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

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

IN2P3/IRFU Machine Learning workshop 2024

Strasbourg, France

C70XP: A Cyclotron with Multiple Activities

1. Production of Radionuclides for Nuclear Medicine

- Imaging
- Therapy

2. Research and Development

- Radiochemistry and Radiobiology
- Physics and Detector Development
- Training and Education

ARRONAX

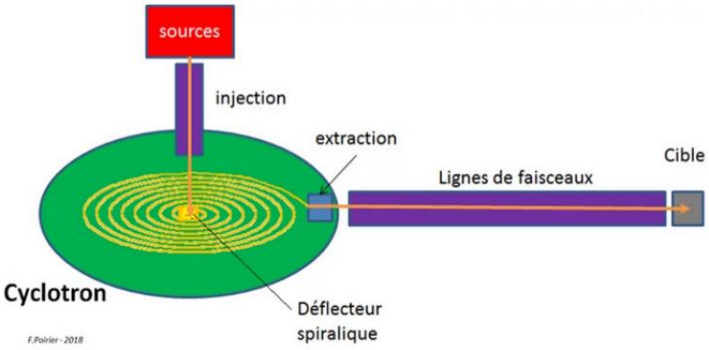
Specificity of the C70XP

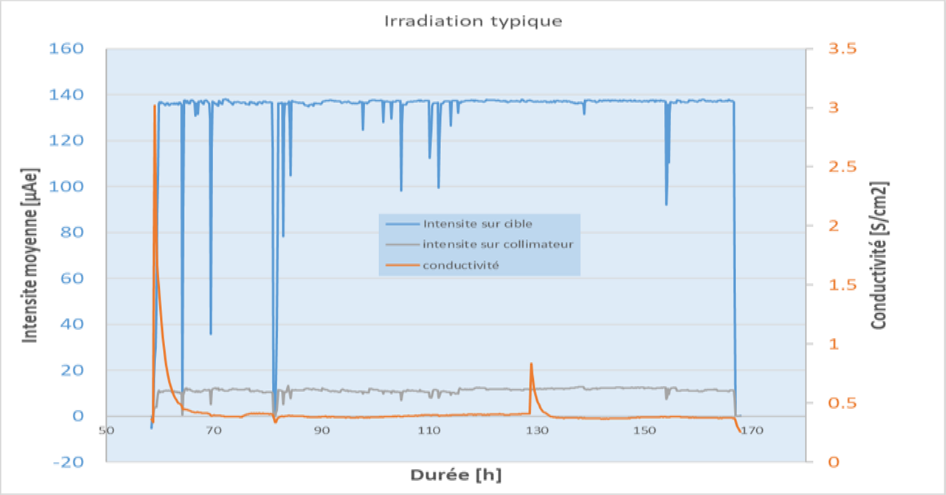
- ARRONAX is able to produce multiple types of particles.
- High-Power Cyclotron for Fixed Target.

Beam Characteristics

Faisceaux de Particules	Énergies (MeV)	Intensités max. (µA)
Protons (H^+)	35-70	375x2
Particules α (He^{2+})	68	70
Dihydrogènes ionisés (HH^+)	35	50
Deutons (D^+)	15-35	50

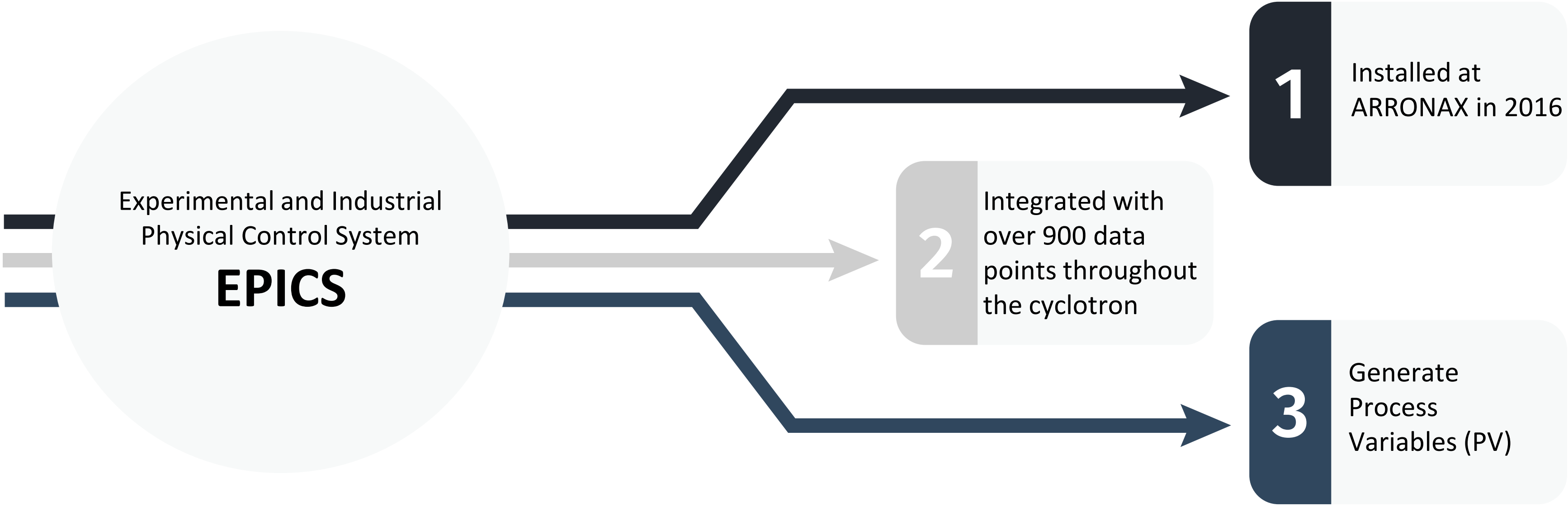
Specificity of the C70XP





- Typical proton intensity over time on a target: Relatively flat with breakdowns, stops and variations.

Anomaly Detection at ARRONAX



At ARRONAX, data exploration and the application of certain algorithms started in 2019.

F.Poirier et al., ” First anomalies exploration from data mining and machine learning at the ARRONAX cyclotron C70XP”, JACoW IPAC2023 (2023) TUPM036, doi: 10.18429/JACoW-IPAC2023-TUPM036

Problem Statement

In order to detect all types of anomalies in our data, why should we turn to machine learning methods rather than relying on statistical approaches?

Comparative Approach

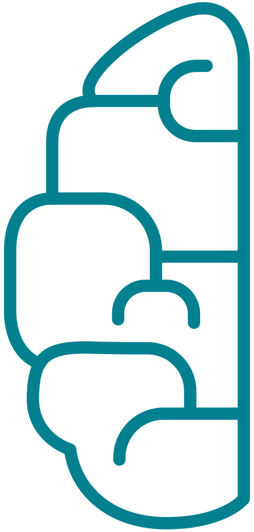
01 Grubbs' test

02 Interquartile method

03 Isolation forest

04 OCSVM

05 Autoencodeur



Statistical Methods: Interquartile Method and Grubbs' Test

Methodology: Grubbs' Test and Interquartile Method

Grubbs' test

Input X={x1 ,x2 ,...,xn } where $x_i \in \mathbb{R}$ is a data point

01 Calculation of the Z-score for each point

02 Calculation of the critical value G

03 Z score > G: Anomaly

Interquartile method

Input X={x1 ,x2 ,...,xn } where each $x_i \in \mathbb{R}^d$ is a feature vector of dimension d

01 Calculation of the quartiles Q1 and Q3 of the sequence means and the interquartile range (IQR)

02 Definition of lower and upper thresholds

03 Mean < lower threshold or > upper threshold: Anomaly

Output: Labeled Points: 1 (anomaly) and 0 (normal)

$$\text{Z score}(xi) = \frac{xi - \mu}{\sigma}$$

- μ and σ: The mean and standard deviation of the dataset

$$G_{\text{critique}} = \frac{N-1}{\sqrt{N}} \times \sqrt{\frac{t^2_{\alpha/(2N),N-2}}{N-2+t^2_{\alpha/(2N),N-2}}}$$

- N: The size of the dataset and α: the test sensitivity (0.05)

- IQR= Q3-Q1
- Lower threshold: $Q1 - 1.5 \times \text{IQR}$
- Upper threshold: $Q3 + 1.5 \times \text{IQR}$

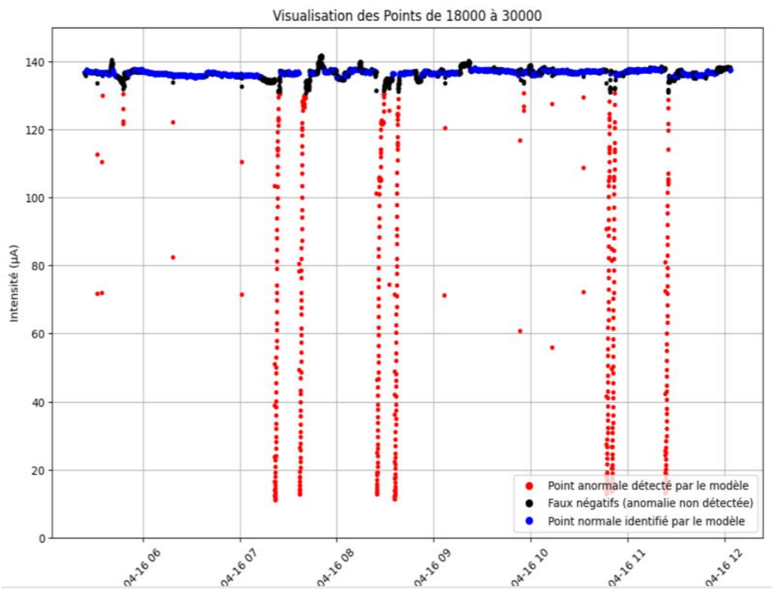
Performance: Grubbs' Test and Interquartile Range Method

Performance evaluation of the IQR method and Grubbs' Test using different metrics

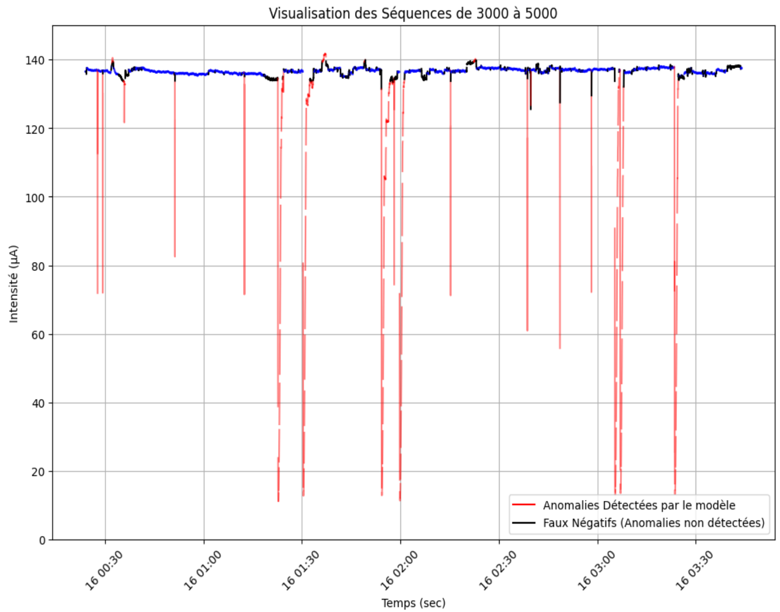
Metrics	Precision	Recall	F1-score	Accuracy
Score with IQ	1.00	0.27	0.42	0.89
Score with Grubbs	1.00	0.15	0.25	0.87

- Precision, recall and F1 score of the poisitive class
- The low F1 score and recall indicate the limitations of these methods in effectively detecting anomalies.

Qualitative Result: Grubbs Test and Interquartile Method



Visualization of a Sample of Target Intensity Data Showing Abnormal Intensities (in Red), Normal Intensities (in Blue), and Undetected Anomalies (in Black) After Applying Grubbs' Test.



Visualization of a Sample of Target Intensity Data Showing Abnormal Intensities (in Red), Normal Intensities (in Blue), and Undetected Anomalies (in Black) After Applying the Interquartile Method.

Machine Learning : Isolation Forest, OCSVM and Auto-encoder

One-Class SVM (OCSVM)

Input:
X={x1 ,x2 ,…,xn } where each $xi \in Rd$ is a feature vector of dimension d, in our case d=6

OCSVM

Output:
Labeled Points: 1 (anomaly) and 0 (normal)

Transformation of the data into a higher-dimensional space F using RBF kernel: $k(xi,xj)$

Encapsulation of normal data within a region far from the origin.

The Lagrange multipliers αi identify the support vectors, thereby defining the decision boundary that outlines the normal region.

Calculation of the decision function:
f(x) > 0 : normal, f(x) < 0 : anormal

$$K(x_i, x_j) = \exp\left(-\gamma \|x_i - x_j\|^2\right)$$

- $K(xi,xj)$ represents the similarity between points xi et xj in the transformed space.
- γ is a parameter that controls the extent of influence of neighboring points on the transformation.

$$f(x) = \text{sign}\left(\sum_{i=1}^{n} \alpha_i K(x_i, x) - \rho\right)$$

- αi represents the Lagrange multipliers.
- ρ represents the decision threshold..

Deterministic autoencoder

1

Encoder

- Input Layer: Receives the sequential data X={x1 ,x2 ,…,xn } where each $xi \in Rd$ is a feature vector of dimension d, in our case, 6.
- Dense Layer: Composed of 4 neurons using the ReLU activation function.
- Extraction of the most relevant features.

2

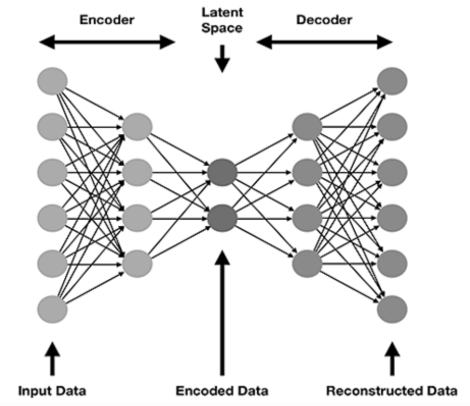
Latent space

- Latent Layer (Dense Layer): Reduces the dimensionality to encoding_dim (encoding_dim=2).

3

Decoder: A mirror of the encoder

- Output: Reconstructed sequences with output_dim = input_dim.
- Restoration of data to its original form from the latent space.



www.query.ai

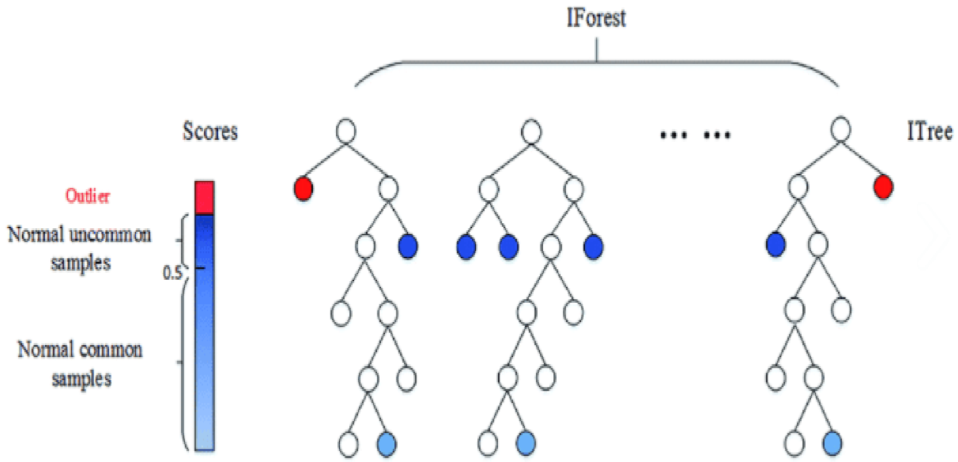
Network architecture for the deterministic autoencoder used

Decision function: Mean Squared Error (MSE).

- MSE > MSE_threshold: Label 1
- MSE < MSE_threshold: Label 0

Isolation forest

- **Input:** X={x1 ,x2 ,…,xn } where each $xi \in Rd$ is a feature vector of dimension d, in this case 6.



www.innova-tsn.com

Mechanism of Isolation Forest for Anomaly Detection

Decision function: f(x)= $2^{h(n)/c(n)}$ – threshold

- $h(x)$: The average path length to isolate x across all the trees in the forest.
- c(n): A normalization function that depends on n, the size of the dataset.

➢ f(x) > 0: Normal
➢ f(x) < 0: Anormal

- **Output**: Labeled Points: 1 (anomaly) and -1 (normal)

Performance: OCSVM, AE et IF

- Input: X={x1 ,x2 ,…,xn } where each $xi \in Rd$ is a feature vector of dimension d, in this case 6.
- Splitting the data into 80/20 for training and testing, with a label 1 rate of 15.06% (April 2019 sample).

Metrics	Precision	Recall	F1-score	Accuracy	AUC ROC	AUC PR
Score with OCSVM	0.84	0.80	0.82	0.95	0.95	0.87
Score with AE	0.95	0.81	0.88	0.97	0.94	0.89
Score with IF	0.68	0.54	0.60	0.88	0.85	0.68

Q1: Why is the performance of the Isolation Forest lower than that of OCSVM and the autoencoder?

March 2021 sample with a label 1 rate of 15.80%.

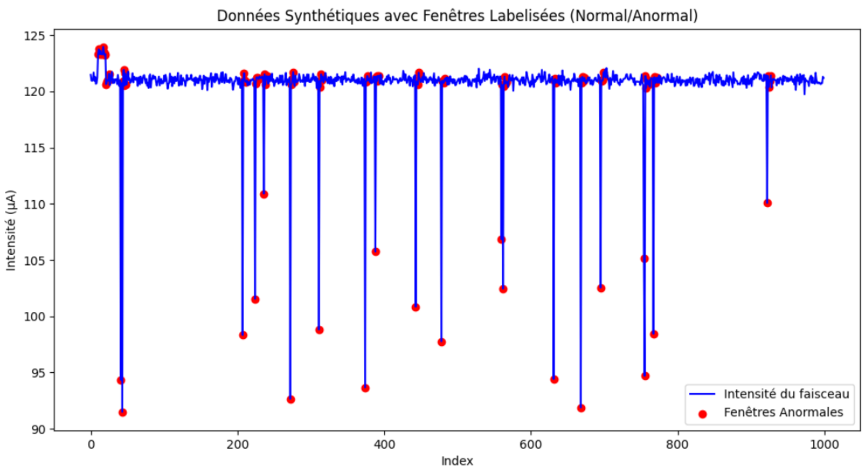
Score with IF 2021 sample	0.51	0.49	0.50	0.85	0.83	0.58

Challenges of Isolation Forest

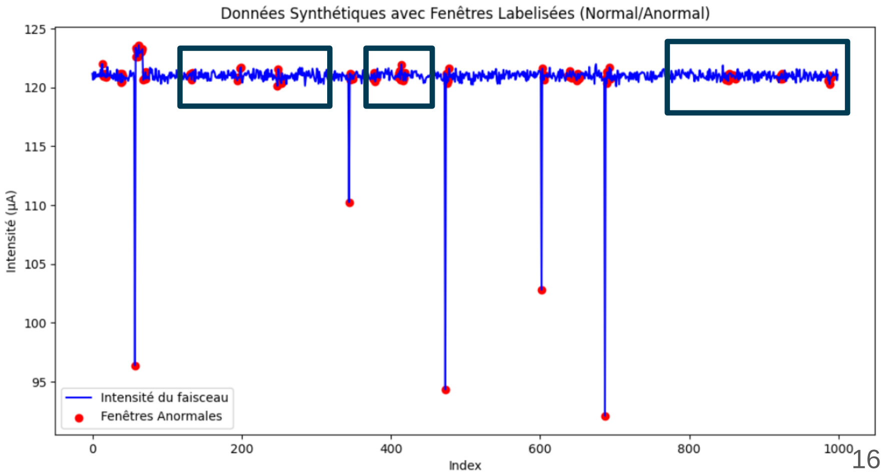
Synthetic Data:

- Normal: With a standard deviation from the mean between 0.2 and 0.4.
- Breakdowns and fluctuations: With standard deviations greater than 2.
- Noise: With standard deviations between 0.45 and 0.65 from the mean.

First application: Normal, breakdowns, and fluctuations (Without noise)

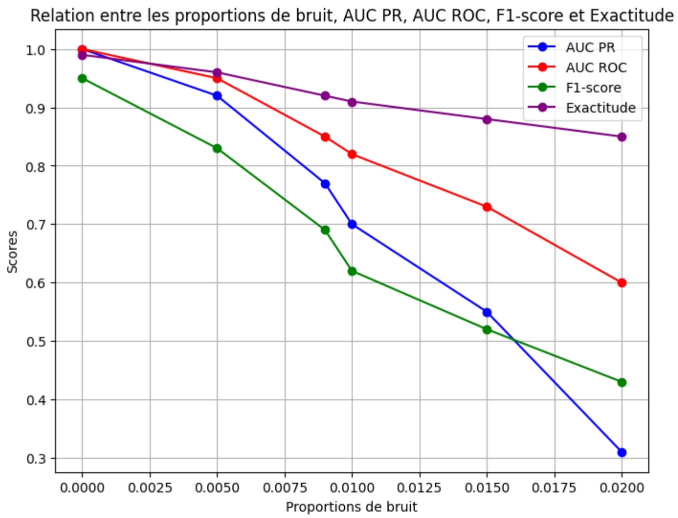


The other applications: Normal, breakdowns, fluctuations, and noise.



Performance of IF on synthetic data

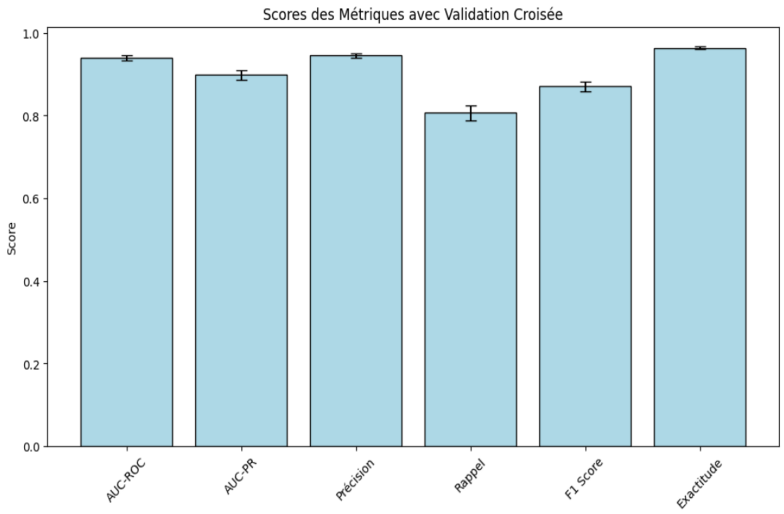
- Limitations of IF: It is mainly effective for detecting anomalies that are far from the mean, but less efficient for those that are more subtle.



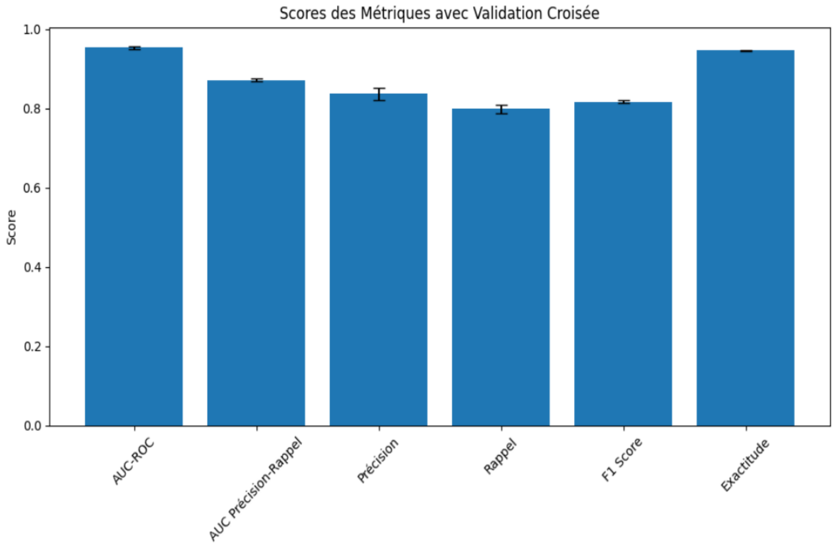
Variation of the values of different performance metrics as a function of the change in the percentage of noise in the synthetic data anomalies.

5-fold cross validation

AE



OCSVM



Metrics	Precision	Recall	F1-score	Accuracy	AUC ROC	AUC PR
Score with OCSVM	0.8367 ± 0.0164	0.7994 ± 0.0105	0.8174 ± 0.0032	0.9462 ± 0.0019	0.9540 ± 0.0031	0.8718 ± 0.0037
Score with AE	0.9454 ± 0.0052	0.8064 ± 0.0175	0.8703 ± 0.0109	0.9639 ± 0.0025	0.9395 ± 0.0064	0.8980 ± 0.0112

Conclusion

Metrics	Precision	Recall	F1-score	Accuracy	AUC ROC	AUC PR
Score with OCSVM	0.84	0.80	0.82	0.95	0.95	0.87
Score with AE	0.95	0.81	0.88	0.97	0.94	0.89
Score with IF	0.68	0.54	0.60	0.88	0.85	0.68
Score with Grubbs' test	1.00	0.15	0.25	0.87	-	-
Score with IQ method	1.00	0.27	0.42	0.89	-	-

- The machine learning methods explored so far show better performance on our data than the two methods tested.
- Our study aims to explore a machine learning method that could surpass these studied algorithms.

Thank you!

Open for your Questions