Uncertainty Estimation for Change Point Detection models

Maria Kovaleva, Alexander Kukharskii, Alexander Stepikin

Supervisor: Evgenia Romanenkova

March, 23
Skoltech

Outline

- Change Points in sequential data
- Natural CPD criteria
- CPD as seq2seq binary classification
- Uncertainty Estimation for deep models
- Evaluation pipeline: datasets & metrics
- Ensembles of deep CP detectors
- CUSUM statistic for CP scores aggregation
- Rejection-based CPD
- Conclusions

Change Points in sequential data

Change Point (CP) is a moment of abrupt regime switch in a data stream.

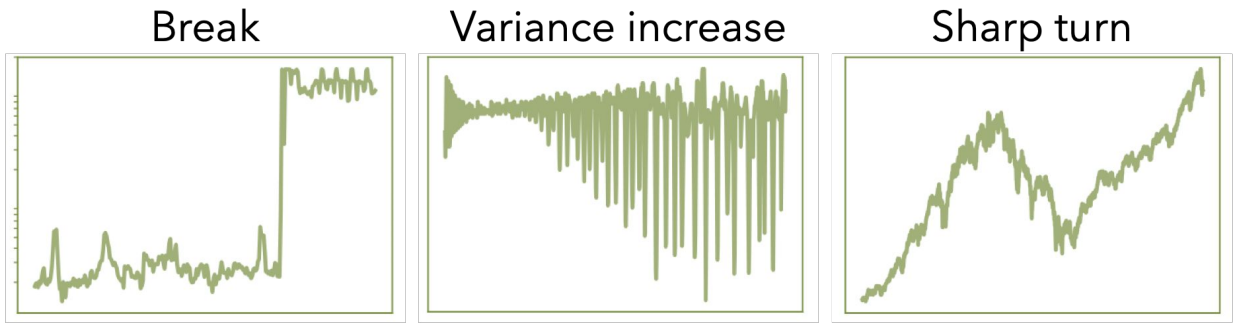

Fig. 1: Examples of simple changes in 1D signals.

Real-world examples:

- Video surveillance
- Health monitoring (fit trackers, ECG, EEG)
- Sensors' readings (oil & gas drilling)
- Financial data analysis
- Control of high-loaded systems (servers, network traffic)
- Text & audio segmentation

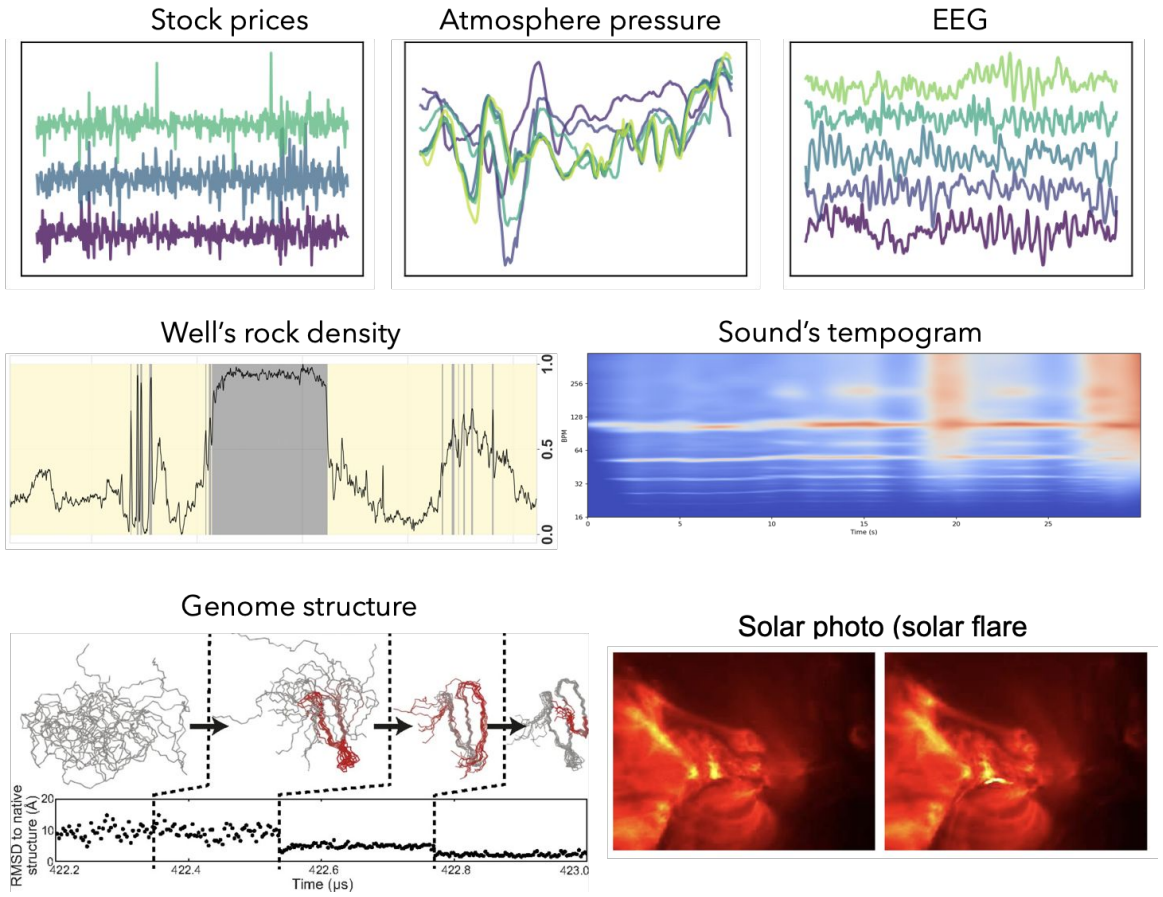

Fig. 2: Examples of real-world CPD problems.

Skoltech

Natural CPD criteria

CP detector is a model that warns about data distribution shifts.

The goal is to minimize Detection Delay and number of False Alarms.

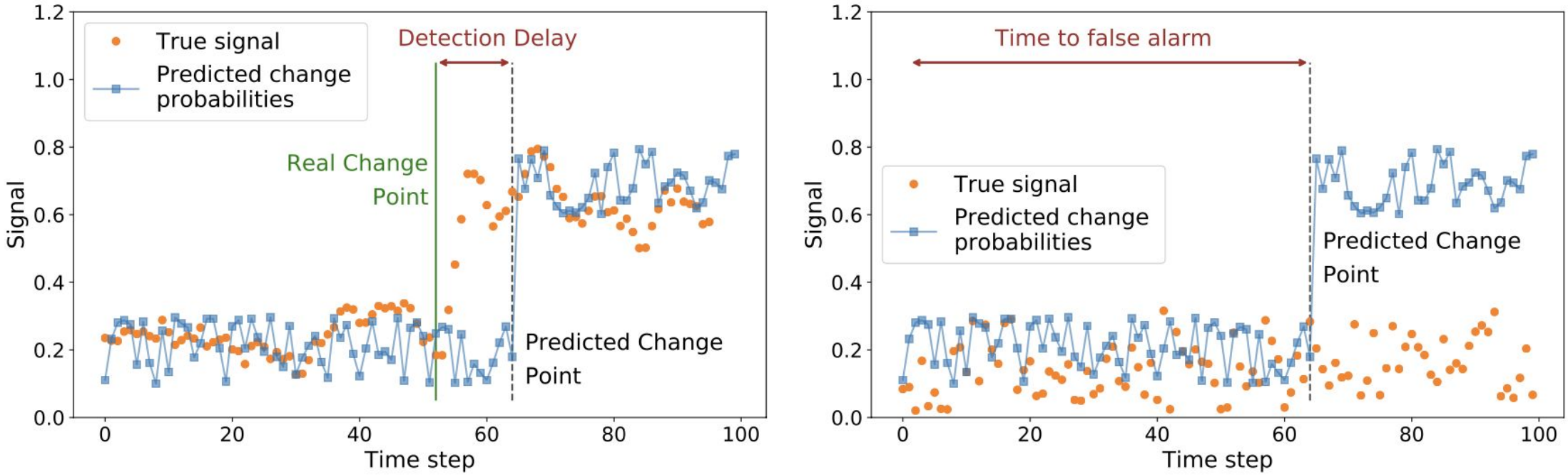

Fig. 3: Natural CPD criteria illustration: low Detection Delay (left) and long Time to a False Alarm (right).

Skoltech

CPD as seq2seq binary classification

Baseline [1]:

- An RNN that predicts change probabilities for each time step: $p_t^i = f_{\mathbf{w}}(X_i^{1:t})$
- CPD is interpreted as a seq2seq binary classification task. The model is trained with binary cross-entropy loss:

$$BCE(l, p) = -\frac{1}{T}\sum_{t=1}^{T}\left(p_t \log l_t + (1 - p_t)\log\left(1 - l_t\right)\right)$$

where l are the true labels, p – predicted CP probabilities.

- The model warns about a CP when the probability exceeds a pre-selected threshold: $p_t \geq s$

Our goal:

- Obtain "confidence" in the model's predictions and benefit from it.

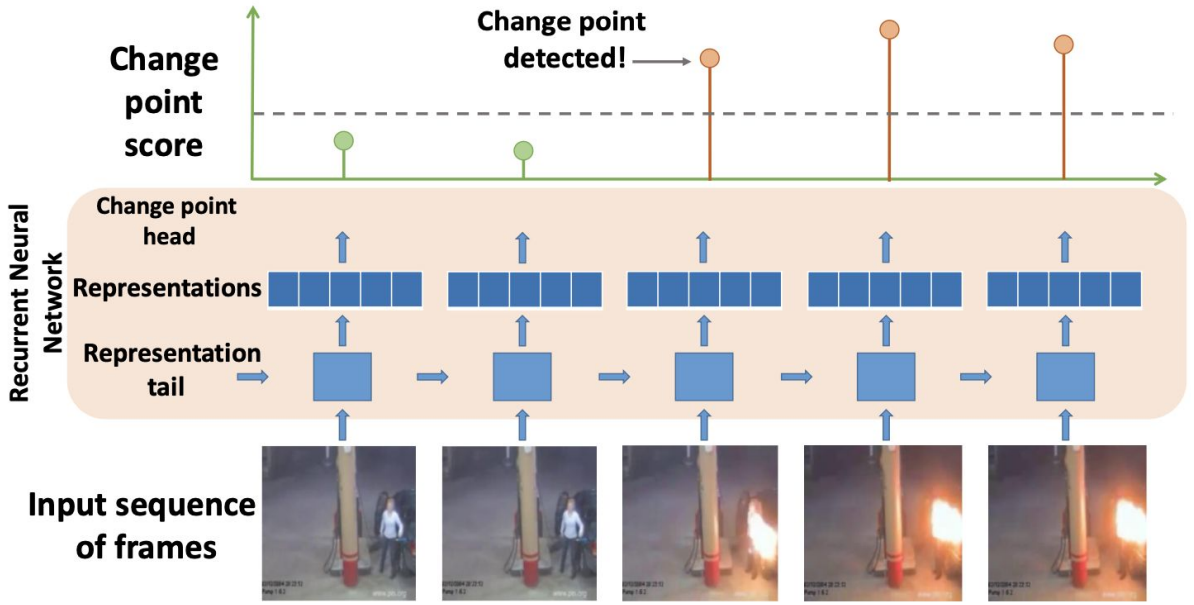

Fig. 4: Scheme of the baseline model. For each time step, the model predicts the change probability.

[1] *Romanenkova E. et al.* InDiD: Instant Disorder Detection via a Principled Neural Network // Proceedings of the 30th ACM International Conference on Multimedia 2022. – 2022. – pp. 3152-3161.

Skoltech

Uncertainty estimation

Ensemble approach: we consider a set of somewhat different models that give different predictions.
Estimate of the uncertainty is a some metric of the difference in predictions of models, for example std.

Ensemble building options:

- Initialization with different weights
- Same models trained on different data subsets
- A set of different models (for example, with different architectures)

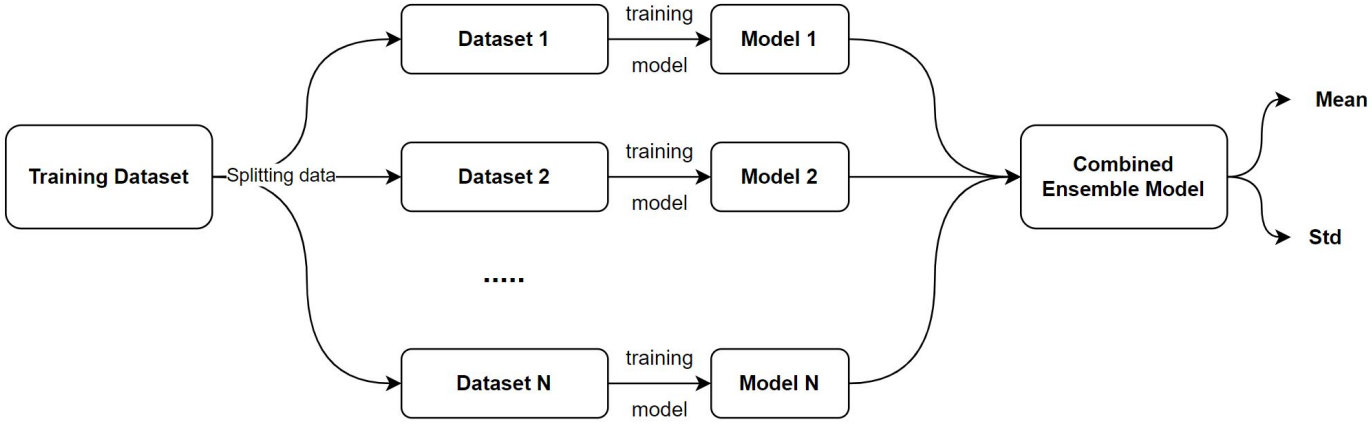

Fig. 5: Scheme of an ensemble.

Skoltech

Evaluation: datasets

We consider:

- Toy example: synthetic sequences of 1D Gaussians with a random change in mean at a random time
- Human Activity Recognition (HAR): sensors' readings of the devices worn by a person
- Sequences of MNIST images with smooth transitions from one digit to another
- Video surveillance dataset: Explosions

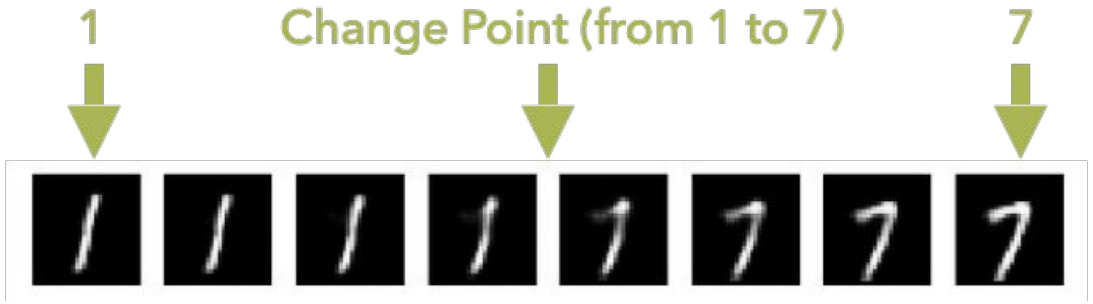

Fig. 6: A sequence of MNIST images with a smooth transition from "1" to "7".

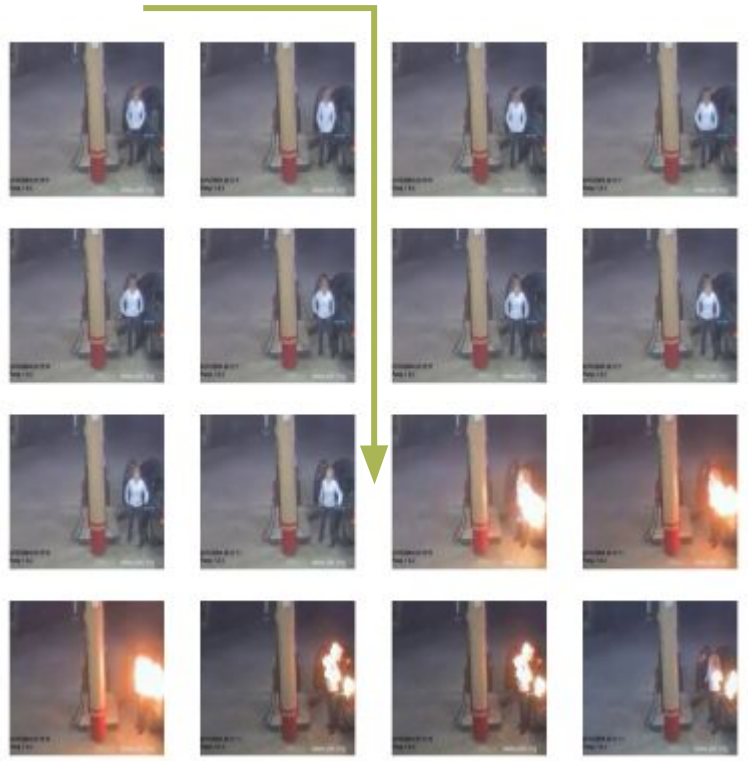

Fig. 7: A video clip with an explosion as a Change Point.

Skoltech

Evaluation: CPD metrics

1. Standard classification quality metrics:
   - Number of TP, TN, FP and FN predictions
   - Precision, Recall, F1-score
2. Specific CPD metrics:
   - Mean Detection Delay
   - Mean Time to a False Alarm
   - Area under the Detection Curve (AUDC)
   - Covering [2]:

$$\text{Covering}(G, G') = \frac{1}{T} \sum_{A \in G} |A| \cdot \max_{A' \in G'} \frac{|A \cap A'|}{|A \cup A'|},$$

where G and G′ — partitions of the sequence made by predicted and true CPs respectively.

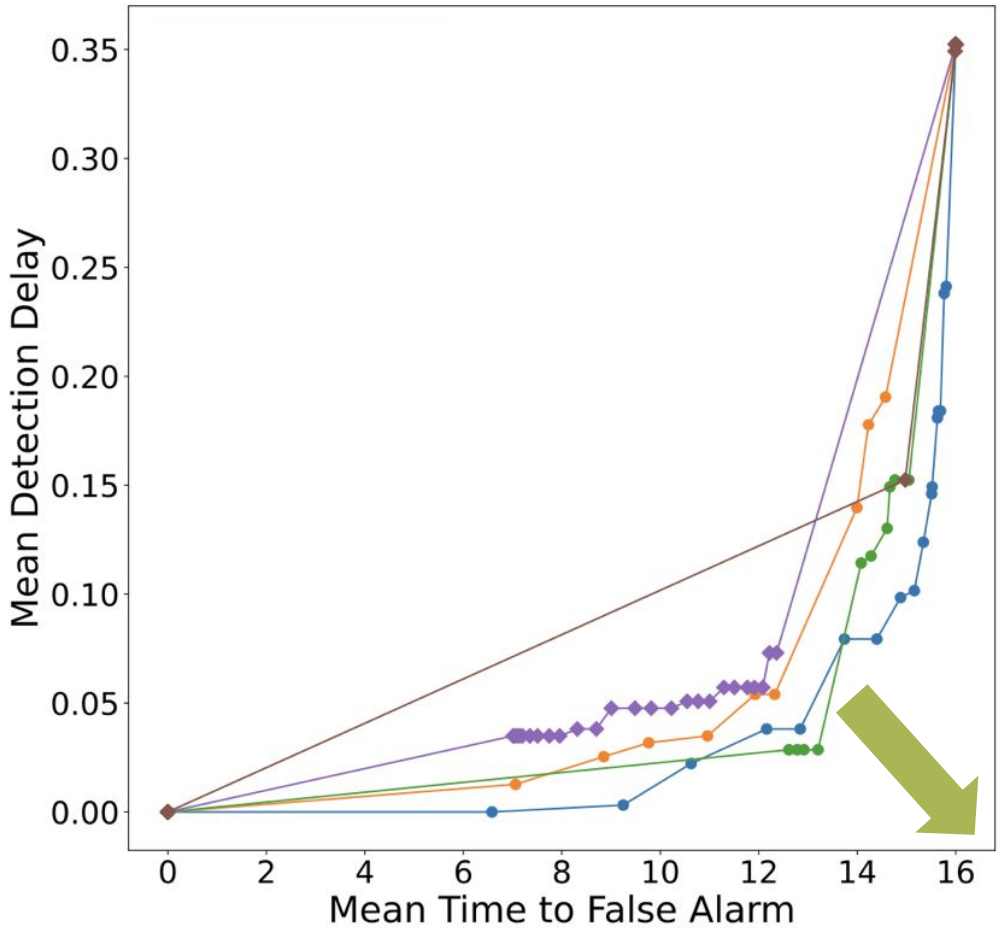

Fig. 8: Examples of the Detection Curves.

[2] Charles Truong, Laurent Oudre, and Nicolas Vayatis. 2020. Selective review of offline change point detection methods. *Signal Processing* 167 (2020), 107299.

Ensembles of deep CP detectors: mean vs quantiles

Model	AUDC	Time to FA	DD	Max F1	Cover
Human Activity Recognition					
Single	43.22	11.00	0.20	0.9886	0.9851
Average	44.95	11.07	0.10	0.9927	0.9938
Quantile (best)	44.25	11.10	0.17	0.9948	0.9898
Sequences of MNIST images					
Single	237.94	44.94	3.37	0.9862	0.9120
Average	175.32	46.87	2.20	0.9893	0.9386
Quantile (best)	164.40	47.21	1.67	0.9965	0.9510
Explosion					
Single	0.82	14.74	0.13	0.3094	0.9728
Average	0.50	15.78	0.23	0.5217	0.9876
Quantile (best)	0.39	15.78	0.23	0.5217	0.9876

Table 1: Result metrics for ensemble of deep CP detectors for 3 different datasets.

CUSUM: aggregating CP scores

- CUSUM [3] involves the calculation of a cumulative sum
  - $S_0 = 0$
  - $S_{n+1} = max(0, S_n + x_{n+1})$
  - stop when S exceeds a certain threshold
- we use $x_n = (\text{mean}_{n+1} - \text{mean}_n)/\text{std}_{n+1}$
- Motivation: if the value of differences in mean grows and std decreases then we are sure that this is change point

- Results:
  - higher cusum threshold leads to less FP predictions but more FN, there is a tradeoff between them
  - in general, the results are worse than in other methods

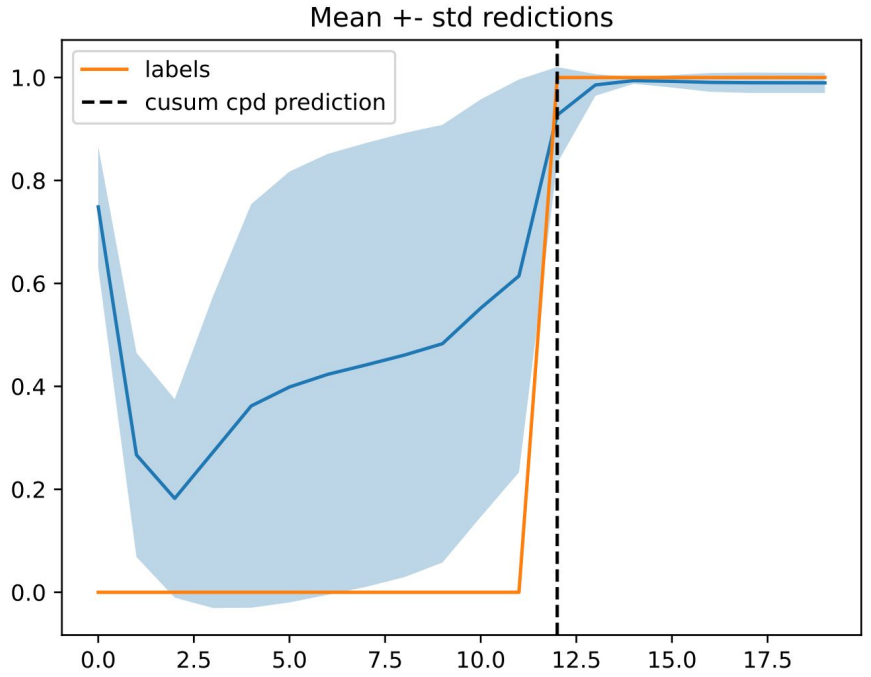

Fig. 9: Example of CUSUM-based prediction for the HAR dataset

[3] AN Shiryaev. 2017. Stochastic Change-Point Detection Problems. Moscow: MCCME (in Russian) (2017).

CPD with rejection: motivation

- Similar to “classification with rejection” [4]: reject CP alarm if uncertainty measure is high.
- This should reduce the number of False Alarms and make model more invariant to the alarm threshold.

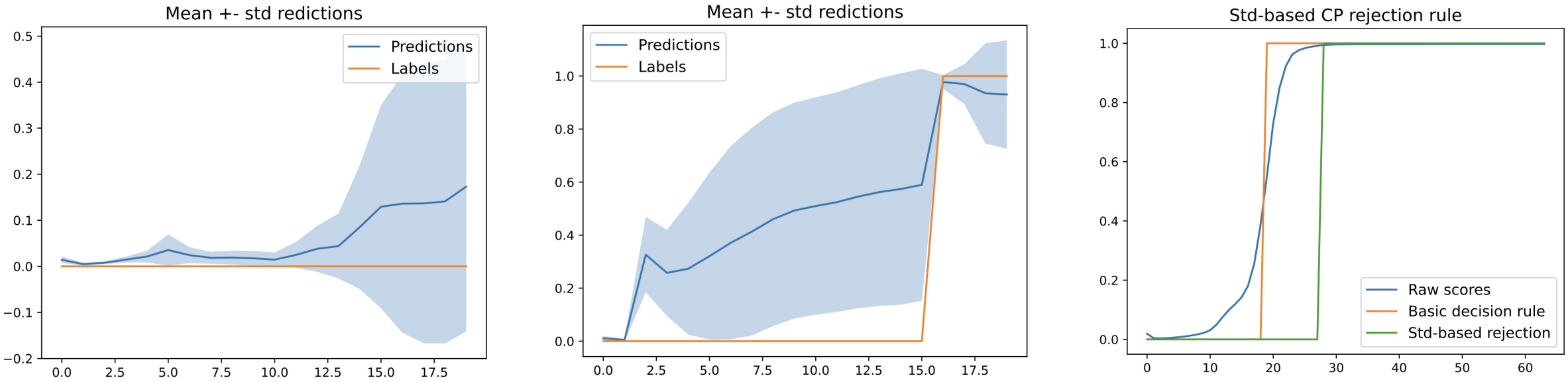

Fig. 10: Motivation of std-based CPD with rejection technique. From left to right: examples of typical ensemble predictions for the HAR dataset and the modification of the model’s score processing.

[4] Eyke Hullermeier, Titouan Lorieul and Willem Waegeman. Uncertainty-aware Classification: from Rejection to Self-valued prediction. Virtual tutorial at ECML/PKDD 2021.

CPD with rejection: results

Model	AUDC	Time to FA	Detection Delay	Max F1	Cover
Sequences of MNIST images					
Standard Ensemble	175.32	46.87	2.20	0.9893	0.9386
CPD with rejection	175.32	46.87	2.20	0.9893	0.9386
Human Activity Recognition					
Standard Ensemble	44.95	11.07	0.10	0.9927	0.9938
CPD with rejection	38.76	10.58	0.10	0.9927	0.9930
Explosion					
Standard Ensemble	0.50	15.78	0.23	0.5217	0.9876
CPD with rejection	0.88	15.25	0.20	0.3889	0.9751

Table 2: Result metrics for std-threshold CPD with rejection for 3 different datasets.

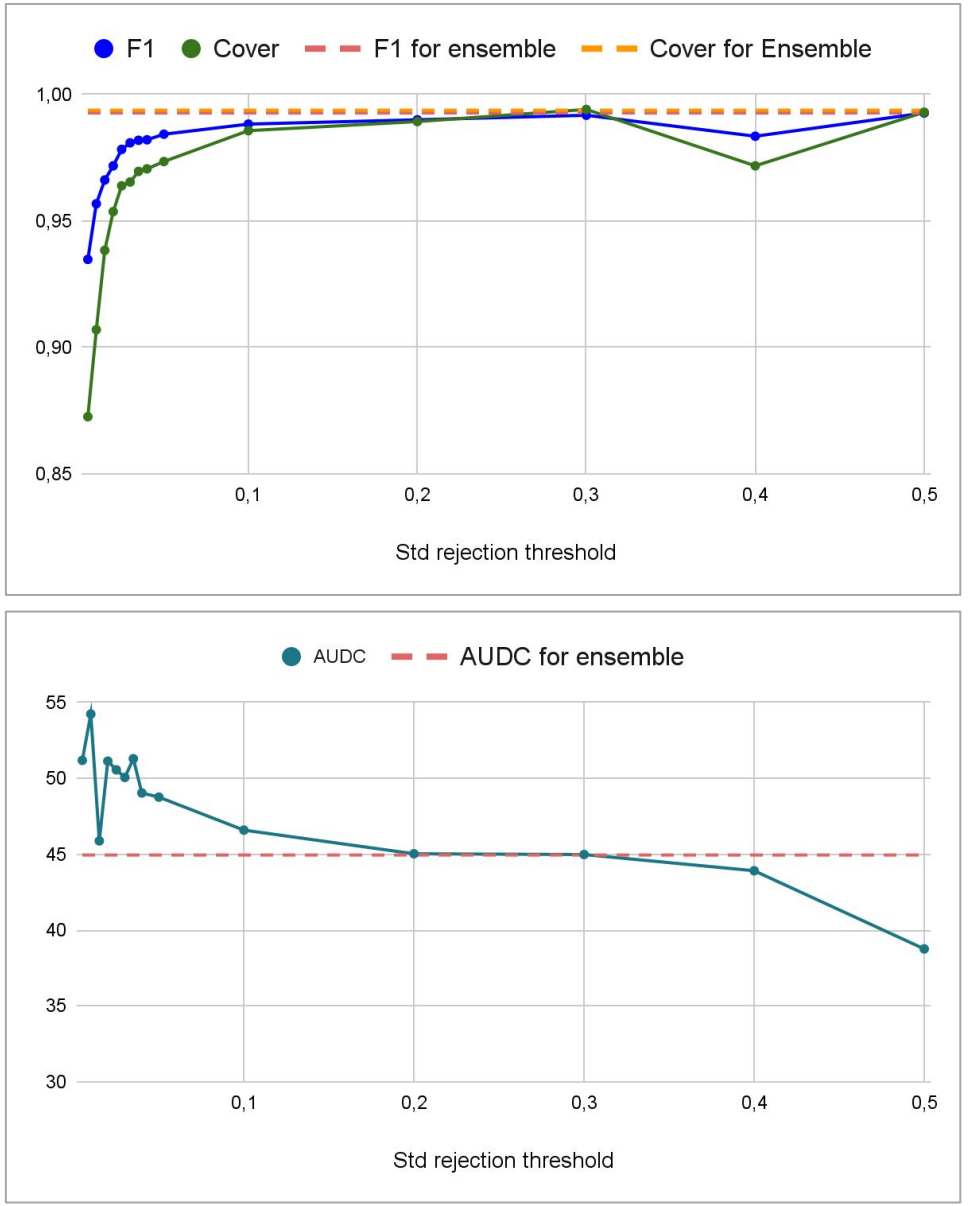

Fig. 11: Metrics dynamics for the rejection-based approach.

Skoltech

Conclusions

In this project, we:

- Explored ensembling methods for CPD task.
- Developed CP scores aggregation technique based on CUSUM-statistic.
- Implemented a basic approach for CPD with rejection.

Results:

- Ensembles of deep CP detectors outperform single CPD baselines in terms of the main quality metrics. Choice of the quantile of the CP scores distribution significantly affects the results.
- CUSUM-based aggregation of CP scores did not improve the metrics but more experiments are to be done.
- CPD with rejection can be beneficial in some cases. However, on average, it does not improve the metrics.