Study and design of interpretable models for detecting anomalies in spectral data stream: Application to the contamination of vacuum process chambers for microelectronics.

N. Rojas Varela, J. Ah-Pine, E. Mephu Nguifo

INTRODUCTION

The continuous flux of data, known as **data streams**, has become a vital source of big data presented in various industries. Using **spectral data** in a data stream typically implies analyzing the spectrum of light emitted or absorbed by materials in real time.

Data stream anomaly detection, involves the process of identifying unusual patterns or events in the continuous flow of data.

Some **challenges** include:

- Full dataset not available in advance.
- High velocity of streams.
- Evolution of data characteristics over time.

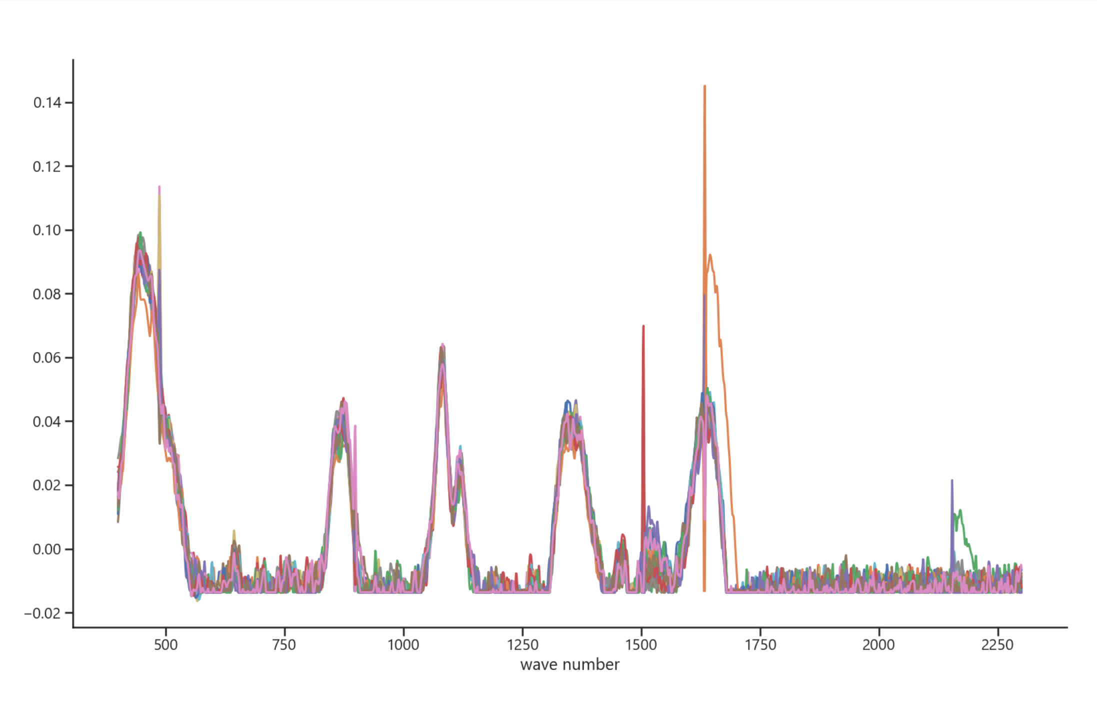


Fig 1: Example of Spectral data.

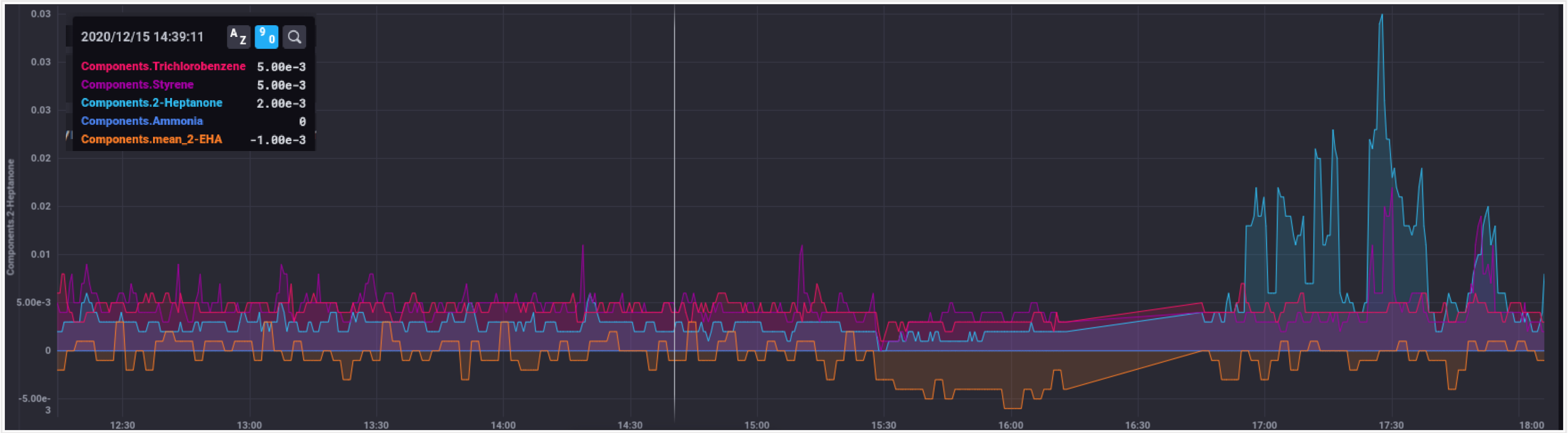


Fig 2: Example of Data Stream. [1]

TYPE OF METHODS

Categories for algorithms based on:

- **Offline learning:** use historical data to learn a model that indicates the anomaly level of the points.
- **Semi-online learning:** the algorithms perform offline learning on a part of the data to obtain a model, then applies real-time anomaly detection on the subsequent data streams.
- **Online learning:** uses incremental learning algorithms to continuously update the model to adapt to changes in the data stream.

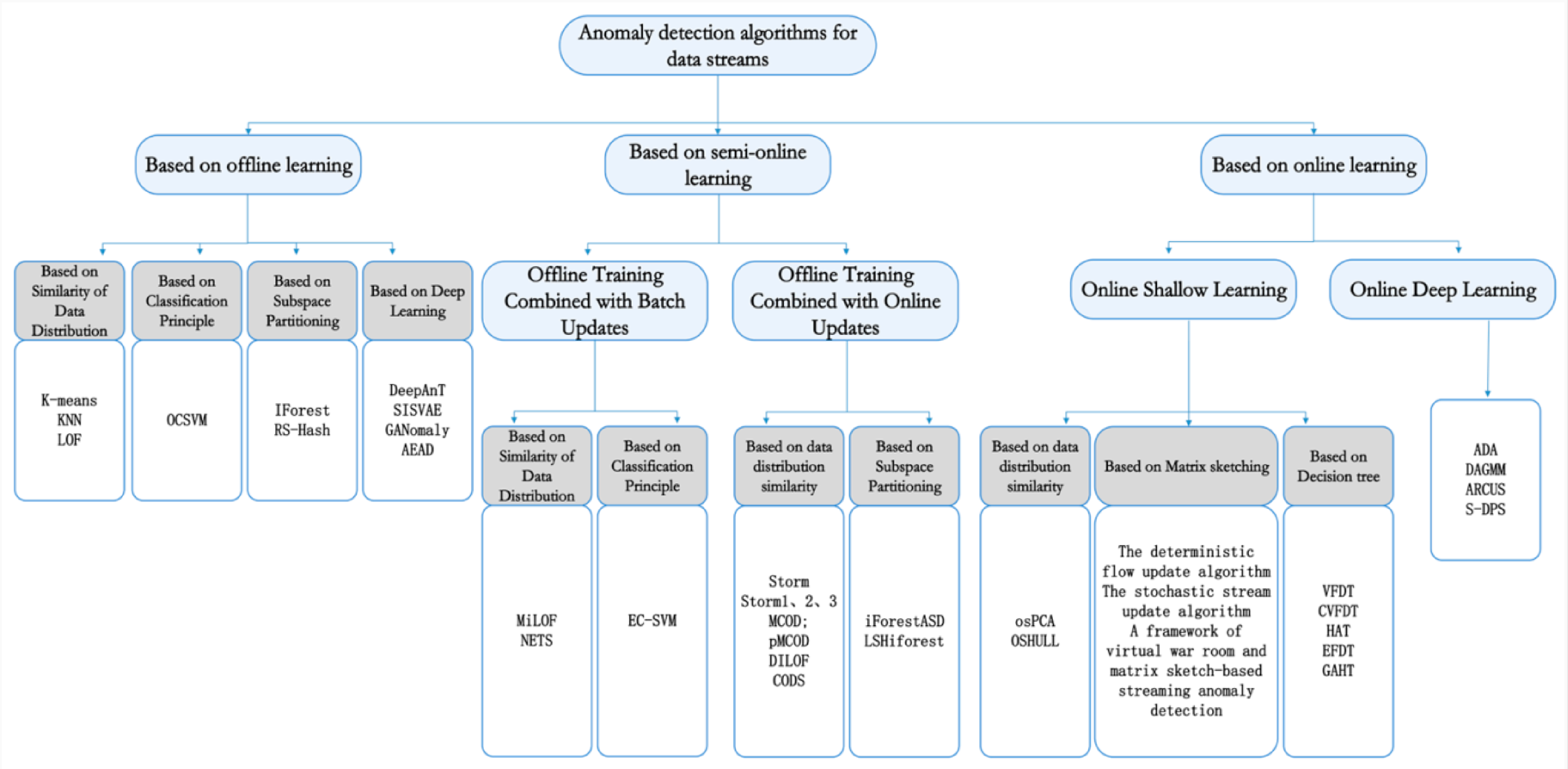
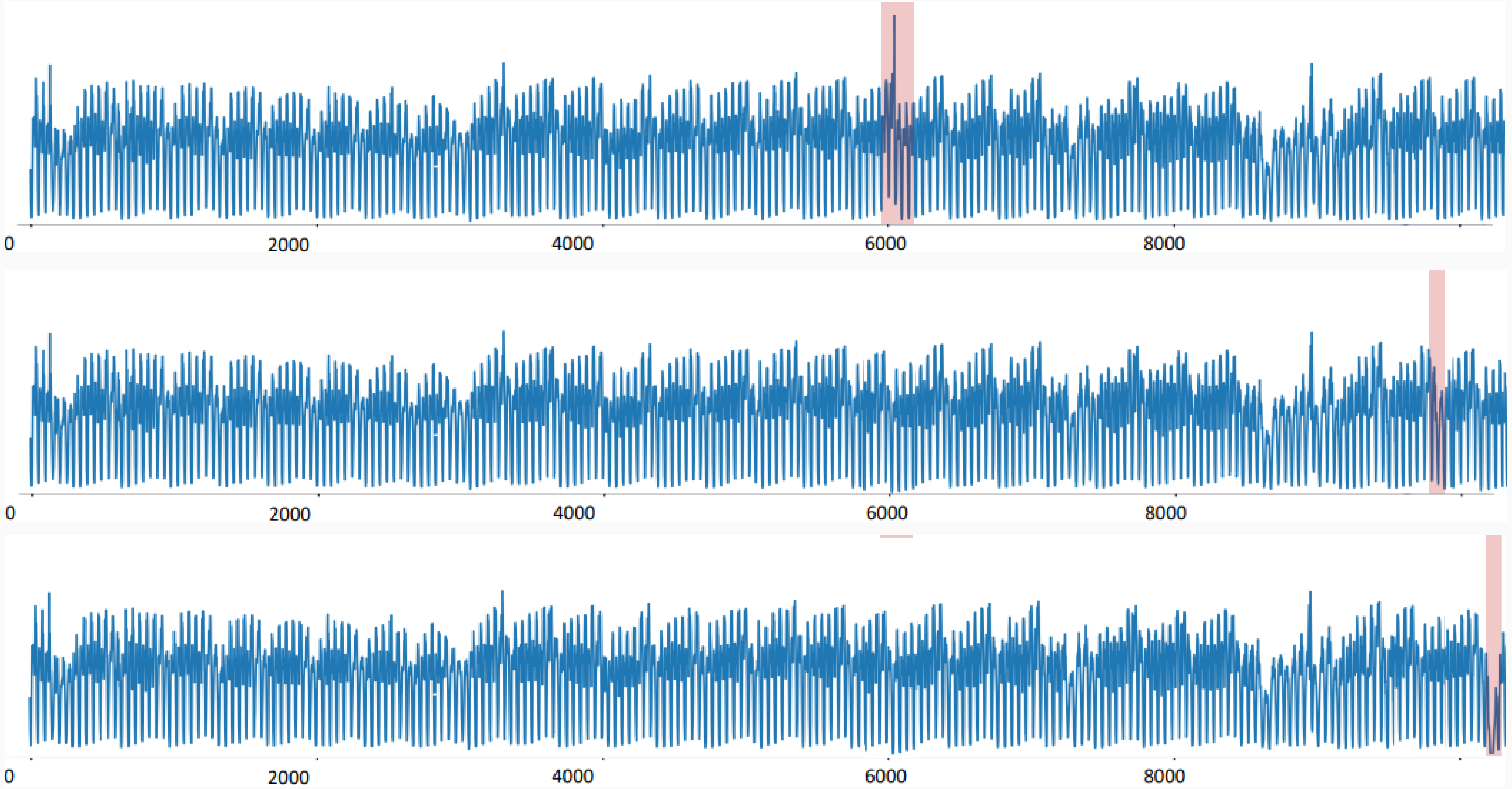


Fig 3. Categories for anomaly detection techniques. [2]

VALIDATION

1. Data Stream



2. Selecting a method

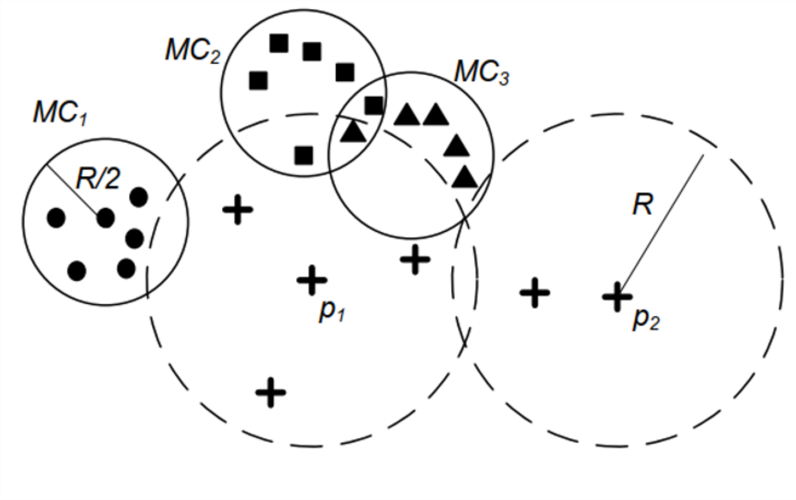


Fig 4. Main ideas MCOD. [4]

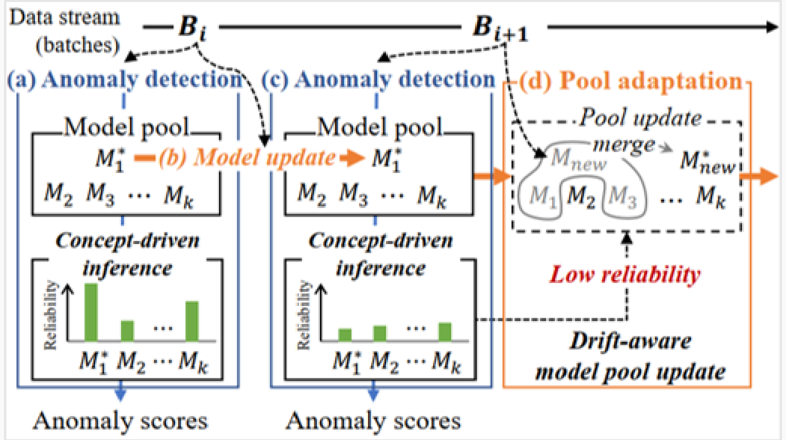
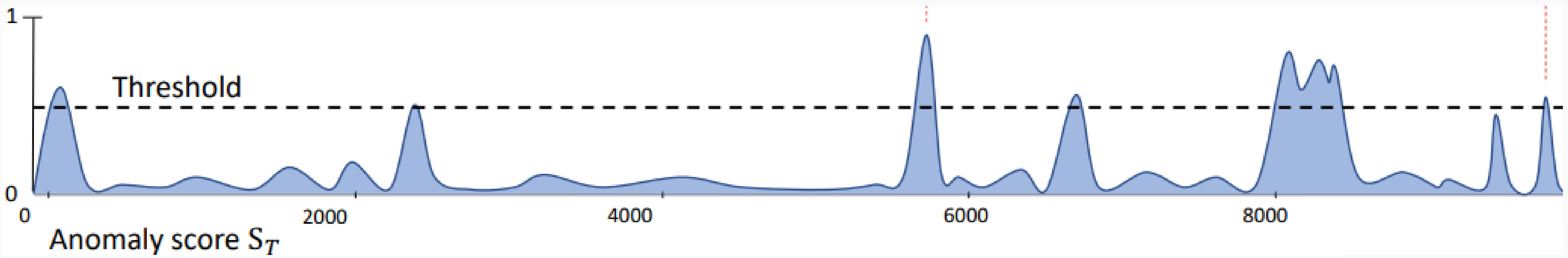


Fig 5. Main ideas ARCUS. [5]

3. Calculating the anomaly score



4. Interpreting the results

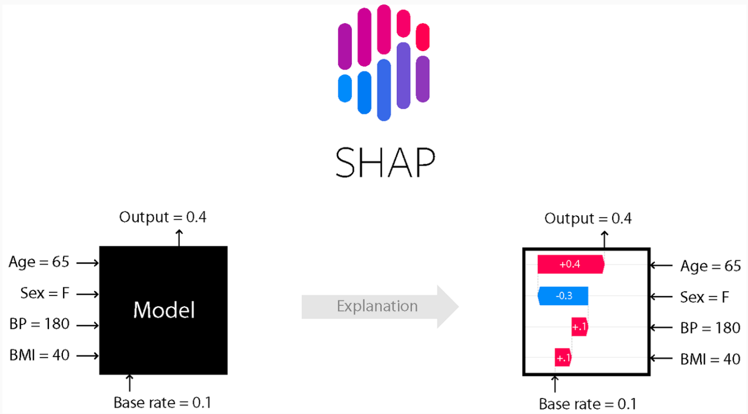


Fig 6. Main ideas SHAP. [7]

REFERENCES

1. Florentin Jiechieu et al. (2024) SEDAF : Prototype d’un Système Explicable de Détection d’Anomalies dans les Flux de Données. Revue des Nouvelles Technologies de l'Information, EGC, Dijon, France, pp. 441-448.
2. Lu, T., Wang, L., & Zhao, X. (2023). Review of anomaly detection algorithms for data streams. Applied Sciences, 13(10), 6353.
3. Liu, F.T.; Ting, K.M.; Zhou, Z.H. Isolation Forest. In Proceedings of the 2008 Eighth IEEE International Conference on Data Mining, Pisa, Italy, 15–19 December 2008; IEEE: New York, NY, USA, 2008; pp. 413–422.
4. Kontaki, M., Gounaris, A., Papadopoulos, A. N., Tsichlas, K., & Manolopoulos, Y. (2011, April). Continuous monitoring of distance-based outliers over data streams. In 2011 IEEE 27th International Conference on Data Engineering (pp. 135-146). IEEE
5. Yoon, S., Lee, Y., Lee, J. G., & Lee, B. S. (2022, August). Adaptive model pooling for online deep anomalydetection from a complexevolving data stream. In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (pp. 2347-2357).
6. Boniol, P., Paparizzos, J., & Palpanas, T. (2023). New Trends in Time Series Anomaly Detection. International Conference on Extending Database Technology.
7. Štrumbelj, E., & Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. Knowledge and information systems, 41, 647-665.

Thank you !

N. Rojas Varela, J. Ah-Pine, E. Mephu Nguifo