RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

FRANCE 2030 PROGRAMME DE RECHERCHE CYBERSÉCURITÉ

PEPR CYBERSÉCURITÉ
Plénière SuperviZ -WP2 - TH2.2

Explainable AI for Network Intrusion Detection Systems in Industrial Control Systems

Léa Astrid KENMOGNE, LIG, Grenoble-INP, UGA
Supervisor : **Stéphane MOCANU**

11-03-2025

anr® agence nationale de la recherche

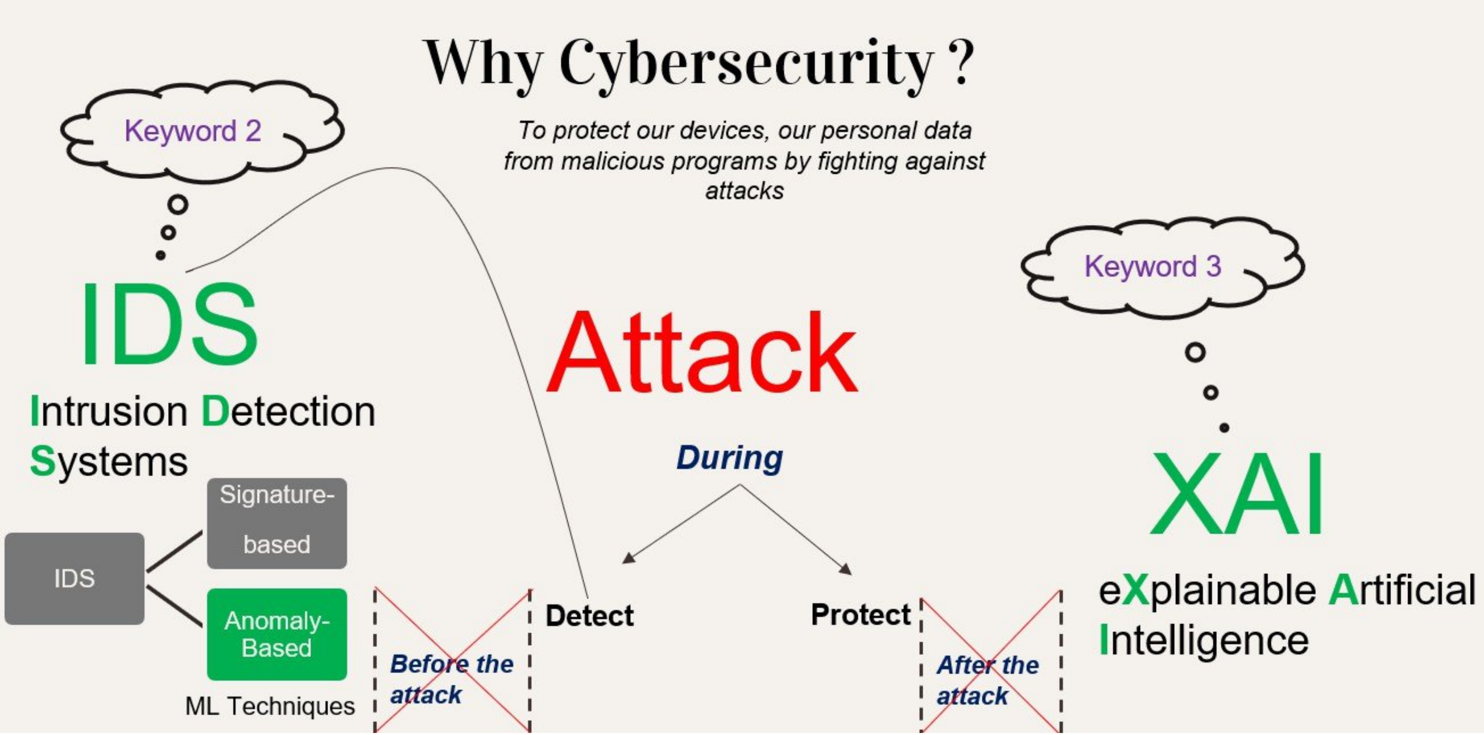
Context

- Cyber attacks target not only IT systems, but also Industrial Control Systems. These are a set of physical and digital elements that interact to ensure the execution of an objective in an industrial environment.

Compared to IT systems, ICS attacks are harder to detect due to:

- Limited resources, restricting additional processes
- Component and technology diversity
- Risk of disrupting system operations.

Context



Why Cybersecurity ?

To protect our devices, our personal data from malicious programs by fighting against attacks

Problem Statement

XAI for IDS in ICS

Establish a system to detect any abnormal behavior in industrial control processes using explainable artificial intelligence techniques.

- **Why is this problem important to deal with ?**
  - **New attacks emerge**
  - **IT systems are different from ICS systems**
  - **Detection during the attack to mitigate consequences**

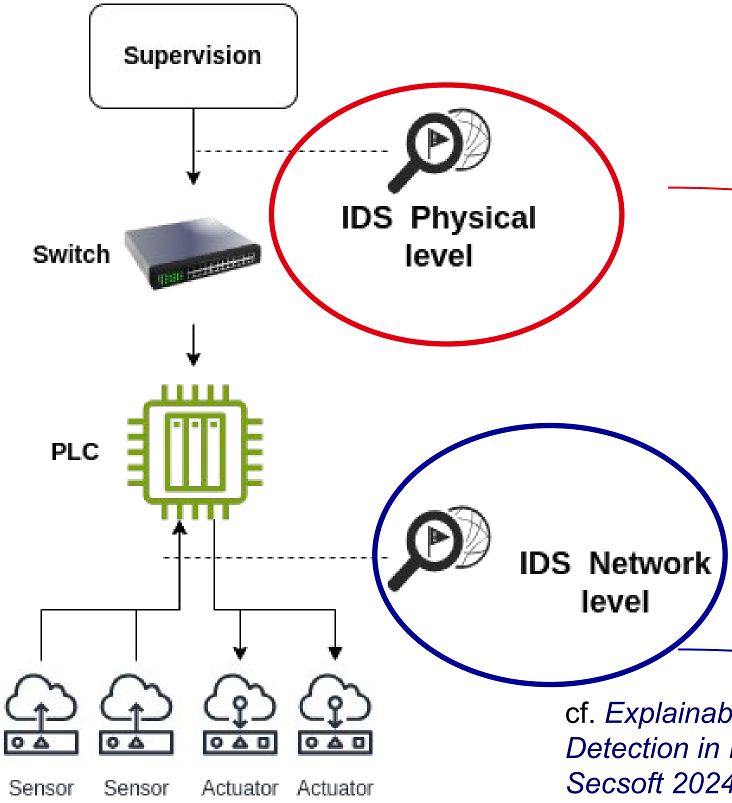
Related Work

- ❑ *Intrusion Detection for ICS, Oualid Koucham, Thesis 2018*
- ❑ *Explaining Anomalies Detected by Autoencoders using SHAP, Liat Antwarg et al., 2019*

- ❑ *Explainable AI for Process-Aware Attack Detection in ICS, Léa Kenmogne and S. Mocanu, Secsoft 2024*

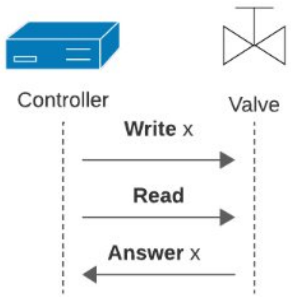
Contribution

- First use-case : overview of the process from the **network point of view** with the sequence of requests.
- Second use-case : study of another process at the **physical level** with the state of the sensors and actuators

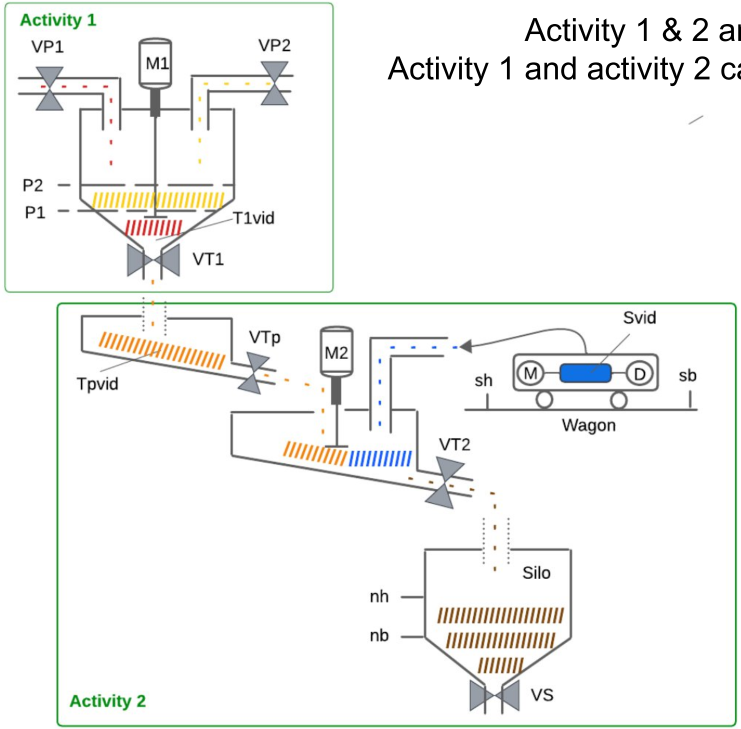


Generated dataset containing a sequence of Write, Read and Answer

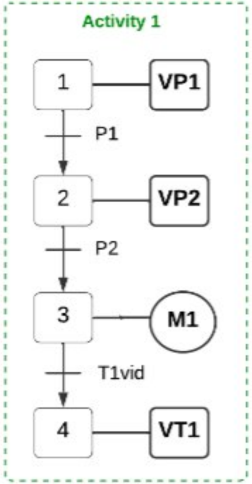
cf. *Explainable AI for Process-Aware Attack Detection in ICS, Léa Kenmogne and S. Mocanu, Secsoft 2024*

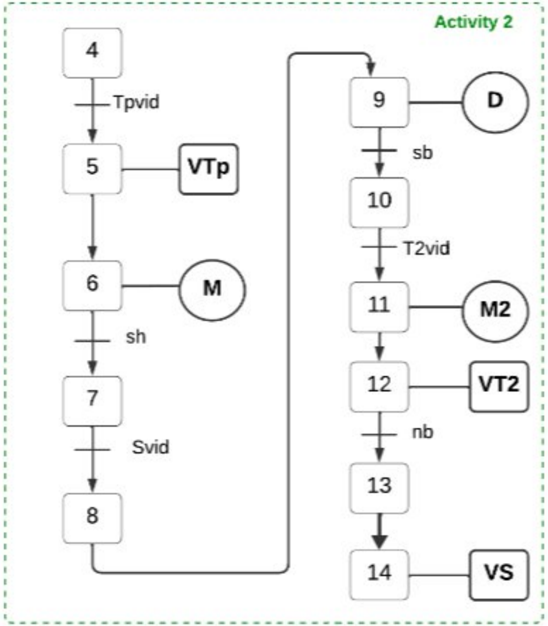


Contribution (Physical Level Detection)

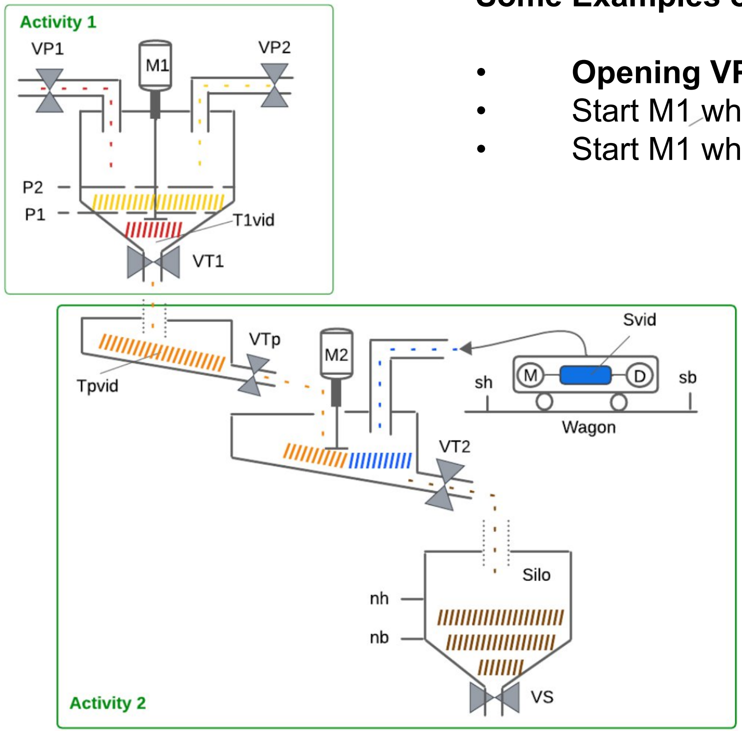


Activity 1 & 2 are sequential
Activity 1 and activity 2 can run simultaneously.





Contribution

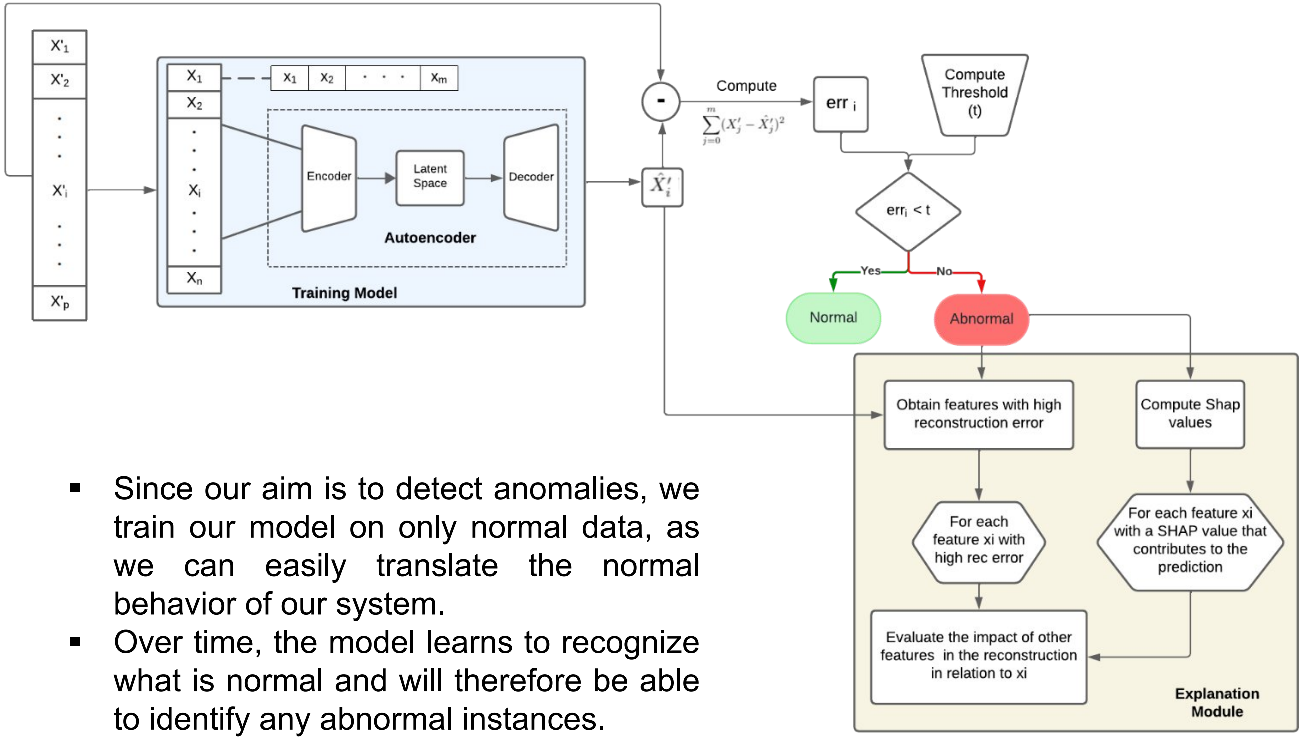


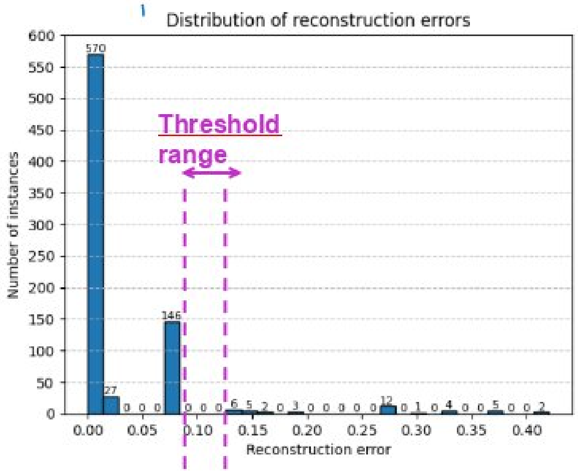
Some Examples of Attacks in Activity 1

- **Opening VP1 when P1 is reached**
- Start M1 when VP2 is open
- Start M1 when T1 is empty

Index	P1	P2	T1vid	VP1	VP2	M1	VT1
1	1	0	0	0	1	0	0
2	1	0	0	0	1	0	0
3	1	0	0	1	1	0	0
4	1	0	0	1	1	0	0
5	1	0	0	1	1	0	0
6	1	0	0	1	1	0	0
7	1	1	0	1	0	1	0
8	1	1	0	1	0	1	0
9	1	1	0	1	0	1	0
10	1	1	0	0	0	1	0

Training Model


X'1
X'2
X'i
X'p
X1
X2
Xi
Xn
x1
x2
xm
Encoder
Latent Space
Decoder
Autoencoder
Training Model
$\hat{X}'_i$
-
Compute
$\sum_{j=0}^{m}(X'_j - \hat{X}'_j)^2$
err i
Compute Threshold (t)
$err_i < t$
Yes
No
Normal
Abnormal
Obtain features with high reconstruction error
Compute Shap values
For each feature xi with high rec error
For each feature xi with a SHAP value that contributes to the prediction
Evaluate the impact of other features in the reconstruction in relation to xi
Explanation Module



- Since our aim is to detect anomalies, we train our model on only normal data, as we can easily translate the normal behavior of our system.
- Over time, the model learns to recognize what is normal and will therefore be able to identify any abnormal instances.

Critical Task:

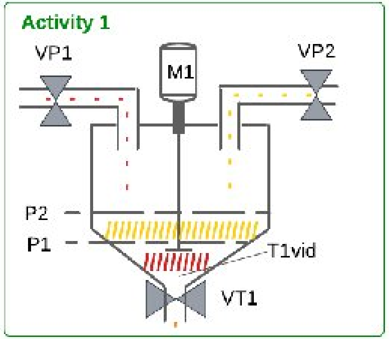
Define the threshold

Evaluation & Explanation (SHAP)

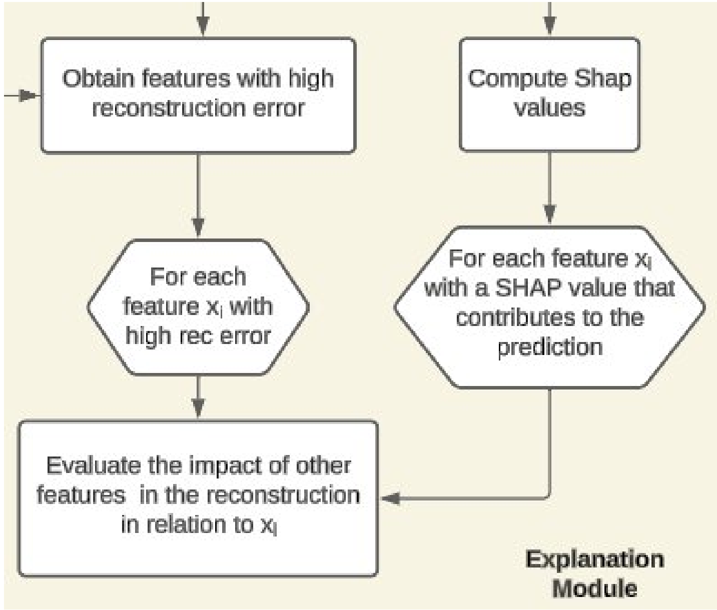
Detection Results of Activity 1

TP	FP
38	2
FN	TN
0	743

Situations that the model did not encounter during the training phase



SHAP is a model-agnostic method based on Shapley values, which are in turn based on game theory and represent the marginal contribution of each feature to model prediction.



Obtain features with high reconstruction error

Compute Shap values

For each feature x_i with high rec error

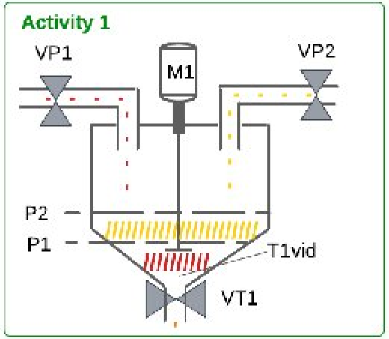
For each feature x_i with a SHAP value that contributes to the prediction

Evaluate the impact of other features in the reconstruction in relation to x_i

Explanation Module

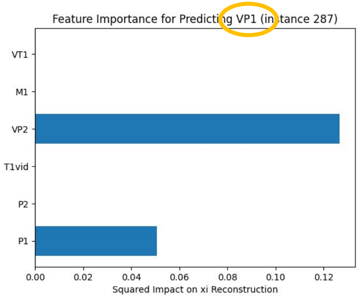
Evaluation & Explanation (SHAP)

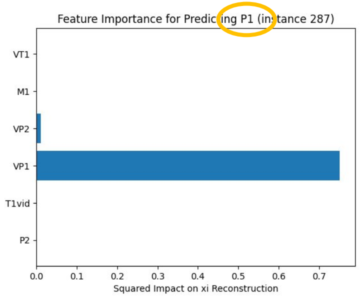
A1 : **Opening VP1 when P1 is reached**

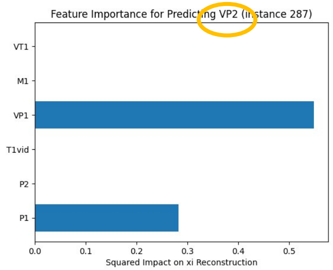


Reconstruction error

P1_c	7.614591e-01
P2_c	3.280796e-07
T1vid_c	6.387896e-05
VP1_c	9.940301e-01
VP2_c	5.767160e-01
M1_c	2.278845e-08
VT1_c	3.555541e-08







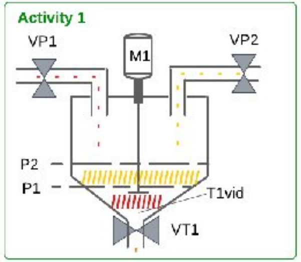
Index	P1	P2	T1vid	VP1	VP2	M1	VT1
1	1	0	0	0	1	0	0
2	1	0	0	0	1	0	0
3	1	0	0	1	1	0	0
4	1	0	0	1	1	0	0
5	1	0	0	1	1	0	0
6	1	0	0	1	1	0	0
7	1	1	0	1	0	1	0
8	1	1	0	1	0	1	0
9	1	1	0	1	0	1	0
10	1	1	0	0	0	1	0

Patterns Detected

- □ Opening VP1 when P1 is reached
- □ Open VP1 and VP2 simultaneously *(Index 3-6)*
- □ Start M1 when VP1 is open *(Index 7-9)*

Classical approach detects only one pattern

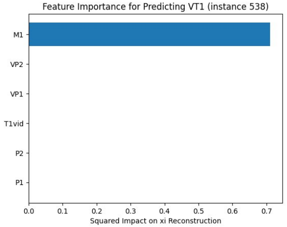
Evaluation & Explanation (SHAP)

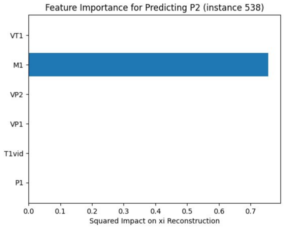


A2 : **Opening VT1 when M1 runs**

Reconstruction error (Index 3-4)

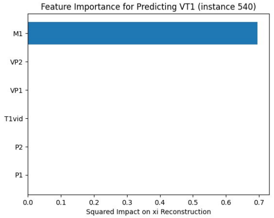
P1_c	2.803364e-07
P2_c	9.950556e-01
T1vid_c	1.543195e-06
VP1_c	4.913005e-06
VP2_c	2.005216e-04
M1_c	9.008750e-05
VT1_c	9.347368e-01

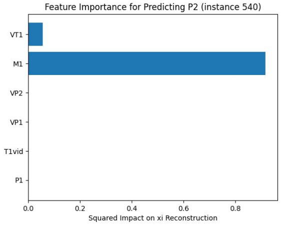


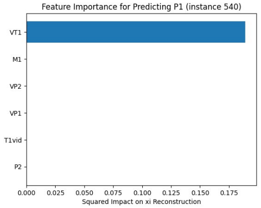


Reconstruction error (Index 5-6)

P1_c	0.997598
P2_c	0.989698
T1vid_c	0.000003
VP1_c	0.000009
VP2_c	0.000167
M1_c	0.000384
VT1_c	0.955338

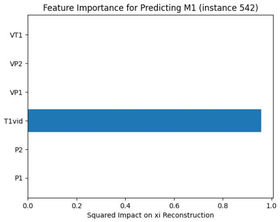






Reconstruction error (Index 7-11)

P1_c	0.596052
P2_c	0.007943
T1vid_c	0.298690
VP1_c	0.000042
VP2_c	0.000030
M1_c	0.997505
VT1_c	0.730999



Patterns Detected

- □ Opening VT1 when M1 runs
- □ Start M1 when P2 is not reached *(Index 3-11)*
- □ VT1 open when P2 is not reached *(Index 3-11)*
- □ VT1 open when P1 is not reached *(Index 5-11)*
- □ M1 runs while T1 is empty *(Index 7-11)*
- □ M1 runs while P1 is not reached *(Index 5-11)*

Classical approach detects only one pattern

In-dex	P1	P2	T1 vid	VP1	VP2	M1	VT1
1	1	1	0	0	0	1	0
2	1	1	0	0	0	1	1
3	1	0	0	0	0	1	1
4	1	0	0	0	0	1	1
5	0	0	0	0	0	1	1
6	0	0	0	0	0	1	1
7	0	0	1	0	0	1	1
8	0	0	1	0	0	1	1
9	0	0	1	0	0	1	1
10	0	0	1	0	0	1	1
11	0	0	1	0	0	1	1
12	0	0	1	0	0	1	0

Conclusion & Future Work

- ➢ Use other explainability methods like LIME to explain results
- ➢ Compare results with SHAP and detect new patterns
- ➢ Work on more consistent datasets from industrial systems (e.g. Singapore datasets)

RÉPUBLIQUE
FRANÇAISE
Liberté
Égalité
Fraternité

FRANCE 2030

PROGRAMME
DE RECHERCHE
CYBERSÉCURITÉ

Thank you