ICSRS 2025
2025 The 9th International Conference on
SYSTEM RELIABILITY AND SAFETY
2025 9th International Conference on Reliability Engineering ICRE

TECHNICAL CO-SPONSORED BY
POLITECNICO MILANO 1863
IEEE
IEEE Reliability Society 75th anniversary

SUPPORTED BY
ESRA
BEIHANG UNIVERSITY
UNIVERSITÉ DE LORRAINE
A. James Clark School of Engineering CALCE

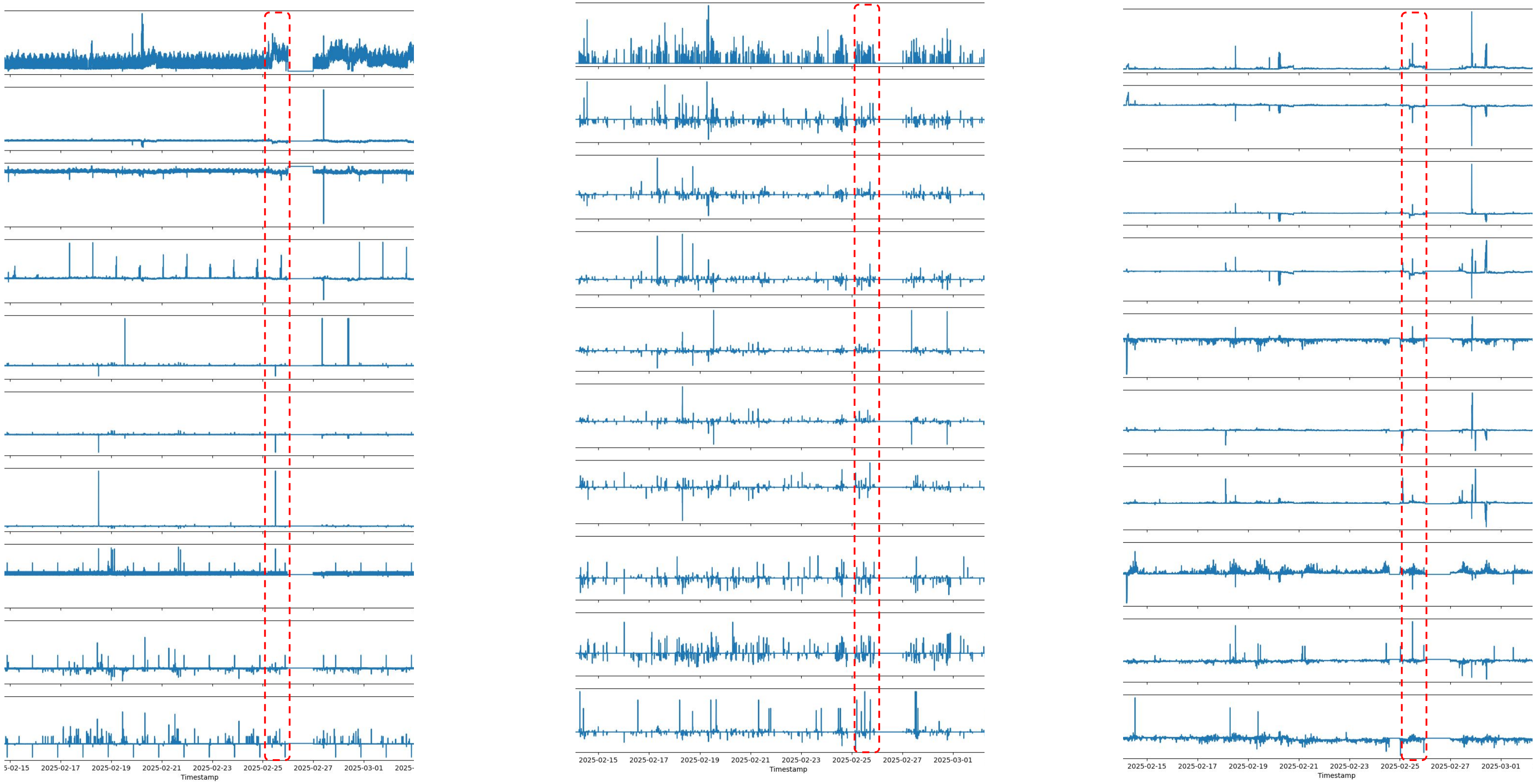
Turin, Italy
November 26-28, 2025

Investigating Explainable AI-enhanced Anomaly Detection Methods for Emergency Shutdown Mitigation in High-Temperature Heat Pump

Dr. -Ing. Baptiste Lambert, Deeksha Singh
The 9th International Conference on System Reliability and Safety Turin, Italy - Nov.26-28.2025

DLR

Motivations

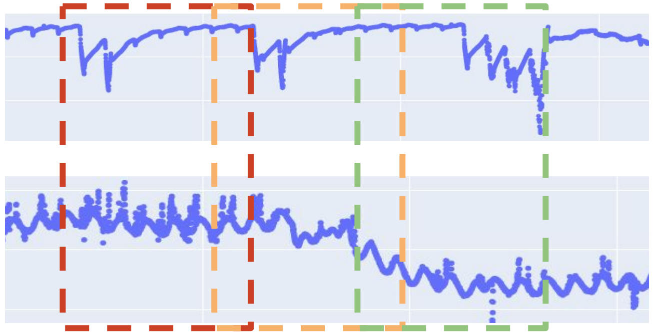


Which feature(s) out of the 30 is failing in the red box?

Methods

Workflow: Preprocessing

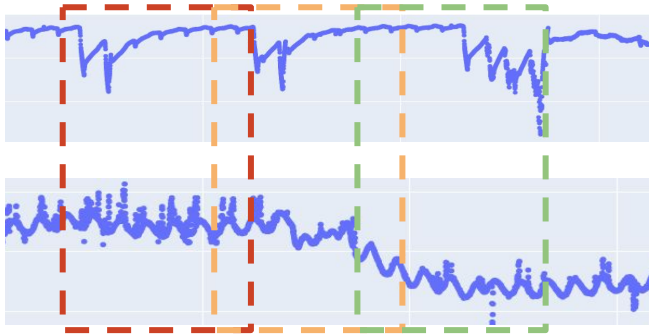
Preprocessing & Sequencing

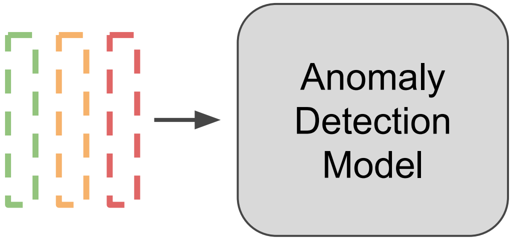


- Evenly sample time series,
- Fully populated,
- Generate 3 distincts datasets of fixed size windows.

Workflow: Training & Calibration

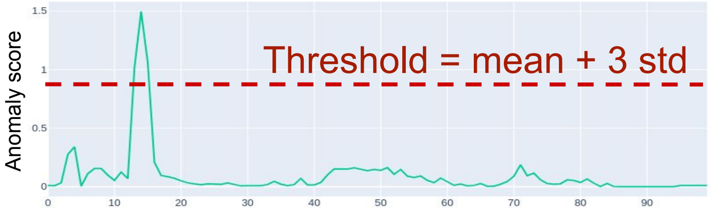
Preprocessing & Sequencing → **Training & Calibration**



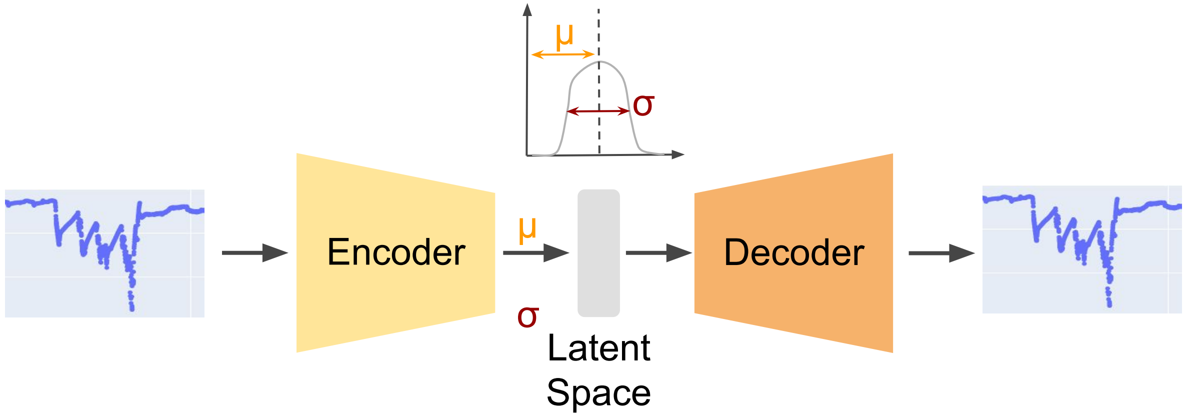


- Evenly sample time series,
- Fully populated,
- Generate 3 distincts datasets of fixed size windows.

- MCHVAE,
- Calibration done using 3-sigma method.

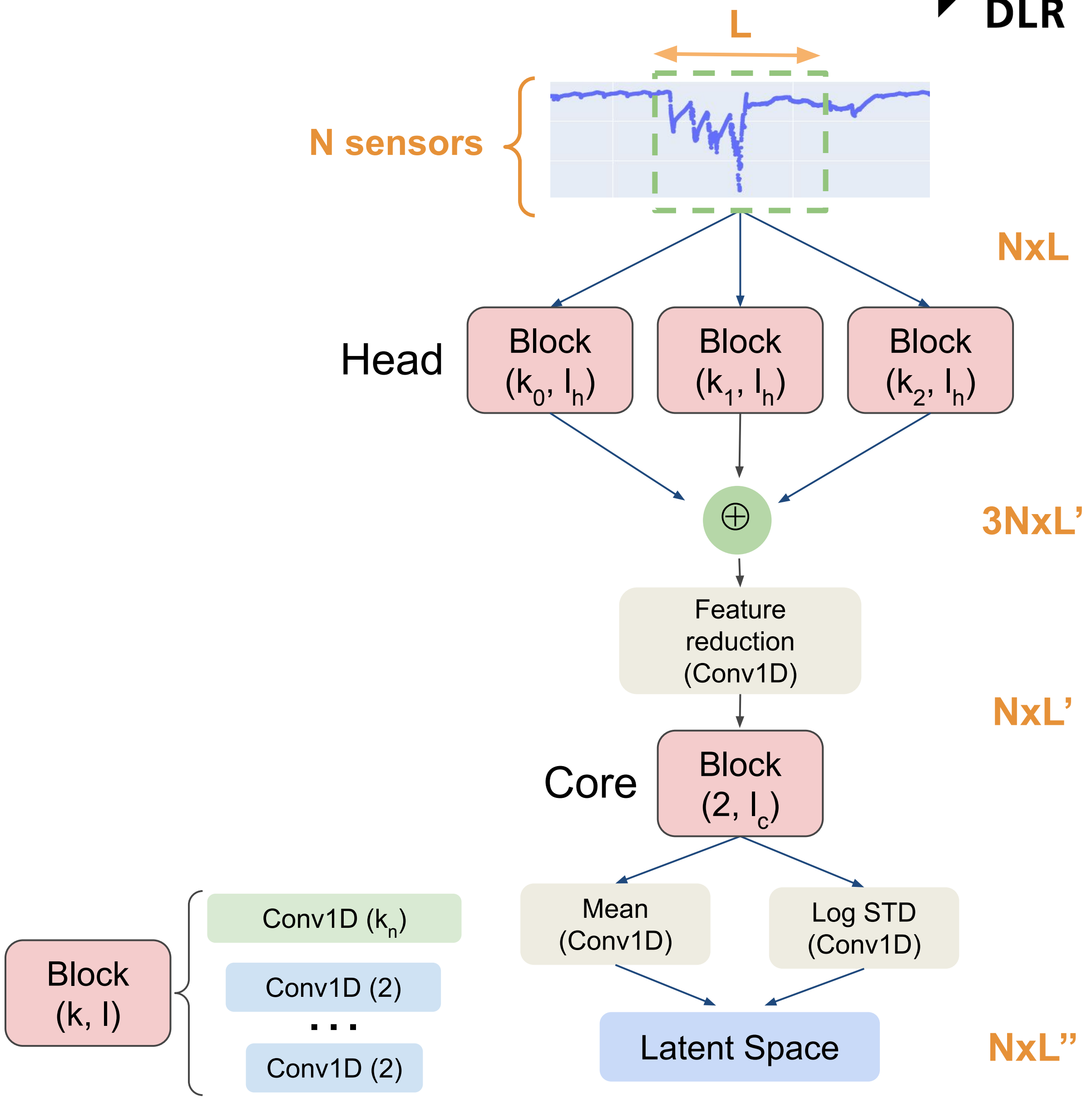


Multi-Convolution Head Variational AutoEncoder (MCHVAE)



Key features:

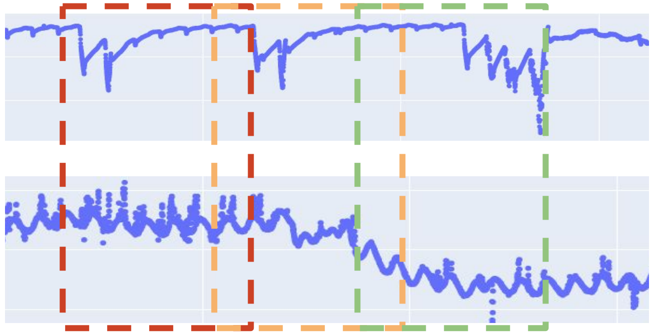
- Fully convolution based network.
- Specific time scales are extracted by the head blocks.
- Core block:
  - reduce the embedding dimension further,
  - learn dependencies between the different time scale embedding.



Code:

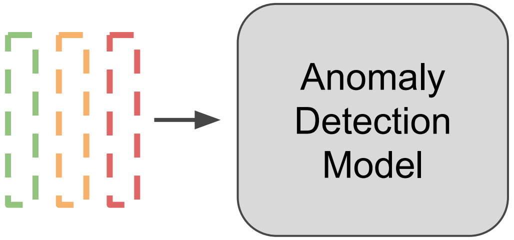
Workflow: Prediction & Explanation

Preprocessing & Sequencing

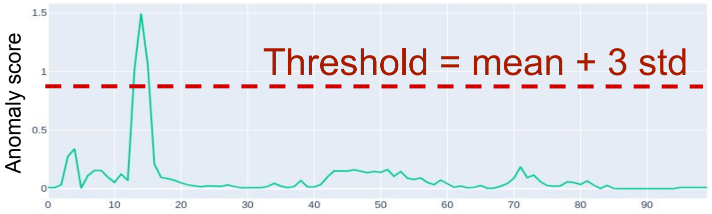


- Evenly sample time series,
- Fully populated,
- Generate 3 distincts datasets of fixed size windows.

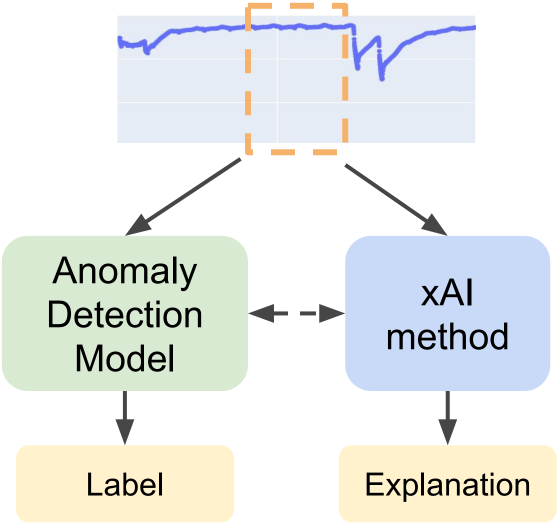
Training & Calibration



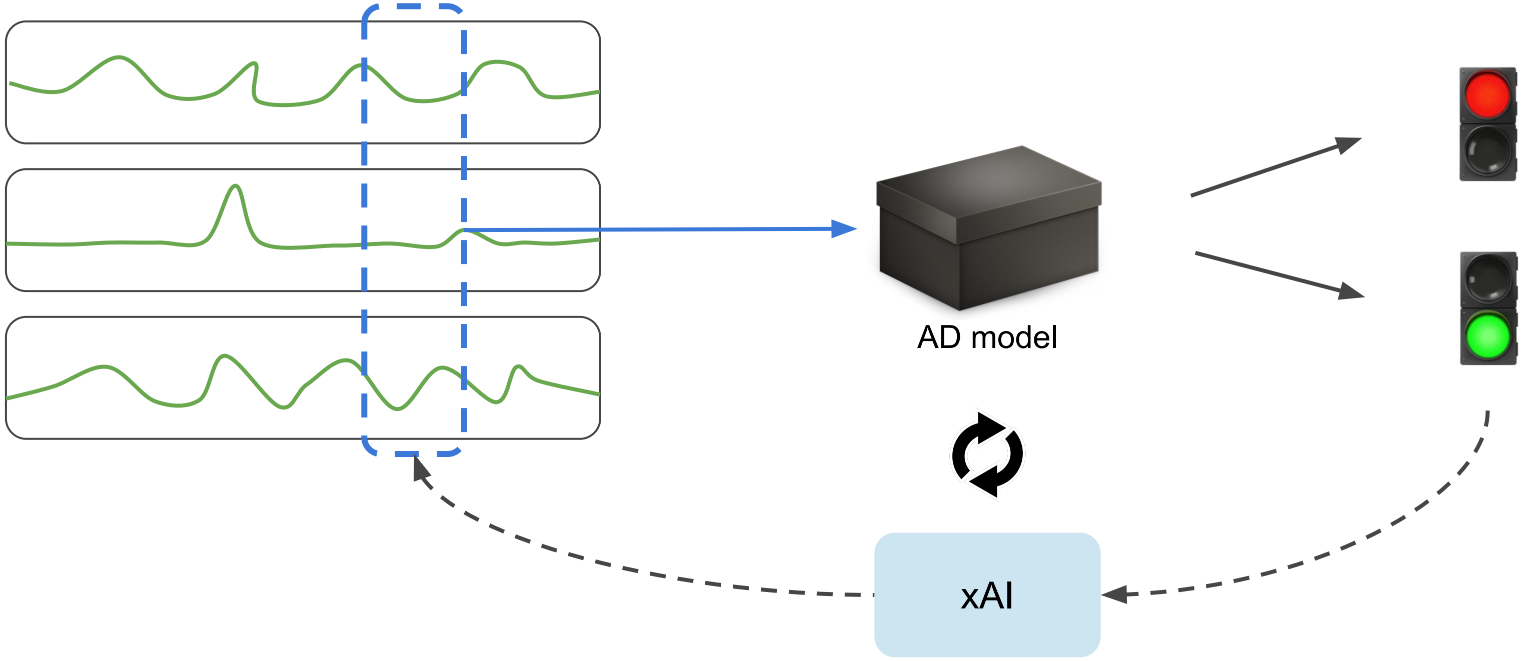
- MCHVAE,
- Calibration done using 3-sigma method.



Prediction & Explanation



Isolation of Failing Sensors via Explainable AI (xAI)



xAI methods:

- WindowSHAP from Nayebi *et al.*
- CounterFactual Explanation (CFE):
  - using normal sequences in the available data.
  - using VAE reconstruction.

Nayebi et al. (2023), *WindowSHAP: An Efficient Framework for Explaining Time-series Classifiers based on Shapley Values*, arXiv:2211.06507.

Results

Cottbus Brayton (CoBra) High-Temperature Heat Pump

CoBra Project:

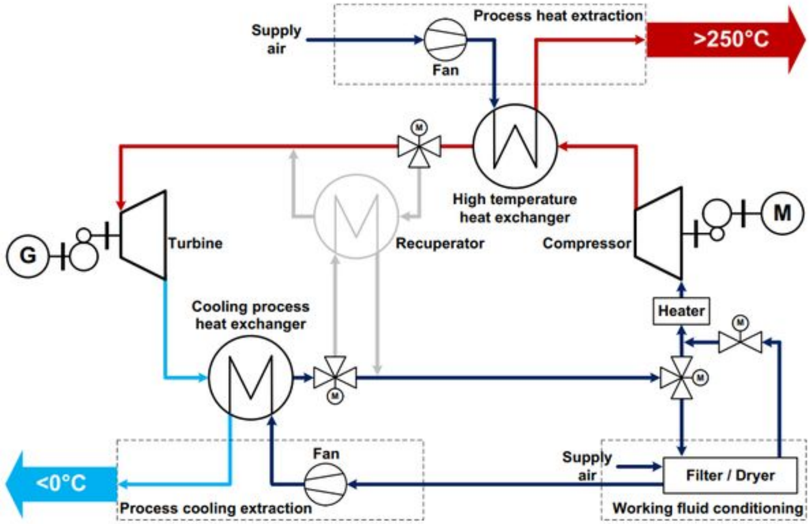
- Objective: CO2-neutral heat generator (150-500°C)
- Target: energy intensive industries (chemicals, iron, steel, paper, …).

Dataset:

- 6 measuring campaigns,
- 115 variables, ~183k observations (1Hz) in total,
- 4 recorded emergency shutdowns.

Our Objective:

- Can the parts of the system leading to shutdowns be identified?
- Reduce downtime and stress on the system by detecting these anomalies earlier.



CoBra Results

Anomaly Detection:

- How well can we detect shutdown?
- Top 4 models found all shutdown,
- Calibration: 3-sigma method.

	F1	**PR-AUC**	**PR-VUS**
LUNAR	**0.833**	0.601	0.121
OCSVM	0.818	**0.914**	**0.470**
MCHVAE	0.421	0.154	0.031
GANF	0.356	0.150	0.146
CBLOF	0.203	0.041	0.031
ECOD	0.154	0.031	0.018
COPOD	0.129	0.043	0.026
MSTVAE	0.129	0.058	0.012
IFOREST	0.080	0.018	0.111
ANOGAN	0.008	0.002	0.011

Alternative models from Zhao, Y. *et. al* (2019). PyOD: A Python toolbox for scalable outlier detection. *Journal of Machine Learning Research, 20*(96), 1-7.

CoBra Results

Anomaly Detection:

- How well can we detect shutdown?
- Top 4 models found all shutdown,
- Calibration: 3-sigma method.

	F1	PR-AUC	PR-VUS
LUNAR	**0.833**	0.601	0.121
OCSVM	0.818	**0.914**	**0.470**
MCHVAE	0.421	0.154	0.031
GANF	0.356	0.150	0.146
CBLOF	0.203	0.041	0.031
ECOD	0.154	0.031	0.018
COPOD	0.129	0.043	0.026
MSTVAE	0.129	0.058	0.012
IFOREST	0.080	0.018	0.111
ANOGAN	0.008	0.002	0.011

Explanation:

- Average rank in the xAI explanation of the variable responsible for the shutdown:

	WSHAP	CFE	CFE reconstruction
LUNAR	9	34	-
OCSVM	2.5	27	-
MCHVAE	2.5	22	2.5

CoBra Results

Anomaly Detection:

- How well can we detect shutdown?
- Top 4 models found all shutdown,
- Calibration: 3-sigma method.

	F1	PR-AUC	PR-VUS
LUNAR	**0.833**	0.601	0.121
OCSVM	0.818	**0.914**	**0.470**
MCHVAE	0.421	0.154	0.031
GANF	0.356	0.150	0.146
CBLOF	0.203	0.041	0.031
ECOD	0.154	0.031	0.018
COPOD	0.129	0.043	0.026
MSTVAE	0.129	0.058	0.012
IFOREST	0.080	0.018	0.111
ANOGAN	0.008	0.002	0.011

Explanation:

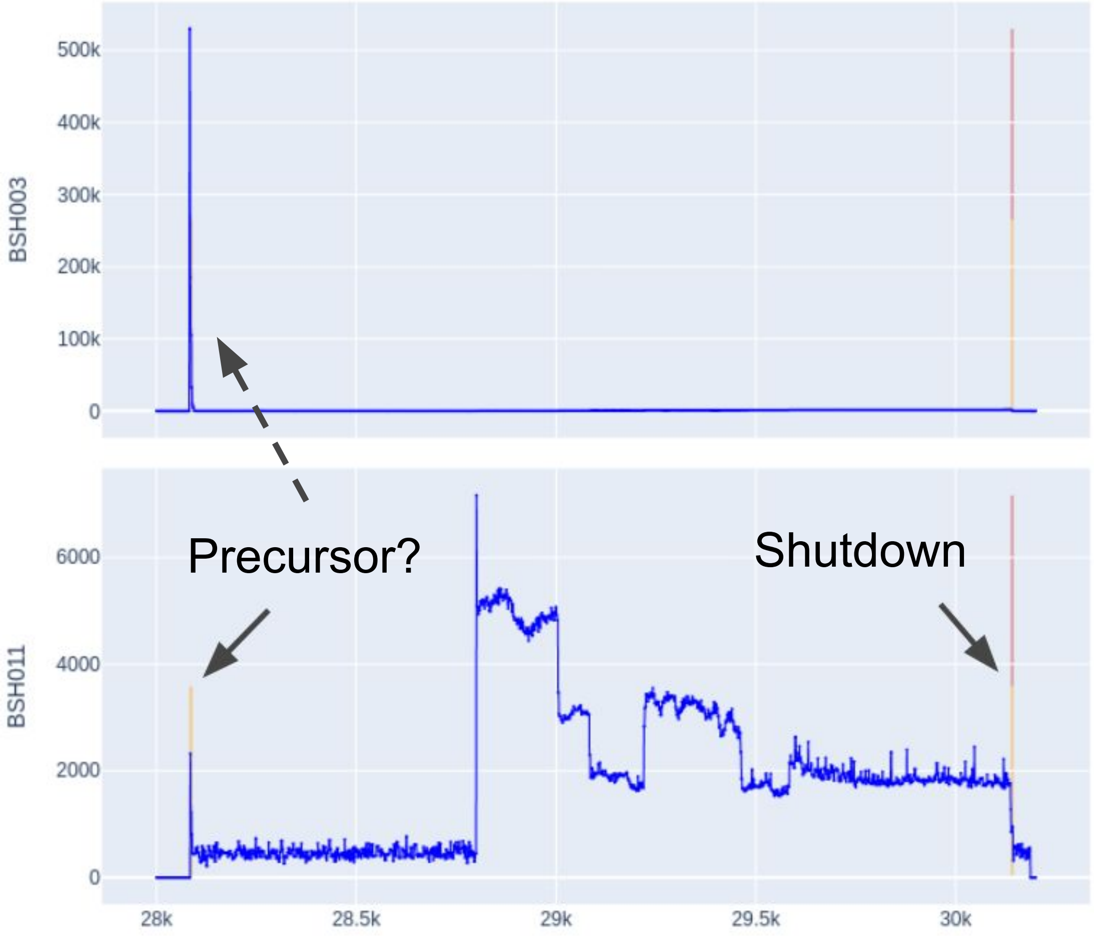
- Average rank in the xAI explanation of the variable responsible for the shutdown:

	WSHAP	CFE	CFE reconstruction
LUNAR	9	34	-
OCSVM	2.5	27	-
MCHVAE	2.5	22	2.5
Run time per explanation:	~21min	~10s	<< 1s

Outlook

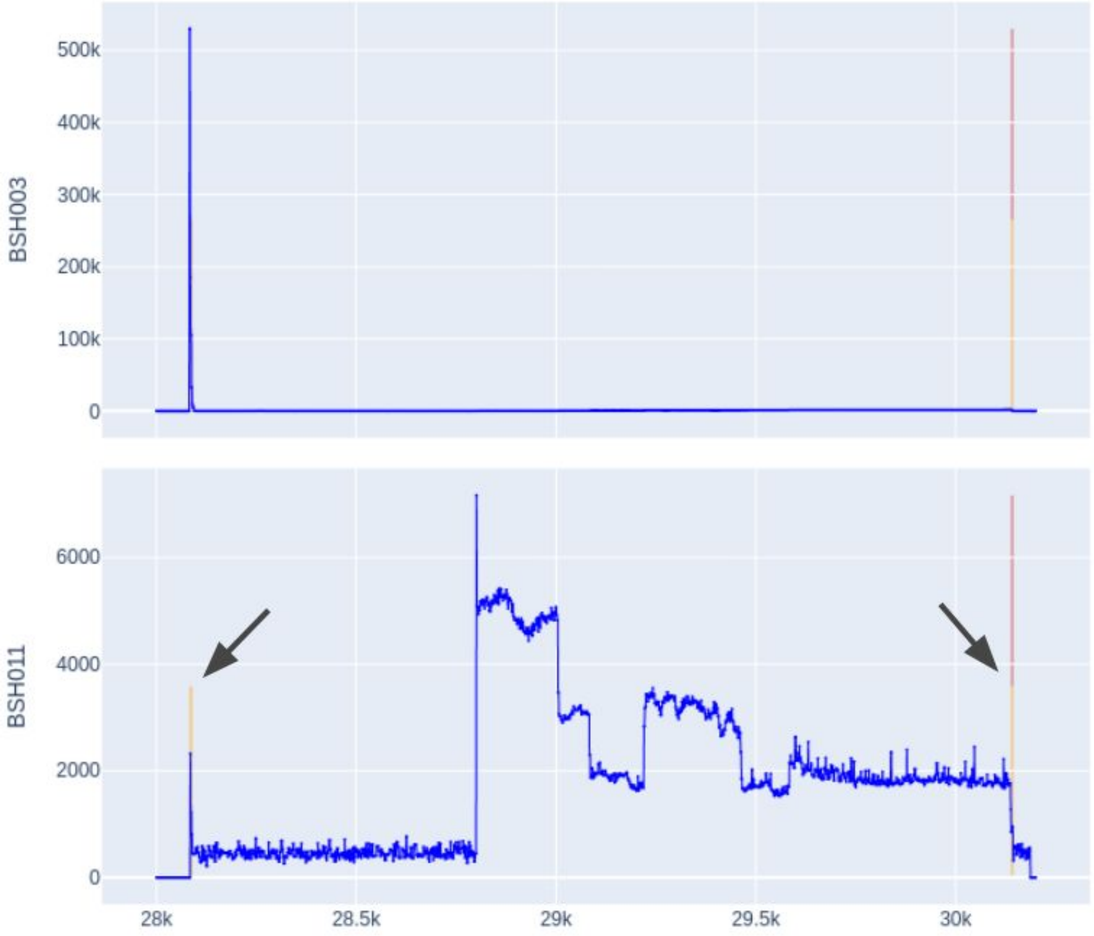
- Investigate the potential precursors

Variable at the origin of the shutdown



Outlook

- Investigate the potential precursors
- Improve counterfactual by leveraging the data and problem nature.



Thank you for your attention

https://www.canva.com/photos/MAFmYPIOP-k/

Impressum

Thema: Improving Fault Diagnosis in Complex Systems: Investigating xAI Techniques for Abnormal Sensor Isolation

Datum: 27.11.2025

Autor: Dr. Baptiste Lambert

baptiste.lambert@dlr.de

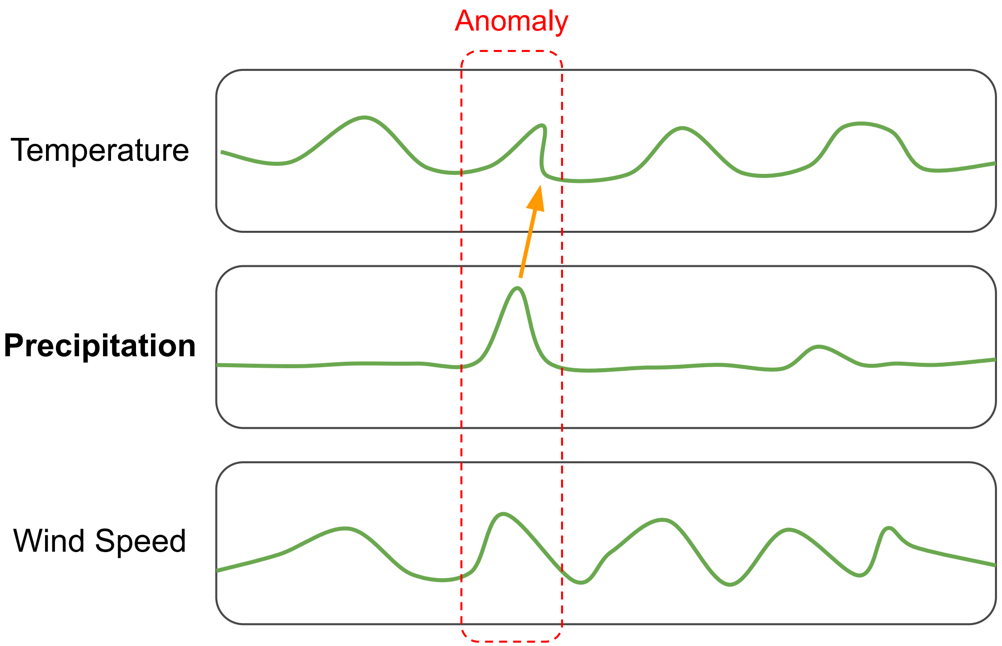
Institut: DLR Institute of Data Science

Bildquellen: All pictures „DLR (CC BY-NC-ND 3.0)“, unless otherwise stated

https://www.canva.com/photos/MAFmYPIOP-k/

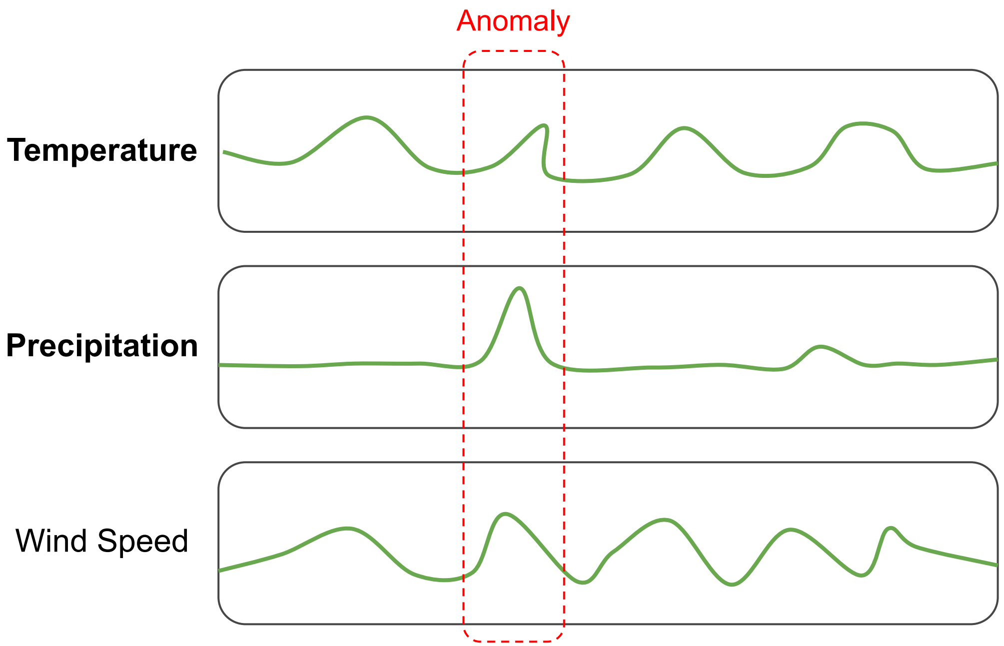
Root Cause Analysis vs Explainable AI (xAI)

Root Cause Analysis



- Find the <u>underlying reasons</u> of the anomaly,
- Root cause(s) of the anomaly can be outside the window detected by the AD model.

AD-xAI Analysis



- Find the variables that <u>influence the AD model</u> prediction at the anomaly,
- Relationship between variables is not considered.

Benchmark Results

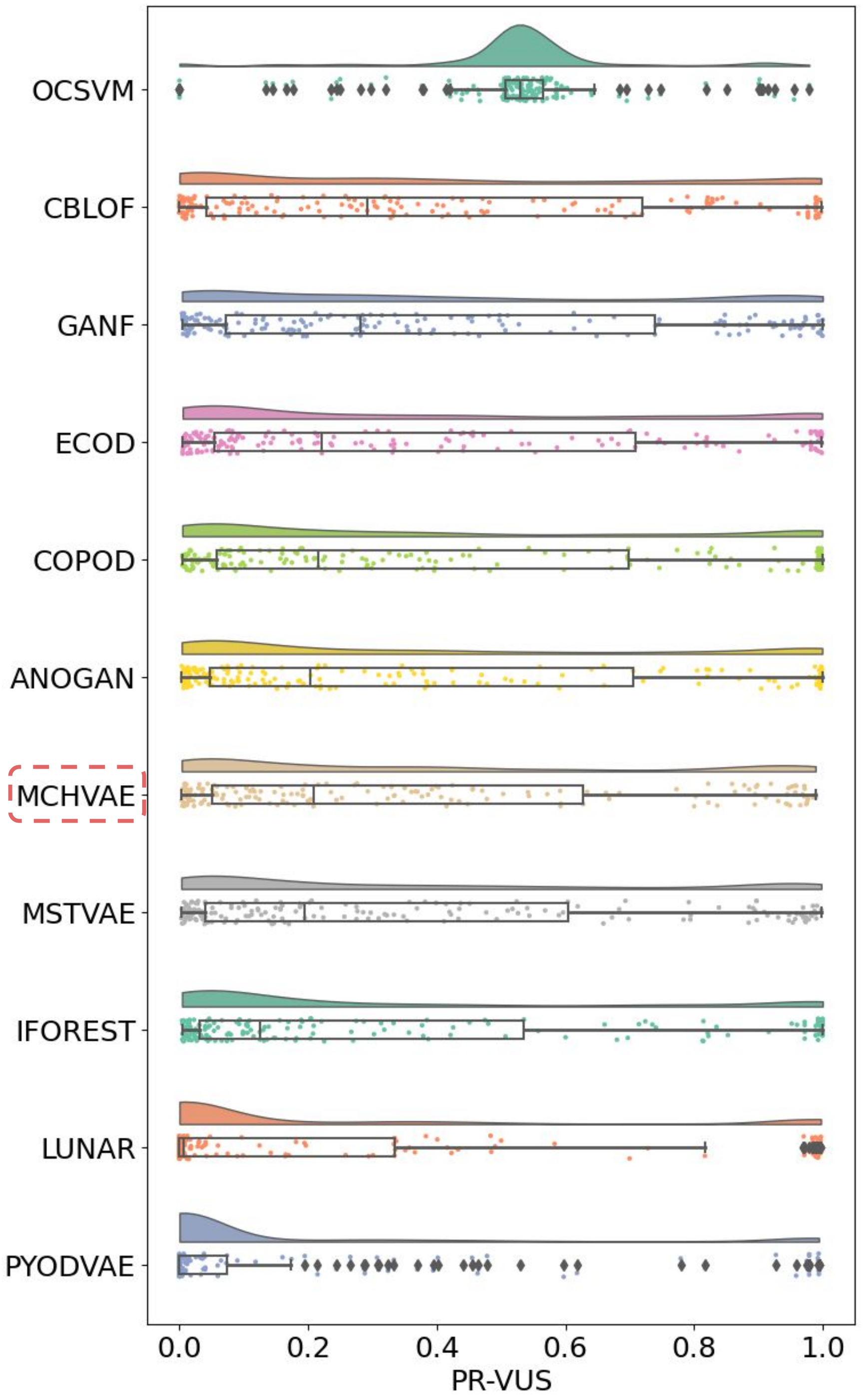
Motivations:

- Get a better assessment of the MCHVAE accuracy.
- Hyperparameter tuning.

Dataset:

- Data and protocol from Lui Q. *et al*
- 200 high quality data from 40 public datasets:
  - 180 for evaluation.
  - 20 for tuning.

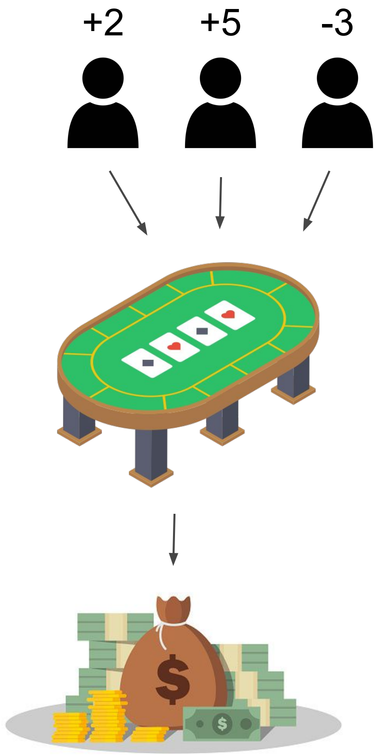
	Type	F1	PR-AUC	PR-VUS	Failing Rate
OCSVM	classic	0.01	0.08	**0.53**	0.0%
CBLOF	classic	**0.21**	0.34	0.39	6.4%
GANF	deep	**0.21**	0.33	0.39	0.0%
ECOD	classic	0.18	0.33	0.38	0.0%
COPOD	classic	0.19	0.33	0.36	0.0%
AnoGAN	deep	0.16	0.32	0.36	0.0%
MCHVAE	deep	0.18	**0.35**	0.36	0.0%
MSTVAE	deep	0.19	0.34	0.35	8.0%
iForest	classic	0.13	0.28	0.31	0.0%
LUNAR	deep	0.13	0.22	0.22	45.1%
DenseVAE	deep	0.11	0.16	0.16	54.3%

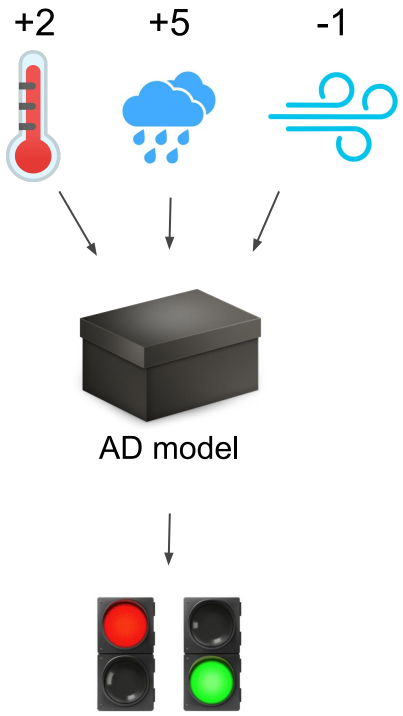


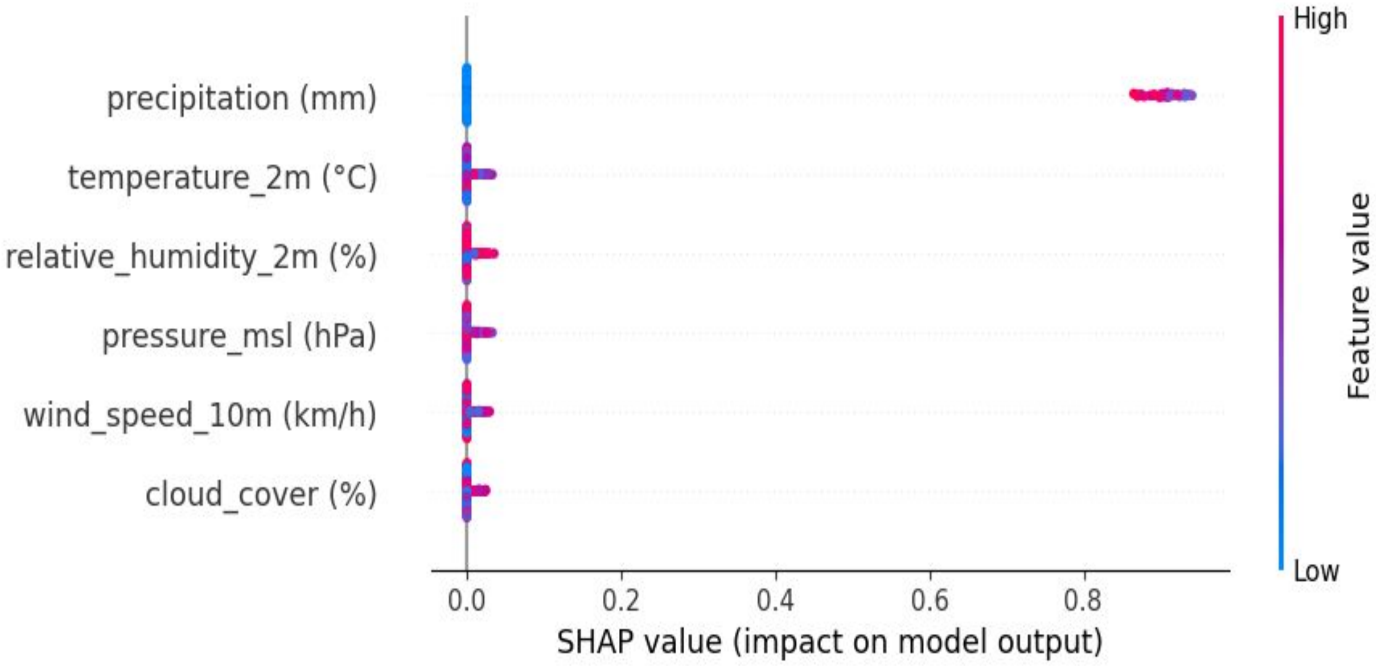
Liu, Q., Paparrizos, J.: The elephant in the room: Towards a reliable time-series anomaly detection benchmark. In: NeurIPS 2024 (2024)
Zhao, Y. *et. al* (2019). PyOD: A Python toolbox for scalable outlier detection. *Journal of Machine Learning Research, 20*(96), 1-7.

xAI: SHapley Additive exPlanations (SHAP)

- Explanation based on Shapley values (game theory) and feature permutation.







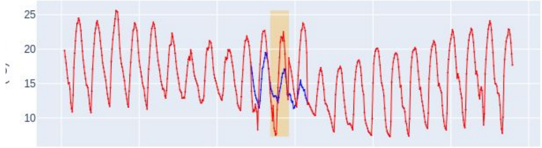
Nayebi et al. (2023), *WindowSHAP: An Efficient Framework for Explaining Time-series Classifiers based on Shapley Values*, arXiv:2211.06507.

xAI: Counterfactual Explanation (CFE)

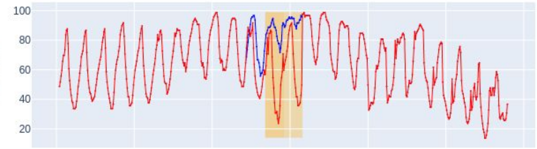
Minimization of:

$$OF_1(x, x_{faulty}) = \underbrace{\|x - x_{faulty}\|_2^2}_{\text{Minimal changes to original}}$$

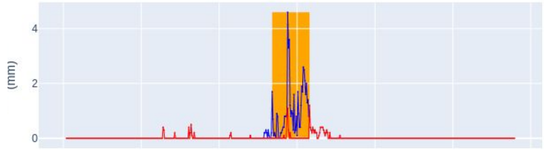
$$OF_2(x, x_{faulty}) = \underbrace{AD_{score}(x)}_{\text{Anomaly score}} + \lambda \underbrace{AD_{label}(x)}_{\text{Abnormal penalization}}$$



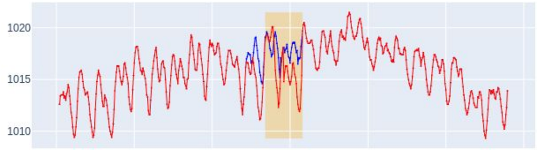
$|V_{1,cf} - V_{1,org}| = 0.3$



$|V_{2,cf} - V_{2,org}| = 0.5$



$|V_{3,cf} - V_{3,org}| = 1.1$



$|V_{4,cf} - V_{4,org}| = 0.1$

Explanation:

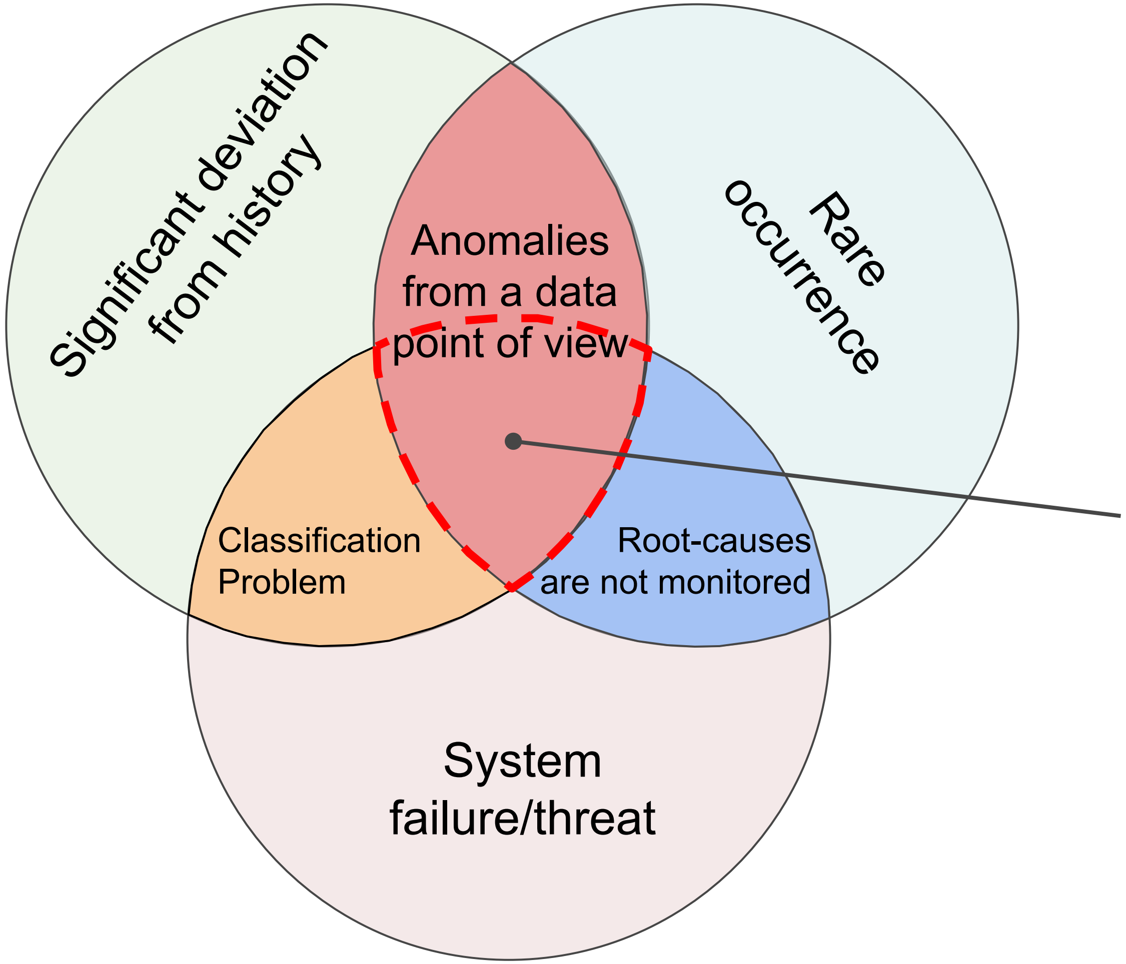
$p(V_1) = 15\%$

$p(V_2) = 25\%$

$\mathbf{p(V_3) = 55\%}$

$p(V_4) = 5\%$

Anomaly?



Anomalies from a operation point of view and accurately detectable only by AD models.