ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

# Drift Localization using Conformal Predictions

Fabian Hinder, Valerie Vaquet, Johannes Brinkrolf, and Barbara Hammer *

Bielefeld University – Faculty of Technology
Inspiration 1, 33619 Bielefeld – Germany

**Abstract**. Concept drift – the change of the distribution over time – poses significant challenges for learning systems and is of central interest for monitoring. Understanding drift is thus paramount, and drift localization – determining which samples are affected by the drift – is essential. While several approaches exist, most rely on local testing schemes, which tend to fail in high-dimensional, low-signal settings. In this work, we consider a fundamentally different approach based on conformal predictions. We discuss and show the shortcomings of common approaches and demonstrate the performance of our approach on state-of-the-art image datasets.

## 1 Introduction

In the classical batch setting of machine learning, data are assumed to originate from a single source and to be independently and identically distributed (i.i.d.). A relevant extension is to allow several different data distributions. Examples include transfer learning, where source and target domains differ; federated learning, where models are trained on data collected across distributed locations; and stream learning, where each time step may follow a different distribution [1, 2]. In all these settings, the fact that the distributions may differ, a phenomenon referred to as *concept drift* [1] or drift for short, plays a central role. In stream learning and system monitoring, this relates to changes over time; in transfer learning, to the differences between source and target distributions, etc.

Gaining understanding of the drift is imperative [2, 3]. Recent work on *model-based drift explanations* [3] has enabled the usage of generic explainable AI (XAI) techniques to obtain such insights. A key step of the approach is the identification of the affected samples, a task referred to as *drift localization* [2, 4]. While there are some methods addressing drift localization [2, 1, 5, 6, 4, 7, 8], they mainly rely on local statistical testing, which tends to fail on high-dimensional data, such as streams of images. In this work, we propose an alternative localization scheme by applying conformal prediction to the underlying idea of model-based drift localization [4]. We first summarize the setup, discuss the shortcomings of the related work, and recap conformal prediction (Section 2). We then present our novel methodology in Section 3 and evaluate it on two established and one novel image stream in Section 4, before we conclude this work in Section 5.

*Funding in the scope of the BMFTR project KI Akademie OWL under grant agreement No. 16IS24057A and the ERC Synergy Grant "Water-Futures" No. 951424 is gratefully acknowledged.

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

## 2 Related Work

### 2.1 Concept Drift and Drift Localization

In this work, we mainly consider concept drift in the context of stream learning and system monitoring; however, the definitions presented below also extend to the other drifting scenarios discussed in the introduction. In stream learning, we model an infinite sequence of independent observations $X_1, X_2, \ldots$, each drawn from a potentially different distribution $X_i \sim \mathcal{D}_i$ [1]. Drift occurs when $\mathcal{D}_i \neq \mathcal{D}_j$ for some $i, j$. A fully probabilistic framework was proposed in [9], augmenting each sample $X_i$ with an observation timestamp $T_i$. In this formulation, concept drift is equivalent to statistical dependence between data $X$ and time $T$.

Based on this idea, drift localization – the task of finding all samples affected by the drift – can be formalized as the local and global temporal distribution differing [4], i.e., $L = \{x : \mathbb{P}_{T|X=x} \neq \mathbb{P}_T\}$. It was shown in [4, Thm. 1 and 2] that for finite time points, e.g., "before drift" and "after drift", this set exists as the unique solution with certain properties and can be approximated using estimators. Since most algorithms detect drift between time windows [2, 1], this justifies many approaches and reduces drift localization to a probabilistic binary classification problem and a statistical test to assess the severity of the mismatch between $\hat{\mathbb{P}}_{T|X=x}$ and $\hat{\mathbb{P}}_T$ relating to the data-point-wise $H_0$-hypothesis "$x$ is not drifting" [4].

### 2.2 Localization Methods and Their Shortcomings

There are a few approaches for drift localization. Nearly all of them are based on comparing the time distribution of local groups of points with the global reference, thus aligning with the considerations of [4]. Those groups are most commonly formed unsupervised, i.e., without taking the time point into account. Classical *kdq*-trees [5] and quad-trees [8] recursively partition the data space along coordinate axes; other approaches use $k$-means variants [7]. Besides those partition-based, there exist $k$-neighborhood-based approaches like LDD-DIS [6]. The model-based approach, introduced in [4], differs from those in so far as they use the time information to choose a better-suited grouping, i.e., by training a decision tree predicting $T$ based on $X$, yet, due to overfitting, data points used to construct the grouping cannot be used for analysis. Furthermore, they suggest a purely heuristic approach based on random forests. While those approaches differ in what statistic and normalization technique they use, all explore the idea of local and global temporal differences.

This induces a triad-off problem: using no (*kdq*-tree, LDD-DIS, etc.) or only a few temporal information, the obtained grouping is sub-optimal, leading to inaccurate estimates $\hat{\mathbb{P}}_{T|x\in G}$; employing much to improve grouping, we are left with little data for the testing, resulting in low per-group test power – as the tests are performed separately on each group, even moderate test-set sizes and group numbers lead to comparably small per-group sample sizes. Both effects lead to an overall low testing power. We thus aim for a technique that allows for a global variance analysis.

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

### 2.3 Conformal Prediction

When training a classifier, the target is usually to minimize the overall expected error. However, such a scheme does not give any guarantees about the correctness of a single prediction. While probabilistic classification improves on this situation in theory by providing class probabilities and thus a measure of uncertainty, models usually over- or underfit, leading to sub-optimal assessments.

Conformal prediction constitutes an alternative scheme returning a set of potential classes $F(x) \subset \mathcal{C}$. This allows one to ensure that the correct class is in the set with arbitrary high certainty, i.e., $\mathbb{P}[Y \in F(X)] \geq \alpha$ for a conformal model $F$. Commonly, one minimizes the expected number of predicted classes.

The most common way to create a conformal model is to wrap a class-wise scoring function, for instance, a probabilistic classifier, into a calibrated model. This calibrated model must be trained on data that has not been used before, yet, since it only processes simple one-dimensional scores, the necessary calibration set can be comparably small. The usage of conformal $p$-values allows choosing $\alpha$ after training calibration, i.e., construct $p_y(x)$ so that $\mathbb{P}[Y \notin \{y : p_y(X) > \alpha\}] > 1 - \alpha$ holds for all $\alpha$.

## 3 Conformal prediction for drift localization

Current state-of-the-art drift localization techniques rely on local statistical testing and either employ unsupervised grouping methods – which are not performant when considering high-dimensional data –, or face a trade-off problem between optimizing grouping quality and test stability. In this work, we propose to leverage conformal prediction to obtain a global testing scheme. In the following, we first describe the general idea and then the algorithmic details.

As discussed before, [4] showed that a point $x$ is non-drifting if and only if the conditional entropy $H(T \mid X = x)$ is maximal, assuming a finite time domain and uniform temporal distribution, i.e., if we are maximally uncertain about the observation time point. Specifically, in [4], this is used as a test statistic and then normalized using a permutation scheme.

To approach this using conformal predictions, recall that if $c \notin F(x)$ then we can be very certain that the correct class is not $c$. Thus, following the ideas of [4], we can reject the hypothesis that $x$ is non-drifting if we can exclude one time point with certainty, i.e., $F(x) \neq \mathcal{T}$. By expressing $F_\alpha$ using conformal $p$-values, i.e., $F_\alpha(x) = \{y \in \mathcal{C} : p_y(x) \geq \alpha\}$, we can reject $H_0$ if the minimal conformal score is smaller $\alpha$, i.e., we obtain the $p$-value $p_{\text{drifting}}(X) = \min_y p_y(X)$. Using a probabilistic classifier as a class-wise scoring function, this relates to the probability of a large deviation of the class-probability from the global probability despite the sample being non-drifting, which is again in line with the considerations of [4].

Using conformal prediction has various advantages compared to the local testing scheme. While conformal prediction still requires a calibration set, since this is only needed to calibrate a one-dimensional signal, it can be chosen much smaller while still offering good performance, allowing us to use more samples for model training. We display this effect in Fig. 1. Here, a main difference, also

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

**Algorithm 1** Conformal Prediction for Drift Localization

**Input:** $(x_i, y_i)_{i=1}^n$, $y_i \in \mathcal{C}$; $n_{\text{boot}} \in \mathbb{N}$
**Output:** $(p_i)_{i=1}^n$
$P_i \leftarrow [\,]$ for all $i$
**for** $t = 1, \ldots, n_{\text{boot}}$ **do**
  $(I_{\text{in}}, I_{\text{oob}}) \leftarrow \textsc{SampleBootstrap}(\{1, \ldots, n\})$
  $\theta \leftarrow \textsc{TrainModel}(X_{I_{\text{in}}}, y_{I_{\text{in}}})$
  **for** $i \in I_{\text{in}}$ **do**
    $$P_i \leftarrow P_i + \min_{c \in \mathcal{C}} \left[ \frac{1 + \sum_{k \in I_{\text{oob}} \,:\, y_k = c} \mathbf{1}[f(c \mid x_k, \theta) \leq f(c \mid x_i, \theta)]}{1 + |\{k \in I_{\text{oob}} \,:\, y_k = c\}|} \right]$$
  **end for**
**end for**
$p_i \leftarrow \text{median}(P_i)$ for all $i$
**Return** $(p_i)_{i=1}^n$

from an efficiency perspective, is that for conformal prediction, we can evaluate whether a sample is drifting on the set used to train the model, not on the calibration set.

Another advantage is that we are no longer limited to specific models. Commonly used local tests require the models to induce some kind of grouping. This is not the case for conformal prediction, thus making it compatible with any scoring function. This not only makes the method more compatible with the usage of supervised trained models but also allows a much larger pool of potential models to choose from.

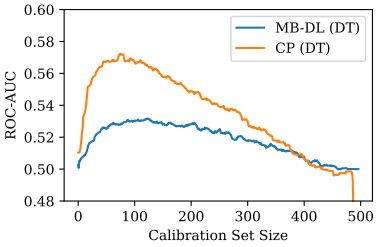


Fig. 1: Effect of grouping-test/calibration-set sizes. (Decision tree, Fish-head dataset, 500 samples, 98 drifting)

The main hurdle for translating our considerations into an algorithm is the need for a calibration set to perform conformal predictions. We propose using bootstrapping as the out-of-bag samples constitute a natural calibration set of decent size, even when oversampling. The model is thus trained on the in-bag samples and then calibrated using the out-of-bag samples. Using the calibrated models, we assign $p$-values to the in-bag samples using the minimal conformal $p$-value. To combine the resulting $p$-values across several bootstraps, we suggest using a median, which is equivalent to an ensemble of tests: we reject $H_0$ at level $\alpha$ if the majority of bootstraps lead to a rejection. The overall scheme is presented in Algorithm 1.

## 4 Experiments

We are evaluating the proposed conformal-prediction-based approach on streams of images. More precisely, we rely on the Fashion-MNIST [10] dataset and the No ImageNet Class Objects (NINCO) [11]. We follow the experimental setup by [4], randomly selecting one class for non-drifting, drifting before, and drifting after.

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

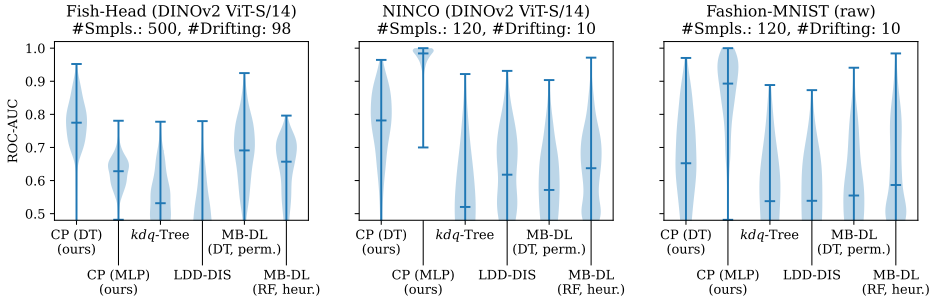


Fig. 2: Experimental results. ROC-AUC (500 runs) for various drift localizes using windows of 250/60 samples with 98/10 drifting samples (in total).

While we use the Fashion-MNIST without further preprocessing, for NINCO, we use deep embeddings (DINOv2 ViT-S/14 [12]).[1]

*Fish-Head Dataset* While this is a simple way to obtain drifting image data streams, we propose to also consider a more subtle drifting scenario where we do not model drift by switching up classes, but induce a more nuanced and realistic type of drift. For our novel benchmark *Fish-Head data stream*, we rely on the ImageNet [13] subset "ImageNette" consisting of the classes "tench," "English springer," "cassette player," "chain saw," "church," "French horn," "garbage truck," "gas pump," "golf ball," and "parachute." We manually split the class "tench" into the two sub-classes where the fish head is turned towards the left or right side, respectively; these are the drifting samples, all other samples are non-drifting. For our analysis, we again use a DINOv2 ViT-S/14 [12] model to embed the images. A simple analysis shows that non-drifting, drifting before, and drifting after are lineally separable in the embedding.

*Evaluation of Method* To evaluate our method, we follow the same procedure as [2][2]. For the simpler NINCO and Fashion-MNIST datasets, we use $2 \times 60$ samples with 10 drifting, for the Fish-Head dataset, we used $2 \times 250$ samples with 98 drifting; here, the increase was necessary as otherwise no method yielded results above random chance. Besides the novel conformal-prediction-based approach (CP) using decision trees (DT) and MLPs as models, we evaluate the model-based approach (MB-DL; [4]) using permutations and decision trees on bootstraps (DT, perm.) similar to Algorithm 1 and heuristic approach based on random forests (RF, heur.), LDD-DIS [6], and *kdq*-trees [5]. Following [2], we use the ROC-AUC as a score. We repeat each experiment 500 times.

The results are shown in Fig. 2. One can clearly observe that for both NINCO and Fashion-MNIST, the proposed conformal-prediction-based localization outperforms the related work, in particular, when using MLPs. Considering the new

[1]DINOv2 (released April 2023) was trained on large-scale image data, whereas NINCO (released June 2023) is a curated out-of-distribution benchmark. Given NINCO's later release and evaluation-focused design, its inclusion in DINOv2's training data is unlikely, though this cannot be confirmed definitively.

[2]See `https://github.com/FabianHinder/Advanced-Drift-Localization` for the code

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

Fish-Head benchmark, we obtain that our method, realized with decision trees, outperforms the other methods. In this setting, the model-based approaches perform at a similar level to our MLP-based version. Overall, the new data benchmark seems to be a more challenging task, making it suitable for further evaluation.

## 5 Conclusion and Future Work

In this work, we proposed a novel strategy for drift localization, replacing the state-of-the-art local testing in data segments with a conformal-prediction-based global strategy allowing for a larger model pool while providing formal guarantees. We experimentally showed the advantage of the proposed methodology on established image data streams, showing that those can essentially be solved by the presented method. Besides, we proposed a novel image stream benchmark containing more subtle drift. As our experiments show, this benchmark is far more challenging, requiring a significantly larger number of samples to ensure results better than random chance. Further investigation and development that is better suited for small-sample size setups, and in particular, work in the case of only few drifting samples, is subject to future work. Furthermore, investigating the relevance of the used deep embedding seems to be a relevant consideration.

## References

[1] J. Lu, A. Liu, F. Dong, F. Gu, J. Gama, and G. Zhang. Learning under concept drift: A review. *IEEE transactions on knowledge and data engineering*, 2018.

[2] F. Hinder, V. Vaquet, and B. Hammer. One or two things we know about concept drift—a survey on monitoring in evolving environments. part b: locating and explaining concept drift. *Frontiers in Artificial Intelligence*, 2024.

[3] F. Hinder, V. Vaquet, J. Brinkrolf, and B. Hammer. Model-based explanations of concept drift. *Neurocomputing*, 2023.

[4] F. Hinder, V. Vaquet, J. Brinkrolf, A. Artelt, and B. Hammer. Localization of concept drift: Identifying the drifting datapoints. In *IJCNN*, 2022.

[5] T. Dasu, S. Krishnan, S. Venkatasubramanian, and K. Yi. An information-theoretic approach to detecting changes in multi-dimensional data streams. 2006.

[6] A. Liu, Y. Song, G. Zhang, and J. Lu. Regional concept drift detection and density synchronized drift adaptation. In *IJCAI*, 2017.

[7] A. Liu, J. Lu, and G. Zhang. Concept drift detection via equal intensity k-means space partitioning. *IEEE transactions on cybernetics*, 2020.

[8] B. A. Ramos, C. L. Castro, T. A. Coelho, and P. P. Angelov. Unsupervised drift detection using quadtree spatial mapping. 2024.

[9] F. Hinder, A. Artelt, and B. Hammer. Towards non-parametric drift detection via dynamic adapting window independence drift detection (dawidd). In *ICML*, 2020.

[10] H. Xiao, K. Rasul, and R. Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

[11] J. Bitterwolf, M. Müller, and M. Hein. In or out? fixing imagenet out-of-distribution detection evaluation. In *ICML*, 2023.

[12] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[13] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.