

Towards Understanding Alerts raised by Unsupervised Network Intrusion Detection Systems

THCON 2024 - April 5th, 2024 - Toulouse

Maxime Lanvin
Frédéric Majorczyk

Pierre-François Gimenez
Ludovic Mé

Yufei Han
Éric Totel

Team



Inria



Introduction on NIDS & motivation

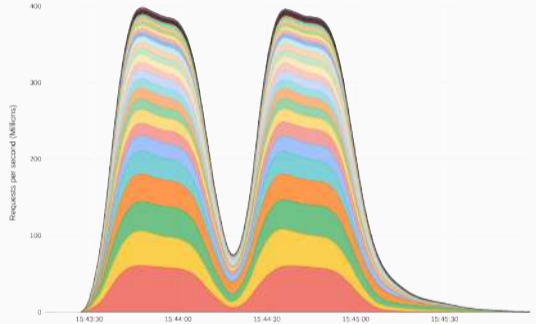
Introduction

Context

Many cyber attacks are conducted with different level of sophistication.

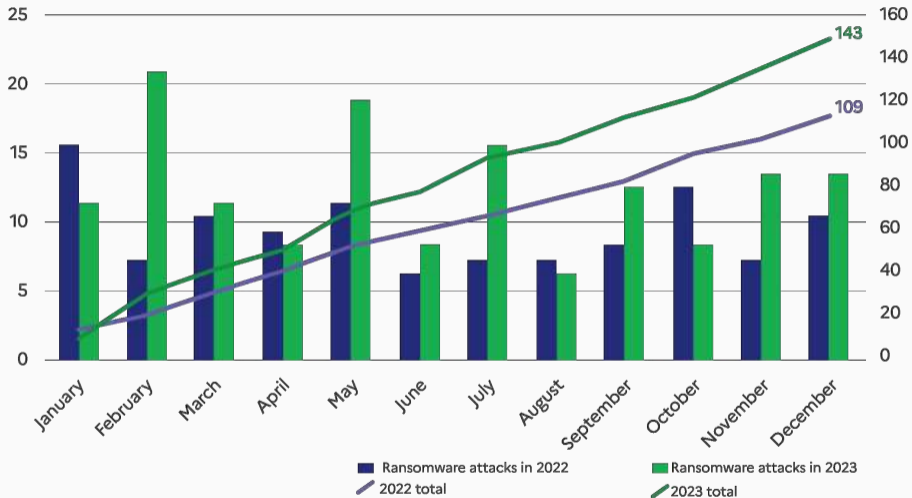
- Low level attacks like scanning and brute forcing performed by bots
- Massive and visible attacks like (D)DoS
- Complex and stealthy attacks (APT)

Requests per second by Metropolitan Area



Example of the largest DDoS conducted against Google Cloud infrastructure in September 2023, **source** : Google Cloud

Introduction



Comparison of ransomware attacks reported to ANSSI in 2022 and 2023, +30% increase, **source** : Cyber Threat Overview 2023, ANSSI

Protection mechanisms

- Password policy, system updates, threat monitoring, firewall filtering, ...
- User awareness of good/bad practices

Intrusion Detection

- Intrusion Detection Systems (**IDS**) offer a way to detect attacks and let operators react according to the alerts. Two possible data sources : system or **network** logs
- We focus in this work on Network IDS (**NIDS**)

Paradigms

- **Signature**-based : detection of signature associated with known attacks
- **Anomaly**-based : detection of deviation from a normal behavior

Comparison of the two paradigms

Signature based alert

Supervision

- ts: 2023-01-19T14:02:46.143Z

- dst_address: "192.168.101.3"
- dst_port: 47426
- src_address: "192.168.101.26"
- src_port: 1389

- signature: "ET ATTACK_RESPONSE Possible
CVE-2021-44228 Payload via LDAPv3
Response"

- category: "Attempted Administrator
Privilege Gain"

- severity: 1

- CVE: CVE_2021_4422



Anomaly based alert

Supervision

- ts: 2023-01-19T14:02:46.143Z

- dst_address: "192.168.101.3"
- dst_port: 47426
- src_address: "192.168.101.26"
- src_port: 1389



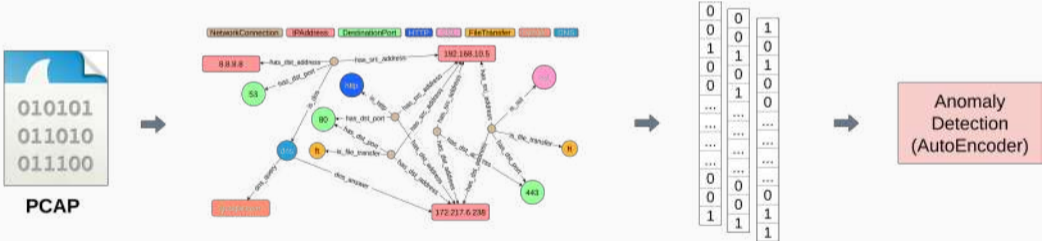
Good Luck ! Enjoy !



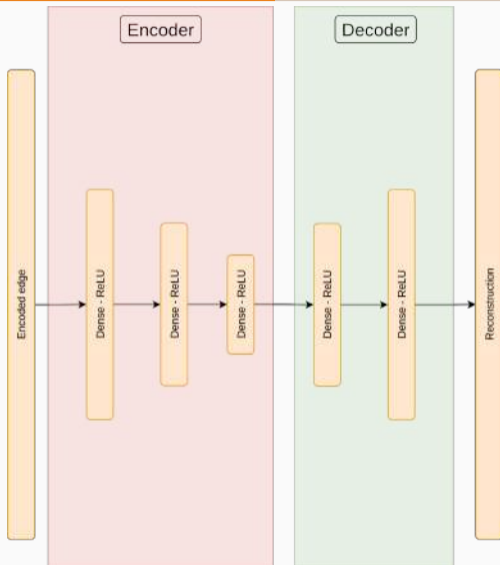
1. Introduction on NIDS & motivation
2. AE-pvalues
3. Benchmark XAI techniques
4. Using explanations on CICIDS2017 dataset
5. Conclusion

Unsupervised anomaly detection

Common Machine Learning pipeline of anomaly-based NIDS



Unsupervised anomaly detection : Autoencoder (AE)



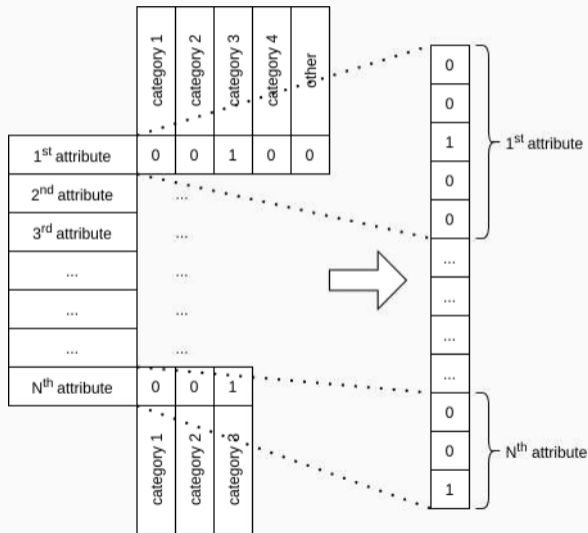
Learning

Minimisation of the reconstruction error between the input vector and its reconstructed version.

Detection

Raise an alert when the reconstruction error is above a threshold.

One Hot Encoding - Meaning of the vectors



Definition

In our context, the **explanations** are an **ordered list of the network attributes** ranked from the most abnormal to the least abnormal.

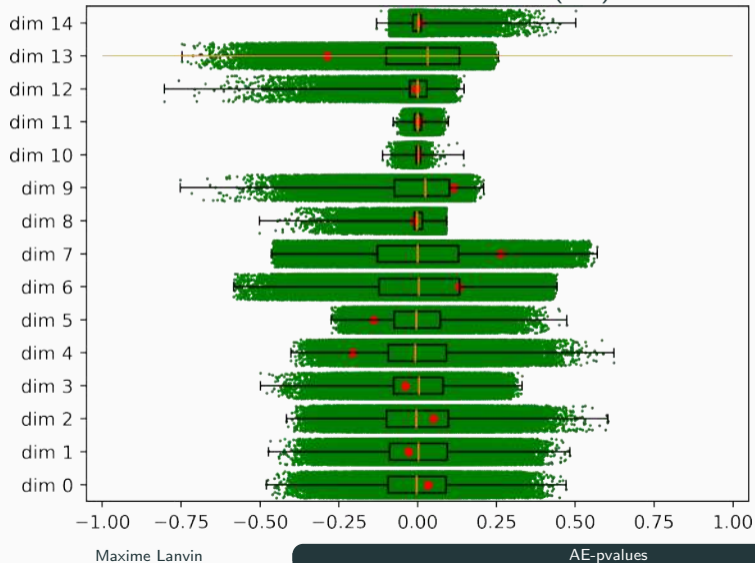
Example

[`connection_duration`, `user_agent`, ..., `http_method`, ..., `destination_port`]

AE-pvalues

XAI techniques for Autoencoders

Reconstruction error distribution (AE)



Possible methods

- Ranking by **absolute** values
- Ranking by **shapley** values
- Ranking by **p-values**

Obervation

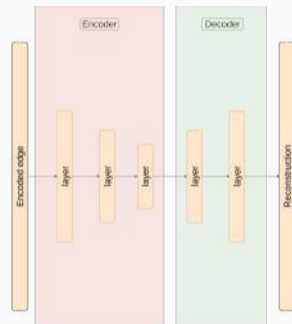
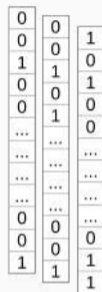
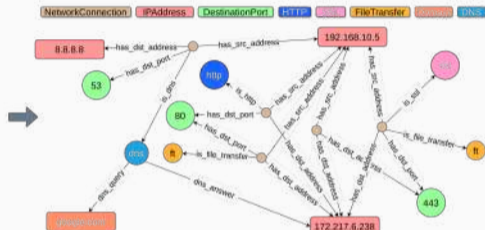
The highest reconstruction error is not always an indication of the most abnormal dimension.

Benchmark XAI techniques

Sec2graph : An anomaly detection NIDS



PCAP



Autoencoder

This NIDS respects the assumptions

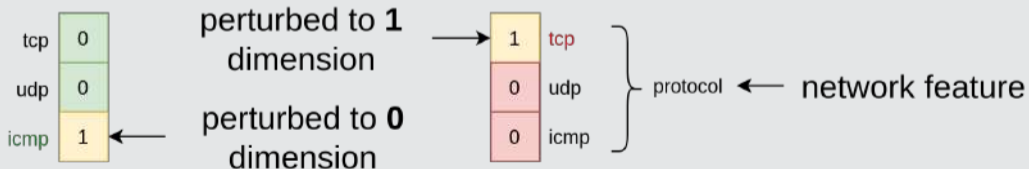
- **Unsupervised** : no attacks used for the training
- **Anomaly-based NIDS** : detect drift from normal behaviours using an **AE**

Methodology for the comparison

Methods

- Inject noise in a known network characteristic of vectors
- Assess ability of XAI methods to find the noisy network characteristic

Exemple of noise insertion in the protocol characteristic



Multiple correct explanations

Statement : $1 + 1 = 0$

What is the right explanation for the mistake?

Multiple correct explanations

Statement : $1 + 1 = 0$

What is the right explanation for the mistake?

- 0 should be 2

Multiple correct explanations

Statement : $1 + 1 = 0$

What is the right explanation for the mistake?

- 0 should be 2
- + should be -

Multiple correct explanations

Statement : $1 + 1 = 0$

What is the right explanation for the mistake?

- 0 should be 2
- + should be -
- 1 should be -1

Multiple correct explanations

Statement : $1 + 1 = 0$

What is the right explanation for the mistake?

- 0 should be 2
- + should be -
- 1 should be -1
- = should be >
- "(mod 2)" is missing
- "is false" is missing

Multiple correct explanations

Statement : $1 + 1 = 0$

What is the right explanation for the mistake?

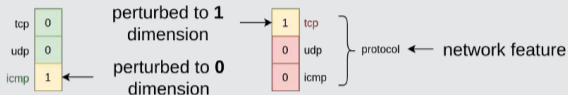
- 0 should be 2
- + should be -
- 1 should be -1
- = should be >
- "(mod 2)" is missing
- "is false" is missing

For network features : correlated attributes

`http_status_code = 200` is equivalent to `http_status_msg = OK`

Benchmark results

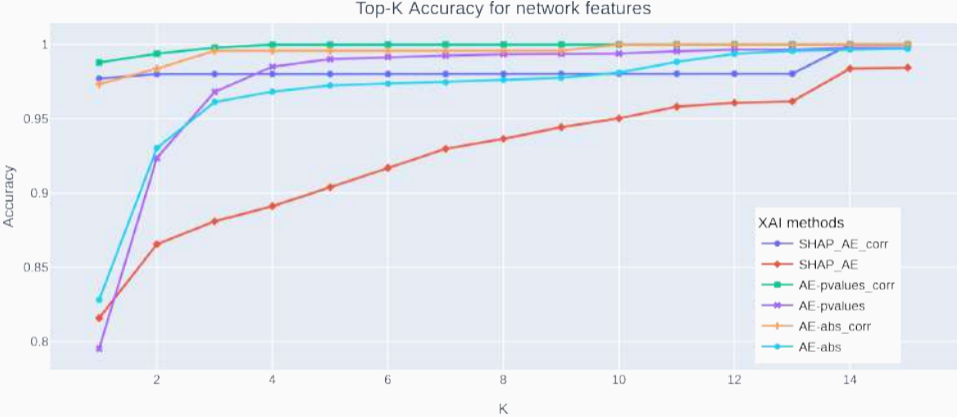
Vocabulary reminder



explaining method	Mean rank of the perturbed to 0 dimension	Mean rank of the perturbed to 1 dimension	Mean rank of the network feature ↓
AE-pvalues_corr	2.96	1.63	1.02
AE-abs_corr	3.89	1.61	1.07
SHAP_AE_corr	4.71	4.44	1.26
Random_corr	5.68	16.3	1.85
AE-pvalues	4.61	3.07	1.39
AE-abs	5.78	4.78	1.49
SHAP_AE	18.96	7.18	2.15
Random	26.93	27.13	7.8

Table of mean ranks of the perturbed to 0 or 1 dimensions, and the network feature where the noise is inserted.

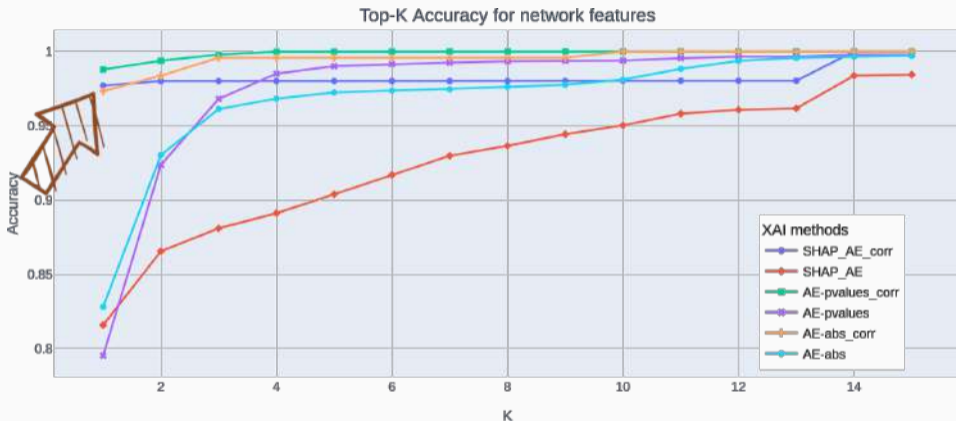
Benchmark results



Top-K accuracy

Proportion of samples for which the right explanation is among the Top-K explanations.

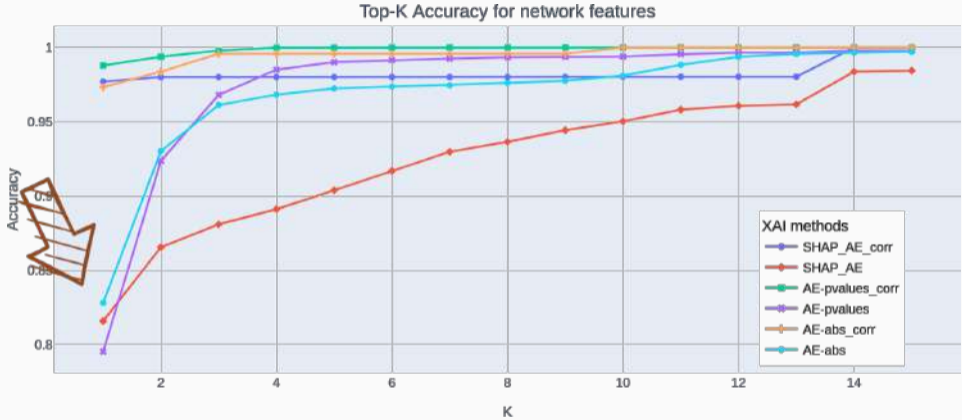
Benchmark results



Top-K accuracy

Proportion of samples for which the right explanation is among the Top-K explanations.

Benchmark results



Top-K accuracy

Proportion of samples for which the right explanation is among the Top-K explanations.

Method	Processing time per sample
SHAP_AE	28 s
AE-pvalues	1.9 ms
AE-abs	1.0 ms

Processing time for one sample for each explaining method

Conclusion


AE-pvalues is approximately 10,000 faster than the SHAP_AE method.

Comparison of the two paradigms

Signature based alert

Supervision

- ts: 2023-01-19T14:02:46.143Z
- dst_address: "192.168.101.3"
- dst_port: 47426
- src_address: "192.168.101.26"
- src_port: 1389
- signature: "ET ATTACK_RESPONSE Possible CVE-2021-44228 Payload via LDAPv3 Response"
- category: "Attempted Administrator Privilege Gain"
- severity: 1
- CVE: CVE_2021_4422




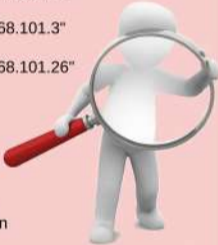
Anomaly based alert

Supervision

- ts: 2023-01-19T14:02:46.143Z
- dst_address: "192.168.101.3"
- dst_port: 47426
- src_address: "192.168.101.26"
- src_port: 1389

Abnormal features:

- connection_duration
- user_agent
- http_method
- http_trans_depth
- http_status_code
- ...

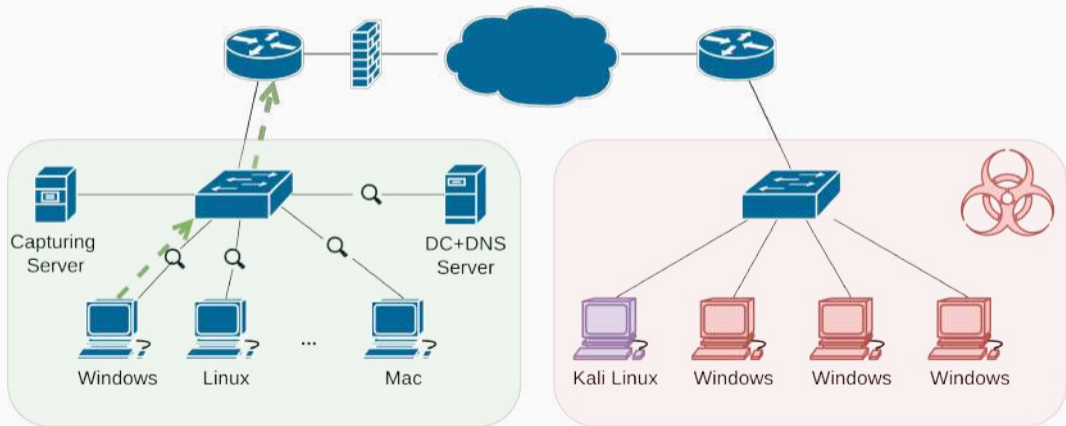


Using explanations on CICIDS2017 dataset

The dataset : CICIDS2017

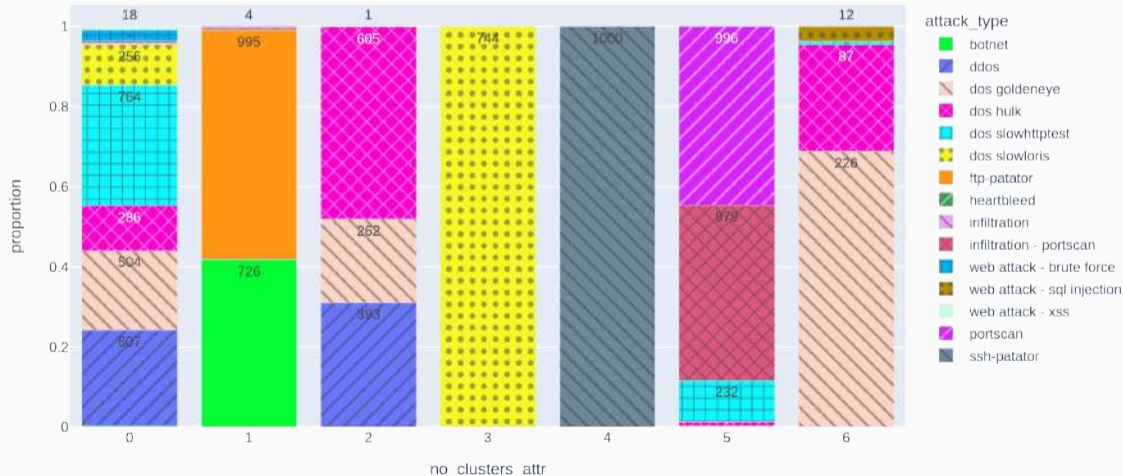
Dataset features

Dataset : **CICIDS17** : 5 days of network traffic, ~ 50 GB, ~ 15 machines



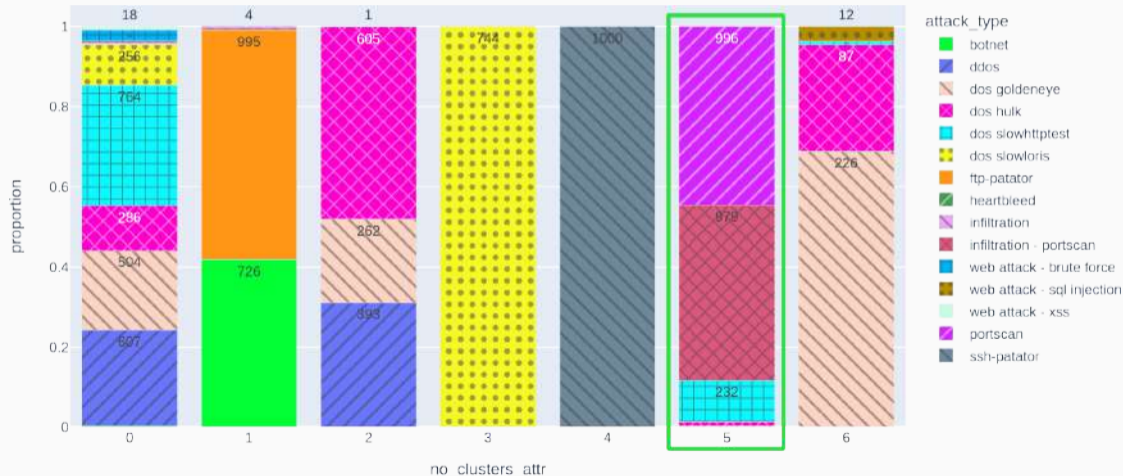
Principle

Clustering of the alerts based on the explanations



Principle

Clustering of the alerts based on the explanations



Applications - Feature contribution to attack types

	http_status_depth	http_status_msg	address_history	port_value	port_value	service	http_status_code	http_method	ua_browser	ua_os	duration	filetransfer_mime_type	conn_state	weird_name	weird_peer	weird_addi	http_info_code	ssh_host_key_msg	ssh_host_key_alg	ssh_host_key	ssh_cipher_algo	ssh_client		
botnet	0.1	0.0	1.8	0.0	20.0	2.4	17.4	0.0	17.5	19.2	18.0	1.8	0.0	0.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0		
heartbleed	0.0	0.0	20.0	20.0	20.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0		
infiltration	0.0	0.0	20.0	2.9	17.1	20.0	0.0	0.0	0.0	0.0	0.0	14.3	17.1	0.0	2.9	2.9	0.0	0.0	0.0	0.0	0.0	0.0		
infiltration - portscan	0.0	0.0	19.9	19.8	19.8	0.2	0.0	0.0	0.0	0.0	0.0	19.8	19.9	0.0	0.1	0.1	0.1	0.0	0.0	0.0	0.0	0.0		
portscan	0.0	0.0	20.0	20.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	20.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0		
ddos	20.0	20.0	7.0	0.0	0.0	0.0	0.3	20.0	20.0	0.0	0.0	0.0	0.0	0.0	12.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0		
dos goldeneye	18.4	14.8	11.2	0.2	0.8	0.3	1.0	18.3	15.3	7.6	8.9	1.1	1.5	0.2	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0		
dos hulk	13.5	14.0	1.9	0.5	3.1	0.0	10.9	13.5	15.9	6.2	6.2	1.0	0.5	5.5	3.2	2.9	1.0	0.0	0.0	0.0	0.0	0.0		
dos slowhttptest	0.4	7.2	5.0	5.1	2.5	0.0	1.7	4.1	3.5	0.1	0.1	12.4	4.6	8.2	13.2	13.2	13.2	1.6	1.6	0.0	0.0	0.0		
dos slowloris	4.3	16.1	0.0	0.8	0.0	0.0	16.9	20.0	3.0	3.1	3.1	0.0	0.0	1.3	0.0	0.0	0.0	15.7	15.7	0.0	0.0	0.0		
ftp-patator	0.0	0.0	20.0	0.1	19.9	20.0	0.0	0.0	0.0	0.0	0.0	20.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0		
ssh-patator	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	20.0	19.9	19.9	19.9	20.0
web attack - brute force	20.0	19.7	0.0	0.0	0.0	0.0	0.3	0.3	0.5	19.7	19.7	0.0	0.0	0.0	0.0	0.0	0.0	19.7	0.0	0.0	0.0	0.0	0.0	
web attack - sql injection	0.0	0.0	3.1	20.0	0.0	20.0	0.0	0.0	0.0	20.0	20.0	0.0	0.0	16.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
web attack - xss	0.0	18.9	0.0	20.0	0.0	0.0	0.0	0.0	0.0	20.0	20.0	0.0	0.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	

Applications - Feature contribution to attack types

	http_status_depth	http_status_msg	address_history	port_value	port_value	service	http_status_code	http_method	ua_browser	ua_os	duration	filetransfer_mime_type	conn_state	weird_type	weird_name	weird_peer	http_info_addi	http_info_code	ssh_host_key_msg	ssh_key_alg	ssh_host_key	ssh_cipher_algo	ssh_client	
botnet	0.1	0.0	1.8	0.0	20.0	2.4	17.4	0.0	17.5	19.2	18.0	1.8	0.0	0.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
heartbleed	0.0	0.0	20.0	20.0	20.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
infiltration	0.0	0.0	20.0	2.9	17.1	20.0	0.0	0.0	0.0	0.0	0.0	14.3	17.1	0.0	2.9	2.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
infiltration - portscan	0.0	0.0	19.9	19.8	19.8	0.2	0.0	0.0	0.0	0.0	0.0	19.8	19.9	0.0	0.1	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	
portscan	0.0	0.0	20.0	20.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	20.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
ddos	20.0	20.0	7.0	0.0	0.0	0.0	0.3	20.0	20.0	0.0	0.0	0.0	0.0	0.0	12.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
dos goldeneye	18.4	14.8	11.2	0.2	0.8	0.3	1.0	18.3	15.3	7.6	8.9	1.1	1.5	0.2	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
dos hulk	13.5	14.0	1.9	0.5	3.1	0.0	10.9	13.5	15.9	6.2	6.2	1.0	0.5	5.5	3.2	2.9	1.0	0.0	0.0	0.0	0.0	0.0	0.0	
dos slowhttptest	0.4	7.2	5.0	5.1	2.5	0.0	1.7	4.1	3.5	0.1	0.1	12.4	4.6	8.2	13.2	13.2	13.2	1.6	1.6	0.0	0.0	0.0	0.0	
dos slowloris	4.3	16.1	0.0	0.8	0.0	0.0	16.9	20.0	3.0	3.1	3.1	0.0	0.0	1.3	0.0	0.0	0.0	15.7	15.7	0.0	0.0	0.0	0.0	
ftp-patator	0.0	0.0	20.0	0.1	19.9	20.0	0.0	0.0	0.0	0.0	0.0	20.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
ssh-patator	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	20.0	19.9	19.9	19.9	20.0
web attack - brute force	20.0	19.7	0.0	0.0	0.0	0.0	0.3	0.3	0.5	19.7	19.7	0.0	0.0	0.0	0.0	0.0	0.0	19.7	0.0	0.0	0.0	0.0	0.0	
web attack - sql injection	0.0	0.0	3.1	20.0	0.0	20.0	0.0	0.0	0.0	20.0	20.0	0.0	0.0	16.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
web attack - xss	0.0	18.9	0.0	20.0	0.0	0.0	0.0	0.0	0.0	20.0	20.0	0.0	0.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	

Applications - Feature contribution to attack types

	http_status_depth	http_status_msg	address_history	port_value	port_value	service	http_status_code	http_method	ua_browser	ua_os	duration	filetransfer_mime_type	conn_state	weird_name	weird_peer	http_info_addi	http_info_code	ssh_host_key_msg	ssh_host_key_alg	ssh_host_key	ssh_cipher_algo	ssh_client		
botnet	0.1	0.0	1.8	0.0	20.0	2.4	17.4	0.0	17.5	19.2	18.0	1.8	0.0	0.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0		
heartbleed	0.0	0.0	20.0	20.0	20.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0		
infiltration	0.0	0.0	20.0	2.9	17.1	20.0	0.0	0.0	0.0	0.0	0.0	14.3	17.1	0.0	2.9	2.9	0.0	0.0	0.0	0.0	0.0	0.0		
infiltration - portscan	0.0	0.0	19.9	19.8	19.8	0.2	0.0	0.0	0.0	0.0	0.0	19.8	19.9	0.0	0.1	0.1	0.1	0.0	0.0	0.0	0.0	0.0		
portscan	0.0	0.0	20.0	20.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	20.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0		
ddos	20.0	20.0	7.0	0.0	0.0	0.0	0.3	20.0	20.0	0.0	0.0	0.0	0.0	0.0	12.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0		
dos goldeneye	18.4	14.8	11.2	0.2	0.8	0.3	1.0	18.3	15.3	7.6	8.9	1.1	1.5	0.2	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0		
dos hulk	13.5	14.0	1.9	0.5	3.1	0.0	10.6	13.5	15.9	6.2	6.2	1.0	0.5	5.5	3.2	2.9	1.0	0.0	0.0	0.0	0.0	0.0		
dos slowhttptest	0.4	7.2	5.0	5.1	2.5	0.0	1.7	4.1	3.5	0.1	0.1	12.4	4.6	8.2	13.2	13.2	13.2	1.6	1.6	0.0	0.0	0.0		
dos slowloris	4.3	16.1	0.0	0.8	0.0	0.0	16.9	20.0	3.0	3.1	3.1	0.0	0.0	1.3	0.0	0.0	0.0	15.7	15.7	0.0	0.0	0.0		
ftp-patator	0.0	0.0	20.0	0.1	19.9	20.0	0.0	0.0	0.0	0.0	0.0	20.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0		
ssh-patator	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	20.0	19.9	19.9	19.9	20.0
web attack - brute force	20.0	19.7	0.0	0.0	0.0	0.0	0.3	0.3	0.5	19.7	19.7	0.0	0.0	0.0	0.0	0.0	0.0	19.7	0.0	0.0	0.0	0.0	0.0	
web attack - sql injection	0.0	0.0	3.1	20.0	0.0	20.0	0.0	0.0	0.0	20.0	20.0	0.0	0.0	16.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
web attack - xss	0.0	18.9	0.0	20.0	0.0	0.0	0.0	0.0	0.0	20.0	20.0	0.0	0.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	

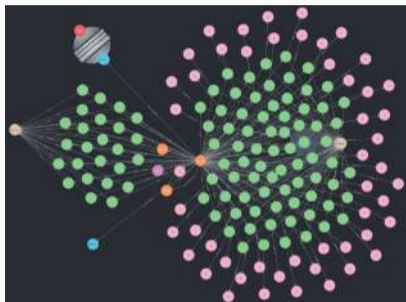
Applications - Feature contribution to attack types

	http_status_depth	http_status_msg	address_history	port_value	port_value	service	http_status_code	http_method	ua_browser	ua_os	duration	filetransfer_mime_type	conn_state	weird_name	weird_peer	http_info_addi	http_info_code	ssh_host_key_msg	ssh_host_key_alg	ssh_host_key	ssh_cipher_algo	ssh_client		
botnet	0.1	0.0	1.8	0.0	20.0	2.4	17.4	0.0	17.5	19.2	18.0	1.8	0.0	0.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0		
heartbleed	0.0	0.0	20.0	20.0	20.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0		
infiltration	0.0	0.0	20.0	2.9	17.1	20.0	0.0	0.0	0.0	0.0	0.0	14.3	17.1	0.0	2.9	2.9	0.0	0.0	0.0	0.0	0.0	0.0		
infiltration - portscan	0.0	0.0	19.9	19.8	19.8	0.2	0.0	0.0	0.0	0.0	0.0	19.8	19.9	0.0	0.1	0.1	0.1	0.0	0.0	0.0	0.0	0.0		
portscan	0.0	0.0	20.0	20.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	20.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0		
ddos	20.0	20.0	7.0	0.0	0.0	0.0	0.3	20.0	20.0	0.0	0.0	0.0	0.0	0.0	12.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0		
dos goldeneye	18.4	14.8	11.2	0.2	0.8	0.3	1.0	18.3	15.3	7.6	8.9	1.1	1.5	0.2	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0		
dos hulk	13.5	14.0	1.9	0.5	3.1	0.0	10.9	13.5	15.9	6.2	6.2	1.0	0.5	5.5	3.2	2.9	1.0	0.0	0.0	0.0	0.0	0.0		
dos slowhttptest	0.4	7.2	5.0	5.1	2.5	0.0	1.7	4.1	3.5	0.1	0.1	12.4	4.6	8.2	13.2	13.2	13.2	1.6	1.6	0.0	0.0	0.0		
dos slowloris	4.3	16.1	0.0	0.8	0.0	0.0	16.9	20.0	3.0	3.1	3.1	0.0	0.0	1.3	0.0	0.0	0.0	15.7	15.7	0.0	0.0	0.0		
ftp-patator	0.0	0.0	20.0	0.1	19.9	20.0	0.0	0.0	0.0	0.0	0.0	20.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0		
ssh-patator	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	20.0	19.9	19.9	19.9	20.0
web attack - brute force	20.0	19.7	0.0	0.0	0.0	0.0	0.3	0.3	0.5	19.7	19.7	0.0	0.0	0.0	0.0	0.0	0.0	19.7	0.0	0.0	0.0	0.0	0.0	
web attack - sql injection	0.0	0.0	3.1	20.0	0.0	20.0	0.0	0.0	0.0	20.0	20.0	0.0	0.0	16.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
web attack - xss	0.0	18.9	0.0	20.0	0.0	0.0	0.0	0.0	0.0	20.0	20.0	0.0	0.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	

Applications - Feature contribution to attack types

	http_status_depth	http_status_msg	address_history	port_value	port_value	service	http_status_code	http_method	ua_browser	ua_os	duration	filetransfer_mime_type	conn_state	weird_name	weird_peer	http_info_addi	http_info_code	ssh_host_key_msg	ssh_host_key_alg	ssh_host_key	ssh_cipher_algo	ssh_client		
botnet	0.1	0.0	1.8	0.0	20.0	2.4	17.4	0.0	17.5	19.2	18.0	1.8	0.0	0.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0		
heartbleed	0.0	0.0	20.0	20.0	20.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0		
infiltration	0.0	0.0	20.0	2.9	17.1	20.0	0.0	0.0	0.0	0.0	0.0	14.3	17.1	0.0	2.9	2.9	0.0	0.0	0.0	0.0	0.0	0.0		
infiltration - portscan	0.0	0.0	19.9	19.8	19.8	0.2	0.0	0.0	0.0	0.0	0.0	19.8	19.9	0.0	0.1	0.1	0.1	0.0	0.0	0.0	0.0	0.0		
portscan	0.0	0.0	20.0	20.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	20.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0		
ddos	20.0	20.0	7.0	0.0	0.0	0.0	0.3	20.0	20.0	0.0	0.0	0.0	0.0	0.0	12.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0		
dos goldeneye	18.4	14.8	11.2	0.2	0.8	0.3	1.0	18.3	15.3	7.6	8.9	1.1	1.5	0.2	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0		
dos hulk	13.5	14.0	1.9	0.5	3.1	0.0	10.9	13.5	15.9	6.2	6.2	1.0	0.5	5.5	3.2	2.9	1.0	0.0	0.0	0.0	0.0	0.0		
dos slowhttptest	0.4	7.2	5.0	5.1	2.5	0.0	1.7	4.1	3.5	0.1	0.1	12.4	4.6	8.2	13.2	13.2	13.2	1.6	1.6	0.0	0.0	0.0		
dos slowloris	4.3	16.1	0.0	0.8	0.0	0.0	16.9	20.0	3.0	3.1	3.1	0.0	0.0	1.3	0.0	0.0	0.0	15.7	15.7	0.0	0.0	0.0		
ftp-patator	0.0	0.0	20.0	0.1	19.9	20.0	0.0	0.0	0.0	0.0	0.0	20.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0		
ssh-patator	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	20.0	19.9	19.9	19.9	20.0
web attack - brute force	20.0	19.7	0.0	0.0	0.0	0.0	0.3	0.3	0.5	19.7	19.7	0.0	0.0	0.0	0.0	0.0	0.0	19.7	0.0	0.0	0.0	0.0	0.0	
web attack - sql injection	0.0	0.0	3.1	20.0	0.0	20.0	0.0	0.0	0.0	20.0	20.0	0.0	0.0	16.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
web attack - xss	0.0	18.9	0.0	20.0	0.0	0.0	0.0	0.0	0.0	20.0	20.0	0.0	0.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	

Applications - True Postive analysis - Web attack : Brute Force



single connection graph

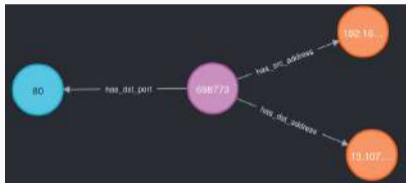
network_feature	value
http_method	POST
http_referrer	http ://205.174.165.68/dv/login.php
http_request_body_len	130
http_status_code	302
http_status_msg	Found
http_trans_depth	84
user_agent_browser	Mozilla/5.0
user_agent_os	Linux x86_64

Top 5 explanations

user_agent_browser - user_agent_os - http_status_msg

http_status_code - http_trans_depth

Applications - Forensic analysis - A False Positive Analysis



single connection graph

network_feature	value
ts	1499254964.698078
src_ip	192.168.10.15
dst_ip	13.107.4.50
src_port	49451
dst_port	80
proto	tcp
history	DadAttr
conn_state	RSTRH
orig_bytes	4226
resp_pkts	8884791

Top 5 explanations

port_value - history - conn_state - resp_pkts - orig_bytes

CICIDS2017 Dataset

- In [1], we manage to identify an unlabelled attack in the CICIDS2017 intrusion detection dataset thanks to the AE-pvalues explanations mechanism
- Many false positives alerts had explanations containing weird conn_state values
- We figured out that a port scan attack was unlabelled as such

. [1] Lanvin, M., Gimenez, P.F., Han, Y., Majorczyk, F., Mé, L., Totel, É. (2023). Errors in the CICIDS2017 Dataset and the Significant Differences in Detection Performances It Makes. In : Kallel, S., Jmaiel, M., Zulkernine, M., Hadj Kacem, A., Cuppens, F., Cuppens, N. (eds) Risks and Security of Internet and Systems. CRiSIS 2022. Lecture Notes in Computer Science, vol 13857. Springer, Cham.
https://doi.org/10.1007/978-3-031-31108-6_2

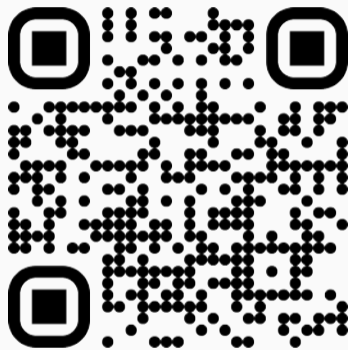
Conclusion

Summary :

- Explanation technique for alerts raised by AutoEncoder-based NIDS
- Clustering alerts based on explanations
- Help manual analysis

Future works

Leverage explanation techniques for the detection and alert triage



gitlab code for *AE-pvalues*
gitlab.inria.fr/mlanvin/ae-pvalues