# On metafeatures' ability of implicit concept identification

MOA Synthetic data streams

Supplementary material

## Data stream description

The research was initially performed for three types of data:

- Synthetic – generated using the generator from the stream-learn library [3]. The streams were described by 10 numerical informative (relevant in classification task) features.
- Semi-synthetic – generated using the procedure described in [2] by interpolating feature projections from real-world datasets to simulate concept drift.
- Real-world data streams.

The research was extended to synthetic streams generated using the MOA [1] framework. Twelve base data streams with a static concept were generated, then a sudden concept drift was simulated by concatenating fragments of streamreplications originating from the same generator. The base streams were generated using four generators from the MOA library:

- RBF – random radial basis function binary data stream, described by 7 numerical attributes.
- LED – problem of predicting the digit displayed on a 7-segment LED display. Consists of 24 binary attributes describing 10 classes.
- HYPERPLANE – binary problem of predicting class of a rotating hyperplane, described by 7 numerical attributes.
- SEA – SEA concepts functions, described in [4]. Stream consists of 3 numerical attributes, that vary from 0 to 10, where only 2 of them are relevant to the classification task.

The evaluated streams consisted of 100,000 objects in 500 chunks (of 200 objects each), and contained one sudden concept drift. The classification results obtained by the *Gaussian Naive Bayes* classifier trained only on the first chunk of data are presented in the Figure 1. There is a visible drop in classification accuracy resulting from concept drift around the 250th data chunk.

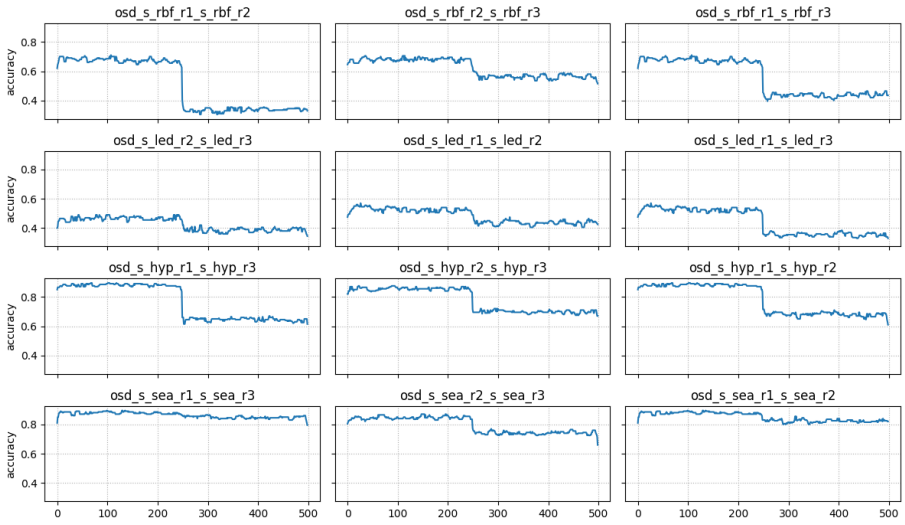


Figure 1: Classification accuracy in generated data streams

# 1 Experiment 1

In the first experiment, all tested metafeatures were calculated within stream chunks. The obtained values are presented in the Figures 2 - 10. Each figure presents the results for the 5 most informative (based on PCA) metafeatures from a given category and for the streams obtained from four generators (RBF, LED, Hyperplane, SEA). The color of the points identifies the concept from which the sample described by the metafeatures comes. Depending on the generator used, metafeature groups allow to notice more or less effective concept separation.

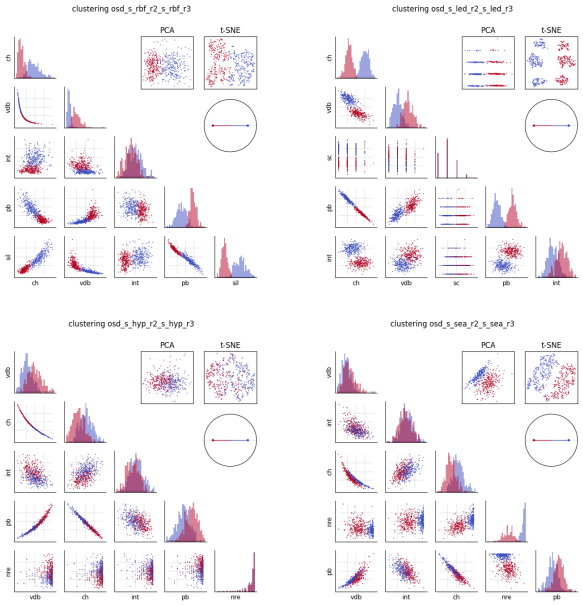


Figure 2: Clustering

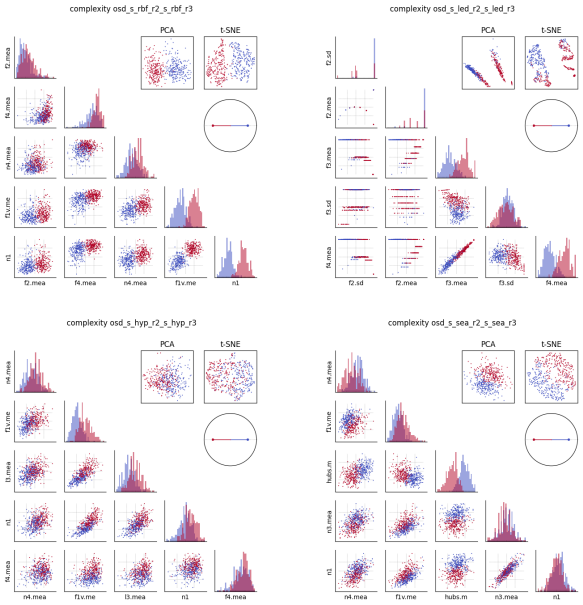


Figure 3: Complexity

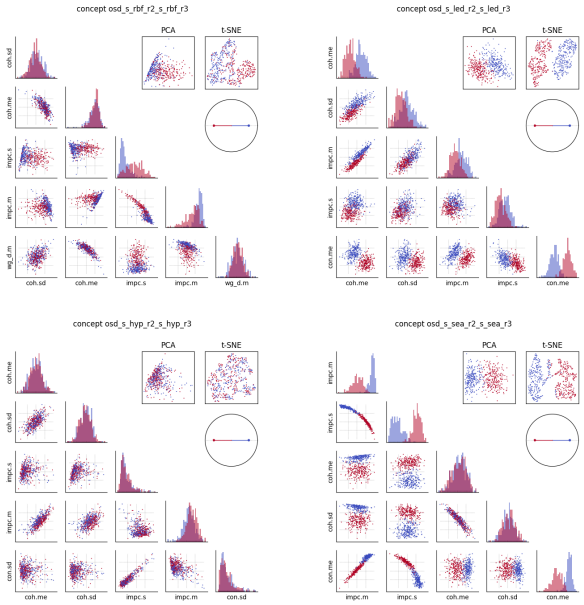


Figure 4: Concept

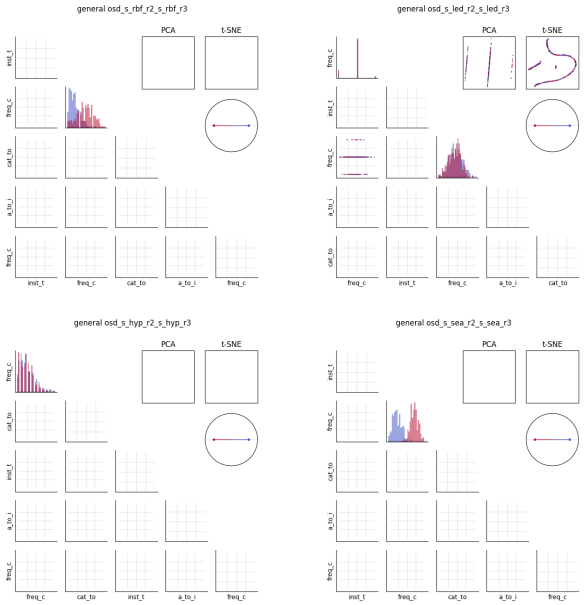


Figure 5: General

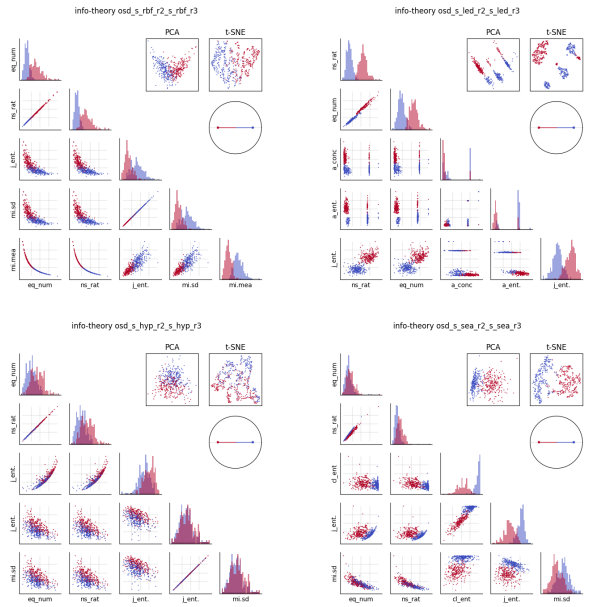


Figure 6: Information Theory

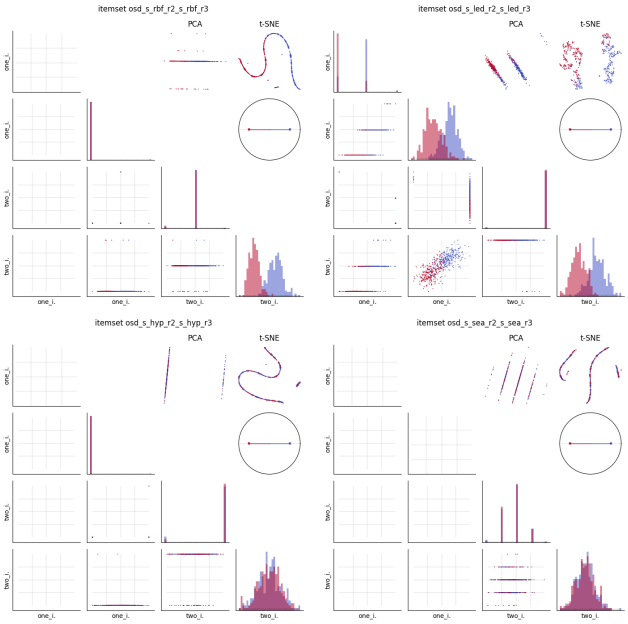


Figure 7: Itemset

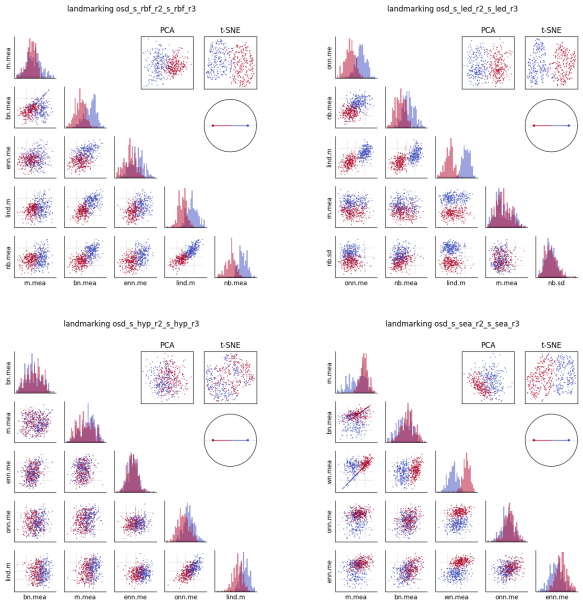


Figure 8: Landmarking

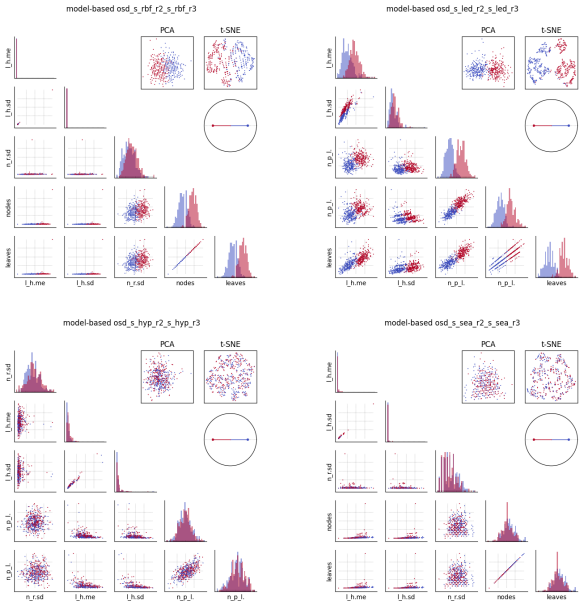


Figure 9: Model-based

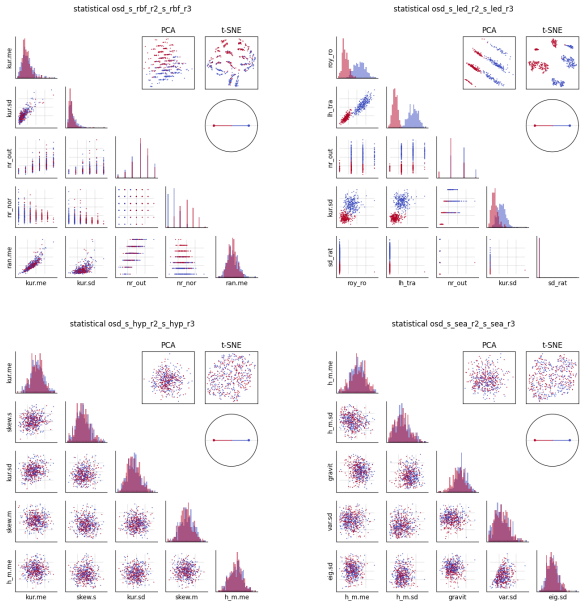


Figure 10: Statistical

## 2 Experiment 2

In the second experiment, metafeatures were used to identify concepts in a classification task. For synthetic streams generated with MOA, characterized by a single sudden concept drift halfway through the stream, this resulted in a balanced binary classification problem.

Figure 11 shows the averaged results for 4 generators in the form of heat maps, where higher color intensity is associated with higher classification quality. The streams generated using the LED generator stand out significantly in this experiment – all metafeature groups except the *general* category allow for effective concept identification. In the case of other streams in general *clustering*, *complexity*, *concept*, *information theory*, *landmarking* and *statistical* categories usually show satisfactory results. The observations are similar to the results for streams obtained using the generator from the stream-learn library, semi-synthetic and real streams described in the main part of the work. A higher classification accuracy resulting from the evaluation of the binary problem is noticeable.

In the case of the steam-learn generator, the *statistical* group shown to be the most important in the concept identification task. In the case of MOA streams, it still allows for effective classification, but its advantage over other groups is not that significant.

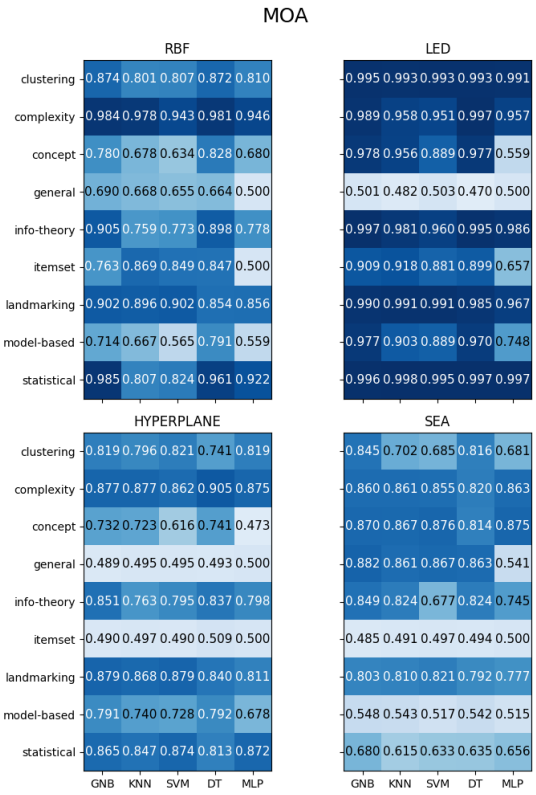


Figure 11: Concept classification in metafeature groups

# 3 Experiment 3

The third experiment concerned the feature selection from promising groups of metafeatures in the concept classification task. The results for all 12 generated streams are shown in Figure 12. Again, a strong dependence of the classification quality on the used generator was noticed – the streams generated with LED show almost 100% of accuracy.

The conclusions for MOA streams are similar to the results for other types of streams – selecting up to about 14 most significant metafeatures brings high classification quality for all tested classifiers. When the number of features increases, the classifiers lose the ability to effectively recognize the given concept. As in the case of the other streams, redundant features (not informative) do not significantly impact classification with the Decision Tree (DT) and Gaussian Naive Bayes (GNB).

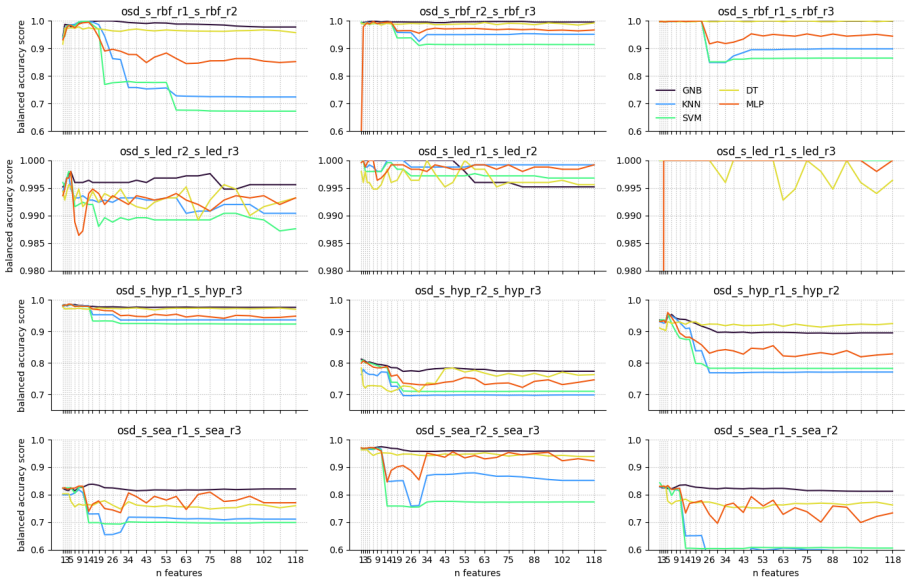


Figure 12: Metaattribute selection

# 4 Experiment 4

In Experiment 4, the variance between metafeature values and the concept identifier described by the F-statistic was analyzed. The accumulated results from the replications are presented in Figure 13.

In case of this experiment, there is a lot of variability depending on the data generator used and the replications themselves. In all four types of streams, metafeatures from the *complexity* group are highly important, which was not as noticeable in the case of the remaining streams (stream-learn generator, semi-synthetic and real streams), where the group of statistical metafeatures was more significant. The obtained results are consistent with the conclusions from Experiment 2, where the importance of metafeatures from many groups, including complexity, was noticed, while for the streams presented in the original work, already in Experiment 2, metafeatures from the statistical category had the highest significance.

Despite some discrepancy with the original results, almost all metafeatures initially identified as promising are visible among the 50 most informative metafeatures (presented in the figure). The exceptions are the *c1* measure from the *complexity* category, describing on the class imbalance - which remained stable in the case of MOA streams, and the *statistical* measure based on the mean feature values. This may be due to the characteristics of MOA generators, which generate data in a predefined, unchanging range, unlike the stream-learn generator and semi-synthetic generator.

It should be emphasized that, contrary to the results presented in the main part of the article, in the case of these generators it was not possible to clearly determine the metafeature ranking – different metafeature labels are presented on the X-axis of each sub-plot. It is also worth emphasizing the variability of the accumulated f-statistic itself. For streams obtained with the LED generator, the accumulated value is at a maximum of 60 thousand, while for the other generators from MOA it reaches a maximum at about 4 - 6 thousand.

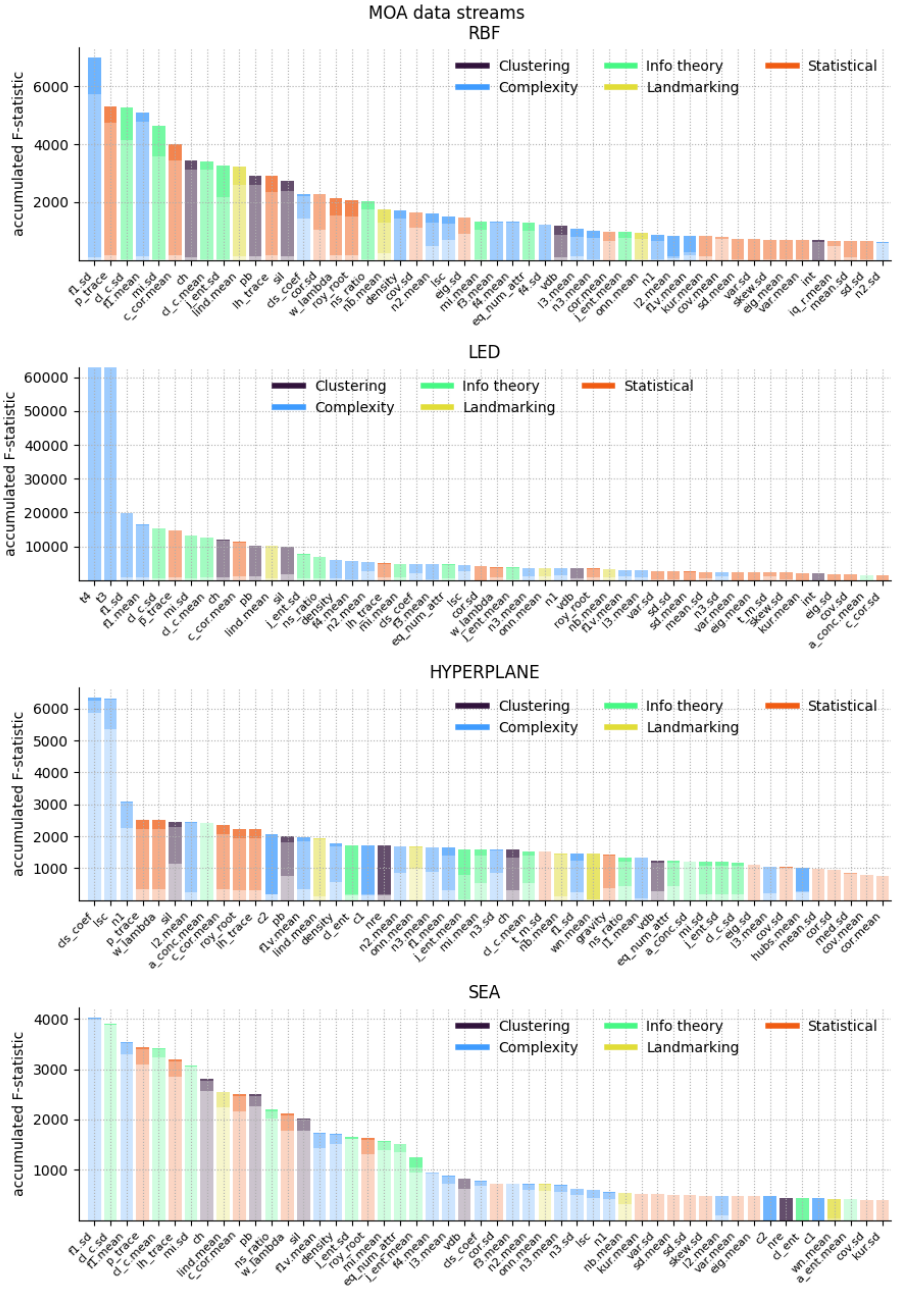


Figure 13: Variance analysis

# 5 Experiment 5

The results from previous experiment show that, depending on the generator and the replication, different factors can effectively identify a concept. Despite the lack of clear compatibility on the promising metafeatures from MOA and other generators, we decided to conduct a covariance analysis of the initially identified pool of 17 promising metafeatures. The results of the fifth experiment are shown in Figure 14.

The results are similar to those for the stream-learn generator and semi-synthetic streams. There is strong covariance in (1) mean-based statistical measures, (2) normalized relative entropy (*nre*), entropy of class proportions (*c1*) and target attribute Shannon's entropy (*cl_ent*) measures, and (3) maximum Fisher's discriminant ratio measures (*f1*), concentration coefficient (*clc*), joined entropy (*j_ent*), mutual information (*mi*).

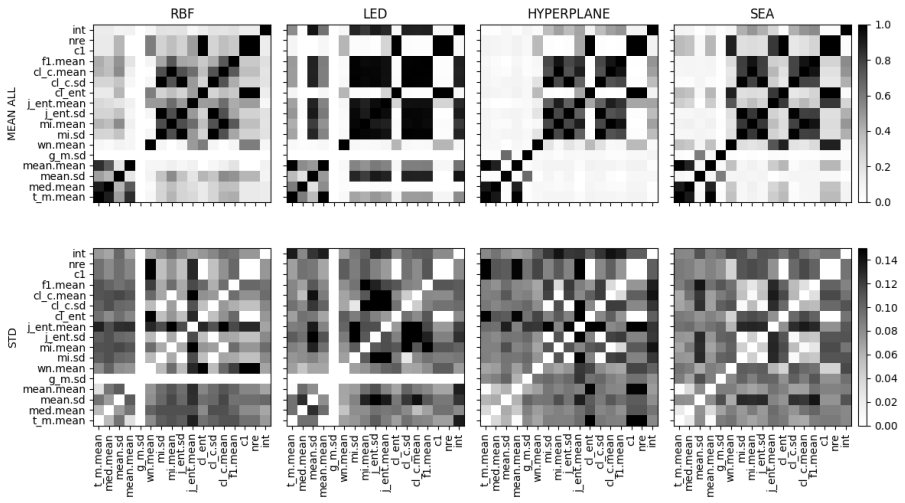


Figure 14: Covariance analysis

# References


[1] A. Bifet, G. Holmes, B. Pfahringer, P. Kranen, H. Kremer, T. Jansen, and T. Seidl. Moa: Massive online analysis, a framework for stream classification and clustering. In *Proceedings of the first workshop on applications of pattern analysis*, pages 44–50. PMLR, 2010.

[2] J. Komorniczak and P. Ksieniewicz. Data stream generation through real concept's interpolation. In *30th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, ESANN*, pages 5–7, 2022.

[3] P. Ksieniewicz and P. Zyblewski. Stream-learn—open-source python library for difficult data stream batch analysis. *Neurocomputing*, 478:11–21, 2022.

[4] W. N. Street and Y. 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*, pages 377–382, 2001.