Univariate Time Series Data Mining and Machine Learning for Anomaly Detection on the ARRONAX Cyclotron Operation Data

Fatima Basbous (Arronax), Diana Mateus (LS2N), Freddy Poirier (Arronax), Ferid Haddad (Arronax)

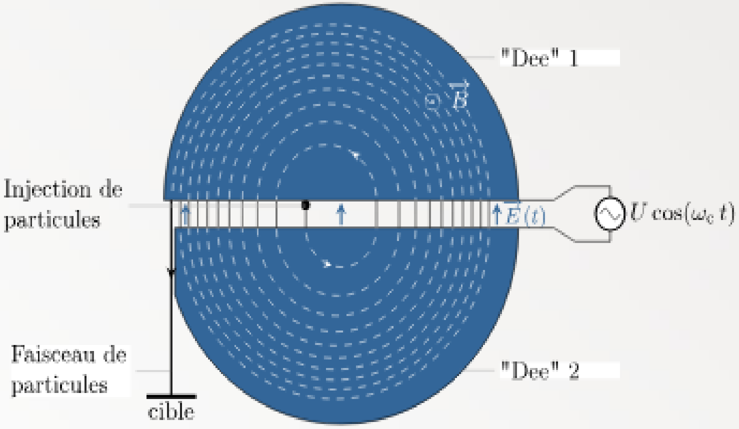
PhD hours presentation

24 April 2025

UNIVERSITÉ DE NANTES

Specificity of the C70XP

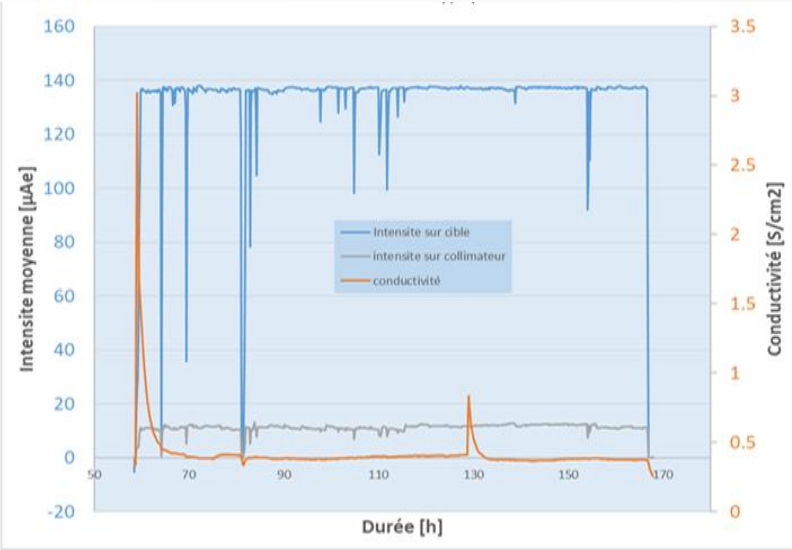
- ARRONAX is able to produce multiple types of particles.
- High-Power cyclotron for fixed target.



femto-physique.fr

Cyclotron: from injection to particle beam

Project Start: Anomaly Detection with Data Mining and ML



Time evolution of target intensity, collimator intensity, and conductivity during irradiation

- Typical proton intensity over time on a target: Relatively flat with breakdowns, stops and variations.
- In 2019, data exploration and the application of certain algorithms for anomaly detection on the operation data started at ARRONAX [1].

[1] F. Poirier et al., 2023, doi: 10.18429/JACoW-IPAC2023-TUPM036.

OBJECTIVE

OUR AIM

Develop an active anomaly detection method to improve reliability and operational efficiency.

COMPARATIVE STUDY

- One-Class SVM (OCSVM)
- Isolation Forest (IF)
- Autoencoder (AE)
- Autoencoder-Isolation Forest (AE-IF)

Data collection

Experimental and Industrial Physical Control System

EPICS

1. Installed at ARRONAX in 2016
2. Integrated with over 900 data points throughout the cyclotron
3. Generate Process Variables (PV)

Dataset Description

- 26 distinct datasets are used, each corresponding to a different experiment.
- Each dataset contains:
  - $X = \{x_1, x_2, \dots, x_n\}$
  - d: number of features per vector
  - n ≈ 75 000 observations per dataset (on average)
- 10% of the training data is used for hyperparameter tuning.

Data Split

Usage	Percentage
Training	80%
Testing	20%
Validation (From training)	10%

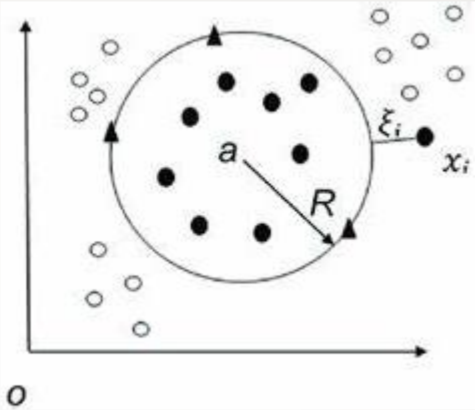
OCSVM

Principle

- Unlike traditional SVMs, OCSVM models only the normal data.
- Anomalies are considered as points close to the origin.
- Uses the kernel trick for high-dimensional projection.
- Support vectors define the decision boundary.
- Classifies data as normal (0) or anomalous (1) based on proximity to the boundary.

Used Parameters

- RBF (Radial basis function) kernel
- Contamination Tuned between 5% and 25% depending on dataset.



doi: https://doi.org/10.1371/journal.pone.0226115.g001

OCSVM: Outlier identification through a boundary

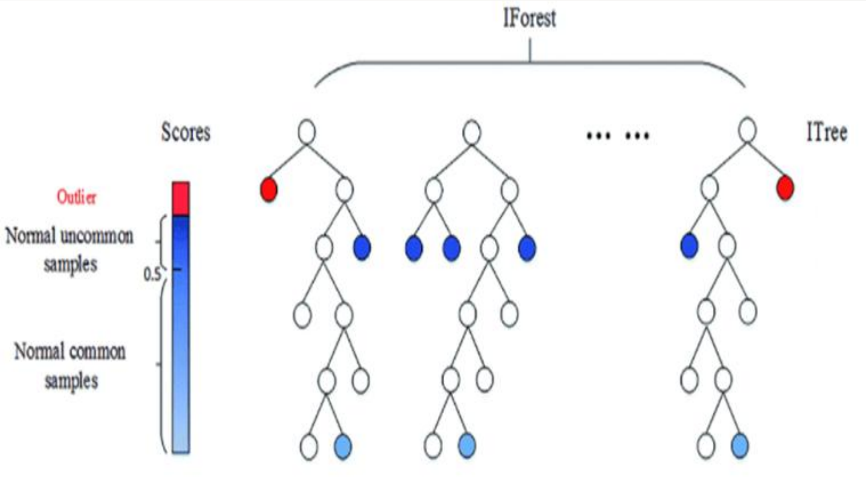
Isolation Forest

Principle

- Assumes anomalies are easier to isolate than normal points
- Builds a forest of randomly generated binary trees
- An anomaly score is computed based on the path length of a data point
  - Shorter path → higher anomaly score

Parameters

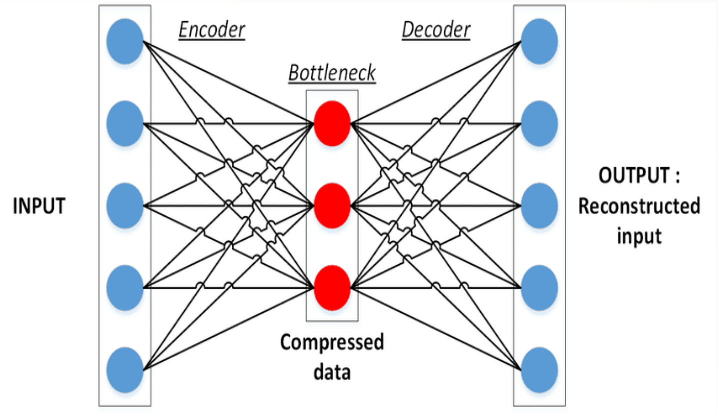
- Number of trees: 250
- Number of features per split: 2
- Proportion of data used per tree: 20%
- Contamination: Tuned between 5% and 25% per dataset



www.innova-tsn.com

Mechanism of Isolation Forest for Anomaly Detection

Autoencoder



www.analyticssindiamag.com

The basic structure of an autoencoder includes an encoder, a bottleneck and a decoder.

Principle

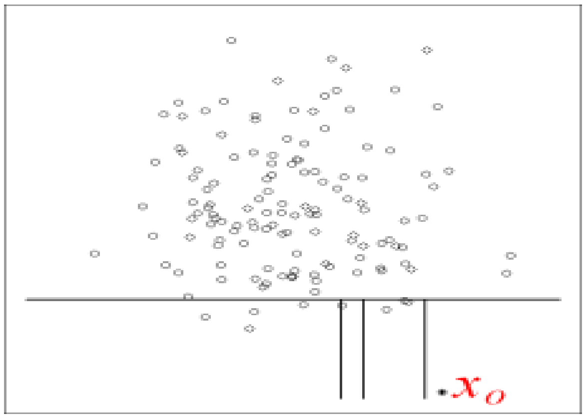
- Unsupervised neural networks designed to reconstruct their input.
- Learn to compress data into a lower-dimensional latent space and then decode it.

Approach

- Train the model only on normal data.
- Learns a compact representation of normal patterns.
- When presented with anomalous data, the model fails to reconstruct it accurately.
- Decision function: Mean Squared Error (MSE), somme des distances au carré entre la sortie et l'entrée du réseau
  - MSE > threshold: Label 1
  - MSE < threshold: Label 0

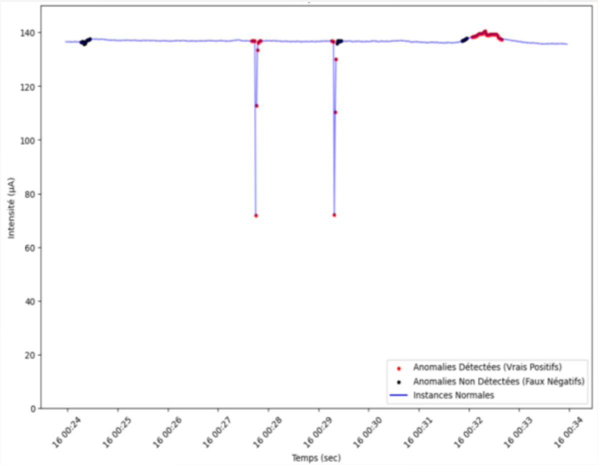
Towards a Hybrid Approach: AE + IF

- **Limitations of IF:** Axis-aligned splits in IF limit the detection of complex, near-mean anomalies.
- **Complex anomalies:** Revealed by small fluctuations around the mean.



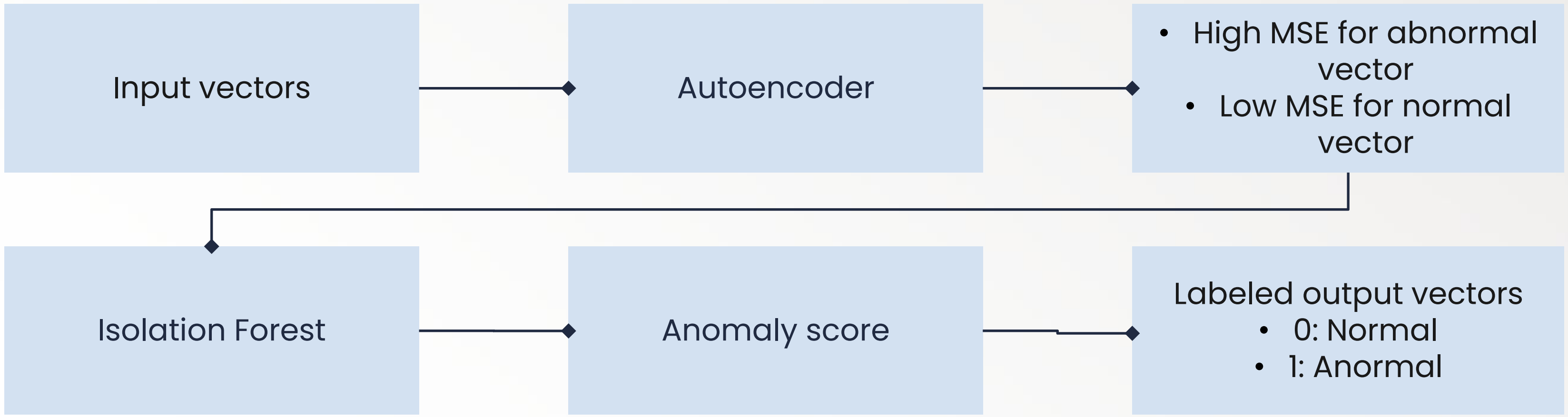
Liu et al., 2008, DOI 10.1109/ICDM.2008.17

Example of anomaly point x_0 isolation from a set of 135 points of a Gaussian distribution by axis-parallel partitioning using Isolation forest



Intensity Data Over Time with Isolation Forest Results: Detected Anomalies (Red), Missed Anomalies (Black), and Normal Data (Blue)

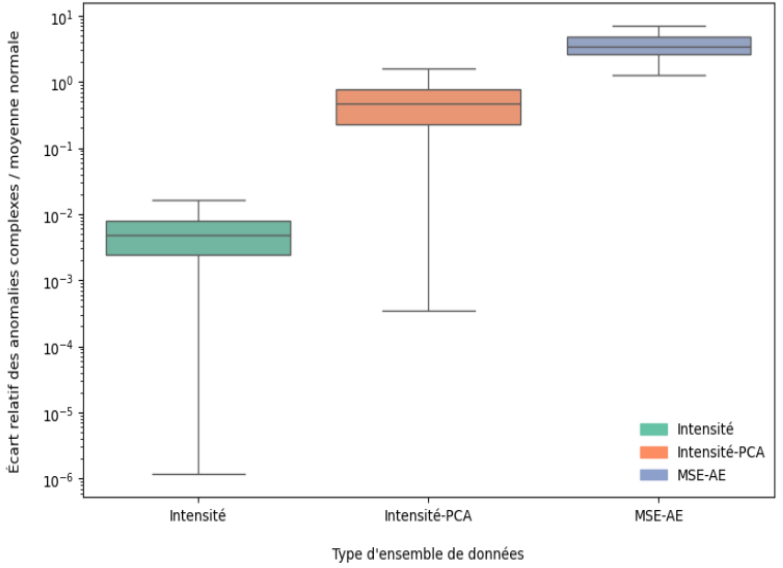
AE-IF approach



The AE-IF Anomaly Detection Pipeline

- **Hypothesis:** Combining the Autoencoder with Isolation Forest helps to shift complex anomalies away from the mean of normal data, making them easier to isolate.

AE-IF approach (Continued)

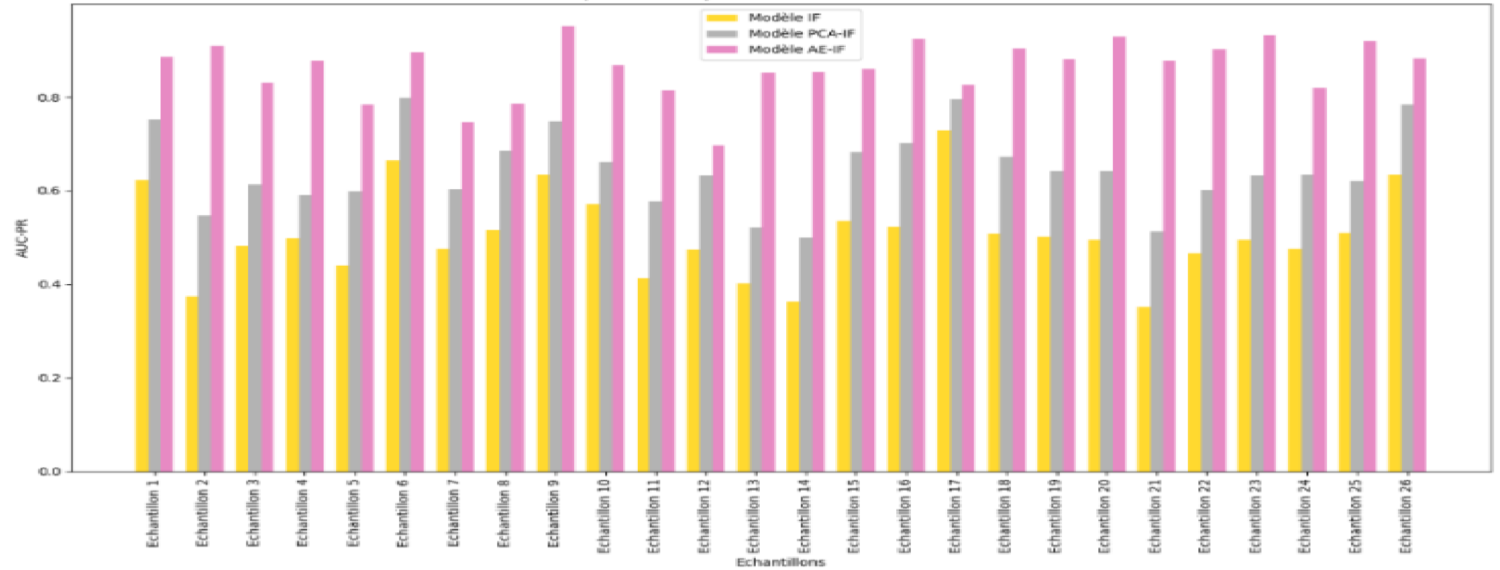


	Intensité	Intensité-PCA	MSE-AE
Min	1.15e-06	3.49e-04	1.27e+00
Q1	2.42e-03	2.26e-01	2.61e+00
Médiane	4.84e-03	4.67e-01	3.46e+00
Q3	7.95e-03	7.71e-01	4.78e+00
Max	1.74e-02	1.75e+00	7.17e+00

Comparative Statistics of Relative Deviations of Complex Anomalies from the Normal Mean in Three Cases: Raw Intensity, Intensity after PCA, and Autoencoder MSE

- PCA (Principal Component Analysis): Dimensionality reduction method.

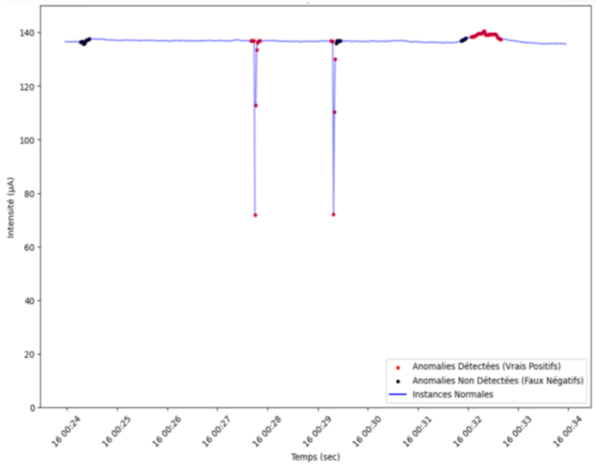
Experimental Results 1



Comparison of AUC-PR performance of the AE-IF, IF, and IF-PCA models on 26 datasets.

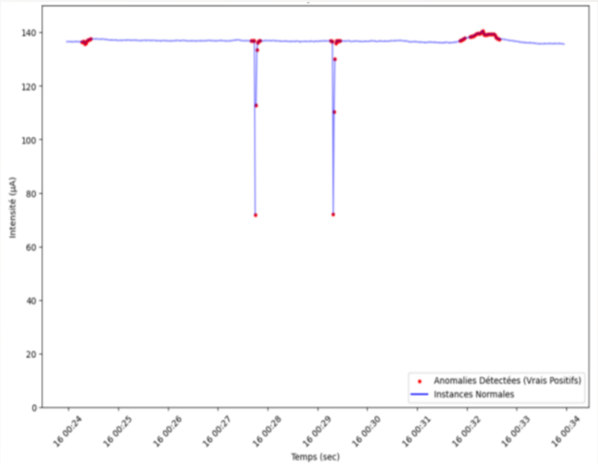
- **AUC-PR (Area Under the Precision-Recall Curve):** Measures the trade-off between precision and recall at different thresholds.

Experimental Results 1 (Continued)



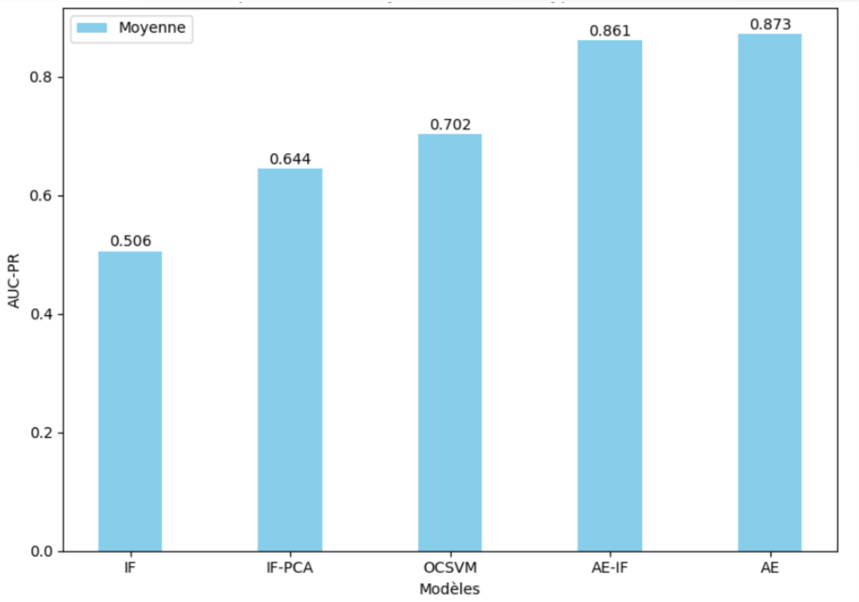
Intensity Data Over Time with Isolation Forest Results: Detected Anomalies (Red), Missed Anomalies (Black), and Normal Data (Blue)

- The AE-IF model effectively detects complex anomalies.



Intensity Data Over Time with AE-IF Results: Detected Complex and Simple Anomalies (Red), and Normal Data (Blue)

Experimental Results 2



Comparison of Average AUC-PR Performance Across Anomaly Detection Algorithms: IF, IF-PCA, OCSVM, AE-IF, and AE

- The top-performing models: AE-IF and AE.

Conclusion

- **The autoencoder proves to be particularly effective for our temporal data.** By learning meaningful latent representations, it reduces noise, highlights normal behaviors, and detects subtle anomalies often missed by traditional methods.

Thank you

Open for your questions