Time to Retrain? Evaluation of the CDSeer drift detection method in noisy and imbalanced minimal-supervision classification problems

Aleksandra Kaminska, Muhammad Zeeshan Babar

University of Leeds, School of Computing, ODL MSc in AI, UK.

Abstract. Machine learning model pipelines in production have been shown to suffer from performance degradation over time due to data distribution and concept drift. Detecting the concept drift aspect can be particularly challenging in streaming, settings such as Internet of Things, especially when faced with scarcity of labelled points. [1] proposed a novel CDSeer concept detection method for such sparsely labelled problems. In this project we extend their evaluation of CDSeer to noisy and imbalanced problems and to wider set of drift type, as well as propose and evaluate adjustments to the CDSeer's label querying strategy and pseudo-label propagation, demonstrating that while these amendments have their merits, they do not outperform the original CDSeer method in such information-poor problem settings.

Keywords: semi-supervised drift detection, concept drift detection, active online learning

1 Introduction

With machine learning models being increasingly deployed in a wide range of real-world contexts, the ongoing high, stable performance of these models becomes a crucial concern. At the same time, performance degradation of deployed models over time is a common occurrence, as data distributions and feature-label relationships evolve. Such degradation makes deployed models progressively less useful and reliable as the time since deployment increases. Periodic retraining and redeployment of models is a generally accepted remedy for these issues yet can be expensive – both computationally and due to the expert labelling required. Consequently, it is desirable to retrain models only when necessary, guided by reliable drift detection that reliably signals changes in model relevance.

Detection of model drift has been an active area of research for a few decades, both in terms of detecting changes over time in input data distribution ("data drift") and in terms of changing relationships between the input features and the model outputs ("concept drift"). Detecting data drift is often an easier problem and can be achieved using statistical techniques. Detection of concept drift is more challenging, and especially so in semi-supervised settings with low label availability (label scarcity), noise or class imbalance. These characteristics are common in Internet-of-Things (IoT) and similar high velocity streaming contexts, in which only a small fraction of instances can be labelled.

[1] note that the leading concept drift detection methods in such high-volume, label-scarce settings often require excessive labelling and are typically not model-agnostic but rather optimised for specific model architectures. These are concerns to industry adoption, relaxing which would make deployment of model drift detectors in Machine Learning pipelines significantly more practical.

With that in mind, [1] propose a novel model-agnostic technique ("CDSeer") for detecting concept drift in semi-supervised settings with high label scarcity and evaluate its performance on a selection of synthetic and real-world datasets. Their CDSeer method introduces a shadow model ("inspector") that is trained in a semi-supervised manner on a recent subset of data and a small sample of recent expert-generated labels. The method relies on clustering of the recent data window to select a representative subset of points for which expert labels will be request. The most recently obtained labels are propagated to the recent data window via label propagation algorithm, and the results are used to train the shadow inspector model in a supervised manner. Difference in predictions between such shadow model are the reference model are then tracked by a standard supervised drift detector and used to detect concept drift change points.

In the CDSeer method, clustering and label propagation form a key part of the methodology, critical for the effectiveness of the drift detection. In their proposal [1] utilise Euclidean distance as a similarity metric for both clustering and for label propagation, without consideration of whether this type of metric is appropriate considering the type of relationship between features and target in a given problem domain. Indeed, this choice introduces implicit assumptions regarding

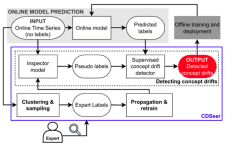


Fig. 1: Overview of CDSeer method ([1]), p.4

the nature of the modelled concept, ones that might not necessarily hold in all the scenarios. It is possible that this assumption may lead to suboptimal clustering and suboptimal training of the shadow model in domains that exhibit more complex relationship between features and target. Furthermore, in evaluating and benchmarking the CDSeer performance, [1] have utilised a set of synthetic datasets which are all noise-free, balanced, exhibit exclusively abrupt concept drift and show no localisation of the drift, i.e. the concept drift generally occurs across entire population rather than for a select minority group. [1] also utilise a small number of real-world datasets in their assessment, however in absence of the ground truth about change points and types of concept drift, these datasets cannot compensate for the selectiveness of the synthetic datasets.

In this project I aim to address these concerns by evaluating the CDSeer method on a wider selection of datasets, with particular focus on synthetic datasets with known drift points and exhibiting noise, class imbalance and varied types of concept drift including gradual and recurrent concept drift. I then propose several approaches to adjusting the CDSeer algorithm that could relax the constraints introduced via its clustering/label propagation approach and evaluate performance of these amendments against the original CDSeer algorithm.

2 Literature Review

Concept drift is a phenomenon in which the statistical properties of a target domain change over time in an arbitrary way (Lu, Zhang and Lu, 2014, p.1). Occurrence of concept drift has been identified as the root cause of decreasing performance in many real-world deployments of machine learning systems. Consequently, a need arises to be able to identify and adapt to such concept drift to avoid degradation in model reliability and reduced reliability of the model results.

Concept drift has been also referred to in literature as data drift, dataset shift, covariate shift and concept shift, often with slightly different meaning. Lu et al. (2018) systematise definition of concept drift as the problem in which $\exists t : P_t(X, y) \neq P_{t+1}(X, y)$ and subcategorise based on sources of the drift – source I caused by input data distribution only ("virtual drift", "feature space drift"), source II caused exclusively by change in relationship between input features and target variables ("actual drift", "decision boundary drift"), and source III representing a mixture of sources I and II. Following [1], the focus for this project is drift owing to source II type, a.k.a. actual drift. Hereon within this report any references to concept drift should be taken to mean the actual drift, unless stated otherwise.

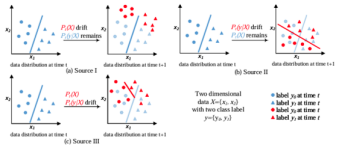


Fig. 2: Sources of concept drift (Lu et al, 2018, p.3)

Concept drift is commonly divided into four types – sudden drift, also referred to as abrupt drift; gradual drift; reoccurring drift, also referred to as recurrent drift; and incremental drift. The four types of concept drift are formally defined in [2] and [3] and illustrated in 3. In their assessment of CDSeer method Pham et al. [1] have focussed solely on abrupt drift type datasets; this project aims to extend their assessment to other types of drift, with particular focus on gradual and recurrent drift.

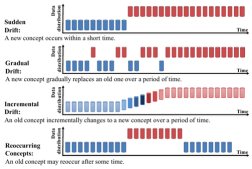


Fig. 3: Types of concept drift (Lu et al, 2018, p.3)

Concept Drift Detection algorithms (CDDs) are categorised as supervised, semi-supervised and unsupervised based on their utilisation of ground truth labels. Unsupervised drift detectors do not utilise any ground truth labels and consequently can only detect drift if changes to input data distribution occur (virtual drift) and hence are not of interest in this report. Supervised drift detectors utilise true labels for all the data points, while semi-supervised methods utilise labels for some (usually few) data of the points. With the focus of [1] and this project on sparsely labelled, high volume streaming problems, the semi-supervised methods are of primary importance, while the supervised detectors serve as one of the building blocks for semi-supervised methods under consideration.

Error Rate Based Detectors, which track changes to the error rate of base classifiers and raise a drift detection alarm when the change in the error rate is deemed to be statistically significant, form an important class of supervised and semi-supervised concept drift detectors. Gama et al. [2] classify error rate based drift detectors into three broad types: *(1)* sequence-based methods, which evaluate prediction results sequentially in relation to a reference value (e.g. running mean) and report drift once a predetermined threshold/significance value is reached; such methods include CUSUM and Page-Hinkley (PH) [4], *(2)* statistical-based methods, which calculate statistical parameters such as mean and standard deviation of predicted results, and report drift when these statics exhibit statistically significant shift; examples include DDM – classic drift detection method [5], EDDM - an early drift detection method

sensitive to gradual drift scenarios [6], ECDD – using exponentially weighted moving average detection [7], or RDDM – reactive DDM [8]; and *(3)* window-based methods, which utilise two different data instance windows for older and newer instances, and monitor divergence between the two windows; examples include ADWIN, HDDM – DDM based method utilising Hoeffding inequality [9], or Paired Learners - a method using two different base learners, a stable one (focussed on entire set of data) and reactive one (focussed on recent data only), and monitoring changes in their relative predictions [10].

Paired Learners can be seen as an inspiration for the CDSeer drift detector defined in [1], which also includes a classifier based on recent data as a gauge of drifting concept. Another recent important work on the topic is [11], utilising a long-term stable classifier and dynamic classifier to detect both sudden and gradual changes for online learning Concept Drift Detection.

Semi-supervised concept drift detection methods are more realistic than supervised ones in the streaming, big-data, high-velocity settings such as IoT environments, in which labels are not inherently available for new data points and need to be procured, subject to certain cost and effort. Semi-supervised CDDs require only a subset of data to be labelled, lessening the extent of the problem. Existing methods achieve that by selecting only most important data instances for labelling, utilising e.g. active learning – e.g. [12], [13], and [14]; class imbalance [15], classifier confidence [16], [17], grid density sampling [18] or pseudo-error from classifier ensemble [19]. However, these drift detection methods tend to be targeted towards reference models with specific ML architectures, as well as rely on very specific metrics to sample labels – both characteristics reduce their usability in general, model agnostic streaming settings.

Online active learning is a paradigm in machine learning that aims to select the most informative data points to label from a data stream and is particularly important topic in context of semi-supervised concept drift detection in general, and CDSeer algorithm [1] in particular. Cacciarelli and Kulahci [20] present a comprehensive survey of recent approaches to active learning for data streams, including the connection between active learning and semi-supervised learning for data streaming scenarios. As they point out, the widely analysed data problem of pool-based active learning [21]; [22] is not practical in most real-world streaming settings which are dynamic and of sequential nature. Instead in such scenarios one must perform online selective sampling [23]. Cacciarelli and Kulahci review the topic of budget constraints as applied to data annotation in this context, types and implications of latency in label availability, as well as different categories of strategies of instance selection criteria. Specifically, the diversity- and density-based approaches are the key category for the problem space outlined in [1]. These methods exploit the structural information of the input data space and aim to select data points representative of the overall distribution of data – including clustering used to label representative data points [24]; [25], graph-based methods employed to explore the structure information of labelled and unlabelled data points [26], and methods building on the semi-supervised label propagation strategy [27].

Explainable AI (XAI) techniques have been explored in context of concept drift detection, both in context of detecting drift occurrences and in context of explaining the driving factors for these occurrences. Lunderberg and Lee prosed SHAP (Shapley additive explanations) method to measure the influence of each feature on model predictions [28]. Lundberg at al [29] further postulates that SHAP metrics allow to spot drift in features in predictive models. The topic of utilising XAI techniques has been an active research field, including [30]; [31]; [32].

3 Methodology

The methodology and decisions utilised in this project fall within three categories: drift detector designs evaluated, datasets utilised, and finally the evaluation approach and metrics used. These are briefly outlined below, alongside the research questions that the project is looking to address.

3.1 Research Questions

This project focuses on the following questions.

RQ1 *Is CDSeer performance maintained on datasets with noise, class imbalance or complex decision boundaries?*

The synthetic datasets used to evaluate the CDSeer performance in drift detection in [1] are all class-balanced and without noise or redundant information. Furthermore, the datasets tend to have straightforward, typically linear decision boundaries. This project addresses this shortcoming by evaluating CDSeer method on a wider set of synthetic datasets exhibiting more realistic features such as noise and feature redundancy.

RQ2 *How does CDSeer perform in the presence of other drift types such as gradual or recurring drifts? Is the performance improved by utilising supervised drift detectors tailored towards such drift types?*

The synthetic datasets in [1] all exhibit abrupt, non-recurrent drift points only. This project examines how the CDSeer method performs on datasets exhibiting broader types of drift, in particular gradual drift and recurrent drift, and whether this performance is improved by using a tailored supervised drift detector such as EDDM [6] instead of general-purpose Page-Hinkley recommended in [1].

RQ3 *At what threshold of expert label availability does CDSeer method become more effective in detecting drift than an outright supervised drift monitoring using only the (scarce) available expertly labelled subset of points?*

CDSeer method aims to detect drift by extrapolating pseudo-labels from the limited set of recent confirmed labels and train a shadow recency model using these pseudo-labels. Such an approach introduces assumptions about the structure of the data as well as certain approximation of results. The alternative would be to utilise a supervised drift detector with ground truth labels only, without resorting to extrapolation of pseudo labels. In this project I aim to explore the threshold of expert-label availability rate at which such a ground-truth only approach exhibits comparable performance to CDSeer, and how that threshold relates to the 0.6%-0.8% label availability rate assumed in [1].

RQ4 *What effect do selected amendments to the clustering and label propagation approach, proposed by the author of this project, have on the overall effectiveness of CDSeer-like drift detector?*

CDSeer clustering and label propagation introduce implicit assumptions about the nature of the modelled problem and the relationship between features and target variable by utilising Euclidean-based similarity metric for both clustering and label propagation. That may not be appropriate for all use case sand may introduce unstable performance, leading to e.g. increased amount of false positive detections and unnecessary training cost. Similarly, the CDSeer method does not consider any preprocessing of data, such as scaling of points, in training the shadow model. While the shadow model is based on Random Forest and hence more resilient to unscaled data, the clustering and propagation itself may be affected by unscaled data, especially bearing in mind reliance on Euclidean distance metric, potentially leading to less effective drift detection. This project introduces a number of variations of the clustering/label propagation method in CDSeer and evaluates their performance against the original CDSeer method.

3.2 Drift Detector Design

CDSeer Drift Detector Pham et al. [1] have not made the implementation of the CDSeer drift detector public; hence the algorithm is reimplemented using the outline in [1]. Where the implementation details have not been provided by Pham et al. [1], reasonable assumptions are made based on related literature and validated against the experiment results in [1].

The CDSeer method is used for online drift detection and relies on tracking error rates between the reference model being monitored and the shadow model trained using a recent small sample of expert-provided labels (referred to in [1] as "inspector" model). For every new data point presented to the drift detector – from now on referred to as a step – a new shadow model is trained using a set of recent data points without labels ("inspector window"), and a very smaller set of recent expert-labelled data ("label memory"). The available ground truth labels are propagated to the rest of the inspector window using a label spreading algorithm, creating a set of pseud-labels. Furthermore, on every step the shadow model selects a subset of points from the inspector window that it considers representative and queries the oracle to obtain expert labels for all or some of the points, which are then used to update the label memory.

In line with Pham et al. [1], clustering, label spreading and training of the shadow model is performed on every step, except for the points used to train (or retrain) the reference model. Whether or not points used to retrain the reference model are also evaluated by the drift detector is dependent on the mode of evaluation being run (see Metrics and evaluation approach for details). The shadow model is an instance of a RandomForest model. The clustering uses DBSCAN method in the first instance, with epsilon parameter estimated using a simplified knee detection algorithm [33]. If the obtained clustering is not found to be sufficiently good, a KMeans clustering method is used instead. The label propagation is performed using Label Spreading algorithm

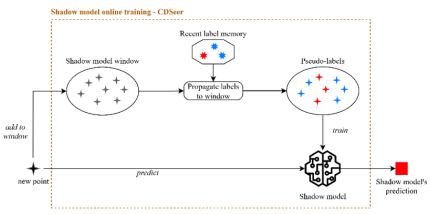


Fig. 4: Shadow model - online training step - original CDSeer

[34]. The RandomForest, DBSCAN, KMeans and LabelSpreading algorithms are implemented using scikit-learn library, with parameters as per Pham et al. [1].

On the topic of procuring ground truth labels, Pham et al. [1] state only that stratified sampling is performed on the identified clusters, and that 15/10 (a hyperparameter) most recent labels are retained at any one time, without commenting on the specific sampling technique, the proportion of labels requested at each step, or whether immediate availability of labels is assumed. Consequently, in this project the expert label availability is controlled by two parameters – the proportion of inspector window that is requested for labelling in every step ("inspector curiosity ratio") and the likelihood that the Oracle will provide a requested label ("compliance factor"), allowing to control for the expert label budget during the online drift detection steps. The stratified sampling utilises simple random sampling, while ensuring that at least one sample from each cluster is selected. It is assumed that there is no systematic delay in availability of labels, but rather they can be provided by Oracle synchronously.

Finally, CDSeer utilises a standard supervised drift detector for monitoring of the error rate. Pham et al. [1] name Page-Hinkley detector as the detector of choice, and for consistency the same detector is utilised in this project unless stated otherwise. It should however be noted that different detectors have been considered and experimented with as part of this project, most notably ADWIN, DDM, and FHDDM (Fast Hoeffding Drift Detection Method). Additionally, EDDM [6] has been used to evaluate **RQ3**. As an implementation note, the drift detectors' implementations for the River machine learning library (`https://riverml.xyz/`) have been used in this project.

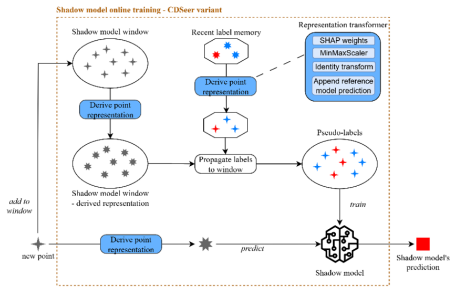


Fig. 5: Shadow model - online training step - CDSeer variants

Proposed CDSeer Adaptations In this project several adaptations of the clustering/label-propagation approach in CDSeer are proposed, hypothesising that through different processing of the feature space, these variations may increase effectiveness and stability of the CDSeer-like drift detection in certain types of domains. These variations are evaluated against the original CDSeer algorithm.

Clustering approaches

Var1. Scaling data prior to clustering and label propagation ("Scaling"). CDSeer's usage of Euclidean distance in DBSCAN, KMeans and LabelSpreading means that differing ranges of different features in dataset may impact the derivation of pseudo-labels and affect stability of the drift detection algorithm. To counter this, the first CDSeer variant performs data scaling prior to clustering and propagation. The scale is based on the initial (offline) training set for the shadow inspector model, and transforms features to $[0, 1]$ range using scikit-learn's MinMaxScaler. Subsequently each point is transformed using this scaler as the first step of training the shadow inspector model.

Var2. Utilise SHAP values as coordinates of each point for clustering and label propagation ("SHAP"). In using Euclidean distance for clustering and label-spreading, CDSeer makes assumptions about shape of decision boundaries, roughly that neighbouring points typically have similar model outcomes. This assumption will not hold for some use cases, especially ones with complex decision boundaries. Var2 ("SHAP") modification explores whether SHAP values of each data point might offer a better representation of the point for the purpose of clustering and pseudo-label derivation. The SHAP (SHapley Additive exPlanations) values are obtained using SHAP explainer, trained on the reference model and its training set. The SHAP explainer is implemented using KernelExplainer from SHAP package `https://shap.readthedocs.io/`.

Var3. Combine coordinates of the point with output of reference model ("RefModel). Assuming the reference model being monitored is of sufficient quality, one could argue that nearby points (in terms of similarity metric used) should not be considered similar if they yield different outcomes of the reference model, as they do not form a cohesive cluster. In that spirit in Var3 adaptation the outcome of (up to date) reference model is considered when assessing points' similarity for clustering and label propagation. Specifically, the model outcome is added as an additional dimension to the point representation. This does not as such eliminate the possibility of points with different outcomes ending up in same cluster, but it does allow the method to take the model outcomes into account when assessing similarity. Note that the same approach could be combine with SHAP representation in Var2.

Var4. Combine Scaling and RefModel – Var1 and Var2 ("Scaling + RefModel"). Scaling of data prior to clustering/propagation can be combined with supplementing the point representation with the outcomes of the (latest) reference model.

Var5. Random selection of points for labelling. Pham et al. [1] on one hand define an involved procedure to cluster and recent points and select the candidates for labelling, while on the other hand do not seem to systematically consider the information gain aspects and work here and here on selecting best labels in active learning. One alternative then is to incorporate this this and this following on from reference and reference – this approach is not explored in this project. On the other end of spectrum one could select the points completely at random from the recent point window. Pham et al. [1] comment that this would be overly simplistic, without justification. In this project the random selection of points for labelling is benchmarked against the original CDSeer method to validate this assumption.

Var6. Supervised drift detection based on ground-truth expert labels only. The derivation of pseudo-labels in the CDSeer method allows to train the shadow model that can be used as a comparison to the reference model in the absence of ground-truth label values, in the hope that the recency of data used for shadow model will allow to spot the drift occurrence. However the derivation of pseudo-labels is necessarily based on certain assumptions and approximate, which may lead to substandard modelling and imprecise drift detection. Alternatively, one could ignore unlabelled points in drift detection and consider in a fully supervised manner only the points for which expert labels are available. Having said that the extreme scarcity of labels available in the setting outlined in CDSeer might mean there are simply not enough ground-truth labels available to spot any drift occurring. To quantify this, this project explores effectiveness of such supervised drift detection versus CDSeer method at different levels of label availability.

Dimensionality Reduction The speed of processing is a key factor in the high-volume, high-velocity settings considered in this project – too slow processing of each point may result in being unable to make decision and detect drift in a timely manner. To enable the appropriate processing speed, this project performs dimensionality reduction through Singular Value Decomposition whenever a representation of a point exceeds a predefined number of dimensions. This is done so as to alleviate the computational burden.

Label Propagation Any changes to the point representation described in variants 1-4 are also applied in label propagation, i.e. the same representation of each point is used both for clustering and for label propagation Variant 5 performs label propagation as outlined in the original CDSeer method, while variant 6 does not perform label propagation.

3.3 Datasets

This project expands on the datasets used in [1] to evaluate CDSeer method, firstly by adding variations of these datasets that exhibit richer characteristics, and secondly by incorporating additional synthetic and real-life datasets. All the basic variants of the datasets below are widely used in concept drift detection research.

Sine A synthetic dataset with a decision boundary based on variations of sine function. Introduced by Gama et al. [5] The basic variant of this dataset used in [1] represents a binary classification problem, with features in the $[0, 1]$ range, balanced and without noise. It exhibits two abrupt concept drifts, at 3,000 and 10,000 steps respectively, following the SINE1-REVERSE SINE1–SINE2 concept pattern.

For this project, variations of the Sine dataset have been generated to add class imbalance, redundant features and scaling of feature subset. Furthermore, the concept order has been changed for some of the dataset variants, to follow the SINE2-REVERSE SINE2-SINE1 pattern to generate dataset with less simplistic decision boundaries – as SINE1's and REVERSE SINE1's boundaries approximate linear function.

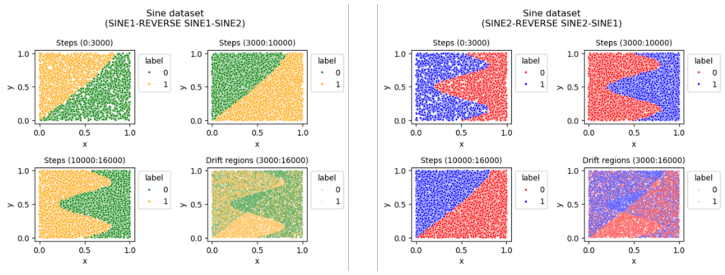


Fig. 6: Illustration of concepts and one of drift regions in Sine datasets

SEA A synthetic dataset introduced by [35], with a decision boundary based on shifting linear function. The basic variant of this dataset used in [1] represents a binary classification problem, with features in

the [0,10] range, balanced and without noise. It exhibits two abrupt concept drifts, at 3,000 and 10,000 steps respectively, following the Variant0-Variant1-Variant2 concept pattern.

For this project, variations of the SEA dataset have been generated to add noise (at 2% level), redundant features and scaling of feature subset.

Chocolate Synthetic dataset, implementing lattice-like decision boundaries. Introduced by [36], sourced from `https://github.com/THUFDD/THU-Concept-Drift-Datasets/tree/main`. Drift points are abrupt and occur after 20%, 40%, 60% and 80% of population. Drift exhibited by versions termed "abrupt" is also recurrent, whereas versions termed "sudden" have a smaller region of drift at each concept drift event due to the rotating nature of drift.

For this project both a multi-label and binary versions of dataset have been generated, with and without noise and redundant features.

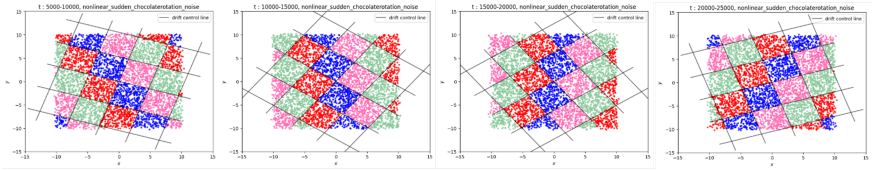


Fig. 7: Multi-class rotating chocolate dataset after consecutive drifts

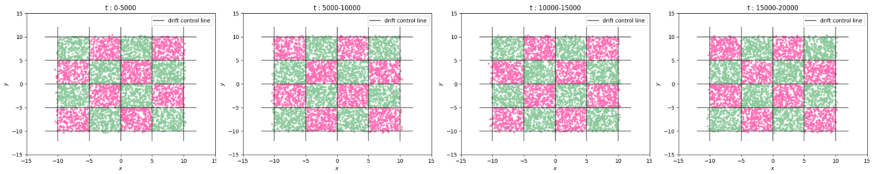


Fig. 8: Binary non-rotating chocolate dataset after consecutive drifts

Harvard Concept Drift Datasets - Gradual Drift A collection of synthetic datasets for concept drift detection testing, developed by [37] and available from `https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/5OWRGB`. The datasets all represent binary classification problems with 40,000 instances, 3 concept drift points and 4 concepts, with either abrupt or gradual drift. In this project only gradual drift versions are used, with drifts at time steps 9,500, 20,000 and 30,500, each with width of 1000. Detailed description of each dataset can be found in [37]. Specifically, the following datasets are used in this project:

- sine_0123_gradual - 2 numerical features, a balanced binary class, and without noise.
- sea_3210_gradual_noise_0.2, sea_0123_gradual_noise_0.2 - 3 numerical features (1 redundant), a balanced binary class, noise in concept with 0.2 probability.
- mixed_1010_gradual, mixed_0101_gradual - ruled by sequence of classification functions. 4 numerical features, a balanced binary class, and without noise.
- rt_8873-9856-7896-2563, rt_2563-7896-9856-8873 - ruled by a sequence of Random Tree functions. 2 numerical features, a balanced binary class, without noise.

It ought to be noted that the 20% level of noise in the sea_3210_gradual_noise_0.2 and sea_0123_gradual_noise_0.2 goes beyond what would normally be considered a reasonable noise threshold and may affect the evaluation results due to challenges of reference model and/or shadow model in incorporating such a high noise threshold. Nonetheless the datasets have been used extensively in past works on drift detection, thus they have been included here for consistency.

NOAA Weather Dataset (NOAA) A subset of the original NOAA (National Oceanic and Atmospheric Administration) weather measurements dataset, containing 50 years of weather data from Offutt Air Force Base in Bellevue, Nebraska. The dataset prepared by [38] and sourced from `https://users.rowan.edu/~polikar/nse.html`. This is a real-world dataset, frequently used in drift detection research. It represents an imbalanced binary classification task, with unknown drift points. Approximate drift points have been estimated using method outlined in [1], for consistency with their evaluation of CDSeer.

Electricity (Elec2) Dataset representing electricity prices in New South Wales, Australia, introduced by [39]. The dataset covers a period of 2 years, with records taken every half an hour, and is a popular benchmark for testing concept drift detectors. It represents a binary classification problem (will the

price go up or down), with real drift points unknown. Approximate drift points have been estimated using method outlined in [1], for consistency with their evaluation of CDSeer. It ought to be noted that concerns have been raised by [40] regarding the merit using this dataset for testing, however we have included it in this project for consistency with original benchmarking of CDSeer.

JiaoLong DSMS v2 (JIAOLONG) A dataset collected by the National Deep Sea Center in Qingdao, Shandong, China, in the exploration task for the JiaoLong Deep-sea Manned Submersible on March 19, 2017. Introduced by [41] and sourced from `https://github.com/THUFDD/JiaolongDSMS_datasets`. Data is the multi-variate time series with 24 features and 3 target classes, exhibiting slight class imbalance. The real concept drift points are unknown, but indicative regions have been described by [41] – see 9. In this project the drift points have been estimated in line with methodology in [1].

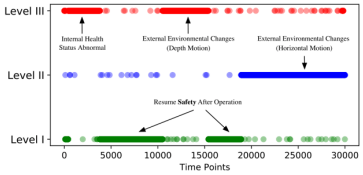


Fig. 9: Indicative drift point ranges in JiaoLong DSMS dataset (from [41])

Datasets Summary Table in 10 summarises the key features of dataset families used in this project. More detailed description can be found in respective references within above description of each dataset, and in references.

Dataset family	Samples	Features	Drift	Drift count	Type	Introduced/key characteristics	Used for CDSeer
Sine	16,000	2-4	known	2	synthetic	Redundant features, class imbalance, scaling of features.	Basic variant
SEA	16,000	2-3	known	2	synthetic	Noise, redundant features, scaling of features.	Basic variant
Chocolate bar	25,000	2-5	known	4	synthetic	Multi-class and binary version, noise, redundant features, complex decision boundary.	No
Harvard Concept Drift datasets	41,000	2-4	known	3	synthetic	Gradual drift versions. Complex decision boundaries.	No
NOAA	18,159	8	estimated	3	real-world	Imbalanced dataset. Features of varying ranges.	Yes
Elec2	45,312	9	estimated	21	real-world	High autocorrelation of labels, see (Zlobaite, 2013). Nonetheless dataset widely used in drift detection research.	Yes
Jiaolong	30,000	24	estimated (indicative ranges known)	6	real-world	Imbalanced dataset. High dimensionality. Features of varying ranges.	No

Fig. 10: Summary of datasets

3.4 Evaluation Approach and Metrics

Metrics Effectiveness of each version of drift detector is quantified using a set of measures related to the effectiveness of the drift detection, as well as to accuracy of the reference model retrained whenever the drift detection signals occur:

1. **BalancedAccuracy and F1 of reference model predictions** Reference model is retrained when drift is detected, as described in Evaluation approach. section. The metrics are by comparing reference model predictions with ground truth labels and are calculated using all points on which reference model predictions have been obtained. For multi-label problems, the macro version of F1 metric is reported.
2. **Drift precision and recall** Drift precision is the percentage of true positives of all the detected concept drifts by a CDD (Concept Drift Detector). Drift recall is the percentage of true positives detected by a CDD among all the GT concept drifts.

Drift detection is treated as correct of it occurs within X steps ("tolerance") after the ground truth drift. For non-gradual drift datasets with known drift points the X values of 500, 1000 and 2000 are used, producing three different sets of measures. For datasets with gradual drift, a tolerance of 3000 is also used to approximately account for width of the drift.

For datasets for which the drift points are not known, Pham et al. [1] estimate the ground truth drift points by using a supervised Page-Hinkley detector on the error rate generated by comparing predictions of reference model (without retraining) to the ground truth label. Such approximations are dependent on the specific reference model being tested, and affected by the quality of this model, and thus are highly approximate. Nonetheless the same approach has been used in this project for the sake of consistency with [1].

For datasets with estimated drift, a 2-sided drift precision and recall metrics are calculated, in addition to metrics outlined above. 2-sided metric in this instance should be taken to mean a metric in which a detected drift is considered to be correct for the value of tolerance X, if such drift is *either before or after* a ground truth drift, with the distance of at most X steps to that ground truth drift.

Furthermore, for selected experiments performance of the model at regular intervals is also examined, to illustrate the evolution of respective metrics, as well as the impact of general model suitability on quality of drift detection.

Finally, it should be noted that a measure corresponding to drift detection lag has been considered in place of fixed-width versions of drift precision and drift recall, as such measure is both relatively simple to understand and conveys a meaningful amount of information. However, this measure has been decided against to maintain consistency with [1] and their metric choices.

Evaluation Approach Each variant of the drift detection algorithm is evaluated using a variation of prequential evaluation approach, whereby following the initial training of reference model and of the shadow model, each subsequent point is presented to the reference model and to drift detection algorithm for assessment, and only it may be used to update the reference model – the (potential) updates of the model are described in more detail below. Furthermore, each set of evaluation results is obtained by averaging 5 consecutive evaluation runs, seeded with known seed values $(0, 1, \ldots, 4)$.

The size of the dataset used for training the reference model ($WINDOW\_SIZE$), the number of most recent points visible (without labels) to the shadow model ($INSP\_SIZE$) and the size of shadow model's memory for known recent labels (MEM_SIZE) are all hyperparameters, as per [1]. Pham et al. [1] evaluate CDSeer for several combinations of these hyperparameters; for consistency with the most prominent of their results and unless indicated otherwise, this project utilizes $WINDOW_SIZ$E of 1000, $INSP\_SIZE$ of 500, and MEM_SIZE of 15.

In prequential evaluation, the reference model would typically be updated with the new information (point) immediately after the point has been evaluated from prediction and drift detection perspective. However, the CDSeer detector is by definition meant to be model agnostic and thus one cannot assume that the reference model can be trained online. Consequently, in this project as in [1] it is assumed that model retraining occurs only once the drift has been detected.

Pham et al. [1] state that the retraining of the reference model is not within the scope of their paper, but that nonetheless as part of the evaluation they performed reference model retraining whenever drift has occurred. The details are not stated; however, the reproducing of their results suggests that the reference models have been retrained backwards, on the most recent points seen up to the moment of drift detection. This approach has the benefit of being able to immediately retrain the models without needing to wait for new points to come in (though one may need to await the derivation of labels for the points for training). It does however mean that the more efficient the drift detector is in terms of speed to detect drift, the more of pre-drift points will be used to train the new reference model. This is likely to lead to increase in false positive drift detections and makes the drift detectors appear less effective in evaluation. The alternative would be to train forward, using the first $WINDOW\_SIZE$ points following the drift. Such an approach would provide better reflection of effectiveness of the drift detector model by increasing the likelihood of each new version of reference model being trained on a coherent concept. It would however mean that not all the available points in the dataset could be used for evaluation of the drift detector.

Consistency with [1] demands that backwards retraining is used, however using the forward retraining approach would give more accurate view of effectiveness of drift detection methods being assessed. As a compromise both retraining methods are utilised in this project, with the specific method stated when discussing results. When comparisons are being made with [1] backward training is used, whereas in more general settings the forward training is preferred.

Lastly, the streaming nature of the problem is simulated by sequentially presenting each consecutive point to the reference model and drift detector algorithms, utilizing the River framework – with the exception of points used for (re)training the reference model, which are processed offline in a batch manner.

Implementation Implementation of this project is available on `https://github.com/kydrysek/CDSeer_Enhancements`, and access to the code can be requested. Below in algorithm 1 we present as pseudocode the key aspects of modified drift detection in CDSeer variants.

Algorithm 1 Orchestration of drift detection using supervised detector and shadow model

Require: Trained reference model $Model$, stream of input data $Stream$ initialised supervised drift detector SDD, shadow model $Inspector$ and oracle $Oracle$

Ensure: Detection of concept drift, i.e. changes in relationship between model output and model inputs

while ($point$ in $Stream$ **and not** SDD detected drift) **do**
  add $point$ to $Oracle$
  $modelPrediction \leftarrow$ Model.predict($point$)

  Inspector prediction:
  add $point$ to $Inspector$ window # **FIFO queue**
  $pointRepresentation \leftarrow Inspector.transform(point)$

  derive $clustering$ following CDSeer, using $pointRepresentations$ for current $window$
  $pointsToRequest \leftarrow$ StratifiedSample($clustering$)
  request $labels(pointsToRequest)$ for $Oracle$ # **Asynchronous, labelling lag**
  $newLabels, labelledPoints \leftarrow Oracle.recentlyLabelled()$ # **From this/previous requests**
  update short term memory of labels in $Inspector$ using $newLabels, labelledPoints$

  memoryRepresentations $\leftarrow$ transformed $labels$ from $memory$
  $pseudoLabels \leftarrow LabelSpreading(memoryRepresentations, memorylabels, windowRepresentations)$

  $Inspector \leftarrow Inspector.retrain(pseudoLabels, windowRepresentations)$

  $inspPrediction \leftarrow$ Inspector.predict($point$) # **Always just trained**

  update SDD detector with **prediction error value** ($modelPred! = inspPrediction$)

end while

return signal for drift detection, so that model can be retrained and experiment reinitialised

The critical differences in algorithm 1 compared to original CDSeer method is consistent use for transformed representations for points, corresponding to derivation of SHAP coordinates, scaled coordinates or another suitable transformation,

These transformations are initialised when the Inspector shadow model is reinitialised, using information on points (not labels) corresponding to the training window for Inspector. These points are provided by the Oracle, and are used to e.g. train MinMaxScaler or SHAP Explainer.

4 Experiment Results

In performing the evaluation outlined above, several observations arise not only about the effectiveness of CDSeer and its variants, but also about the drift detection in general. Firstly, it is important to note that very fast and accurate detection of drift can worsen the performance of reference model, unless the appropriate steps are taken to shield the reference model from retraining on a significant amount of pre-drift data. This is demonstrated in 11 using one of the Chocolate datasets – a supervised drift detector with full label knowledge is so fast at spotting drift points that a model retrained on past data retrains mostly on pre-drift data, lowering its future performance. This does not present a problem if model is retrained on post-drift data ("forwards" version), but it does pose limitations on how quickly the new model can be available due to need to procure new labelled data.

Furthermore, our experiments highlight another fundamental observation regarding reliance of drift detection performance on how well the reference model selected fits the target problem. If a particular model architecture is not particularly well suited to the task to hand and exhibits substandard performance even pre-drift, then a drift detector – CDSeer or otherwise – may well

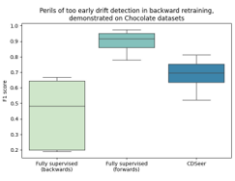
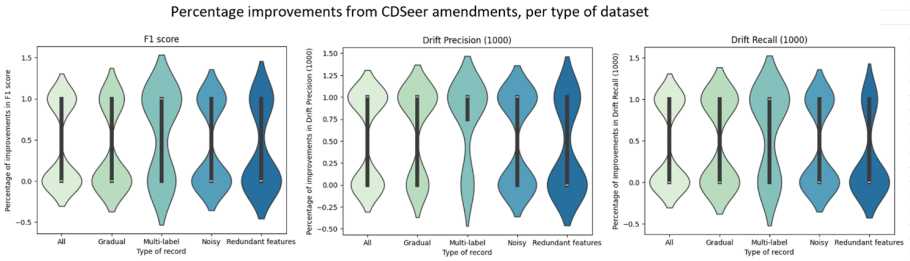


Fig. 11: Accurate drift detection can, with substandard pipeline setup, lead to worse models

struggle to reliably detect drifting in the reference model as the model never predicted data accurately to start with. This also applies to the architecture used for shadow model, as in cases where it is not suitable to model the problem the inherent noise in the shadow model may obscure concept drifts. This is to an extent illustrated in the Harvard SEA dataset, owing to large proportion of noise in the dataset – see Appendix 2 7. The good fit of the reference model for the use case is thus paramount, though realistically this is not a major concern as the performance would have been tested before deployment. Likewise, one also needs to consider for each problem the extent to which the RandomForest architecture with the sample size described in CDSeer is capable of effectively modelling the use case in question.

Additionally, we need to remember that while F1 (binary/macro) and Accuracy of the reference model with retraining are useful and relevant datapoints, they can be counterintuitive in that a poorly performing drift detector which generates a lot of false positive events would lead to frequent retraining of the model and hence artificially increased Accuracy / F1 of the reference model. This is illustrated for example in the performance of backwards-retrained CDSeer variants on the SEA dataset (see Appendix 2 7). That performance would however come at the cost, both timewise and budgetary, of needing to retrain the model unnecessarily, and is something we explicitly want to avoid in our setting due to sparsely supervised nature of the data.

Regarding **RQ1** and **RQ2**, in line with expectations, the CDSeer drift detection generally worsens in performance in presence of noisy datasets and of redundant variables – see Appendix 1 6. It can also experience instability, in the sense of sometimes performing better and sometimes worse, for corresponding pairs of datasets e.g. with stretched input features (see the SEA datasets, stretched versions; Appendix 2 7), or when faced with matching datasets with reversed order of concept (see Harvard dataset pairs, Appendix 2 7). However, it needs to be noted that the supervised drift detection with full knowledge also drops in performance on these same datasets, as do the CDSeer variants introduced in this project – and crucially none of the proposed variants systematically remediates the drop in performance observed on such types of data in CDSeer. Nor is incorporation of EDDM strongly beneficial for gradual drift sets in our particular evaluations, The one exception to this is potentially performance of improvements on non-binary datasets. However, bearing in mind that majority of this datapoint is driven by performance of the MinMaxScaler amendment, this should nearly certainly be attributed to the specific characteristics of the multi-label dataset in this evaluation, namely the JiaoLong dataset and the Chocolate multi-label variants. Interestingly also, the querying of points at random for labels does not perform as badly as Pham et al. (2024) have assumed, and in a number of instances performs better than the original CDSeer – this minor effect can be observed across a number of different dataset characteristics.



On **RQ4**, none of the proposed improvements seem to systematically improve on performance of the original CDSeer. Some of them perform than others in particular situations, such as MinMaxScaler offering better stability and more consistent results for stretched versions of the same dataset, or models with more complex decision boundaries, like Chocolate datasets or Harvard Mixed and RT datasets, benefiting for appending of reference model predictions prior to clustering and label spreading – see SVC (Support Vector Classifier) results in Appendix 2 7. Overall, however no single variant offers meaningful increase in drift detection performance for any of the types of datasets. Nonetheless the results show that all these adaptions have merit and possibly could be combined in a classifier ensemble – see section 5 for further discussion.

It is worth noting that usage of SHAP weights does seem to hold promise, as in a5 number of runs across three different families of datasets (Harvard,SEA and SINE), the drift recall, accuracy and F1 all showed signs of improving for a significant number of datasets. Specifically the SHAP-only coordinates fared better in over 50% of runs, and were also close to 50% of instances improved

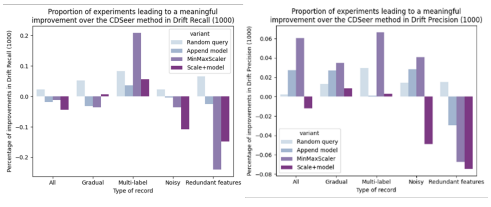


Fig. 12: Random querying for labels as a valid procurement strategy

in case of drift precision. The version that combines SHAP coordinates with reference model output fared a bit worse, particularly in terms of precision. It did however still provide a level of improvement that ought to be further analysed.

Variant	Total runs	F1	Drift precision (1000)	Drift recall (1000)	Drift precision (2000)	Drift recall (2000)
Both SHAP	44	63.636%	40.909%	65.909%	38.636%	59.091%
SHAP only	26	65.385%	46.154%	38.462%	50.000%	50.000%
SHAP+model	18	50.000%	16.667%	61.111%	11.111%	27.778%

Table 1: Performance metrics across different SHAP variants

Lastly on the **RQ3**, CDSeer does appear to be meaningfully more effective than a method which does not attempt to produce pseudo-labels but rather perform supervised drift detection with only the sparse labels available. That is true even at meaningful amounts of label availability, such as 10% availability. This makes sense instinctively, as derivation of sparse data points lead to slow and very inefficient drift detection; while at the same time the pseudo-labels generated by CDSeer method encode mostly realistic assumptions about similarity of samples and their outcomes, creating a sufficiently good proxy dataset on which to apply frequent drift detection.

5 Conclusion and Future Work

We have filled in the gap in [1] by performing a more comprehensive evaluation of CDSeer detector's performance on datasets with challenging characteristics, such as noise, imbalance or gradual drift. Further, we have proposed several amendments to the architecture of the shadow model in CDSeer method and evaluated performance of these amendments against the performance of original CDseer method. We have ascertained that the CDSeer method [1] does actually perform much more reliably than originally expected even in noisy or imbalanced environments, and to an extent in presence of gradual drift, despite its assumptions regarding suitability of Euclidean distance similarity for comparing model outcomes. This conclusion makes sense in the context of the very scarce availability of the labels in the problem outlined – in the region of 0.6%-0.8%. With such extreme label scarcity, the label querying method introduced in CDSeer is just as good as most others, though as we've observed random querying would also be a suitable alternative – in a nutshell, in the problem setting there is so little information available that the set of assumptions introduced through the clustering and pseudo-label derivation in CDSeer are not any more detrimental than assumptions of any other label querying strategy with similar level of information utilisation. Having said, there are other avenues that we think are worth following, in particular in context of multiple hypothesis test drift detection and further adaptions for gradual concept drift – see below.

As we have observed in this project, different label querying strategies lend are most suitable to different use cases and datasets. Thus, us a follow up it would be worth exploring how best multiple hypothesis test drift detection [3] can be utilised in sparsely supervised concept drift detection methods such as CDSeer, particularly focusing on the parallel multiple hypothesis tests. Furthermore, in defining CDSeer Pham et al. [1] state that shadow model is implemented by a random forest. In future work it would be worthwhile to explore other models as a base for the shadow model: *(1)* similarity based model such as KNearestNeighbours, acknowledging the similarity assumptions already implicit in CDSeer label querying and propagation, and *(2)* online models, such as Mondrian Forests [42] or Adaptive Random Forests [43], to avoid the need to train shadow model from scratch for every prediction.

To further the work on suitability of CDSeer for gradual drift, we propose to evaluate wider range of detectors targeted towards gradual drift, such as FW-DDM [44], ECDD [7], or ensemble of classifiers methods [13]. In addition to gradual drift suitability enhancements, one also ought to investigate performance of CDSeer and variants on problems with drift restricted to a minority subset, a problem that has been shown by [45] to be a challenge for many drift detectors. Furthermore, future work should perform a more comprehensive analysis and approach to label budgeting, in particular accounting for lag in expert label availability and assessing impact that has on efficiency of label propagation and shadow model training.

Finally, it would be interesting to look at applying Explainable AI techniques to try to achieve causal explainability of detected drifts, which could lead to actionable insights, following on from ideas presented by [46].

References

[1] Tri Minh Triet Pham et al. "Time to Retrain? Detecting Concept Drifts in Machine Learning Systems". In: *2025 IEEE/ACM 47th International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*. IEEE. 2025, pp. 260–271. DOI: 10.48550/arXiv.2410.09190.

[2] João Gama et al. "A survey on concept drift adaptation". In: *ACM computing surveys (CSUR)* 46.4 (2014), pp. 1–37. DOI: doi.org/10.1145/2523813.

[3] Jie Lu et al. "Learning under concept drift: A review". In: *IEEE transactions on knowledge and data engineering* 31.12 (2018), pp. 2346–2363.

[4] Raquel Sebastião and José Maria Fernandes. "Supporting the page-hinkley test with empirical mode decomposition for change detection". In: *International symposium on methodologies for intelligent systems*. Springer. 2017, pp. 492–498.

[5] Joao Gama et al. "Learning with drift detection". In: *Brazilian symposium on artificial intelligence*. Springer. 2004, pp. 286–295. DOI: 10.1007/978-3-540-28645-5_29.

[6] Manuel Baena-Garcıa et al. "Early drift detection method". In: *Fourth international workshop on knowledge discovery from data streams*. Vol. 6. 2006, pp. 77–86.

[7] Gordon J Ross et al. "Exponentially weighted moving average charts for detecting concept drift". In: *Pattern recognition letters* 33.2 (2012), pp. 191–198. DOI: 10.1016/j.patrec.2011.08.019.

[8] Roberto SM Barros et al. "RDDM: Reactive drift detection method". In: *Expert Systems with Applications* 90 (2017), pp. 344–355.

[9] Isvani Frias-Blanco et al. "Online and non-parametric drift detection methods based on Hoeffding's bounds". In: *IEEE Transactions on Knowledge and Data Engineering* 27.3 (2014), pp. 810–823. DOI: 10.1109/tkde.2014.2345382.

[10] Stephen H Bach and Marcus A Maloof. "Paired learners for concept drift". In: *2008 Eighth IEEE International Conference on Data Mining*. IEEE. 2008, pp. 23–32.

[11] Hang Zhang et al. "Online active learning paired ensemble for concept drift and class imbalance". In: *IEEE Access* 6 (2018), pp. 73815–73828.

[12] Wei Fan et al. "Active mining of data streams". In: *Proceedings of the 2004 SIAM International Conference on Data Mining*. SIAM. 2004, pp. 457–461. DOI: 10.1137/1.9781611972740.46.

[13] Xingquan Zhu et al. "Active learning from data streams". In: *Seventh IEEE International Conference on Data Mining (ICDM 2007)*. IEEE. 2007, pp. 757–762.

[14] Mohammad Masud et al. "Classification and novel class detection in concept-drifting data streams under time constraints". In: *IEEE Transactions on knowledge and data engineering* 23.6 (2010), pp. 859–874.

[15] Edwin Lughofer et al. "Recognizing input space and target concept drifts in data streams with scarcely labeled and unlabelled instances". In: *Information Sciences* 355 (2016), pp. 127–151.

[16] Ahsanul Haque, Latifur Khan, and Michael Baron. "Semi supervised adaptive framework for classifying evolving data stream". In: *Pacific-Asia conference on knowledge discovery and data mining*. Springer. 2015, pp. 383–394. DOI: 10.1007/978-3-319-18032-8_30.

[17] Ahsanul Haque et al. "Efficient handling of concept drift and concept evolution over stream data". In: *2016 IEEE 32nd International Conference on Data Engineering (ICDE)*. IEEE. 2016, pp. 481–492. DOI: https://doi.org/10.1109/icde.2016.7498264.

[18] Tegjyot Singh Sethi, Mehmed Kantardzic, and Hanquing Hu. "A grid density based framework for classifying streaming data in the presence of concept drift". In: *Journal of Intelligent Information Systems* 46.1 (2016), pp. 179–211. DOI: 10.1007/s10844-015-0358-3.

[19] Felipe Pinage, Eulanda M dos Santos, and Joao Gama. "A drift detection method based on dynamic classifier selection". In: *Data Mining and Knowledge Discovery* 34.1 (2020), pp. 50–74.

[20] Davide Cacciarelli and Murat Kulahci. "Active learning for data streams: a survey". In: *Machine Learning* 113.1 (2024), pp. 185–239. DOI: 10.1007/s10994-023-06454-2.

[21] Charu C Aggarwal et al. "Active learning: A survey". In: *Data classification*. Chapman and Hall/CRC, 2014, pp. 599–634. DOI: 10.1201/b17320-27.

[22] Punit Kumar and Atul Gupta. "Active learning query strategies for classification, regression, and clustering: A survey". In: *Journal of Computer Science and Technology* 35.4 (2020), pp. 913–945. DOI: 10.1007/s11390-020-9487-4.

[23] Steve Hanneke and Liu Yang. "Toward a general theory of online selective sampling: Trading off mistakes and queries". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2021, pp. 3997–4005. URL: https://proceedings.mlr.press/v130/hanneke21a.html.

[24] Fan Min et al. "Three-way active learning through clustering selection". In: *International Journal of Machine Learning and Cybernetics* 11.5 (2020), pp. 1033–1046. DOI: 10.1007/s13042-020-01099-2.

[25] Dino Ienco, Indrė Žliobaitė, and Bernhard Pfahringer. "High density-focused uncertainty sampling for active learning over evolving stream data". In: *Proceedings of the 3rd international workshop on big data, streams and heterogeneous source mining: algorithms, systems, programming models and applications*. PMLR. 2014, pp. 133–148. URL: https://proceedings.mlr.press/v36/ienco14.html.

[26] Hongjing Zhang, SS Ravi, and Ian Davidson. "A graph-based approach for active learning in regression". In: *Proceedings of the 2020 SIAM International Conference on Data Mining*. SIAM. 2020, pp. 280–288. DOI: 10.1137/1.9781611976236.32.

[27] Jun Long et al. "Graph-based active learning based on label propagation". In: *International Conference on Modeling Decisions for Artificial Intelligence*. Springer. 2008, pp. 179–190.

[28] Scott M Lundberg and Su-In Lee. "A unified approach to interpreting model predictions". In: *Advances in neural information processing systems* 30 (2017). DOI: 10.1038/s42256-019-0138-9.

[29] Scott M Lundberg et al. "From local explanations to global understanding with explainable AI for trees". In: *Nature machine intelligence* 2.1 (2020), pp. 56–67. DOI: 10.1016/j.ins.2016.03.034.

[30] Johannes Haug et al. "Change detection for local explainability in evolving data streams". In: *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 2022, pp. 706–716. URL: https://arxiv.org/abs/2209.02764.

[31] Yongsoo Lee et al. "Explainable Artificial Intelligence-Based Model Drift Detection Applicable to Unsupervised Environments." In: *Computers, Materials & Continua* 76.2 (2023).

[32] Jayesh Tripathi, Heitor Gomes, and Marcus Botacin. "Towards Explainable Drift Detection and Early Retrain in ML-Based Malware Detection Pipelines". In: *International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment*. Springer. 2025, pp. 3–24. DOI: 10.1007/978-3-031-97623-0_1.

[33] Ville Satopaa et al. "Finding a" kneedle" in a haystack: Detecting knee points in system behavior". In: *2011 31st international conference on distributed computing systems workshops*. IEEE. 2011, pp. 166–171. DOI: 10.1109/ICDCSW.2011.20.

[34] Yoshua Bengio, Nicolas Le Roux, and Olivier Delalleau. *Efficient non-parametric function induction in semi-supervised learning*. CIRANO, 2004.

[35] W Nick Street and YongSeog Kim. "A streaming ensemble algorithm (SEA) for large-scale classification". In: *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*. 2001, pp. 377–382. DOI: 10.1145/502512.502568.

[36] Songqiao Hu et al. "CADM+: Confusion-based learning framework with drift detection and adaptation for real-time safety assessment". In: *IEEE Transactions on Neural Networks and Learning Systems* 36.3 (2024), pp. 5126–5139. DOI: 10.1109/tnnls.2024.3369315.

[37] Jesús López Lobo. "Synthetic datasets for concept drift detection purposes". In: *Harv. Dataverse* (2020). URL: https://doi.org/10.7910/dvn/5owrgb.

[38] Gregory Ditzler and Robi Polikar. "Incremental learning of concept drift from streaming imbalanced data". In: *IEEE transactions on knowledge and data engineering* 25.10 (2012), pp. 2283–2301.

[39] Michael Harries, New South Wales, et al. "Splice-2 comparative evaluation: Electricity pricing". In: (1999).

[40] Indre Zliobaite. "How good is the electricity benchmark for evaluating concept drift adaptation". In: *arXiv preprint arXiv:1301.3524* (2013). URL: https://arxiv.org/abs/1301.3524.

[41] Zeyi Liu et al. "An online active broad learning approach for real-time safety assessment of dynamic systems in nonstationary environments". In: *IEEE Transactions on Neural Networks and Learning Systems* 34.10 (2022), pp. 6714–6724. DOI: 10.1109/tnnls.2022.3222265.

[42] Balaji Lakshminarayanan, Daniel M Roy, and Yee Whye Teh. "Mondrian forests: Efficient online random forests". In: *Advances in neural information processing systems* 27 (2014). URL: https://arxiv.org/abs/1406.2673.

[43] Heitor M Gomes et al. "Adaptive random forests for evolving data stream classification". In: *Machine Learning* 106.9 (2017), pp. 1469–1495. DOI: 10.1007/s10994-017-5642-8.

[44] Anjin Liu, Guangquan Zhang, and Jie Lu. "Fuzzy time windowing for gradual concept drift adaptation". In: *2017 IEEE international conference on fuzzy systems (FUZZ-IEEE)*. IEEE. 2017, pp. 1–6.

[45] Flavio Giobergia et al. "A Synthetic Benchmark to Explore Limitations of Localized Drift Detections". In: *International Workshop on Discovering Drift Phenomena in Evolving Landscapes*. Springer. 2024, pp. 101–110.

[46] David Komnick et al. "Causal Explanation of Concept Drift–A Truly Actionable Approach". In: *arXiv preprint arXiv:2507.23389* (2025).

6 Appendix 1: Further Illustration of Results

This appendix contains further comparisons of the CDSeer variants introduced in this project and their relative performance on different types of datasets. Figure 13 presents the results at the overall level of types of datasets characteristics, while figure 14 breaks down results by the specific CDSeer amendment considered.

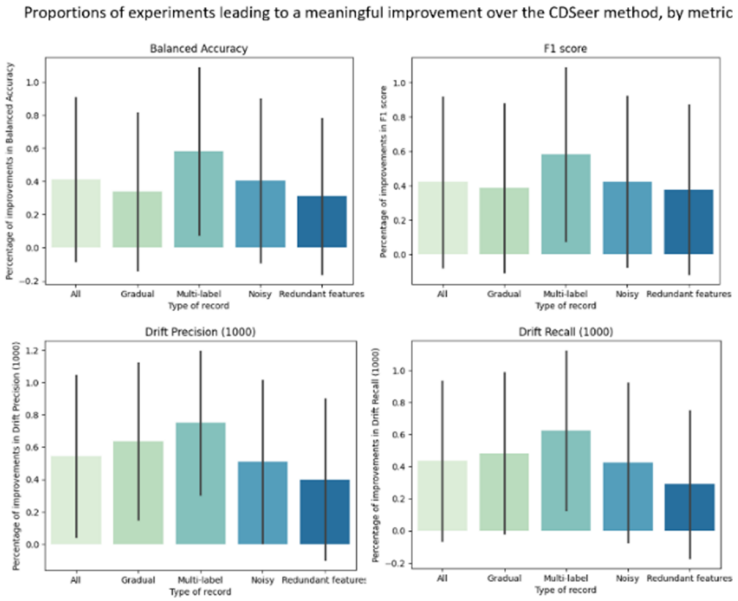


Fig. 13: Types of datasets benefitting from one of the proposed amendments (non-SHAP)

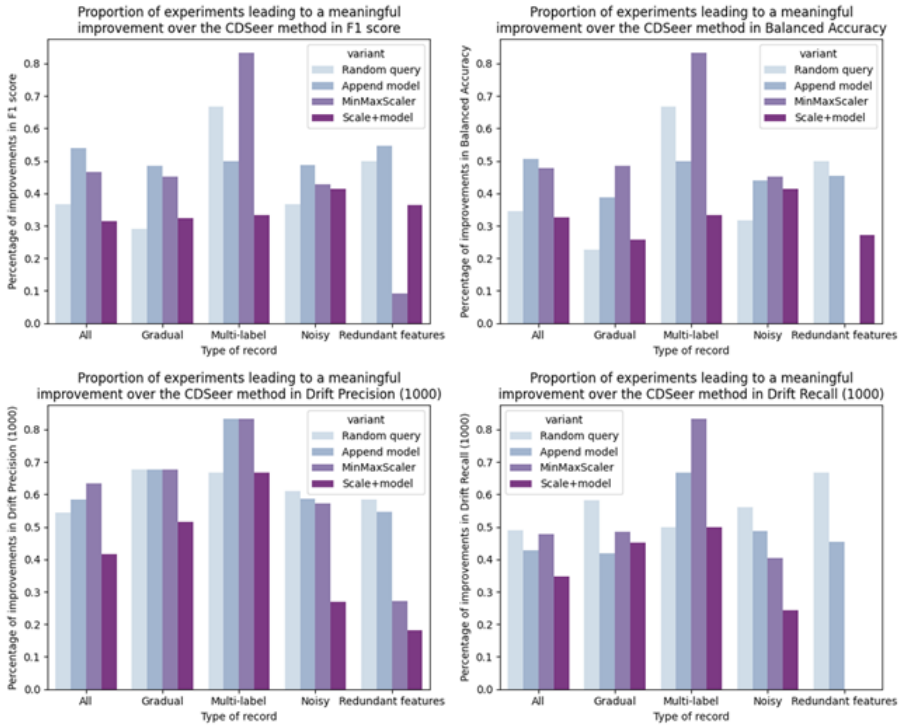


Fig. 14: Proportion of experiments showing minor benefits from adjusting querying strategy (non-SHAP)

Further figure 15 illustrates performance of a sample reference model with retraining due to drift detection – the spikes in accuracy are due to retraining of the model and restart of accuracy calculation (for the purposes of this illustration).

Finally figure 16 presents a breakdown of performance of SHAP-variant runs benchmarked to the corresponding CDSeer results.

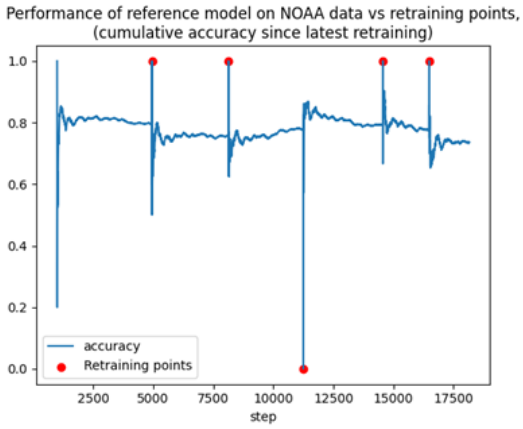


Fig. 15: Sample performance of reference model with retraining, measured by accuracy since last retraining

Type of Run	SHAP version	Family	File	BalancedAccuracy	F1	Drift precision (1000)	Drift recall (1000)	Drift precision (2000)	Drift recall (2000)
PH back	SHAP only	Harv	Mixed 1010	89.241%	89.224%	6.786%	46.667%	22.045%	100.000%
PH back	SHAP only	Harv	Sine 0123	90.518%	90.519%	10.079%	33.333%	35.238%	100.000%
PH forw	SHAP only	Harv	Mixed 1010	90.233%	90.233%	9.790%	40.000%	21.399%	86.667%
PH forw	SHAP only	Harv	Sine 0123	93.090%	93.095%	33.048%	53.333%	60.571%	100.000%
PH back	SHAP only	SEA	Redund. cols, stretched (v1)	94.243%	96.036%	11.667%	30.000%	22.381%	60.000%
PH back	SHAP only	SEA	Redund. cols, stretched (v2)	95.037%	96.659%	17.000%	40.000%	34.000%	73.333%
PH back	SHAP only	SEA	Original	95.284%	96.946%	5.000%	10.000%	32.000%	70.000%
PH back	SHAP only	SEA	Original, stretched (v1)	94.700%	96.255%	14.667%	30.000%	25.333%	50.000%
PH back	SHAP only	SEA	Original, stretched (v2)	95.227%	96.860%	16.667%	40.000%	35.000%	70.000%
PH back	SHAP only	SEA	Original, noise, redund. cols, stretched (v2)	91.597%	94.331%	14.667%	30.000%	26.000%	60.000%
PH back	SHAP only	SEA	Original, noise	93.461%	95.735%	19.667%	50.000%	39.333%	90.000%
PH back	SHAP only	SEA	Original, noise, stretched (v1)	93.144%	95.734%	19.000%	40.000%	40.667%	83.333%
PH back	SHAP only	SEA	Original, noise, stretched (v2)	92.716%	95.177%	20.000%	40.000%	33.333%	70.000%
PH forw	SHAP only	SINE	Balanced, no noise	86.084%	85.996%	63.333%	70.000%	86.667%	100.000%
PH forw	SHAP only	SINE	Balanced, noise	86.177%	85.996%	80.000%	80.000%	93.333%	100.000%
PH forw	SHAP only	SINE	Imbalanced, no noise, stretched (v2)	89.631%	88.919%	50.000%	60.000%	73.333%	90.000%
PH forw	SHAP only	SINE	Imbalanced, noise, stretched (v1)	89.176%	88.494%	48.333%	80.000%	53.333%	90.000%
PH back	SHAP only	SINE	Balanced, no noise	85.163%	85.048%	23.000%	60.000%	58.857%	100.000%
PH back	SHAP only	SINE	Balanced, no noise, stretched (v1)	86.875%	86.804%	23.159%	80.000%	51.048%	100.000%
PH back	SHAP only	SINE	Balanced, no noise, stretched (v2)	84.612%	84.556%	25.389%	80.000%	55.619%	100.000%
PH back	SHAP only	SINE	Balanced, noise	86.752%	86.623%	32.667%	90.000%	44.325%	100.000%
PH back	SHAP only	SINE	Balanced, noise, stretched (v2)	85.088%	84.969%	19.405%	70.000%	44.429%	100.000%
PH back	SHAP only	SINE	Imbalanced, no noise, stretched (v1)	82.810%	81.567%	30.381%	100.000%	52.857%	100.000%
PH back	SHAP only	SINE	Imbalanced, no noise, stretched (v2)	86.099%	85.081%	28.857%	80.000%	51.571%	100.000%
PH back	SHAP only	SINE	Imbalanced, noise, stretched (v1)	85.217%	84.173%	29.048%	100.000%	42.103%	100.000%
PH back	SHAP only	SINE	Imbalanced, noise, stretched (v2)	83.819%	82.740%	25.714%	80.000%	40.167%	100.000%
PH back	SHAP+model	Harv	Mixed 1010	89.260%	89.246%	15.493%	68.333%	40.239%	100.000%
PH back	SHAP+model	Harv	Sea 0123, noise 0.2	76.389%	76.451%	8.891%	60.000%	15.994%	80.000%
PH back	SHAP+model	Harv	Sine 0123	89.900%	89.891%	8.000%	26.667%	39.273%	95.000%
PH forw	SHAP+model	Harv	Mixed 1010	90.725%	90.713%	17.494%	53.333%	36.097%	100.000%
PH forw	SHAP+model	Harv	Sea 0123, noise 0.2	76.058%	76.159%	7.752%	33.333%	15.852%	66.667%
PH forw	SHAP+model	Harv	Sine 0123	92.756%	92.770%	30.857%	53.333%	57.571%	100.000%
PH back	SHAP+model	SEA	Redund. cols, stretched (v1)	94.248%	96.240%	18.048%	50.000%	30.381%	30.381%
PH back	SHAP+model	SEA	Original	94.569%	96.241%	6.667%	10.000%	31.667%	31.667%
PH back	SHAP+model	SEA	Original, noise, redund. cols, stretched (v2)	92.001%	94.757%	23.000%	50.000%	34.667%	34.667%
PH back	SHAP+model	SEA	Original, noise	91.709%	94.352%	0.000%	0.000%	10.667%	10.667%
PH forw	SHAP+model	SINE	Balanced, no noise	89.183%	89.092%	86.667%	90.000%	93.333%	100.000%
PH forw	SHAP+model	SINE	Balanced, noise	88.569%	88.437%	66.667%	80.000%	80.000%	100.000%
PH forw	SHAP+model	SINE	Imbalanced, no noise, stretched (v2)	86.616%	85.805%	70.000%	90.000%	76.667%	100.000%
PH forw	SHAP+model	SINE	Imbalanced, noise, stretched (v1)	88.722%	88.029%	50.000%	70.000%	50.000%	70.000%
PH back	SHAP+model	SINE	Balanced, no noise	85.889%	85.826%	38.000%	58.190%	100.000%	100.000%
PH back	SHAP+model	SINE	Balanced, noise	86.060%	85.983%	34.286%	37.381%	100.000%	100.000%
PH back	SHAP+model	SINE	Imbalanced, no noise, stretched (v2)	86.772%	85.804%	36.000%	53.571%	100.000%	100.000%
PH back	SHAP+model	SINE	Imbalanced, noise, stretched (v1)	85.392%	84.408%	22.222%	46.071%	100.000%	100.000%

Fig. 16: Sample performance of reference model with retraining, measured by accuracy since last retraining

7 Appendix 2: Results by Dataset Family

This appendix showcases selected results of evaluations performed. It is organised by the family of the dataset, and in one instance by type of reference model. In most cases both results for "backwards" retraining and the "forwards" mode are presented.

Note that results highlighted in bright green background represent metrics where a proposed CDSeer variant improves upon the original proposed in [1]. Results highlighted with light blue background represent metrics of the same value as those for corresponding (original) CDSeer method.

7.1 Real-World Datasets - NOAA, JIAOLONG, ELEC

Reference model retraining mode: backwards			Reviewed revisions			
File	Supervised CDD	CDSeer - original	Random selection of label candidates	MinMaxScaling	Append model prediction	MinMaxScaling + append model prediction
Accuracy	77.34%	77.90%	78.02%	77.56%	77.92%	77.32%
BalancedAccuracy	71.37%	71.79%	71.45%	71.82%	71.34%	71.97%
F1	84.11%	84.56%	84.77%	84.21%	84.69%	83.94%
Drift precision (500)	0.00%	6.32%	6.67%	12.42%	3.89%	4.73%
Drift precision (1000)	0.00%	21.12%	18.64%	16.43%	19.30%	12.39%
Drift precision (2000)	0.00%	29.38%	29.09%	26.15%	21.30%	21.23%
Drift precision (3000)	40.00%	35.83%	37.58%	41.99%	27.49%	34.15%
Drift recall (500)	0.00%	20.00%	20.00%	50.00%	13.33%	20.00%
Drift recall (1000)	0.00%	58.33%	51.67%	62.67%	60.00%	43.33%
Drift recall (2000)	0.00%	72.67%	81.67%	78.81%	66.67%	54.33%
Drift recall (3000)	66.67%	80.33%	88.33%	88.78%	73.33%	69.81%
Drift precision - 2 sided (500)	20.00%	12.76%	15.30%	18.23%	15.13%	11.80%
Drift precision - 2 sided (1000)	20.00%	46.87%	41.21%	28.63%	47.96%	27.61%
Drift precision - 2 sided (2000)	20.00%	63.89%	61.29%	52.91%	65.93%	50.22%
Drift recall - 2 sided (500)	33.33%	40.00%	41.67%	70.00%	46.67%	43.33%
Drift recall - 2 sided (1000)	33.33%	84.67%	80.00%	85.00%	83.67%	68.67%
Drift recall - 2 sided (2000)	33.33%	90.95%	90.83%	96.67%	87.79%	85.00%

Fig. 17: Results for NOAA dataset across all variants, in backwards retraining mode

Reference model retraining mode: forwards			Reviewed revisions			
File	Supervised CDD	CDSeer - original	Random selection of label candidates	MinMaxScaling	Append model prediction	MinMaxScaling + append model prediction
Accuracy	77.38%	78.07%	78.34%	78.35%	77.92%	77.79%
BalancedAccuracy	71.07%	71.12%	71.53%	72.94%	70.23%	71.36%
F1	84.25%	84.92%	85.08%	84.76%	84.99%	84.56%
Drift precision (500)	0.00%	11.86%	0.00%	10.00%	0.00%	6.19%
Drift precision (1000)	0.00%	17.57%	8.33%	17.22%	6.19%	12.38%
Drift precision (2000)	25.00%	27.29%	22.33%	27.50%	16.38%	25.38%
Drift precision (3000)	50.00%	39.14%	32.33%	38.33%	19.71%	40.57%
Drift recall (500)	0.00%	20.00%	0.00%	26.67%	0.00%	13.33%
Drift recall (1000)	0.00%	33.33%	20.00%	46.67%	13.33%	26.67%
Drift recall (2000)	33.33%	53.33%	53.33%	73.33%	33.33%	46.67%
Drift recall (3000)	66.67%	68.33%	73.33%	86.67%	40.00%	68.33%
Drift precision - 2 sided (500)	25.00%	11.86%	7.33%	12.50%	19.67%	9.52%
Drift precision - 2 sided (1000)	25.00%	36.29%	30.67%	32.50%	34.90%	24.71%
Drift precision - 2 sided (2000)	50.00%	60.71%	56.00%	51.11%	64.10%	51.43%
Drift recall - 2 sided (500)	33.33%	20.00%	13.33%	33.33%	33.33%	20.00%
Drift recall - 2 sided (1000)	33.33%	61.67%	61.67%	75.00%	61.67%	46.67%
Drift recall - 2 sided (2000)	66.67%	83.33%	86.00%	95.00%	81.33%	77.67%

Fig. 18: Results for NOAA dataset across all variants, in forwards retraining mode

Reference model retraining mode: backwards			Reviewed revisions			
File	Supervised CDD	CDSeer - original	Random selection of label candidates	MinMaxScaling	Append model prediction	MinMaxScaling + append model prediction
Accuracy	95.63%	85.04%	87.33%	90.38%	84.89%	91.86%
BalancedAccuracy	95.59%	85.56%	87.49%	90.51%	85.59%	91.88%
MacroF1	95.59%	84.65%	87.43%	90.25%	84.43%	91.73%
Drift precision (500)	46.15%	23.53%	29.30%	20.35%	24.67%	25.56%
Drift precision (1000)	69.23%	47.94%	49.27%	39.06%	50.57%	42.71%
Drift precision (2000)	76.92%	82.06%	78.24%	80.06%	85.56%	72.89%
Drift precision (3000)	92.31%	91.43%	96.36%	96.92%	94.29%	92.33%
Drift recall (500)	75.00%	40.00%	46.19%	40.00%	38.10%	56.67%
Drift recall (1000)	81.82%	62.26%	60.60%	65.83%	61.19%	75.56%
Drift recall (2000)	83.33%	80.59%	74.50%	82.42%	79.95%	84.14%
Drift recall (3000)	92.31%	88.71%	88.97%	91.92%	89.86%	92.36%
Drift precision - 2 sided (500)	61.54%	23.53%	29.30%	20.35%	24.67%	25.56%
Drift precision - 2 sided (1000)	84.62%	47.94%	49.27%	39.06%	50.57%	45.50%
Drift precision - 2 sided (2000)	100.00%	90.00%	83.70%	89.83%	88.41%	88.42%
Drift recall - 2 sided (500)	88.89%	40.00%	46.19%	40.00%	38.10%	56.67%
Drift recall - 2 sided (1000)	91.67%	62.26%	60.60%	65.83%	61.19%	79.13%
Drift recall - 2 sided (2000)	92.86%	83.87%	76.17%	84.07%	81.92%	89.11%

Fig. 19: Results for JiaoLong dataset across all variants, in backwards retraining mode

7.2 Synthetic Datasets

Reference model retraining mode: forwards			Reviewed revisions			
File	Supervised CDD	CDSeer - original	Random selection of label candidates	MinMaxScaling	Append model prediction	MinMaxScaling + append model prediction
Accuracy	98.36%	87.45%	89.19%	86.45%	89.28%	87.32%
BalancedAccuracy	98.15%	87.08%	89.35%	87.13%	89.62%	88.03%
MacroF1	98.22%	87.26%	89.35%	86.68%	89.58%	87.43%
Drift precision (500)	83.33%	49.71%	46.86%	38.00%	52.00%	46.67%
Drift precision (1000)	83.33%	49.71%	56.57%	54.00%	56.00%	54.67%
Drift precision (2000)	83.33%	67.24%	75.43%	100.00%	78.00%	77.33%
Drift precision (3000)	83.33%	73.43%	78.29%	100.00%	78.00%	84.67%
Drift recall (500)	83.33%	46.67%	40.00%	33.33%	46.67%	40.00%
Drift recall (1000)	83.33%	46.67%	50.00%	46.67%	50.00%	46.67%
Drift recall (2000)	83.33%	63.33%	66.67%	72.14%	67.62%	66.67%
Drift recall (3000)	100.00%	78.00%	80.67%	83.81%	80.67%	81.33%
Drift precision - 2 sided (500)	100.00%	67.24%	65.71%	38.00%	70.67%	62.00%
Drift precision - 2 sided (1000)	100.00%	67.24%	78.29%	54.00%	74.67%	70.00%
Drift precision - 2 sided (2000)	100.00%	90.95%	97.14%	100.00%	100.00%	92.67%
Drift recall - 2 sided (500)	100.00%	63.33%	56.67%	33.33%	63.33%	53.33%
Drift recall - 2 sided (1000)	100.00%	63.33%	67.62%	46.67%	66.67%	60.00%
Drift recall - 2 sided (2000)	100.00%	84.29%	83.81%	72.14%	87.14%	80.00%

Fig. 20: Results for JiaoLong dataset across all variants, in forwards retraining mode

Reference model retraining mode: backwards			Reviewed revisions			
File	Supervised CDD	CDSeer - original	Random selection of label candidates	MinMaxScaling	Append model prediction	MinMaxScaling + append model prediction
Accuracy	75.21%	76.87%	77.47%	77.50%	77.19%	77.51%
BalancedAccuracy	74.77%	76.04%	76.71%	76.75%	76.40%	76.56%
F1	71.05%	72.08%	72.94%	72.99%	72.55%	72.59%
Drift precision (500)	15.38%	24.34%	24.19%	23.24%	25.21%	23.35%
Drift precision (1000)	43.59%	47.88%	48.41%	50.96%	49.34%	47.20%
Drift precision (2000)	79.49%	84.03%	81.85%	86.06%	86.21%	84.43%
Drift precision (3000)	100.00%	94.82%	93.56%	96.64%	95.64%	95.18%
Drift recall (500)	28.57%	66.98%	72.06%	76.79%	65.20%	76.29%
Drift recall (1000)	73.91%	92.09%	95.52%	97.03%	93.80%	97.24%
Drift recall (2000)	93.94%	99.61%	100.00%	99.71%	100.00%	100.00%
Drift recall (3000)	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
Drift precision - 2 sided (500)	48.72%	47.33%	45.37%	44.20%	45.30%	46.92%
Drift precision - 2 sided (1000)	87.18%	84.56%	83.70%	84.78%	83.96%	83.35%
Drift precision - 2 sided (2000)	97.44%	99.66%	99.70%	99.24%	98.65%	98.92%
Drift recall - 2 sided (500)	79.17%	92.27%	96.22%	98.30%	91.56%	96.19%
Drift recall - 2 sided (1000)	97.14%	100.00%	100.00%	100.00%	99.59%	100.00%
Drift recall - 2 sided (2000)	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%

Fig. 21: Results for ELEC dataset across all variants, in backwards retraining mode

Reference model retraining mode: forwards			Reviewed revisions			
File	Supervised CDD	CDSeer - original	Random selection of label candidates	MinMaxScaling	Append model prediction	MinMaxScaling + append model prediction
Accuracy	77.35%	77.00%	77.51%	78.01%	76.63%	76.64%
BalancedAccuracy	77.05%	76.51%	77.11%	77.14%	75.73%	75.19%
F1	73.22%	73.23%	73.66%	73.48%	71.78%	70.43%
Drift precision (500)	28.57%	22.53%	25.67%	25.92%	25.81%	27.57%
Drift precision (1000)	42.86%	43.31%	48.08%	42.39%	50.04%	46.48%
Drift precision (2000)	85.71%	84.13%	83.72%	86.11%	84.59%	91.17%
Drift precision (3000)	95.24%	94.45%	92.20%	95.32%	93.98%	98.00%
Drift recall (500)	28.57%	27.62%	31.43%	31.43%	28.57%	26.67%
Drift recall (1000)	42.86%	53.33%	59.05%	51.43%	55.24%	44.76%
Drift recall (2000)	90.00%	93.79%	93.10%	96.51%	93.42%	82.95%
Drift recall (3000)	100.00%	98.30%	99.09%	99.26%	99.13%	95.99%
Drift precision - 2 sided (500)	57.14%	49.65%	48.27%	56.22%	49.09%	50.52%
Drift precision - 2 sided (1000)	76.19%	82.88%	84.37%	82.82%	84.37%	85.44%
Drift precision - 2 sided (2000)	100.00%	100.00%	99.23%	100.00%	99.13%	100.00%
Drift recall - 2 sided (500)	57.14%	60.00%	59.05%	68.57%	54.29%	48.57%
Drift recall - 2 sided (1000)	84.21%	94.65%	95.56%	96.27%	91.48%	83.62%
Drift recall - 2 sided (2000)	100.00%	100.00%	100.00%	100.00%	100.00%	98.00%

Fig. 22: Results for ELEC dataset across all variants, in forwards retraining mode

Reference model retraining mode: backwards		Dataset version											
		Balanced, no noise	Balanced, no noise, stretched (v1)	Balanced, no noise, stretched (v2)	Balanced, noise	Balanced, noise, stretched (v1)	Balanced, noise, stretched (v2)	Imbalanced, no noise	Imbalanced, no noise, stretched (v1)	Imbalanced, no noise, stretched (v2)	Imbalanced, noise	Imbalanced, noise, stretched (v1)	Imbalanced, noise, stretched (v2)
Accuracy	Supervised CDD	25.11%	25.11%	25.11%	25.87%	25.87%	25.87%	25.45%	25.45%	25.45%	25.59%	25.59%	25.59%
	CDSeer - original	85.66%	86.69%	84.08%	83.28%	85.00%	83.90%	84.97%	82.86%	82.69%	82.79%	83.47%	83.17%
	Random selection of label candidates	84.32%	84.39%	83.59%	82.07%	80.15%	86.41%	83.09%	81.27%	83.25%	84.69%	81.93%	82.64%
	MinMaxScaling	84.17%	76.33%	83.32%	85.11%	84.27%	84.56%	82.03%	77.39%	76.44%	83.92%	85.36%	81.85%
	Append model prediction	85.50%	82.65%	85.73%	85.47%	85.47%	87.00%	86.85%	87.14%	85.12%	82.54%	85.93%	85.60%
	MinMaxScaling + append model prediction	85.49%	85.40%	83.20%	86.42%	87.83%	84.70%	86.92%	77.72%	77.73%	85.14%	85.44%	83.89%
	SHAP_noScale_noModel	85.16%	86.87%	84.61%	86.75%		85.09%		82.86%	86.14%		85.28%	83.87%
	SHAP_noScale_Model	85.89%			86.06%					86.80%		85.43%	
BalancedAccuracy	Supervised CDD	25.11%	25.11%	25.11%	25.87%	25.87%	25.87%	25.33%	25.33%	25.33%	25.60%	25.60%	25.60%
	CDSeer - original	85.66%	86.69%	84.08%	83.28%	85.00%	83.90%	84.92%	82.81%	82.63%	82.75%	83.46%	83.16%
	Random selection of label candidates	84.32%	84.39%	83.59%	82.07%	80.15%	86.41%	83.05%	81.20%	83.21%	84.63%	81.90%	82.57%
	MinMaxScaling	84.17%	76.33%	83.32%	85.11%	84.27%	84.56%	82.00%	77.34%	76.37%	83.88%	85.34%	81.89%
	Append model prediction	85.50%	82.65%	85.73%	85.47%	85.47%	87.00%	86.82%	87.10%	85.06%	82.61%	85.88%	85.55%
	MinMaxScaling + append model prediction	85.49%	85.40%	83.20%	86.42%	87.83%	84.70%	86.89%	77.68%	77.64%	85.11%	85.41%	83.89%
	SHAP_noScale_noModel	85.16%	86.87%	84.61%	86.75%		85.09%		82.81%	86.10%		85.22%	83.82%
	SHAP_noScale_Model	85.89%			86.06%					86.77%		85.39%	
F1	Supervised CDD	22.62%	22.62%	22.62%	23.19%	23.19%	23.19%	22.73%	22.73%	22.73%	24.32%	24.32%	24.32%
	CDSeer - original	85.59%	86.61%	84.00%	83.28%	84.84%	83.81%	83.81%	81.57%	81.34%	81.62%	82.43%	82.19%
	Random selection of label candidates	84.20%	84.32%	83.53%	81.93%	79.93%	86.35%	81.85%	79.83%	82.01%	83.58%	80.79%	81.36%
	MinMaxScaling	84.14%	76.51%	83.33%	84.97%	84.12%	84.43%	80.77%	75.83%	74.79%	82.82%	84.39%	80.98%
	Append model prediction	85.42%	82.54%	85.67%	85.38%	85.38%	86.90%	85.87%	86.14%	83.95%	81.97%	84.93%	84.56%
	MinMaxScaling + append model prediction	85.42%	85.39%	83.20%	86.36%	87.71%	84.67%	85.94%	76.27%	76.08%	84.14%	84.43%	82.92%
	SHAP_noScale_noModel	85.05%	86.80%	84.56%	86.62%		84.97%		81.57%	85.08%		84.17%	82.74%
	SHAP_noScale_Model	85.83%			85.98%					85.80%		84.41%	
Drift precision (1000)	Supervised CDD	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	CDSeer - original	30.74%	18.93%	26.72%	22.30%	11.94%	22.26%	20.60%	30.38%	29.05%	22.22%	18.69%	23.52%
	Random selection of label candidates	19.84%	26.03%	17.52%	19.88%	11.86%	25.90%	20.24%	24.55%	26.83%	16.21%	22.67%	21.52%
	MinMaxScaling	22.86%	45.71%	38.26%	13.33%	14.52%	16.05%	19.24%	40.48%	43.00%	11.90%	14.72%	28.17%
	Append model prediction	16.86%	28.44%	19.68%	27.29%	19.17%	39.62%	22.88%	25.71%	21.43%	28.10%	22.76%	18.69%
	MinMaxScaling + append model prediction	31.00%	36.16%	35.05%	31.43%	34.67%	25.14%	24.00%	48.67%	38.38%	27.14%	28.28%	28.21%
	SHAP_noScale_noModel	23.00%	23.16%	25.39%	32.67%		19.40%		30.38%	28.86%		29.05%	25.71%
	SHAP_noScale_Model	38.00%			34.29%					36.00%		22.22%	
Drift recall (1000)	Supervised CDD	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	CDSeer - original	90.00%	70.00%	80.00%	80.00%	40.00%	70.00%	70.00%	100.00%	80.00%	70.00%	60.00%	70.00%
	Random selection of label candidates	70.00%	90.00%	60.00%	70.00%	30.00%	90.00%	70.00%	70.00%	90.00%	50.00%	60.00%	60.00%
	MinMaxScaling	70.00%	100.00%	100.00%	40.00%	43.33%	60.00%	60.00%	90.00%	90.00%	40.00%	60.00%	90.00%
	Append model prediction	60.00%	73.33%	70.00%	80.00%	70.00%	100.00%	70.00%	90.00%	70.00%	90.00%	73.33%	60.00%
	MinMaxScaling + append model prediction	70.00%	100.00%	100.00%	100.00%	100.00%	80.00%	70.00%	90.00%	90.00%	80.00%	90.00%	90.00%
	SHAP_noScale_noModel	60.00%	80.00%	80.00%	90.00%		70.00%		100.00%	80.00%		100.00%	80.00%
	SHAP_noScale_Model	58.86%	51.05%	55.62%	44.33%	46.79%	44.43%	48.00%	52.86%	51.57%	42.86%	42.10%	40.17%
Drift precision (2000)	Supervised CDD		100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	CDSeer - original	58.19%	51.43%	42.17%	37.38%	42.32%	45.71%	46.07%	42.24%	53.57%	41.67%	46.07%	46.93%
	Random selection of label candidates	44.13%	46.35%	37.05%	41.79%	43.71%	48.00%	52.86%	41.33%	44.13%	41.79%	53.33%	45.43%
	MinMaxScaling	58.86%	51.05%	55.62%	44.33%	46.79%	44.43%	48.00%	52.86%	51.57%	42.86%	42.10%	40.17%
	Append model prediction	55.57%	53.11%	48.89%	54.59%	47.50%	61.52%	58.26%	60.00%	42.86%	49.52%	56.19%	58.93%
	MinMaxScaling + append model prediction	59.00%	65.62%	51.90%	49.52%	65.33%	52.00%	66.00%	60.33%	58.76%	50.00%	49.22%	60.95%
	SHAP_noScale_noModel	46.33%	57.11%	43.50%	41.67%		45.00%		42.24%	59.90%		46.90%	46.07%
	SHAP_noScale_Model	58.86%	51.05%	55.62%	44.33%	46.79%	44.43%	48.00%	52.86%	51.57%	42.86%	42.10%	40.17%
Drift recall (2000)	Supervised CDD	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	CDSeer - original	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	93.33%	100.00%	100.00%
	Random selection of label candidates	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	MinMaxScaling	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	Append model prediction	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	90.00%	100.00%	100.00%
	MinMaxScaling + append model prediction	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	90.00%	100.00%	100.00%	100.00%	100.00%
	SHAP_noScale_noModel	90.00%	100.00%	100.00%	100.00%		100.00%		100.00%	100.00%		100.00%	100.00%
	SHAP_noScale_Model	100.00%			100.00%					100.00%		100.00%	

Fig. 23: Results for SINE datasets across all variants, in backwards retraining mode

Reference model retraining mode: forwards		Dataset version											
		Balanced, no noise	Balanced, no noise, stretched (v1)	Balanced, no noise, stretched (v2)	Balanced, noise	Balanced, noise, stretched (v1)	Balanced, noise, stretched (v2)	Imbalanced, no noise	Imbalanced, no noise, stretched	Imbalanced, no noise, stretched	Imbalanced, noise	Imbalanced, noise, stretched (v1)	Imbalanced, noise, stretched (v2)
Accuracy	Supervised CDD	96.79%	96.79%	96.79%	96.69%	96.69%	96.69%	97.49%	97.49%	97.49%	96.57%	96.57%	96.57%
	CDSeer - original	89.37%	83.92%	86.14%	84.32%	84.39%	87.78%	86.45%	84.68%	86.47%	87.16%	87.51%	85.07%
	Random selection of label candidates	83.76%	87.22%	76.27%	83.17%	86.88%	80.76%	87.86%	89.92%	86.73%	81.37%	83.37%	83.02%
	MinMaxScaling	85.66%	92.81%	85.37%	84.88%	85.15%	84.23%	86.09%	84.51%	90.88%	78.61%	84.18%	87.29%
	Append model prediction	85.80%	85.68%	87.04%	89.00%	89.34%	88.90%	84.69%	86.97%	88.83%	87.30%	86.66%	88.03%
	MinMaxScaling + append model prediction	85.90%	92.09%	92.03%	87.39%	88.00%	88.82%	86.36%	90.13%	90.08%	87.59%	88.58%	89.09%
	SHAP_noScale_noModel	86.08%			86.17%					89.65%		89.19%	
	SHAP_noScale_Model	89.18%			88.57%					86.62%		88.72%	
BalancedAccuracy	Supervised CDD	96.79%	96.79%	96.79%	96.69%	96.69%	96.69%	97.48%	97.48%	97.48%	96.52%	96.52%	96.52%
	CDSeer - original	89.37%	83.92%	86.14%	84.32%	84.40%	87.78%	86.44%	84.64%	86.42%	87.17%	87.47%	85.01%
	Random selection of label candidates	83.76%	87.22%	76.28%	83.17%	86.88%	80.76%	87.83%	89.90%	86.67%	81.31%	83.35%	83.05%
	MinMaxScaling	85.66%	92.82%	85.37%	84.88%	85.15%	84.24%	86.08%	84.46%	90.86%	78.68%	84.21%	87.26%
	Append model prediction	85.80%	85.68%	87.04%	89.00%	89.34%	88.90%	84.66%	86.97%	88.83%	87.27%	86.63%	87.98%
	MinMaxScaling + append model prediction	85.90%	92.10%	92.03%	87.40%	88.01%	88.82%	86.34%	90.15%	90.07%	87.57%	88.54%	89.06%
	SHAP_noScale_noModel	86.08%			86.18%					89.63%		89.18%	
	SHAP_noScale_Model	89.18%			88.57%					86.62%		88.72%	
F1	Supervised CDD	96.77%	96.77%	96.77%	96.68%	96.68%	96.68%	97.29%	97.29%	97.29%	96.31%	96.31%	96.31%
	CDSeer - original	89.27%	83.92%	86.12%	84.14%	84.25%	87.66%	85.61%	83.70%	85.43%	86.41%	86.74%	84.08%
	Random selection of label candidates	83.59%	87.13%	76.43%	83.16%	86.76%	80.86%	87.04%	89.25%	85.78%	80.17%	82.42%	82.25%
	MinMaxScaling	85.61%	92.76%	85.41%	84.70%	85.04%	84.15%	85.23%	83.29%	90.22%	77.76%	83.19%	86.57%
	Append model prediction	85.71%	85.62%	86.92%	88.87%	89.27%	88.79%	83.66%	86.11%	88.10%	86.49%	85.88%	87.20%
	MinMaxScaling + append model prediction	85.77%	92.04%	91.97%	87.25%	87.87%	88.71%	85.45%	89.50%	89.37%	86.82%	87.86%	88.39%
	SHAP_noScale_noModel	86.00%			86.00%					88.92%		88.49%	
	SHAP_noScale_Model	89.09%			88.44%					85.80%		88.03%	
Drift precision (1000)	Supervised CDD	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	CDSeer - original	38.00%	35.00%	28.33%	41.33%	32.00%	38.33%	36.67%	26.67%	36.67%	20.67%	25.67%	29.00%
	Random selection of label candidates	29.00%	41.67%	26.67%	29.33%	15.67%	24.00%	29.00%	20.67%	22.33%	24.00%	34.67%	43.33%
	MinMaxScaling	35.00%	52.00%	39.67%	16.67%	21.67%	25.00%	39.00%	28.67%	55.00%	20.67%	20.67%	36.33%
	Append model prediction	60.00%	50.00%	66.67%	56.67%	40.00%	56.67%	36.67%	70.00%	66.67%	50.00%	51.67%	63.33%
	MinMaxScaling + append model prediction	46.67%	100.00%	90.00%	73.33%	73.33%	66.67%	63.33%	86.67%	73.33%	71.67%	56.67%	56.67%
	SHAP_noScale_noModel	63.33%			80.00%					50.00%		48.33%	
	SHAP_noScale_Model	66.67%			66.67%					70.00%		50.00%	
Drift recall (1000)	Supervised CDD	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	CDSeer - original	80.00%	70.00%	50.00%	80.00%	70.00%	60.00%	70.00%	50.00%	50.00%	40.00%	50.00%	60.00%
	Random selection of label candidates	60.00%	80.00%	50.00%	60.00%	30.00%	50.00%	60.00%	40.00%	40.00%	40.00%	70.00%	70.00%
	MinMaxScaling	60.00%	90.00%	80.00%	30.00%	40.00%	50.00%	60.00%	70.00%	90.00%	40.00%	40.00%	70.00%
	Append model prediction	70.00%	60.00%	70.00%	70.00%	60.00%	70.00%	40.00%	80.00%	70.00%	60.00%	80.00%	80.00%
	MinMaxScaling + append model prediction	50.00%	100.00%	90.00%	80.00%	80.00%	90.00%	70.00%	100.00%	90.00%	90.00%	70.00%	80.00%
	SHAP_noScale_noModel	70.00%			80.00%					60.00%		80.00%	
	SHAP_noScale_Model	90.00%			80.00%					90.00%		70.00%	
Drift precision (2000)	Supervised CDD	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	CDSeer - original	42.00%	40.00%	43.33%	49.33%	37.00%	51.67%	46.67%	48.33%	46.67%	30.67%	48.00%	42.00%
	Random selection of label candidates	48.00%	48.33%	38.33%	37.33%	44.67%	35.67%	44.00%	40.67%	40.33%	48.00%	51.33%	56.67%
	MinMaxScaling	50.67%	56.00%	39.67%	47.33%	47.33%	34.00%	48.00%	32.67%	60.00%	36.33%	34.67%	49.33%
	Append model prediction	86.67%	80.00%	86.67%	80.00%	60.00%	73.33%	93.33%	86.67%	86.67%	73.33%	63.33%	80.00%
	MinMaxScaling + append model prediction	93.33%	100.00%	90.00%	93.33%	80.00%	73.33%	93.33%	86.67%	80.00%	76.67%	73.33%	63.33%
	SHAP_noScale_noModel	86.67%			93.33%					73.33%		53.33%	
	SHAP_noScale_Model	93.33%			80.00%					76.67%		50.00%	
Drift recall (2000)	Supervised CDD	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	CDSeer - original	90.00%	80.00%	80.00%	100.00%	80.00%	80.00%	90.00%	90.00%	60.00%	60.00%	90.00%	90.00%
	Random selection of label candidates	100.00%	90.00%	70.00%	80.00%	80.00%	70.00%	90.00%	80.00%	80.00%	80.00%	100.00%	90.00%
	MinMaxScaling	90.00%	100.00%	80.00%	90.00%	90.00%	70.00%	80.00%	80.00%	100.00%	70.00%	70.00%	100.00%
	Append model prediction	100.00%	100.00%	100.00%	100.00%	90.00%	90.00%	100.00%	100.00%	90.00%	90.00%	100.00%	100.00%
	MinMaxScaling + append model prediction	100.00%	100.00%	90.00%	100.00%	90.00%	100.00%	100.00%	100.00%	100.00%	100.00%	90.00%	90.00%
	SHAP_noScale_noModel	100.00%			100.00%					90.00%		90.00%	
	SHAP_noScale_Model	100.00%			100.00%					100.00%		70.00%	

Fig. 24: Results for SINE datasets across all variants, in forwards retraining mode

Reference model retraining mode: backwards		Dataset version											
		Redund. cols	Redund. cols, stretched (v1)	Redund. cols, stretched (v2)	Original	Original, stretched (v1)	Original, stretched (v2)	Original, noise, redund. cols	Original, noise, redund. cols, stretched (v1)	Original, noise, redund. cols, stretched (v2)	Original, noise	Original, noise, stretched (v1)	Original, noise, stretched (v2)
Accuracy	Supervised CDD	95.07%	95.07%	95.07%	94.88%	94.88%	94.88%	92.31%	92.31%	92.31%	92.30%	92.30%	92.30%
	CDSeer - original	94.67%	95.07%	95.49%	95.50%	95.73%	95.36%	93.55%	93.21%	93.98%	94.07%	94.23%	93.62%
	Random selection of label candidates	94.77%	94.40%	95.84%	94.82%	96.38%	94.94%	93.13%	93.69%	93.18%	92.90%	92.79%	93.99%
	MinMaxScaling	94.95%	94.95%	94.95%	95.44%	95.44%	95.44%	92.95%	93.18%	93.18%	93.49%	93.49%	93.49%
	Append model prediction	94.89%	94.85%	94.53%	94.57%	95.65%	94.49%	93.36%	93.61%	93.02%	93.49%	94.16%	94.11%
	MinMaxScaling + append model prediction	94.48%	94.48%	94.48%	93.52%	93.52%	93.52%	91.48%	91.73%	91.73%	93.06%	93.06%	93.06%
	SHAP_noScale_noModel		94.67%	95.49%	95.86%	94.97%	95.74%			92.41%	94.26%	94.23%	93.54%
	SHAP_noScale_Model		94.92%		94.94%					92.97%	92.46%		
BalancedAccuracy	Supervised CDD	95.14%	95.14%	95.14%	95.26%	95.26%	95.26%	91.98%	91.98%	91.98%	92.17%	92.17%	92.17%
	CDSeer - original	94.20%	94.45%	95.02%	95.29%	95.22%	94.97%	92.44%	92.25%	92.70%	93.04%	93.14%	92.73%
	Random selection of label candidates	94.39%	93.77%	94.98%	93.92%	95.97%	93.65%	92.49%	92.51%	92.27%	92.58%	92.09%	92.81%
	MinMaxScaling	94.51%	94.51%	94.51%	95.34%	95.34%	95.34%	92.00%	92.24%	92.24%	92.65%	92.65%	92.65%
	Append model prediction	94.36%	94.49%	93.83%	94.31%	95.33%	94.30%	92.02%	92.63%	92.20%	92.78%	93.34%	93.32%
	MinMaxScaling + append model prediction	93.57%	93.57%	93.57%	92.58%	92.58%	92.58%	91.02%	90.36%	90.36%	91.61%	91.61%	91.61%
	SHAP_noScale_noModel		94.24%	95.04%	95.28%	94.70%	95.23%			91.60%	93.46%	93.14%	92.72%
	SHAP_noScale_Model		94.25%		94.57%					92.00%	91.71%		
F1	Supervised CDD	96.31%	96.31%	96.31%	96.15%	96.15%	96.15%	94.21%	94.21%	94.21%	94.17%	94.17%	94.17%
	CDSeer - original	96.04%	96.35%	96.66%	96.65%	96.85%	96.57%	95.23%	94.95%	95.56%	95.61%	95.73%	95.24%
	Random selection of label candidates	96.12%	95.85%	96.95%	96.20%	97.33%	96.32%	94.86%	95.34%	94.92%	94.64%	94.61%	95.56%
	MinMaxScaling	96.25%	96.25%	96.25%	96.60%	96.60%	96.60%	94.75%	94.92%	94.92%	95.16%	95.16%	95.16%
	Append model prediction	96.22%	96.17%	95.96%	95.95%	96.78%	95.89%	95.10%	95.26%	94.79%	95.14%	95.66%	95.62%
	MinMaxScaling + append model prediction	95.93%	95.93%	95.93%	95.20%	95.20%	95.20%	93.57%	93.87%	93.87%	94.89%	94.89%	94.89%
	SHAP_noScale_noModel		96.04%	96.66%	96.95%	96.25%	96.86%			94.33%	95.73%	95.73%	95.18%
	SHAP_noScale_Model		96.24%		96.24%					94.76%	94.35%		
Drift precision (1000)	Supervised CDD	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	50.00%	50.00%	50.00%	50.00%	50.00%	50.00%
	CDSeer - original	32.00%	13.33%	19.00%	11.67%	19.67%	18.00%	16.00%	5.36%	21.00%	19.67%	19.00%	13.00%
	Random selection of label candidates	17.33%	14.19%	15.86%	4.00%	33.19%	6.67%	32.86%	8.00%	20.38%	26.67%	20.33%	31.67%
	MinMaxScaling	12.33%	12.33%	12.33%	18.33%	18.33%	18.33%	8.33%	8.33%	8.33%	5.00%	5.00%	5.00%
	Append model prediction	7.86%	14.05%	9.00%	11.67%	7.33%	16.67%	5.71%	19.75%	22.33%	3.33%	17.94%	10.67%
	MinMaxScaling + append model prediction	0.00%	0.00%	0.00%	10.00%	10.00%	10.00%	21.67%	16.67%	16.67%	0.00%	0.00%	0.00%
	SHAP_noScale_noModel		11.67%	17.00%	5.00%	14.67%	16.67%			14.67%	19.67%	19.00%	20.00%
	SHAP_noScale_Model		18.05%		6.67%					23.00%	0.00%		
Drift recall (1000)	Supervised CDD	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	50.00%	50.00%	50.00%	50.00%	50.00%	50.00%
	CDSeer - original	60.00%	40.00%	40.00%	20.00%	40.00%	40.00%	40.00%	20.00%	50.00%	50.00%	40.00%	30.00%
	Random selection of label candidates	30.00%	40.00%	40.00%	10.00%	70.00%	10.00%	70.00%	20.00%	60.00%	60.00%	50.00%	80.00%
	MinMaxScaling	30.00%	30.00%	30.00%	30.00%	30.00%	30.00%	20.00%	20.00%	20.00%	10.00%	10.00%	10.00%
	Append model prediction	20.00%	40.00%	20.00%	20.00%	20.00%	30.00%	20.00%	50.00%	50.00%	10.00%	50.00%	30.00%
	MinMaxScaling + append model prediction	0.00%	0.00%	0.00%	10.00%	10.00%	10.00%	30.00%	20.00%	20.00%	0.00%	0.00%	0.00%
	SHAP_noScale_noModel		30.00%	40.00%	10.00%	30.00%	40.00%			30.00%	50.00%	40.00%	40.00%
	SHAP_noScale_Model		50.00%		10.00%					50.00%	0.00%		
Drift precision (2000)	Supervised CDD	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	CDSeer - original	47.67%	24.00%	39.67%	23.33%	30.33%	28.00%	44.00%	16.21%	32.33%	39.33%	40.67%	34.67%
	Random selection of label candidates	24.67%	21.05%	36.19%	17.00%	43.90%	26.67%	41.19%	30.00%	28.38%	30.67%	29.33%	41.67%
	MinMaxScaling	24.67%	24.67%	24.67%	27.33%	27.33%	27.33%	31.33%	33.33%	33.33%	20.67%	20.67%	20.67%
	Append model prediction	21.38%	22.05%	17.33%	41.67%	32.71%	28.33%	28.76%	28.83%	34.33%	19.67%	24.16%	26.67%
	MinMaxScaling + append model prediction	15.67%	15.67%	15.67%	26.67%	26.67%	26.67%	26.67%	26.67%	26.67%	30.00%	30.00%	30.00%
	SHAP_noScale_noModel		22.38%	34.00%	32.00%	25.33%	35.00%			26.00%	39.33%	40.67%	33.33%
	SHAP_noScale_Model		30.38%		31.67%					34.67%	10.67%		
Drift recall (2000)	Supervised CDD	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	CDSeer - original	70.00%	60.00%	83.33%	40.00%	60.00%	53.33%	80.00%	50.00%	73.33%	90.00%	83.33%	70.00%
	Random selection of label candidates	50.00%	60.00%	73.33%	40.00%	90.00%	50.00%	80.00%	60.00%	80.00%	63.33%	63.33%	80.00%
	MinMaxScaling	60.00%	60.00%	60.00%	50.00%	50.00%	50.00%	70.00%	70.00%	70.00%	40.00%	40.00%	40.00%
	Append model prediction	50.00%	60.00%	40.00%	60.00%	70.00%	50.00%	70.00%	80.00%	70.00%	43.33%	70.00%	60.00%
	MinMaxScaling + append model prediction	40.00%	40.00%	40.00%	30.00%	30.00%	30.00%	40.00%	30.00%	30.00%	40.00%	40.00%	40.00%
	SHAP_noScale_noModel		60.00%	73.33%	70.00%	50.00%	70.00%			60.00%	90.00%	83.33%	70.00%
	SHAP_noScale_Model		30.38%		31.67%					34.67%	10.67%		

Fig. 25: Results for SEA datasets across all variants, in backwards retraining mode

Reference model retraining mode: backwards		Harvard gradual drift datasets				Chocolate datasets			
		Mixed 0101	Mixed 1010	RT 873985678962563	Sine 0123	Abrupt recurrent, binary	Abrupt recurrent, multiclass	Abrupt rotating, binary	Abrupt rotating, multiclass
Accuracy	Supervised CDD	78.41%	79.12%	65.69%	77.50%	21.38%	19.30%	63.58%	62.44%
	CDSeer - original	84.75%	84.71%	74.70%	85.92%	73.07%	75.28%	79.29%	82.14%
	Random selection of label candidates	84.34%	84.32%	74.80%	84.21%	71.73%	75.77%	79.17%	82.01%
	MinMaxScaling	84.19%	84.52%	74.75%	84.93%	57.67%	71.07%	75.02%	81.01%
	Append model prediction	84.97%	84.06%	74.68%	85.77%	74.52%	75.77%	79.65%	79.64%
	MinMaxScaling + append model prediction	85.16%	84.85%	74.95%	85.65%	79.10%	78.01%	84.64%	83.57%
BalancedAccuracy	Supervised CDD	78.40%	79.12%	65.43%	77.50%	21.38%	19.28%	63.58%	62.76%
	CDSeer - original	84.75%	84.71%	74.46%	85.92%	73.06%	75.27%	79.29%	82.24%
	Random selection of label candidates	84.34%	84.32%	74.54%	84.21%	71.73%	75.77%	79.17%	82.05%
	MinMaxScaling	84.19%	84.52%	74.48%	84.93%	57.67%	71.07%	75.02%	81.13%
	Append model prediction	84.97%	84.06%	74.43%	85.77%	74.52%	75.77%	79.65%	79.75%
	MinMaxScaling + append model prediction	85.16%	84.85%	74.70%	85.65%	79.10%	78.01%	84.64%	83.62%
F1 (binary/macro)	Supervised CDD	78.45%	78.34%	61.63%	75.99%	21.33%	19.28%	63.49%	62.76%
	CDSeer - original	84.32%	84.32%	71.79%	84.09%	71.41%	75.27%	79.24%	82.21%
	Random selection of label candidates	84.21%	84.58%	71.62%	84.82%	57.29%	71.07%	75.09%	81.11%
	MinMaxScaling	84.21%	84.58%	71.62%	84.82%	57.29%	71.07%	75.09%	81.11%
	Append model prediction	84.97%	84.04%	71.72%	85.72%	74.27%	75.77%	79.80%	79.75%
	MinMaxScaling + append model prediction	85.20%	84.86%	72.06%	85.60%	78.91%	78.01%	84.76%	83.65%
Drift precision (1000)	Supervised CDD	66.67%	66.67%	60.00%	60.00%	100.00%	100.00%	100.00%	80.00%
	CDSeer - original	7.13%	9.98%	6.81%	4.56%	24.02%	16.84%	29.02%	16.43%
	Random selection of label candidates	9.13%	11.26%	11.06%	11.94%	24.44%	24.79%	23.69%	33.33%
	MinMaxScaling	10.79%	7.72%	8.45%	7.98%	12.86%	24.71%	28.29%	16.65%
	Append model prediction	9.15%	9.33%	7.86%	8.69%	23.11%	26.19%	13.64%	21.18%
	MinMaxScaling + append model prediction	9.03%	9.76%	9.65%	2.67%	21.27%	20.29%	21.09%	21.53%
Drift recall (1000)	Supervised CDD	100.00%	100.00%	75.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	CDSeer - original	46.67%	66.67%	53.33%	26.67%	53.00%	45.00%	55.00%	40.00%
	Random selection of label candidates	55.00%	80.00%	81.67%	73.33%	50.00%	71.00%	40.00%	77.00%
	MinMaxScaling	66.67%	53.33%	66.67%	46.67%	20.00%	65.00%	40.00%	40.00%
	Append model prediction	66.67%	60.00%	50.00%	41.67%	60.00%	85.00%	35.00%	45.00%
	MinMaxScaling + append model prediction	66.67%	66.67%	73.33%	13.33%	65.00%	77.00%	56.00%	64.00%
Drift precision (2000)	Supervised CDD	88.89%	100.00%	80.00%	90.00%	100.00%	100.00%	100.00%	80.00%
	CDSeer - original	25.48%	24.93%	15.93%	23.52%	48.98%	45.54%	48.89%	37.00%
	Random selection of label candidates	21.09%	22.59%	19.88%	24.04%	54.44%	47.66%	39.40%	52.00%
	MinMaxScaling	24.77%	25.00%	18.09%	24.73%	27.74%	39.51%	40.14%	38.54%
	Append model prediction	25.37%	22.74%	18.45%	28.33%	43.56%	49.14%	30.34%	40.40%
	MinMaxScaling + append model prediction	27.10%	28.17%	21.89%	30.70%	41.94%	37.94%	46.70%	37.90%
Drift recall (2000)	Supervised CDD	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	CDSeer - original	100.00%	100.00%	95.00%	100.00%	95.00%	96.67%	90.00%	82.00%
	Random selection of label candidates	95.00%	100.00%	96.67%	93.33%	95.00%	100.00%	66.00%	93.33%
	MinMaxScaling	100.00%	100.00%	92.00%	93.33%	42.00%	90.00%	55.00%	79.00%
	Append model prediction	100.00%	91.00%	95.00%	100.00%	100.00%	100.00%	67.00%	76.00%
	MinMaxScaling + append model prediction	100.00%	100.00%	96.00%	100.00%	100.00%	100.00%	96.00%	89.14%

Fig. 26: Results for Harvard and Chocolate datasets with SVC reference model, in backwards retraining mode

Reference model retraining mode: forwards		Harvard gradual drift datasets				Chocolate datasets			
		Mixed 0101	Mixed 1010	RT 873985678962563	Sine 0123	Abrupt recurrent, binary	Abrupt recurrent, multiclass	Abrupt rotating, binary	Abrupt rotating, multiclass
Accuracy	**Supervised CDD**	78.37%	78.78%	74.61%	81.16%	93.26%	93.23%	91.48%	92.19%
	CDSeer - original	84.70%	86.03%	74.66%	86.93%	77.50%	74.68%	76.50%	85.66%
	Random selection of label candidates	84.78%	83.90%	75.29%	84.94%	67.82%	77.03%	76.16%	80.52%
	MinMaxScaling	85.60%	84.97%	74.83%	86.25%	69.71%	75.10%	73.31%	79.41%
	Append model prediction	85.65%	85.12%	74.69%	87.18%	68.85%	79.19%	81.72%	82.07%
	MinMaxScaling + append model prediction	84.75%	85.16%	75.68%	87.89%	77.40%	78.72%	84.23%	84.43%
BalancedAccuracy	**Supervised CDD**	78.37%	78.78%	74.20%	81.16%	93.27%	93.22%	91.49%	92.23%
	CDSeer - original	84.70%	86.03%	74.49%	86.93%	77.50%	74.69%	76.51%	85.79%
	Random selection of label candidates	84.78%	83.90%	75.03%	84.94%	67.82%	77.03%	76.16%	80.71%
	MinMaxScaling	85.60%	84.97%	74.62%	86.25%	69.72%	75.10%	73.31%	79.55%
	Append model prediction	85.65%	85.12%	74.53%	87.18%	68.87%	79.19%	81.72%	82.09%
	MinMaxScaling + append model prediction	84.75%	85.16%	75.48%	87.89%	77.40%	78.72%	84.23%	84.56%
F1 (binary/macro)	**Supervised CDD**	77.87%	79.33%	70.90%	80.04%	93.24%	93.22%	91.51%	92.21%
	CDSeer - original	84.70%	86.00%	71.42%	86.97%	77.34%	74.68%	76.65%	85.73%
	Random selection of label candidates	84.81%	83.85%	72.36%	84.92%	67.64%	77.02%	76.35%	80.66%
	MinMaxScaling	85.53%	84.86%	71.89%	86.27%	69.52%	75.10%	73.50%	79.53%
	Append model prediction	85.57%	85.05%	72.34%	87.20%	68.96%	79.18%	81.89%	82.14%
	MinMaxScaling + append model prediction	84.75%	85.18%	73.14%	87.87%	77.14%	78.72%	84.36%	84.51%
Drift precision (1000)	**Supervised CDD**	100.00%	100.00%	60.00%	60.00%	100.00%	100.00%	100.00%	100.00%
	CDSeer - original	5.00%	14.65%	10.71%	9.45%	23.24%	27.71%	23.33%	23.57%
	Random selection of label candidates	10.76%	5.17%	10.62%	6.69%	9.52%	19.29%	14.71%	16.67%
	MinMaxScaling	13.49%	13.55%	5.71%	12.79%	18.50%	30.36%	21.33%	19.17%
	Append model prediction	14.30%	8.64%	7.34%	12.22%	10.00%	28.93%	10.71%	0.00%
	MinMaxScaling + append model prediction	13.03%	8.10%	6.23%	15.02%	40.76%	46.64%	25.71%	36.19%
Drift recall (1000)	**Supervised CDD**	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	CDSeer - original	20.00%	53.33%	46.67%	33.33%	35.00%	40.00%	30.00%	40.00%
	Random selection of label candidates	40.00%	20.00%	46.67%	26.67%	15.00%	35.00%	20.00%	30.00%
	MinMaxScaling	53.33%	53.33%	26.67%	46.67%	25.00%	55.00%	25.00%	35.00%
	Append model prediction	53.33%	33.33%	33.33%	33.33%	15.00%	55.00%	20.00%	0.00%
	MinMaxScaling + append model prediction	46.67%	33.33%	26.67%	40.00%	60.00%	75.00%	40.00%	60.00%
Drift precision (2000)	**Supervised CDD**	100.00%	100.00%	60.00%	60.00%	100.00%	100.00%	100.00%	100.00%
	CDSeer - original	24.70%	27.84%	19.55%	31.39%	49.14%	41.90%	40.67%	51.43%
	Random selection of label candidates	25.00%	22.24%	19.20%	23.83%	54.57%	40.36%	24.90%	39.17%
	MinMaxScaling	23.31%	22.49%	18.79%	23.27%	47.50%	52.86%	38.00%	35.71%
	Append model prediction	28.27%	23.79%	19.36%	37.94%	35.24%	36.43%	42.70%	30.50%
	MinMaxScaling + append model prediction	22.89%	25.84%	21.76%	36.48%	61.33%	55.33%	57.14%	54.76%
Drift recall (2000)	**Supervised CDD**	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	CDSeer - original	93.33%	93.33%	86.67%	100.00%	75.00%	60.00%	55.00%	85.00%
	Random selection of label candidates	93.33%	81.67%	86.67%	81.67%	80.00%	75.00%	35.00%	70.00%
	MinMaxScaling	93.33%	80.00%	86.67%	86.67%	70.00%	95.00%	45.00%	65.00%
	Append model prediction	100.00%	93.33%	86.67%	93.33%	55.00%	70.00%	80.00%	60.00%
	MinMaxScaling + append model prediction	80.00%	100.00%	88.33%	100.00%	90.00%	90.00%	90.00%	86.00%

Fig. 27: Results for Harvard and Chocolate datasets with SVC reference model, in forwards retraining mode

	Accuracy			F1		
	Fully supervised - backwards retraining of reference model	Fully supervised - forwards retraining of reference model	CDSeer - original backwards retrainign of reference model	Fully supervised - backwards retraining of reference model	Fully supervised - forwards retraining of reference model	CDSeer - original backwards retrainign of reference model
Abrupt recurrent, binary	20.20%	96.53%	72.33%	19.84%	96.50%	72.05%
Abrupt recurrent, multiclass	18.85%	97.49%	76.37%	18.62%	97.45%	76.21%
Abrupt recurrent, binary, noise, redundance	31.08%	77.99%	54.73%	28.70%	77.79%	53.55%
Abrupt recurrent, multiclass, noise, redundance	19.96%	93.45%	57.65%	21.05%	93.69%	57.52%
Abrupt recurrent, binary, noise	20.99%	95.57%	72.23%	20.74%	95.56%	72.11%
Abrupt recurrent, multiclass, noise	18.94%	96.19%	74.74%	18.97%	96.41%	74.98%
Abrupt recurrent, binary, redundance	34.14%	84.10%	52.65%	33.16%	83.75%	52.18%
Abrupt recurrent, multiclass, redundance	19.59%	95.03%	64.61%	18.46%	95.11%	64.05%
Abrupt rotating, binary	66.57%	91.53%	79.70%	66.53%	91.61%	79.70%
Abrupt rotating, multiclass	65.63%	91.94%	81.08%	62.74%	91.69%	80.08%
Abrupt rotating, binary, noise, redundance	64.00%	80.66%	64.33%	63.41%	80.53%	63.66%
Abrupt rotating, multiclass, noise, redundance	64.80%	86.43%	67.21%	62.03%	85.78%	66.16%
Abrupt rotating, binary, noise	66.57%	90.42%	75.17%	66.77%	90.51%	75.08%
Abrupt rotating, multiclass, noise	64.10%	91.25%	78.75%	62.00%	91.23%	77.94%
Abrupt rotating, binary, redundance	64.33%	80.14%	63.25%	64.18%	79.75%	62.95%
Abrupt rotating, multiclass, redundance	64.13%	87.00%	67.00%	63.09%	87.41%	66.06%

Fig. 28: Results for Chocolate datasets - BalancedAccuracy and F1

	Drift precision (1000)			Drift recall (1000)		
	Fully supervised - backwards retraining of reference model	Fully supervised - forwards retraining of reference model	CDSeer - original backwards retrainign of reference model	Fully supervised - backwards retraining of reference model	Fully supervised - forwards retraining of reference model	CDSeer - original backwards retrainign of reference model
Abrupt recurrent, binary	100.00%	100.00%	22.38%	100.00%	100.00%	50.00%
Abrupt recurrent, multiclass	100.00%	100.00%	19.24%	100.00%	100.00%	52.00%
Abrupt recurrent, binary, noise, redundance	100.00%	75.00%	0.00%	100.00%	75.00%	0.00%
Abrupt recurrent, multiclass, noise, redundance	100.00%	100.00%	6.67%	100.00%	100.00%	10.00%
Abrupt recurrent, binary, noise	100.00%	100.00%	20.83%	100.00%	100.00%	45.00%
Abrupt recurrent, multiclass, noise	100.00%	100.00%	21.45%	100.00%	100.00%	60.00%
Abrupt recurrent, binary, redundance	100.00%	100.00%	40.00%	100.00%	100.00%	40.00%
Abrupt recurrent, multiclass, redundance	100.00%	100.00%	8.33%	100.00%	100.00%	10.00%
Abrupt rotating, binary	80.00%	100.00%	23.67%	100.00%	100.00%	40.00%
Abrupt rotating, multiclass	80.00%	100.00%	20.34%	100.00%	100.00%	45.00%
Abrupt rotating, binary, noise, redundance	80.00%	100.00%	22.33%	100.00%	100.00%	25.00%
Abrupt rotating, multiclass, noise, redundance	80.00%	100.00%	22.33%	100.00%	100.00%	20.00%
Abrupt rotating, binary, noise	100.00%	100.00%	8.21%	100.00%	100.00%	15.00%
Abrupt rotating, multiclass, noise	66.67%	100.00%	22.96%	100.00%	100.00%	55.00%
Abrupt rotating, binary, redundance	66.67%	100.00%	10.00%	100.00%	100.00%	5.00%
Abrupt rotating, multiclass, redundance	80.00%	100.00%	16.67%	100.00%	100.00%	20.00%

Fig. 29: Results for Chocolate datasets - Drift Precision and Recall at 1000 step tolerance

	Drift precision (2000)			Drift recall (2000)		
	Fully supervised - backwards retraining of reference model	Fully supervised - forwards retraining of reference model	CDSeer - original backwards retrainign of reference model	Fully supervised - backwards retraining of reference model	Fully supervised - forwards retraining of reference model	CDSeer - original backwards retrainign of reference model
Abrupt recurrent, binary	100.00%	100.00%	45.21%	100.00%	100.00%	91.00%
Abrupt recurrent, multiclass	100.00%	100.00%	46.57%	100.00%	100.00%	91.00%
Abrupt recurrent, binary, noise, redundance	100.00%	75.00%	21.67%	100.00%	75.00%	20.00%
Abrupt recurrent, multiclass, noise, redundance	100.00%	100.00%	36.67%	100.00%	100.00%	45.00%
Abrupt recurrent, binary, noise	100.00%	100.00%	41.39%	100.00%	100.00%	85.00%
Abrupt recurrent, multiclass, noise	100.00%	100.00%	52.18%	100.00%	100.00%	100.00%
Abrupt recurrent, binary, redundance	100.00%	100.00%	43.33%	100.00%	100.00%	45.00%
Abrupt recurrent, multiclass, redundance	100.00%	100.00%	63.33%	100.00%	100.00%	70.00%
Abrupt rotating, binary	80.00%	100.00%	47.05%	100.00%	100.00%	85.00%
Abrupt rotating, multiclass	80.00%	100.00%	43.99%	100.00%	100.00%	82.00%
Abrupt rotating, binary, noise, redundance	80.00%	100.00%	48.00%	100.00%	100.00%	50.00%
Abrupt rotating, multiclass, noise, redundance	80.00%	100.00%	40.67%	100.00%	100.00%	35.00%
Abrupt rotating, binary, noise	100.00%	100.00%	29.79%	100.00%	100.00%	47.00%
Abrupt rotating, multiclass, noise	66.67%	100.00%	36.35%	100.00%	100.00%	75.00%
Abrupt rotating, binary, redundance	66.67%	100.00%	50.00%	100.00%	100.00%	30.00%
Abrupt rotating, multiclass, redundance	80.00%	100.00%	26.67%	100.00%	100.00%	30.00%

Fig. 30: Results for Chocolate datasets - Drift Precision and Recall at 2000 step tolerance

		Harvard gradual datasets - PH					Harvard gradual datasets - EDDM				
Reference model retraining mode: backwards		Mixed 0101	Mixed 1010	Sea 0123, noise 0.2	Sea 3210, noise 0.2	Sine 0123	Mixed 0101	Mixed 1010	Sea 0123, noise 0.2	Sea 3210, noise 0.2	Sine 0123
Accuracy	**Supervised CDD**	84.95%	82.62%	74.15%	74.64%	71.55%	82.98%	87.79%	75.45%	75.87%	80.35%
	CDSeer - original	88.14%	86.78%	76.37%	76.26%	90.11%	80.32%	70.14%	76.05%	75.73%	81.65%
	Random selection of label candidates	88.06%	84.37%	76.35%	76.27%	88.75%	83.22%	82.61%	75.64%	75.49%	65.89%
	MinMaxScaling	86.94%	85.03%	76.44%	76.54%	88.05%	77.52%	82.00%	75.35%	75.45%	69.27%
	Append model prediction	88.40%	87.96%	76.59%	76.14%	88.90%	72.87%	76.27%	75.06%	75.18%	78.09%
	MinMaxScaling + append model prediction	87.98%	86.61%	76.26%	76.27%	90.32%	66.59%	74.26%	75.46%	75.03%	80.02%
BalancedAccuracy	**Supervised CDD**	84.95%	82.62%	74.15%	74.64%	71.55%	82.98%	87.79%	75.45%	75.87%	80.35%
	CDSeer - original	88.14%	86.78%	76.37%	76.26%	90.11%	80.32%	70.14%	76.05%	75.72%	81.65%
	Random selection of label candidates	88.06%	84.37%	76.36%	76.27%	88.75%	83.22%	82.61%	75.64%	75.48%	65.89%
	MinMaxScaling	86.94%	85.03%	76.44%	76.54%	88.05%	77.52%	82.00%	75.35%	75.45%	69.27%
	Append model prediction	88.40%	87.96%	76.59%	76.14%	88.90%	72.87%	76.27%	75.07%	75.17%	78.09%
	MinMaxScaling + append model prediction	87.98%	86.61%	76.26%	76.27%	90.32%	66.59%	74.26%	75.47%	75.03%	80.02%
F1	**Supervised CDD**	84.91%	82.50%	74.59%	74.73%	71.13%	82.91%	87.76%	76.03%	75.51%	80.15%
	CDSeer - original	88.13%	86.77%	76.29%	76.12%	90.11%	80.25%	70.01%	76.02%	74.94%	81.56%
	Random selection of label candidates	88.04%	84.35%	76.49%	76.01%	88.74%	83.21%	82.59%	75.68%	74.53%	65.69%
	MinMaxScaling	86.95%	85.02%	76.43%	76.36%	88.02%	77.48%	82.02%	75.62%	75.40%	69.09%
	Append model prediction	88.38%	87.93%	76.60%	76.08%	88.89%	72.86%	76.14%	76.04%	74.22%	77.99%
	MinMaxScaling + append model prediction	87.99%	86.61%	76.25%	76.15%	90.33%	66.32%	74.08%	75.99%	74.51%	79.90%
Drift precision (1000)	**Supervised CDD**	50.00%	50.00%	50.00%	33.33%	54.55%	47.06%	46.67%	0.00%	0.00%	50.00%
	CDSeer - original	7.53%	13.05%	8.63%	5.11%	4.73%	8.80%	10.55%	3.40%	6.22%	6.54%
	Random selection of label candidates	11.92%	13.85%	7.68%	6.07%	14.47%	10.58%	12.81%	1.27%	6.97%	24.23%
	MinMaxScaling	14.82%	11.31%	7.83%	11.27%	10.64%	5.82%	0.00%	2.71%	7.47%	13.94%
	Append model prediction	13.14%	10.20%	6.19%	6.24%	10.54%	14.52%	11.43%	6.02%	1.54%	11.00%
	MinMaxScaling + append model prediction	15.12%	16.97%	6.49%	8.94%	8.99%	34.67%	20.22%	7.44%	8.14%	5.79%
Drift recall (1000)	**Supervised CDD**	100.00%	100.00%	100.00%	66.67%	100.00%	100.00%	87.50%	0.00%	0.00%	100.00%
	CDSeer - original	40.00%	66.67%	60.00%	33.33%	26.67%	70.35%	59.22%	38.81%	47.03%	16.67%
	Random selection of label candidates	66.67%	80.00%	53.33%	40.00%	73.33%	65.22%	82.81%	13.33%	32.06%	63.57%
	MinMaxScaling	75.00%	55.00%	53.33%	66.67%	60.00%	13.33%	0.00%	20.00%	28.33%	33.33%
	Append model prediction	56.67%	53.33%	40.00%	40.00%	46.67%	26.67%	23.33%	21.67%	6.67%	32.00%
	MinMaxScaling + append model prediction	61.67%	80.00%	46.67%	60.00%	40.00%	23.33%	26.67%	35.33%	40.00%	18.67%
Drift precision (2000)	**Supervised CDD**	83.33%	83.33%	66.67%	33.33%	81.82%	70.59%	66.67%	0.00%	0.00%	68.75%
	CDSeer - original	26.39%	29.49%	17.26%	15.48%	23.38%	16.59%	23.04%	7.04%	14.09%	31.64%
	Random selection of label candidates	24.96%	25.97%	15.10%	16.48%	28.10%	36.29%	27.33%	1.27%	12.73%	37.72%
	MinMaxScaling	28.18%	29.18%	18.04%	17.41%	25.56%	35.64%	7.33%	3.71%	19.47%	27.46%
	Append model prediction	35.51%	29.36%	14.44%	13.07%	30.33%	52.38%	45.71%	12.78%	1.54%	44.47%
	MinMaxScaling + append model prediction	32.68%	38.50%	11.30%	16.83%	29.24%	53.24%	60.00%	12.94%	11.99%	34.70%
Drift recall (2000)	**Supervised CDD**	100.00%	100.00%	100.00%	66.67%	100.00%	100.00%	100.00%	0.00%	0.00%	100.00%
	CDSeer - original	100.00%	100.00%	85.33%	80.00%	91.67%	84.64%	82.67%	54.11%	64.38%	75.33%
	Random selection of label candidates	100.00%	100.00%	88.33%	84.33%	100.00%	96.98%	96.40%	13.33%	35.00%	74.82%
	MinMaxScaling	100.00%	100.00%	93.33%	88.33%	93.33%	51.67%	13.33%	26.67%	42.67%	57.33%
	Append model prediction	100.00%	100.00%	86.67%	73.33%	96.00%	62.00%	73.33%	35.00%	6.67%	82.67%
	MinMaxScaling + append model prediction	100.00%	100.00%	65.00%	90.00%	95.00%	51.67%	70.00%	47.22%	50.33%	72.14%

Fig. 31: Results for Harvard datasets with PageHinkley and EDDM detectors, in backwards retraining mode

Reference model retraining mode:forwards		Harvard gradual datasets - PH					Harvard gradual datasets - EDDM				
		Mixed 0101	Mixed 1010	Sea 0123, noise 0.2	Sea 3210, noise 0.2	Sine 0123	Mixed 0101	Mixed 1010	Sea 0123, noise 0.2	Sea 3210, noise 0.2	Sine 0123
Accuracy	**Supervised CDD**	81.60%	78.91%	76.13%	76.82%	93.05%	95.45%	76.85%	75.23%	76.13%	96.47%
	CDSeer - original	89.21%	89.15%	76.28%	76.81%	89.25%	90.81%	92.43%	76.74%	76.54%	86.74%
	Random selection of label candidates	86.78%	88.73%	76.20%	76.74%	91.79%	90.75%	87.13%	75.57%	74.98%	89.61%
	MinMaxScaling	90.10%	88.89%	76.52%	76.38%	90.78%	92.58%	87.88%	75.87%	76.03%	92.45%
	Append model prediction	91.71%	90.65%	76.33%	76.44%	93.04%	92.73%	92.98%	76.45%	76.24%	93.14%
	MinMaxScaling + append model prediction	91.68%	92.48%	76.33%	76.51%	91.35%	93.39%	93.86%	76.58%	76.70%	92.71%
BalancedAccuracy	**Supervised CDD**	81.60%	78.91%	76.13%	76.82%	93.05%	95.45%	76.85%	75.24%	76.13%	96.47%
	CDSeer - original	89.21%	89.15%	76.28%	76.81%	89.25%	90.81%	92.43%	76.74%	76.54%	86.74%
	Random selection of label candidates	86.78%	88.73%	76.20%	76.74%	91.79%	90.75%	87.13%	75.58%	74.99%	89.61%
	MinMaxScaling	90.10%	88.89%	76.52%	76.38%	90.78%	92.58%	87.88%	75.87%	76.03%	92.45%
	Append model prediction	91.71%	90.65%	76.33%	76.44%	93.04%	92.73%	92.98%	76.45%	76.24%	93.14%
	MinMaxScaling + append model prediction	91.68%	92.48%	76.33%	76.51%	91.35%	93.39%	93.86%	76.58%	76.70%	92.71%
F1	**Supervised CDD**	81.57%	78.83%	76.11%	76.52%	92.99%	95.45%	76.68%	75.89%	75.82%	96.46%
	CDSeer - original	89.19%	89.14%	76.21%	76.54%	89.23%	90.80%	92.44%	76.71%	76.11%	86.58%
	Random selection of label candidates	86.77%	88.72%	76.34%	76.41%	91.80%	90.74%	87.12%	75.93%	74.96%	89.45%
	MinMaxScaling	90.09%	88.88%	76.58%	76.11%	90.74%	92.58%	87.87%	75.63%	75.56%	92.43%
	Append model prediction	91.70%	90.62%	76.32%	76.43%	93.05%	92.72%	92.97%	76.39%	75.77%	93.15%
	MinMaxScaling + append model prediction	91.69%	92.48%	76.29%	76.32%	91.36%	93.38%	93.86%	76.58%	76.53%	92.72%
Drift precision (1000)	**Supervised CDD**	50.00%	50.00%	50.00%	20.00%	40.00%	20.00%	14.29%	0.00%	9.09%	25.00%
	CDSeer - original	15.79%	15.44%	12.41%	12.15%	6.22%	8.61%	8.04%	9.09%	4.94%	7.24%
	Random selection of label candidates	17.71%	14.11%	6.30%	14.14%	9.56%	7.46%	8.42%	8.04%	8.05%	7.94%
	MinMaxScaling	13.11%	12.57%	9.49%	7.97%	15.90%	8.63%	6.91%	8.66%	6.75%	8.17%
	Append model prediction	21.11%	21.67%	7.87%	7.14%	24.67%	20.68%	21.15%	7.00%	6.76%	31.33%
	MinMaxScaling + append model prediction	19.52%	26.90%	5.73%	8.81%	24.00%	17.11%	13.74%	6.20%	9.97%	23.24%
Drift recall (1000)	**Supervised CDD**	100.00%	100.00%	66.67%	33.33%	66.67%	66.67%	33.33%	0.00%	33.33%	66.67%
	CDSeer - original	53.33%	53.33%	53.33%	53.33%	20.00%	86.67%	80.00%	93.33%	46.67%	66.67%
	Random selection of label candidates	53.33%	46.67%	26.67%	60.00%	33.33%	66.67%	80.00%	66.67%	66.67%	73.33%
	MinMaxScaling	40.00%	46.67%	40.00%	33.33%	53.33%	80.00%	66.67%	86.67%	60.00%	73.33%
	Append model prediction	60.00%	60.00%	33.33%	33.33%	46.67%	73.33%	86.67%	73.33%	66.67%	60.00%
	MinMaxScaling + append model prediction	46.67%	66.67%	26.67%	40.00%	33.33%	73.33%	53.33%	33.33%	53.33%	46.67%
Drift precision (2000)	**Supervised CDD**	66.67%	66.67%	50.00%	60.00%	80.00%	50.00%	28.57%	0.00%	18.18%	50.00%
	CDSeer - original	28.58%	23.09%	18.25%	19.77%	21.57%	14.45%	14.03%	16.97%	13.26%	14.52%
	Random selection of label candidates	32.29%	29.67%	16.06%	18.77%	25.44%	15.99%	14.88%	14.67%	14.08%	15.82%
	MinMaxScaling	31.45%	25.14%	18.85%	15.66%	31.58%	16.32%	13.83%	15.89%	13.65%	15.70%
	Append model prediction	36.90%	33.33%	19.93%	14.19%	59.00%	29.41%	34.04%	14.61%	14.09%	48.67%
	MinMaxScaling + append model prediction	40.48%	40.24%	15.74%	18.10%	62.00%	28.95%	27.28%	15.06%	16.16%	54.57%
Drift recall (2000)	**Supervised CDD**	100.00%	100.00%	66.67%	100.00%	100.00%	100.00%	50.00%	0.00%	66.67%	100.00%
	CDSeer - original	93.33%	80.00%	80.00%	80.00%	73.33%	95.00%	93.33%	100.00%	91.00%	88.33%
	Random selection of label candidates	100.00%	100.00%	66.67%	80.00%	86.67%	95.00%	88.33%	75.00%	76.00%	95.00%
	MinMaxScaling	100.00%	86.67%	80.00%	66.67%	100.00%	100.00%	92.00%	96.00%	85.00%	92.00%
	Append model prediction	100.00%	93.33%	86.67%	66.67%	100.00%	100.00%	100.00%	100.00%	86.00%	93.33%
	MinMaxScaling + append model prediction	100.00%	100.00%	73.33%	80.00%	86.67%	100.00%	100.00%	80.00%	80.00%	100.00%

Fig. 32: Results for Harvard dataset with PageHinkley detector, in forwards retraining mode