIN seem to perform much better with NB than with HT, whereas DDM and EDDM are the opposite. In common, none of these methods delivered particularly good performances with any of the tested base classifiers.

It is worth adding that, despite the best two methods *not* showing such performance dependency, the highest differences in ranks between them seem to be due to the choice of base learner, with RDDM being slightly better than HDDM_A when using NB whereas HDDM_A was better with HT. In both cases, these differences were statistically indistinguishable between them and also to some other methods.

5. Drift detections results and analysis

The last section analyzed the results of the experiments based on the accuracy performance of the tested concept drift detectors. Analyzing the concept drift identifications of the methods can provide a different perspective on their performances.

For each *abrupt* dataset configuration, considering the number of repetitions adopted in the experiments, the mean distance (μD) to the exact drift positions of the true positive concept drift detections and the total number of false negatives (FN), false positives (FP), true negatives (TN), and true positives (TP) of each method were recorded.

The drifts detected within 2% of the concept size after the correct drift positions were computed as true positives, the same procedure adopted in [4–6]. For instance, in the 500K datasets, the concepts last for 100K instances and, thus, detections occurred in up to 2K instances after the perfect points were considered true positives.

It is worth emphasizing that this analysis includes only the *abrupt* datasets because the exact positions of the concept drifts are known. In the gradual drifts datasets, there are no single change points and it is therefore not clear when the drift identifications should be considered as positive or negative.

In addition, the results regarding the evaluation of the methods using *Precision* and *Recall* [41], as well as MCC [34], were also calculated. In all of them, higher values mean that the corresponding methods delivered better results. *Precision*, defined as $TP \mid (TP + FP)$, returns the proportion of predicted concept drifts that are existing drifts, whereas *Recall*, given by $TP \mid (TP + FN)$, represents the proportion of the existing drifts that were correctly detected by each detection method.

The MCC criterion is defined in Eq. (1) and was included because many other criteria "are severely influenced by the imbalance ratio between the numbers of positive and negative samples" [32]. It returns values in the [-1, 1] interval and is based on the four values of the confusion matrix: TP, TN, FP, and FN.

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP) \times (TP+FN) \times (TN+FP) \times (TN+FN)}} \quad (1)$$

Tables 10 and 11 summarize the *mean* concept drift identifications of the 14 tested configurations of the methods using NB as base learner, aggregating the results of different datasets by size. Notice that, in these aggregations, the mean distance

Table 10
Detectors mean drift identifications in abrupt datasets using NB (Part 1).

Size	Detector	μD	FN	FP	TN	TP	Precision	Recall	MCC
	DDM	N/A	91.29	79.43	299801	28.71	0.21536081	0.23928571	0.22599404
	EDDM	21.02	97.86	333.86	299546	22.14	0.05670214	0.18452381	0.09936743
	ADWIN	18.09	97.71	283.14	299597	22.29	0.08013609	0.18571429	0.12069909
	ECDD	**15.98**	51.43	343.57	299536	68.57	0.15696841	0.57142857	0.29493260
	STEPD	21.39	54.29	244.14	299636	65.71	0.29938773	0.54761905	0.38919120
	SeqDr2	N/A	110.71	244.14	299636	9.29	0.01935747	0.07738095	0.03416036
10K	SEED	27.62	76.86	327.71	299552	43.14	0.14147592	0.35952381	0.22242970
	$HDDM_A$	25.19	66.14	42.14	299838	53.86	0.48893070	0.44880952	0.46626826
	$HDDM_W$	23.01	**50.00**	62.57	299817	**70.00**	0.55023945	**0.58333333**	**0.56203330**
	FHDDM	25.37	55.29	50.57	299829	64.71	**0.55408500**	0.53928571	0.54095336
	FTDD	26.75	68.00	**30.43**	**299850**	52.00	0.47492104	0.43333333	0.44750519
	$RDDM_{30}$	28.90	92.43	65.14	299815	27.57	0.23239479	0.22976190	0.23032441
	RDDM	29.04	78.43	76.14	299804	41.57	0.32076519	0.34642857	0.33190985
	WSTD	25.91	58.57	53.86	299826	61.43	0.50611196	0.51190476	0.49968641
	DDM	N/A	86.86	109.29	599771	33.14	0.21034528	0.27619048	0.23856965
	EDDM	46.20	94.14	426.71	599453	25.86	0.05489346	0.21547619	0.10528560
	ADWIN	52.00	68.71	397.86	599482	51.29	0.16843144	0.42738095	0.26113942
	ECDD	**19.49**	46.43	728.00	599152	73.57	0.08446048	0.61309524	0.22482403
	STEPD	29.91	41.29	360.00	599520	78.71	0.24276765	0.65595238	0.38613255
	SeqDr2	N/A	112.00	344.57	599535	8.00	0.00588235	0.06666667	0.01952594
20K	SEED	49.09	60.57	550.43	599330	59.43	0.17514522	0.49523810	0.28565994
	$HDDM_A$	38.68	56.00	43.00	599837	64.00	0.55897106	0.53333333	0.54434613
	$HDDM_W$	30.58	**36.00**	104.14	599776	**84.00**	0.55438053	**0.70000000**	**0.61134384**
	FHDDM	32.41	43.29	73.29	599807	76.71	**0.58254900**	0.63928571	0.60202648
	FTDD	35.42	60.00	**30.29**	**599850**	60.00	0.54729776	0.50000000	0.51680121
	$RDDM_{30}$	48.46	85.43	73.14	599807	34.57	0.29776399	0.28809524	0.29254773
	RDDM	51.76	68.43	84.71	599795	51.57	0.36887128	0.42976190	0.39678713
	WSTD	30.34	47.71	57.43	599823	72.29	0.55316171	0.60238095	0.56839378
	DDM	120.97	76.00	131.71	1499748	44.00	0.24898461	0.36666667	0.29517924
	EDDM	74.28	97.14	542.57	1499337	22.86	0.03567264	0.19047619	0.07907518
	ADWIN	75.60	47.43	665.57	1499214	72.57	0.22125501	0.60476190	0.35153168
	ECDD	**31.05**	38.86	1824.43	1498055	81.14	0.03974312	0.67619048	0.16182681
	STEPD	42.87	34.57	740.43	1499140	85.43	0.13586017	0.71190476	0.30204413
	SeqDr2	193.88	32.14	415.57	1499464	87.86	0.36924018	0.73214286	0.49002033
50K	SEED	76.53	39.43	1124.29	1498756	80.57	0.23643196	0.67142857	0.38215474
	$HDDM_A$	79.65	42.14	41.29	1499839	77.86	**0.64212082**	0.64880952	**0.64422746**
	$HDDM_W$	45.39	**26.00**	267.71	1499612	**94.00**	0.44454349	**0.78333333**	0.56050015
	FHDDM	49.34	29.86	172.86	1499707	90.14	0.52275883	0.75119048	0.60283718
	FTDD	52.48	49.86	**32.00**	**1499848**	70.14	0.61925646	0.58452381	0.59810310
	$RDDM_{30}$	122.12	72.14	78.71	1499801	47.86	0.37626047	0.39880952	0.38685027
	RDDM	102.97	52.43	88.29	1499792	67.57	0.46634617	0.56309524	0.51008267
	WSTD	46.92	36.71	87.14	1499793	83.29	0.55822341	0.69404762	0.61091974
	DDM	204.85	72.14	174.71	2999705	47.86	0.25723574	0.39880952	0.30860128
	EDDM	202.43	98.43	601.86	2999278	21.57	0.02897979	0.17976190	0.06905262
	ADWIN	131.62	31.71	1122.14	2998758	88.29	0.25343920	0.73571429	0.41092081
	ECDD	**46.53**	35.71	3754.57	2996125	84.29	0.02027861	0.70238095	0.11780856
	STEPD	65.74	30.86	1398.86	2998481	89.14	0.07710190	0.74285714	0.23292182
	SeqDr2	211.75	25.86	546.29	2999334	94.14	0.37590183	0.78452381	0.51138103
100K	SEED	106.42	**21.29**	2034.86	2997845	**98.71**	0.22961983	**0.82261905**	0.40957478
	$HDDM_A$	134.13	33.00	44.71	2999835	87.00	**0.66887949**	0.72500000	**0.69497828**
	$HDDM_W$	56.88	22.71	538.57	2999341	97.29	0.37526402	0.81071429	0.50084294
	FHDDM	64.47	26.14	350.57	2999529	93.86	0.44972252	0.78214286	0.55133365
	FTDD	92.38	42.14	**35.00**	**2999845**	77.86	0.65231578	0.64880952	0.64748698
	$RDDM_{30}$	197.97	65.14	85.57	2999794	54.86	0.41192585	0.45714286	0.43368141
	RDDM	174.91	41.14	132.14	2999748	78.86	0.44586342	0.65714286	0.53468070
	WSTD	66.78	31.29	134.86	2999745	88.71	0.52532686	0.73928571	0.60647479

Table 11
Detectors mean drift identifications in abrupt datasets using NB (Part 2).

Size	Detector	μD	FN	FP	TN	TP	Precision	Recall	MCC
	DDM	741.68	18.57	69.57	4999890	21.43	0.34907419	0.53571429	0.41338053
	EDDM	N/A	34.57	209.29	4999751	5.43	0.02176624	0.13571429	0.05277863
	ADWIN	247.38	3.57	1230.86	4998729	36.43	0.23485506	0.91071429	0.42346639
	ECDD	192.27	7.14	6329.29	4993631	32.86	0.00485784	0.82142857	0.06231523
	STEPD	194.68	5.86	2259.29	4997701	34.14	0.01892386	0.85357143	0.12385034
	SeqDr2	272.54	6.00	174.43	4999786	34.00	0.36859607	0.85000000	0.53225076
500K	SEED	192.94	**3.14**	2925.43	4997035	**36.86**	0.14569687	**0.92142857**	0.33424945
	$HDDM_A$	361.69	5.14	23.71	4999936	34.86	**0.65789220**	0.87142857	**0.75179353**
	$HDDM_W$	125.42	4.71	919.00	4999041	35.29	0.21741224	0.88214286	0.34484252
	FHDDM	167.86	5.57	585.29	4999375	34.43	0.32494310	0.86071429	0.43671224
	FTDD	276.01	10.00	**19.29**	**4999941**	30.00	0.61669329	0.75000000	0.67812226
	$RDDM_{30}$	575.60	12.86	127.29	4999833	27.14	0.17513516	0.67857143	0.34338293
	RDDM	327.37	5.86	236.14	4999724	34.14	0.13289249	0.85357143	0.33460783
	WSTD	**105.06**	8.57	197.71	4999762	31.43	0.33088496	0.78571429	0.46875125
	DDM	1553.96	16.14	64.71	9999895	23.86	0.38814302	0.59642857	0.45883160
	EDDM	2137.13	35.57	197.14	9999763	4.43	0.01696406	0.11071429	0.04190268
	ADWIN	336.89	1.86	2107.71	9997852	38.14	0.24230037	0.95357143	0.43113149
	ECDD	254.36	5.86	12628.71	9987331	34.14	0.00252568	0.85357143	0.04579650
	STEPD	368.54	4.14	4417.43	9995543	35.86	0.01017397	0.89642857	0.09322290
	SeqDr2	316.60	5.14	186.86	9999773	34.86	0.37357613	0.87142857	0.54299562
1M	SEED	244.02	**1.57**	4979.57	9994980	**38.43**	0.11111597	**0.96071429**	0.29664331
	$HDDM_A$	564.46	3.00	36.71	9999923	37.00	**0.58364139**	0.92500000	**0.72423195**
	$HDDM_W$	252.11	3.71	1846.57	9998113	36.29	0.18839435	0.90714286	0.29787809
	FHDDM	222.35	5.29	1190.86	9998769	34.71	0.28216304	0.86785714	0.38308980
	FTDD	452.36	8.14	**27.71**	**9999932**	31.86	0.55942488	0.79642857	0.66396386
	$RDDM_{30}$	1045.05	8.29	249.71	9999710	31.71	0.11466734	0.79285714	0.29940348
	RDDM	480.14	4.14	466.00	9999494	35.86	0.07599551	0.89642857	0.25917964
	WSTD	**194.02**	8.71	375.29	9999585	31.29	0.26597309	0.78214286	0.39900764
	DDM	2546.81	15.57	68.00	19999892	24.43	0.39353465	0.61071429	0.46604690
	EDDM	N/A	35.14	199.71	19999760	4.86	0.01909436	0.12142857	0.04610653
	ADWIN	380.57	**1.00**	2725.00	19997235	**39.00**	0.22769047	**0.97500000**	0.41922193
	ECDD	429.25	5.71	25271.71	19974688	34.29	0.00126957	0.85714286	0.03252877
	STEPD	683.54	**1.00**	8782.71	19991177	**39.00**	0.00551117	**0.97500000**	0.07163527
	SeqDr2	584.09	4.00	241.00	19999719	36.00	0.34956295	0.90000000	0.53356740
2M	SEED	**311.33**	1.71	8782.43	19991178	38.29	0.07732602	0.95714286	0.24387862
	$HDDM_A$	845.14	1.86	66.71	19999893	38.14	0.48050850	0.95357143	0.65521805
	$HDDM_W$	548.57	1.86	3670.00	19996290	38.14	0.16971330	0.95357143	0.26430638
	FHDDM	499.68	3.57	2368.57	19997591	36.43	0.27191729	0.91071429	0.36107458
	FTDD	659.11	5.43	**40.29**	**19999920**	34.57	**0.52839275**	0.86428571	**0.66775503**
	$RDDM_{30}$	1959.20	5.71	443.57	19999516	34.29	0.07520331	0.85714286	0.25165899
	RDDM	544.72	3.71	944.29	19999016	36.29	0.03906119	0.90714286	0.18692039
	WSTD	799.99	5.86	704.43	19999256	34.14	0.22829732	0.85357143	0.36352025
	DDM	1033.65	53.80	99.63	5771243	31.92	0.29466833	0.43197279	0.34380046
	EDDM	496.21	70.41	358.73	5770984	15.31	0.03343895	0.16258503	0.07050981
	ADWIN	177.45	36.00	1218.90	5770124	49.71	0.20401538	0.68469388	0.34544440
	ECDD	**141.28**	27.31	7268.61	5764074	58.41	0.04430053	0.72789116	0.13429036
	STEPD	200.95	24.57	2600.41	5768742	61.14	0.11281807	0.76904762	0.22842831
	SeqDr2	315.77	42.27	307.55	5771035	43.45	0.26601671	0.61173469	0.38055735
Mean	SEED	143.99	29.22	2960.67	5768382	56.49	0.15954454	0.74115646	0.31065579
	$HDDM_A$	292.70	29.61	42.61	5771300	56.10	**0.58299202**	0.72942177	**0.64015195**
	$HDDM_W$	154.57	**20.71**	1058.37	5770284	**65.00**	0.35713534	**0.80289116**	0.44882103
	FHDDM	151.64	24.14	684.57	5770658	61.57	0.42687697	0.76445578	0.49686104
	FTDD	227.79	34.80	**30.71**	**5771312**	50.92	0.57118599	0.65391156	0.60281966
	$RDDM_{30}$	568.18	48.86	160.45	5771182	36.86	0.24047870	0.52891156	0.31969275
	RDDM	244.42	36.31	289.67	5771053	49.41	0.26425646	0.66479592	0.36488117
	WSTD	181.29	28.20	230.10	5771113	57.51	0.42399704	0.70986395	0.50239341

was calculated only when the corresponding method detected at least one TP in at least five of the seven datasets considered in this procedure. The reason for the aggregation was the overwhelming amount of results. Finally, in each dataset size, the best results are shown in **bold**.

It is worth pointing out that the numbers of the TN and TP detections could also be easily calculated. Given *rep* is the number of repetitions of the experiments, $TN = (size - 4) \times rep - FP$ and $TP = 4 \times rep - FN$.

Tables 12 and 13, below, are similar to Tables 10 and 11 except that they detail the *mean* results of the 14 tested methods using HT as base learner, instead of NB. Once again, in each dataset size, the best results are shown in **bold**. Accordingly, the results of different dataset generators were aggregated by their sizes and using the same criteria.

Table 12
Detectors mean drift identifications in abrupt datasets using HT (Part 1).

Size	Detector	μD	FN	FP	TN	TP	Precision	Recall	MCC
	DDM	N/A	90.00	78.57	299801	30.00	0.20602088	0.25000000	0.22516814
	EDDM	21.06	99.29	321.14	299559	20.71	0.05625867	0.17261905	0.09638524
	ADWIN	25.28	96.86	262.71	299617	23.14	0.08148672	0.19285714	0.12407030
	ECDD	**15.63**	50.00	344.14	299536	70.00	0.15983688	0.58333333	0.30080832
	STEPD	21.49	54.29	252.71	299627	65.71	0.29195356	0.54761905	0.38530818
	SeqDr2	N/A	108.43	246.57	299633	11.57	0.03061970	0.09642857	0.04812685
10K	SEED	24.38	78.00	325.86	299554	42.00	0.13413172	0.35000000	0.21381496
	$HDDM_A$	23.97	65.29	41.14	299839	54.71	0.48818199	0.45595238	0.46980016
	$HDDM_W$	23.16	**49.86**	66.00	299814	**70.14**	0.54903187	**0.58452381**	**0.56110853**
	FHDDM	25.63	54.29	52.57	299827	65.71	**0.55272244**	0.54761905	0.54322347
	FTDD	26.86	67.57	**36.43**	**299844**	52.43	0.46170401	0.43690476	0.44347415
	$RDDM_{30}$	N/A	89.00	56.86	299823	31.00	0.26621177	0.25833333	0.26137608
	RDDM	28.85	77.71	78.14	299802	42.29	0.30434020	0.35238095	0.32450253
	WSTD	25.66	57.86	60.57	299819	62.14	0.50347745	0.51785714	0.50034705
	DDM	N/A	82.14	101.14	599779	37.86	0.21145381	0.31547619	0.25666822
	EDDM	45.54	92.00	401.00	599479	28.00	0.06657544	0.23333333	0.12229883
	ADWIN	47.90	67.14	388.14	599492	52.86	0.15184113	0.44047619	0.25347437
	ECDD	**19.33**	45.43	741.57	599138	74.57	0.08457360	0.62142857	0.22632111
	STEPD	28.64	43.57	397.14	599483	76.43	0.20534792	0.63690476	0.35203806
	SeqDr2	N/A	111.14	342.86	599537	8.86	0.01086303	0.07380952	0.02609195
20K	SEED	47.76	61.14	596.00	599284	58.86	0.14657153	0.49047619	0.25929747
	$HDDM_A$	38.17	51.29	37.29	**599843**	68.71	**0.58967891**	0.57261905	0.57958938
	$HDDM_W$	31.81	**34.00**	104.71	599775	**86.00**	0.55281674	**0.71666667**	**0.61815446**
	FHDDM	32.86	41.14	76.86	599803	78.86	0.56889366	0.65714286	0.60314751
	FTDD	35.08	59.00	**37.14**	**599843**	61.00	0.52160225	0.50833333	0.51022193
	$RDDM_{30}$	48.08	86.57	68.71	599811	33.43	0.30325164	0.27857143	0.29012274
	RDDM	39.75	70.43	94.43	599786	49.57	0.31092582	0.41309524	0.35500303
	WSTD	30.66	48.43	72.29	599808	71.57	0.51587105	0.59642857	0.54532241
	DDM	107.53	67.43	133.57	1499746	52.57	0.27091152	0.43809524	0.33454889
	EDDM	83.04	100.29	485.29	1499395	19.71	0.03218228	0.16428571	0.07103979
	ADWIN	86.16	51.29	770.29	1499110	68.71	0.13514842	0.57261905	0.26843579
	ECDD	**30.05**	39.00	1848.43	1498032	81.00	0.03945008	0.67500000	0.16096799
	STEPD	41.33	35.43	814.57	1499065	84.57	0.11898320	0.70476190	0.28225088
	SeqDr2	192.12	39.14	447.57	1499432	80.86	0.31032494	0.67380952	0.42380236
50K	SEED	77.97	44.71	1386.43	1498494	75.29	0.10936499	0.62738095	0.24713071
	$HDDM_A$	66.13	39.57	**36.29**	**1499844**	80.43	**0.64264792**	0.67023810	**0.65515935**
	$HDDM_W$	42.44	**28.29**	263.43	1499617	**91.71**	0.42950086	**0.76428571**	0.54419086
	FHDDM	48.61	32.57	172.29	1499708	87.43	0.47227060	0.72857143	0.56821663
	FTDD	47.57	49.86	45.00	1499835	70.14	0.55544102	0.58452381	0.56600695
	$RDDM_{30}$	119.54	66.00	74.00	1499806	54.00	0.40963105	0.45000000	0.42883322
	RDDM	94.89	46.43	107.71	1499772	73.57	0.40462797	0.61309524	0.49456281
	WSTD	42.18	38.43	114.00	1499766	81.57	0.47206154	0.67976190	0.55482955
	DDM	167.74	62.86	156.00	2999724	57.14	0.28929211	0.47619048	0.35961077
	EDDM	189.17	101.57	513.29	2999367	18.43	0.02527367	0.15357143	0.06105986
	ADWIN	126.12	37.71	1429.71	2998450	82.29	0.10589051	0.68571429	0.25563339
	ECDD	**46.70**	34.43	3747.57	2996132	85.57	0.02058745	0.71309524	0.11962575
	STEPD	64.14	32.29	1491.71	2998388	87.71	0.06960340	0.73095238	0.22027472
	SeqDr2	205.43	31.14	607.43	2999273	88.86	0.29852841	0.74047619	0.43508194
100K	SEED	112.06	30.00	2722.57	2997157	90.00	0.07804062	0.75000000	0.22586651
	$HDDM_A$	103.87	32.71	**41.86**	**2999838**	87.29	**0.64862699**	0.72738095	**0.68563483**
	$HDDM_W$	52.51	**23.29**	525.00	2999355	**96.71**	0.37004855	**0.80595238**	0.49394895
	FHDDM	64.96	25.43	345.71	2999534	94.57	0.41494827	0.78809524	0.53325143
	FTDD	67.67	40.14	65.00	2999815	79.86	0.56212360	0.66547619	0.60392649
	$RDDM_{30}$	194.27	58.29	86.71	2999793	61.71	0.45083724	0.51428571	0.48094653
	RDDM	156.54	37.00	144.57	2999735	83.00	0.41384667	0.69166667	0.52781647
	WSTD	61.17	31.86	176.14	2999704	88.14	0.44398637	0.73452381	0.55349065

Considering the mean distance of the true positive detections, ECDD and STEPD achieved the best results in most tested datasets, especially in the lower sizes (up to 100K). However, these good results often came at the cost of many FP detections, impairing their accuracies, and this phenomenon was more severe in the case of ECDD. The other good methods in this metric were FHDDM, WSTD and $HDDM_W$: their results were usually close to those of the previous two methods and were often the best in the larger datasets.

Regarding the false negatives (and consequently true positives), several methods presented reasonably similar results. In no particular order, the best methods were ECDD, STEPD, SEED, $HDDM_A$, $HDDM_W$, FHDDM, and WSTD. In the larger datasets (from 100K), ADWIN and $SEQDRIFT_2$ also returned good results in this metric.

Table 13
Detectors mean drift identifications in abrupt datasets using HT (Part 2).

Size	Detector	μD	FN	FP	TN	TP	Precision	Recall	MCC
	DDM	756.68	18.43	57.43	4999903	21.57	0.30949461	0.53928571	0.39816468
	EDDM	1164.43	35.71	162.43	4999798	4.29	0.01818983	0.10714286	0.04237768
	ADWIN	268.07	6.29	2334.14	4997626	33.71	0.03257126	0.84285714	0.15245273
	ECDD	179.70	7.00	6281.57	4993678	33.00	0.00488386	0.82500000	0.06263832
	STEPD	211.14	5.57	2252.29	4997708	34.43	0.02029965	0.86071429	0.12795476
	SeqDr2	312.02	7.29	606.71	4999353	32.71	0.29114818	0.81785714	0.43796389
500K	SEED	213.45	4.86	4311.00	4995649	35.14	0.02444516	0.87857143	0.12122201
	$HDDM_A$	242.10	**3.43**	**43.00**	**4999917**	**36.57**	**0.59002590**	**0.91428571**	**0.71551232**
	$HDDM_W$	**129.87**	5.29	866.14	4999094	34.71	0.30080653	0.86785714	0.39234853
	FHDDM	140.46	5.57	549.00	4999411	34.43	0.32733381	0.86071429	0.41459078
	FTDD	249.79	10.43	60.29	4999900	29.57	0.48771276	0.73928571	0.57320856
	$RDDM_{30}$	488.13	10.71	103.86	4999856	29.29	0.25532757	0.73214286	0.42797920
	RDDM	306.54	6.14	208.86	4999751	33.86	0.16135976	0.84642857	0.36595654
	WSTD	163.13	8.71	234.29	4999726	31.29	0.34552041	0.78214286	0.46900111
	DDM	1486.87	16.43	**62.57**	**9999897**	23.57	0.31251823	0.58928571	0.41384020
	EDDM	N/A	35.71	171.43	9999789	4.29	0.01574987	0.10714286	0.04100340
	ADWIN	334.22	4.57	4515.86	9995444	35.43	0.02014000	0.88571429	0.12201692
	ECDD	226.67	6.00	12525.86	9987434	34.00	0.00253749	0.85000000	0.04580970
	STEPD	396.59	3.00	4410.00	9995550	37.00	0.01110977	0.92500000	0.09813752
	SeqDr2	389.61	6.29	1041.86	9998918	33.71	0.27249993	0.84285714	0.42434076
1M	SEED	284.63	4.29	8338.14	9991622	35.71	0.01286284	0.89285714	0.09748752
	$HDDM_A$	332.52	**2.57**	84.29	9999876	**37.43**	**0.52774438**	**0.93571429**	**0.66750007**
	$HDDM_W$	305.61	3.57	1725.57	9998234	36.43	0.29418438	0.91071429	0.36827915
	FHDDM	**209.80**	5.43	1091.14	9998869	34.57	0.30763450	0.86428571	0.39083444
	FTDD	385.94	10.57	89.43	9999871	29.43	0.46406127	0.73571429	0.55293846
	$RDDM_{30}$	712.96	8.86	221.43	9999739	31.14	0.14667128	0.77857143	0.33303400
	RDDM	370.05	4.29	414.14	9999546	35.71	0.09276471	0.89285714	0.28465911
	WSTD	273.31	7.29	419.14	9999541	32.71	0.30933549	0.81785714	0.43285460
	DDM	2659.79	13.43	**69.43**	**19999891**	26.57	0.36061355	0.66428571	0.47449882
	EDDM	N/A	34.43	225.71	19999734	5.57	0.01413643	0.13928571	0.04435342
	ADWIN	498.41	4.29	8969.29	19990991	35.71	0.01034245	0.89285714	0.08766960
	ECDD	596.54	5.57	25052.86	19974907	34.43	0.00842336	0.86071429	0.03772006
	STEPD	620.27	**1.00**	8723.57	19991236	**39.00**	0.00585089	**0.97500000**	0.07306538
	SeqDr2	690.25	4.29	2157.43	19997803	35.71	0.26243320	0.89285714	0.41923402
2M	SEED	494.84	3.43	16532.72	19983427	36.57	0.00665489	0.91428571	0.07139668
	$HDDM_A$	429.85	2.29	162.43	19999798	37.71	0.44575759	0.94285714	**0.59581415**
	$HDDM_W$	547.97	1.43	3431.86	19996528	38.57	0.28925720	0.96428571	0.35113726
	FHDDM	**423.23**	3.57	2104.57	19997855	36.43	0.30210304	0.91071429	0.37505288
	FTDD	784.17	8.57	159.86	19999800	31.43	**0.45487524**	0.78571429	0.55862944
	$RDDM_{30}$	1134.93	6.00	444.29	19999516	34.00	0.08411549	0.85000000	0.26343735
	RDDM	538.99	3.71	836.29	19999124	36.29	0.04796635	0.90714286	0.20607038
	WSTD	836.56	5.43	750.00	19999210	34.57	0.30167125	0.86428571	0.41457287
	DDM	1035.72	50.10	94.10	5771249	35.61	0.28004353	0.46751701	0.35178567
	EDDM	300.65	71.29	325.76	5771017	14.43	0.03262374	0.15391156	0.06835974
	ADWIN	198.02	38.31	2667.16	5768676	47.41	0.07677436	0.64472789	0.18053616
	ECDD	159.23	26.78	7220.29	5764123	58.94	0.04575610	0.73265306	0.13627018
	STEPD	197.66	25.02	2620.29	5768723	60.69	0.10330691	0.76870748	0.21986136
	SeqDr2	357.89	43.96	778.63	5770564	41.76	0.21091677	0.59115646	0.31637739
Mean	SEED	179.30	32.35	4887.53	5766455	53.37	0.07315311	0.70051020	0.17660226
	$HDDM_A$	176.66	28.16	**63.76**	**5771279**	57.55	**0.56180910**	0.74557823	**0.62414432**
	$HDDM_W$	161.91	**20.82**	997.53	5770345	**64.90**	0.39794945	**0.80204082**	0.47559539
	FHDDM	**135.08**	24.00	627.45	5770715	61.71	0.42084376	0.76530612	0.48975959
	FTDD	228.16	35.16	70.45	5771272	50.55	0.50107430	0.63656463	0.54405800
	$RDDM_{30}$	449.65	46.49	150.84	5771192	39.22	0.27372086	0.55170068	0.35510416
	RDDM	219.38	35.10	269.16	5771074	50.61	0.24797592	0.67380952	0.36551012
	WSTD	204.67	28.29	260.92	5771082	57.43	0.41313194	0.71326531	0.49577403

In both aforementioned metrics, the results of FTDD and RDDM were usually worse than the best results in each dataset but they were often reasonably close to them, especially in the larger datasets.

In the case of false positives (and consequently true negatives), FTDD, $HDDM_A$, and DDM were clearly the best methods. FTDD was a distinct winner in the tests using NB and in the very small datasets (10K and 20K) with HT, with $HDDM_A$ being second best most of the time. In the other datasets using HT, $HDDM_A$ and DDM split the best results. Despite being regularly inferior to the best three detectors in this metric, other configurations returned good results consistently, including $RDDM_{30}$, RDDM, and WSTD.

Analysing the results of *Precision*, the best methods were $HDDM_A$, FTDD, FHDDM, and WSTD. They provided the very best results in many scenarios and generally strong results in most other situations. $HDDM_W$ (and to some extent FHDDM) delivered very strong results in the smaller datasets but not such good results when the size of the datasets increased. $HDDM_W$ was also generally better using HT than using NB. On the other hand, the results of FTDD and DDM were the opposite, progressively stronger with the increase in size of the datasets and often better using NB than using HT. Finally, $RDDM_{30}$ and RDDM were rarely among the very best results, but were consistently among the best 40% configurations in most datasets.

In the case of *Recall*, the differences were reasonably small in the results of a fairly large proportion of the tested configurations with both base learners, but the best methods were $HDDM_W$, STEPD, FHDDM, ECDD, WSTD, $HDDM_A$, and SEED, whereas EDDM was clearly the worst.

To a considerable extent, the results of most configurations of the methods in the MCC criterion were directly related to their *Precision* results, which is probably a consequence of their close results in *Recall*. However, in MCC, $SEQDRIFT_2$ was much closer to the best methods than it was in *Precision*.

Proceeding to answer **RQ2**, its description can be repeated: *What are the best concept drift detectors in terms of detections, measured by* Precision, Recall, *and the MCC metric, in the abrupt datasets?*

Based on the experiments, the answer to **RQ2** is: although there were minor variations in the data regarding the two base learners (NB and HT), the best concept drift detectors overall in terms of detections of concept drifts were $HDDM_A$, FTDD, WSTD, FHDDM, and $HDDM_W$. RDDM and $SEQDRIFT_2$ were generally the next best methods but reasonably far behind the very best configurations in most datasets. Once again, the most well-known methods were rarely among the best configurations in the detections of concept drift.

Fig. 8 corroborates this answer; it captures the evaluation of the MCC results of the methods aggregating all the executed tests using both base classifiers.

To conclude this section, Table 14 summarizes the rank results of the 14 methods in all the evaluated metrics based on all the results previously presented. Once again, the best result in each metric is shown in **bold**.

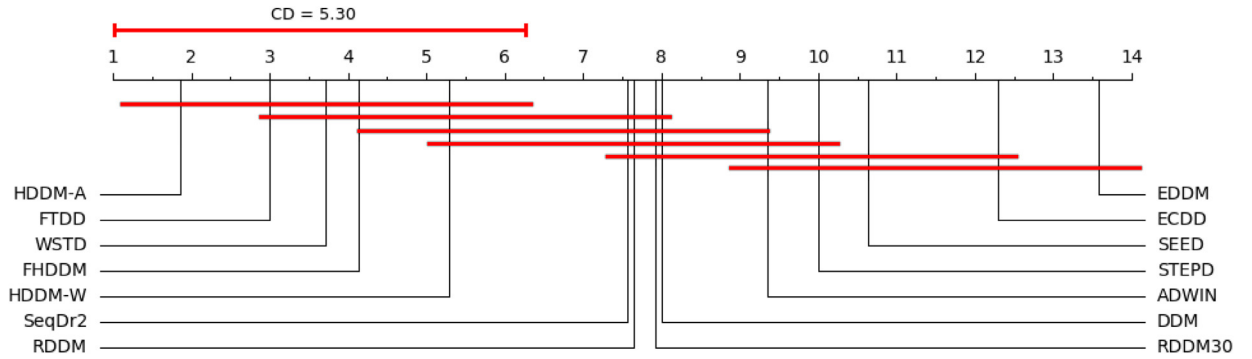

Fig. 8. Comparison of MCC results of Detectors using the Nemenyi test based on the results of Tables 10 to 13 with a 95% confidence interval.

Table 14
Mean rank of the methods in each metric presented in Tables 2 to 13.

Detector	Accuracy	μD	FN/TP	FP/TN	Precision	Recall	MCC
DDM	8.29592	12.82140	12.42860	4.78571	6.57143	12.42860	8.00000
EDDM	9.68112	10.42860	13.71430	8.28571	12.57140	13.71430	13.57140
ADWIN	9.42347	7.35714	7.57143	11.07140	9.85714	7.57143	9.35714
ECDD	10.94640	**2.35714**	6.75000	14.00000	12.85710	6.75000	12.28570
STEPD	9.37500	5.00000	3.60714	11.32140	11.14290	3.60714	10.00000
SeqDr2	7.43367	11.46430	8.60714	8.10714	8.50000	8.60714	7.57143
SEED	9.73214	6.42857	4.82143	12.85710	10.85710	4.82143	10.64290
$HDDM_A$	**4.50255**	7.57143	5.03571	1.85714	**1.64286**	5.03571	**1.85714**
$HDDM_W$	6.29847	3.71429	**2.03571**	8.57143	5.57143	**2.03571**	5.28571
FHDDM	5.59694	4.07143	4.14286	6.85714	3.57143	4.14286	4.14286
FTDD	6.04847	7.64286	9.64286	**1.50000**	2.64286	9.64286	3.00000
$RDDM_{30}$	6.69898	11.92860	11.39290	4.21429	7.42857	11.39290	7.92857
RDDM	4.81633	9.42857	7.89286	6.14286	7.64286	7.89286	7.64286
WSTD	6.15051	4.78571	7.35714	5.42857	4.14286	7.35714	3.71429

6. Additional research questions

This section examines and answers the remaining three research questions proposed, i.e., **RQ3**, **RQ4**, and **RQ5**.

The description of **RQ3** was: *Do the answers to **RQ1** and **RQ2** vary with the different dataset generators used in the experiments? If so, to what extent?*

As expected, the answer to **RQ3** regarding accuracies (**RQ1**) is *yes*; there were considerable differences in the results when the data of the different dataset generators were separated and this phenomenon was more severe in the tests using HT. However, in general, the best two methods (HDDM$_A$ and RDDM) delivered strong accuracy results in most dataset generators, the exception being randomRBF. In fact, the best methods in the randomRBF datasets were ECDD, EDDM, SEED, and SEQDRIFT$_2$, which did *not* present good results in most of the other datasets.

The answer to **RQ3** regarding the detections (**RQ2**) is also *yes*; there are variations in the results of different dataset generators, but these are much more limited than they were in the case of the accuracies. In general, there were numerous changes of order within the best five methods (HDDM$_A$, FTDD, WSTD, FHDDM, and HDDM$_W$), but they remained the best in most dataset generators. As expected, the most notable exception was again randomRBF: in these datasets, the best detections were those of SEQDRIFT$_2$, HDDM$_W$, SEED, HDDM$_A$, and RDDM.

The description of **RQ4** was: *Do the answers of **RQ1** and **RQ2** depend on the size of the concepts included in the datasets? If so, to what extent?*

The answer to **RQ4** regarding accuracies (**RQ1**) is once again *yes*; there were substantial differences in the results of some configurations when the datasets of different sizes were separated. Note SEQDRIFT$_2$ and FTDD are the ones most affected by this phenomenon and both of them improved their results dramatically with the increase in size of the datasets. On the other hand, the tendency for HDDM$_W$ and to some extent FHDDM was to consistently present worse ranks when the size of the datasets increased, though these variations were *not* nearly as large as those of SEQDRIFT$_2$ and FTDD.

The answer to **RQ4** regarding the detections (**RQ2**) is, one more time, *yes*; there were ample differences in the results of some methods when the datasets were separated by size. FTDD and, to a lesser extent, SEQDRIFT$_2$ again improved their detections when the size of the datasets increased, especially using NB as base classifier, whereas HDDM$_W$ became progressively worse.

Finally, the description of **RQ5** was: *In the same datasets, are the best methods of **RQ1** and **RQ2** the same? If so, to what extent?*

The answer is *no*. Looking exclusively at the results of the experiments using the abrupt datasets, it is clear that, in spite of some intersection between the best methods regarding their accuracies and detections, the very best methods are *not* exactly the same in the two criteria. The reason seems to be that false positive detections help to improve the accuracy results of some methods in many datasets, notably the two configurations of RDDM, rather than impairing them, provided the numbers are not too big. However, this issue obviously needs to be further investigated for a more conclusive answer.

7. Conclusion

This article presents an extensive empirical comparison and evaluation of concept drift detection methods and, as such, this work can also be regarded as a research review of existing concept drift detectors. As was to be expected, no single drift detector is better than all the others in all situations.

Further, this article set out to verify/challenge a common belief in the area, namely that the best concept drift detectors are necessarily the ones that detect all the existing concept drifts closest to their correct positions, ideally detecting only them. In addition, to analyze this issue, this article introduced and answered *five* research questions.

Moreover, to answer these research questions, a large-scale set of experiments was carried out to evaluate and compare 14 configurations of concept drift detection methods in a fully supervised setting. The concept drift detectors were compared in terms of both their final accuracies and the quality of their detections of concept drift, and the results were the basis for answering the proposed research questions. Note that the experiments reported here were based on and extended the ones reported by Barros [4].

It is worth mentioning that these large-scale experiments were run in the MOA framework [10], release 2014.11, using a very large number of artificial datasets, with *abrupt* and *gradual* concept drift versions of several sizes. Furthermore, these experiments were executed using two different base classifiers, namely Naive Bayes (NB) and Hoeffding Tree (HT). To the best of our knowledge, this is the largest comparison evaluation ever reported in the area of data stream mining.

The results of these large-scale experiments provide explicit indications of the best concept drift detectors, in terms of accuracy and detection. They also laid the basis for making it possible to analyze the influence of the *type of concept drift*, the *dataset generators*, and the *size of the concepts* on the performance of the methods.

It should be emphasized that HDDM$_A$ and RDDM were consistently among the very best concept drift detection configurations in terms of *accuracy* with both base learners. To some extent, this is also true for FHDDM and FTDD. Regarding the quality of their detections of concept drifts, the top methods were HDDM$_A$, FTDD, WSTD, FHDDM, and HDDM$_W$. In addition, it was observed that the most well-known and cited drift detectors, namely DDM, EDDM, ADWIN, ECDD, and STEPD, as well as SEED, were consistently among the worst configurations in both evaluations.

To conclude, it is also important to emphasize that the answers to the research questions addressed in this article indicated that the aforementioned common belief often does not correspond to reality. In particular, the top accuracy results

of RDDM combined with its unremarkable detections of concept drift suggest that some degree of false positive detections may improve the accuracy results in many datasets rather than harming them. Nevertheless, to be conclusive, this issue demands further investigation.

Despite being very large, the evaluation reported here could be incrementally expanded with other methods and/or base classifiers. In fact, recent methods such as EMZD [12], as well as FPDD and FSDD [12,13], could be added in the comparison of concept drift detection methods in the near future. Methods for semi-supervised and/or unsupervised scenarios as well as methods that address multiclass, evolving, and imbalanced data streams could also be the subject of future evaluations.

Acknowledgments

Silas G. T. C. Santos is supported by a postgraduate grant from CNPq. The authors thank Bruno Maciel for his MOA-Manager script generator and results extraction tool, which greatly helped speed up the generation of the scripts and the analysis of the results of the experiments, despite being still under development.

References

[1] R. Agrawal, T. Imielinski, A.N. Swami, Database mining: a performance perspective, IEEE Trans. Knowl. Data Eng. 5 (6) (1993) 914–925.
[2] S.H. Bach, M.A. Maloof, Paired learners for concept drift, in: 8th IEEE International Conference on Data Mining (ICDM'08), Pisa, Italy, 2008, pp. 23–32.
[3] M. Baena-Garcia, J. Del Campo-Ávila, R. Fidalgo, A. Bifet, R. Gavaldà, R. Morales-Bueno, Early drift detection method, in: International Workshop on Knowledge Discovery from Data Streams, 2006, pp. 77–86.
[4] R.S.M. Barros, Advances in data stream mining with concept drift, Professorship (Full) Thesis, Centro de Informática, Universidade Federal de Pernambuco, Brazil, 2017.
[5] R.S.M. Barros, D.R.L. Cabral, P.M. Gonçalves Jr., S.G.T.C. Santos, RDDM: Reactive drift detection method, Expert Syst. Appl. 90 (C) (2017) 344–355.
[6] R.S.M. Barros, J.I.G. Hidalgo, D.R.L. Cabral, Wilcoxon rank sum test drift detector, Neurocomputing 275 (C) (2018) 1954–1963.
[7] R.S.M. Barros, S.G.T.C. Santos, P.M. Gonçalves Jr., A boosting-like online learning ensemble, in: Proceedings of IEEE International Joint Conference on Neural Networks (IJCNN), Vancouver, Canada, 2016, pp. 1871–1878.
[8] S. Bernstein, The theory of probabilities, 1946, Gastehizdat Publishing House, Moscow.
[9] A. Bifet, R. Gavaldà, Learning from time-changing data with adaptive windowing, in: Proceedings of the 7th SIAM International Conference on Data Mining (SDM'07), Minneapolis, MN, USA, 2007, pp. 443–448.
[10] A. Bifet, G. Holmes, R. Kirkby, B. Pfahringer, MOA: Massive online analysis, J. Mach. Learn. Res. 11 (2010) 1601–1604.
[11] A.G. Bluman, Elementary Statistics: a Step by Step Approach, Ninth, McGraw-Hill, New York, USA, 2014.
[12] D.R.L. Cabral, Statistical tests and detection of concept drifts in data streams, 2017, M.Sc. Dissertation, Centro de Informática, Universidade Federal de Pernambuco. In Portuguese.
[13] D.R.L. Cabral, R.S.M. Barros, Concept drift detection based on fisher's exact test, Inf. Sci. 442-443 (C) (2018) 220–234.
[14] A.P. Dawid, Present position and potential developments: Some personal views: Statistical theory: The prequential approach, J. R. Stat. Soc. Series A 147 (2) (1984) 278–292.
[15] J. Demšar, Statistical comparisons of classifiers over multiple data sets, J. Mach. Learn. Res. 7 (2006) 1–30.
[16] L. Du, Q. Song, L. Zhu, X. Zhu, A selective detector ensemble for concept drift detection, Comput. J. 58 (3) (2014) 457–471.
[17] R. Fisher, On the interpretation of χ^2 from contingency tables, and the calculation of p, J. R. Stat. Soc. 85 (1) (1922) 87–94.
[18] I. Frías-Blanco, J. del Campo-Ávila, G. Ramos-Jiménez, R. Morales-Bueno, A. Ortiz-Díaz, Y. Caballero-Mota, Online and non-parametric drift detection methods based on hoeffdings bounds, IEEE Trans. Knowl. Data Eng. 27 (3) (2015) 810–823.
[19] I. Frías-Blanco, A. Verdecia-Cabrera, A. Ortiz-Díaz, A. Carvalho, Fast adaptive stacking of ensembles, in: Proceedings of the 31st ACM Symposium on Applied Computing (SAC'16), Pisa, Italy, 2016, pp. 929–934.
[20] J. Gama, Knowledge Discovery from Data Streams, Chapman & Hall/CRC, Boca Raton, FL, USA, 2010.
[21] J. Gama, P. Indré, A. Bifet, M. Pechenizkiy, A. Bouchachia, A survey on concept drift adaptation, ACM Comput. Surv. 46 (4) (2014) 44:1–37.
[22] J. Gama, P. Medas, G. Castillo, P. Rodrigues, Learning with drift detection, in: Advances in Artificial Intelligence: SBIA 2004, in: LNCS, volume 3171, Springer, 2004, pp. 286–295.
[23] P.M. Gonçalves Jr., R.S.M. Barros, RCD: A recurring concept drift framework, Pattern Recognit. Lett. 34 (9) (2013) 1018–1025.
[24] P.M. Gonçalves Jr., S.G.T.C. Santos, R.S.M. Barros, D.C.L. Vieira, A comparative study on concept drift detectors, Expert Syst. Appl. 41 (18) (2014) 8144–8156.
[25] W. Hoeffding, Probability inequalities for sums of bounded random variables, J. Am. Stat. Assoc. 58 (1963) 13–30.
[26] D.T.J. Huang, Y.S. Koh, G. Dobbie, R. Pears, Detecting volatility shift in data streams, in: Proceedings of 2014 IEEE International Conference on Data Mining (ICDM), Shenzhen, China, 2014, pp. 863–868.
[27] G. Hulten, L. Spencer, P. Domingos, Mining time-changing data streams, in: Proceedings of the Seventh ACM SIGKDD Intern. Conference on Knowledge Discovery and Data Mining, in: KDD '01, New York, USA, 2001, pp. 97–106.
[28] G.H. John, P. Langley, Estimating continuous distributions in bayesian classifiers, in: Eleventh Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann, San Mateo, 1995, pp. 338–345.
[29] I. Katakis, G. Tsoumakas, I. Vlahavas, Tracking recurring contexts using ensemble classifiers: an application to email filtering, Knowl. Inf. Syst. 22 (2010) 371–391.
[30] J.Z. Kolter, M.A. Maloof, Dynamic weighted majority: An ensemble method for drifting concepts, J. Mach. Learn. Res. 8 (2007) 2755–2790.
[31] Y. Lee, L. Wang, K. Ryu, A system architecture for monitoring sensor data stream, in: Proc. of 7th IEEE Intern. Conference on Computer and Information Technology, 2007, pp. 1026–1031.
[32] J. Liu, Q. Miao, Y. Sun, J. Song, Y. Quan, Fast structural ensemble for one-class classification, Pattern Recognit. Lett. 80 (2016) 179–187.
[33] B.I.F. Maciel, S.G.T.C. Santos, R.S.M. Barros, A lightweight concept drift detection ensemble, in: Proceedings of 27th IEEE International Conference on Tools with Artificial Intelligence (ICTAI), Vietri sul Mare, Italy, 2015, pp. 1061–1068.
[34] B.W. Matthews, Comparison of the predicted and observed secondary structure of t4 phage lysozyme, Biochimica et Biophysica Acta (BBA) - Protein Structure 405 (2) (1975) 442–451.
[35] L.L. Minku, X. Yao, DDD: A new ensemble approach for dealing with concept drift, IEEE Trans. Knowl. Data Eng. 24 (4) (2012) 619–633.
[36] B. Mirza, Z. Lin, N. Liu, Ensemble of subset online sequential extreme learning machine for class imbalance and concept drift, Neurocomputing 149 (2015) 316–329.
[37] T. Mitchell, Machine Learning, McGraw-Hill, New York, NY, USA, 1997.
[38] K. Nishida, K. Yamauchi, Detecting concept drift using statistical testing, in: Proceedings of 10th International Conference on Discovery Science (DS'07), in: LNCS, volume 4755, Springer, 2007, pp. 264–269.
[39] E.S. Page, Continuous inspection schemes, Biometrika 41 (1/2) (1954) 100–115.
[40] R. Pears, S. Sakthithasan, Y. Koh, Detecting concept change in dynamic data streams, Mach. Learn. 97 (3) (2014) 259–293.

[41] J. Perry, A. Kent, M. Berry, Machine literature searching X. machine language; factors underlying its design and development, Am. Document. 6 (4) (1955) 242–254.
[42] A. Pesaranghader, H. Viktor, Fast hoeffding drift detection method for evolving data streams, in: Machine Learning and Knowledge Discovery in Databases, in: LNCS, volume 9852, Springer, 2016, pp. 96–111.
[43] M. Pratama, J. Lu, E. Lughofer, G. Zhang, M. Er, An incremental learning of concept drifts using evolving type-2 recurrent fuzzy neural networks, IEEE Trans. Fuzzy Syst. 25 (5) (2017) 1175–1192.
[44] S. Roberts, Control chart tests based on geometric moving averages, Technometrics 1 (3) (1959) 239–250.
[45] G.J. Ross, N.M. Adams, D.K. Tasoulis, D.J. Hand, Exponentially weighted moving average charts for detecting concept drift, Pattern Recognit. Lett. 33 (2) (2012) 191–198.
[46] S. Sakthithasan, R. Pears, Y. Koh, One pass concept change detection for data streams, in: Advances in Knowledge Discov. and Data Mining, in: LNCS, volume 7819, Springer, 2013, pp. 461–472.
[47] S.G.T.C. Santos, R.S.M. Barros, P.M. Gonçalves Jr., Optimizing the parameters of drift detection methods using a genetic algorithm, in: Proceedings of 27th IEEE Intern. Conf. on Tools with Artificial Intelligence (ICTAI'15), Vietri sul Mare, Italy, 2015, pp. 1077–1084.
[48] S.G.T.C. Santos, P.M. Gonçalves Jr., G.D.S. Silva, R.S.M. Barros, Speeding up recovery from concept drifts, in: Machine Learning and Knowledge Discovery in Databases, in: LNCS, volume 8726, Springer, 2014, pp. 179–194.
[49] A. Shaker, E. Lughofer, Self-adaptive and local strategies for a smooth treatment of drifts in data streams, Evolv. Syst. 5 (4) (2014) 239–257.
[50] F. Wilcoxon, Individual comparisons by ranking methods, Biom. Bull. 1 (6) (1945) 80–83.
[51] I. Žliobaitė, M. Pechenizkiy, J. Gama, An overview of concept drift applications, in: Big Data Analysis: New Algorithms for a New Society, in: Studies in Big Data, 16, Springer, 2016, pp. 91–114.