Creachlab 23/05/2025

LIRMM

SERMA SAFETY & SECURITY

Anomalies Mitigation for Horizontal Side Channel Attacks with Unsupervised Neural Networks

Gauthier Cler, Sébastien Ordas, Philippe Maurine

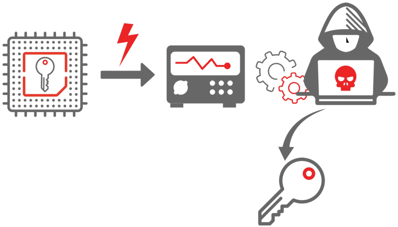
Outline

SECURE YOUR FUTURE

Horizontal Attacks

Side Channel Attacks

An attacker can recover sensitive information by listening on side-channels on a target device (Power, EM, timing, ...).



Several attacks:

- **Profiled**: Able to characterize the leakage before the attack (Templates, Deep Learning, ...)
- **Unprofiled**: Attack directly carried on target (SPA, DPA, ...).

SECURE YOUR FUTURE

Horizontal Attacks

- Single trace attack
- No profiling on open device possible, no leakage assessment, black box
- Usually applied on asymmetric implementations (RSA, ECC, ...).
- Commonly used clustering approach:
  1. Divide trace into patterns, preprocessing steps (cutting, alignment, filtering, ...)
  2. **Points of Interest (PoI) selection with univariate clustering** or dimensionality reduction
  3. Multidimensional clustering

Attack success highly relies on the quality of the trace.

Impact of anomalies on PoI selection

Univariate anomalies model

Outliers (interquantile range)

Distribution tails

$$x \notin R = [Q_1 - \alpha\,\mathrm{IQR}, Q_3 + \alpha\,\mathrm{IQR}]$$
$$\mathrm{IQR} = Q_3 - Q_1$$

Saturated values

Min/max values of digital sampling vertical resolution, for 8bit:

$$x \in \xi(8) = \{-128, 127\}$$

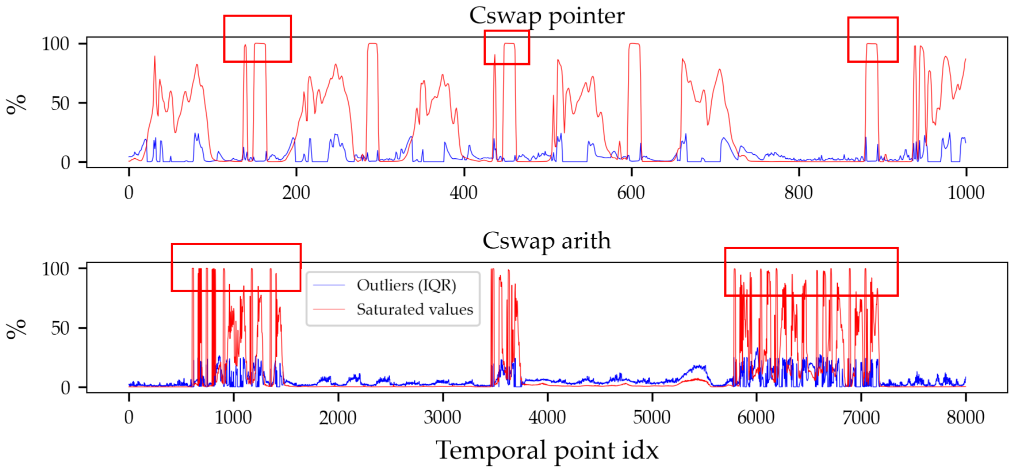
Considered Datasets

Cswap Pointer and Arith public datasets: ECC Scalar multiplication

- Arith dataset: Arithmetic swapping
- Pointer datset: Pointers swapping instead of values

We define the BRR as the percentage of correctly identified bits of the exponent scalar during the clustering process.

Anomalies in data



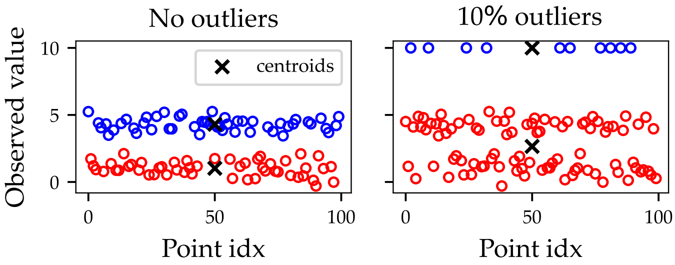
[1] Average anomalies Pointer:33.3%, Arith:16.5%

Impact of anomalies on PoI selection

Clustering is **not robust** to anomalies in data, can cause centroids shift, singularities,...

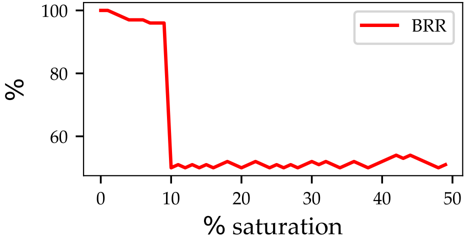
Impact of anomalies on PoI selection

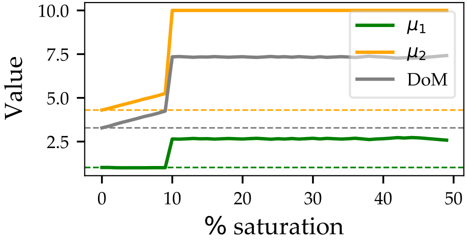
Clustering is **not robust** to anomalies in data, can cause centroids shift, singularities,...



Impact of anomalies on PoI selection

Clustering is **not robust** to anomalies in data, can cause centroids shift, singularities,...





1

[1]BRR: Bit Recovery Rate

Anomalies mitigation

Limits of simple mitigation

Mitigation by ablation

- Remove time points based on anomalies threshold
- Possibly loosing information about the leakage

Limits of simple mitigation

Mitigation by ablation

- Remove time points based on anomalies threshold
- Possibly loosing information about the leakage

Mitigation by replacement

- Replace anomalies points with mean/median of non anomalies for each time point
- Decrease separability of mixture components

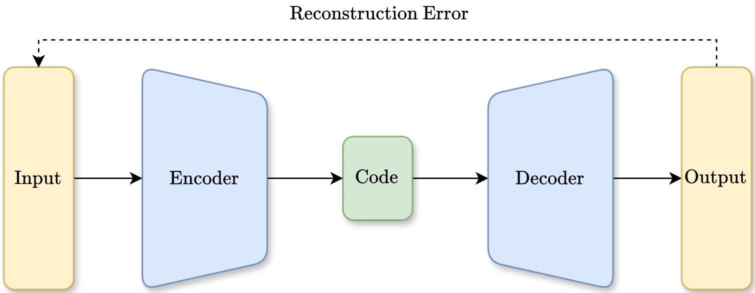
Contribution - Mitigation with neural networks

In this work, alternative methods are studied:

- Able to be trained in an unsupervised manner
- Leakage/information conservation
- Two approaches are considered:
  - Unsupervised mitigation: **Robust auto-encoder**
  - Selfsupervised mitigation: **Cycle generative adversarial networks**

Auto-encoder

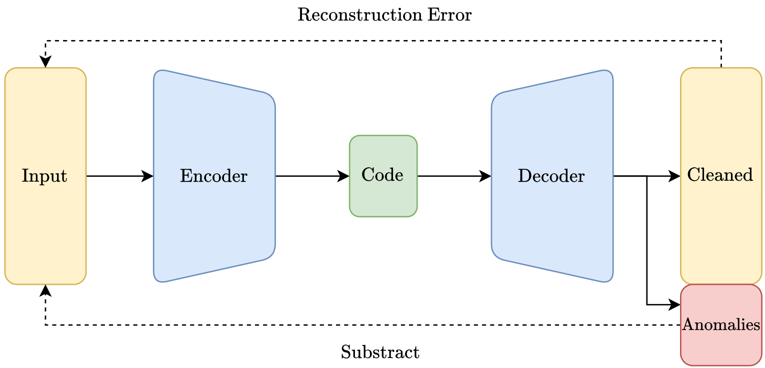
Built from a encoder/decoder $(\mathcal{E}_\phi, \mathcal{F}_\theta)$ network pair. Trained for input reconstruction.



$$\mathcal{L}(\theta, \phi) = ||\mathbf{X} - \mathcal{F}_\theta(\mathcal{E}_\phi(\mathbf{X}))||_2$$

Robust auto-encoder unsupervised mitigation

Decomposition of input data to **cleaned** and **anomalies** matrices.
Prior on the anomalies amount.



Robust auto-encoder unsupervised mitigation

The RAE aims at achieving the following decomposition:

$$\mathbf{X} = \mathbf{L} + \mathbf{S} \tag{1}$$

where:

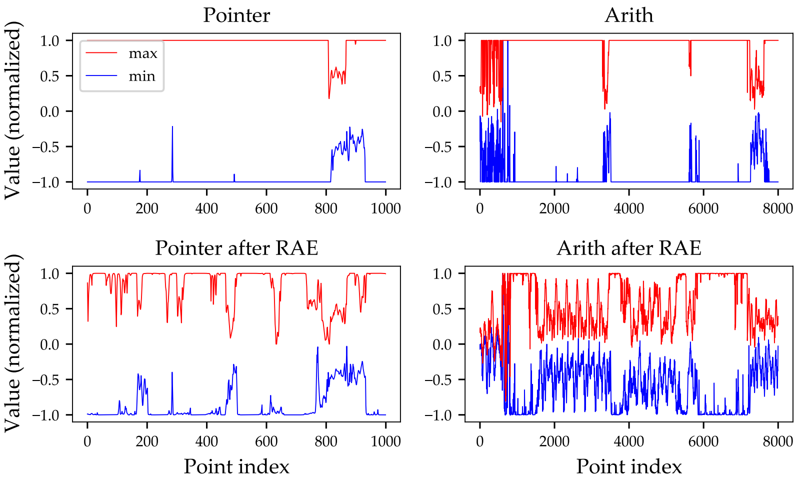
- $\mathbf{X}$: input patterns
- $\mathbf{L}$: cleaned patterns
- $\mathbf{S}$: extracted anomalies

The complete objective is given by:

$$\mathcal{L}(\theta, \phi) = ||\mathbf{L} - \mathcal{F}_\theta(\mathcal{E}_\phi(\mathbf{L}))||_2 + \tau||\mathbf{S}||_1 \tag{2}$$

Left term is optimized through gradient descent while right term is minized with a proximal operator.

Impact on patterns



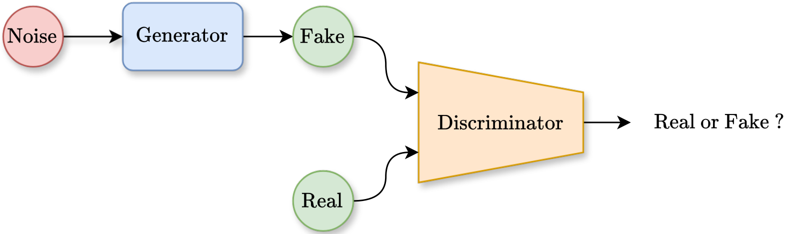
Limits of the RAE

While relevant, the RAE can suffer from some drawbacks:

- The RAE **generates new synthetic patterns**, this can cause **side effects on non anomalies points**.
- In addition, it does not exploit any anomalies model. It is fully unsupervised

An alternative method is proposed to to include the anomalies models, based on **generative adversarial networks**.

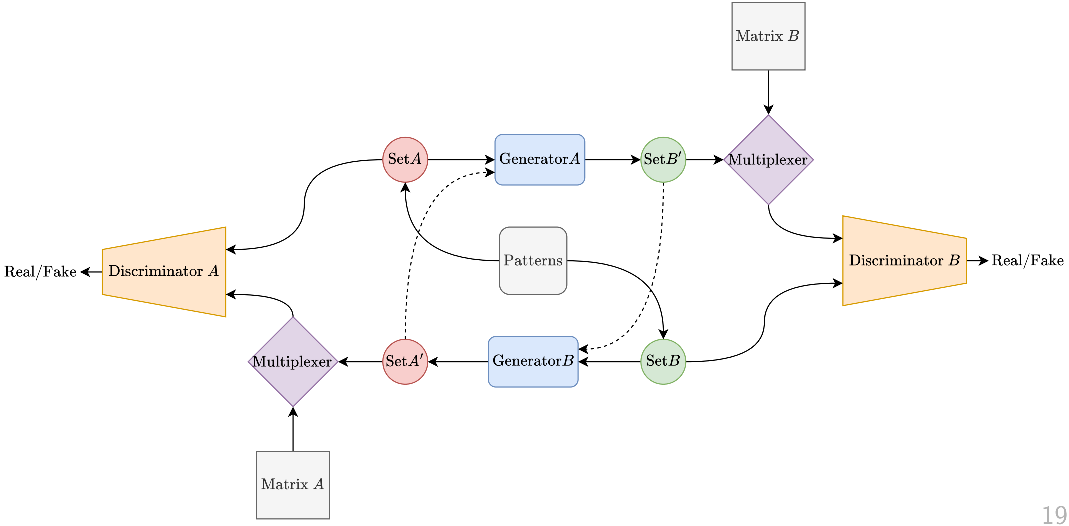
Generative Adversarial Networks



$$\min_G \max_D \mathcal{L}_{\mathsf{GAN}}(G, D, X, Z) = \mathbb{E}_{x\sim X} \log D(x) + \mathbb{E}_{z\sim Z} \log[1 - D(G(z))]$$

SECURE YOUR FUTURE

Multiplexer CycleGAN self-supervised mitigation



Multiplexer CycleGAN self-supervised mitigation

Complete loss is given by:

$$
\begin{aligned}
\mathcal{L}(G_A, G_B, D_A, D_B) &= \mathcal{L}_{\mathsf{GAN}}(G_A, D_B, A, B) \\
&+ \mathcal{L}_{\mathsf{GAN}}(G_B, D_A, B, A) \\
&+ \lambda \mathcal{L}_{\mathsf{cyc}}(G_A, G_B)
\end{aligned} \tag{3}
$$

SECURE YOUR FUTURE

Multiplexer CycleGAN self-supervised mitigation

Complete loss is given by:

$$\begin{aligned}\mathcal{L}(G_A, G_B, D_A, D_B) &= \mathcal{L}_{\mathsf{GAN}}(G_A, D_B, A, B) \\ &+ \mathcal{L}_{\mathsf{GAN}}(G_B, D_A, B, A) \\ &+ \lambda\mathcal{L}_{\mathsf{cyc}}(G_A, G_B)\end{aligned} \tag{3}$$

with consistency loss:

$$\begin{aligned}\mathcal{L}_{\mathsf{cyc}}(G_A, G_B) &= \mathbb{E}_{a\sim A}||G_B(G_A(a)) - a||_1 \\ &+ \mathbb{E}_{b\sim B}||G_A(G_B(b)) - b||_1\end{aligned} \tag{4}$$

Multiplexer CycleGAN self-supervised mitigation

Complete loss is given by:

$$\begin{aligned}\mathcal{L}(G_A, G_B, D_A, D_B) &= \mathcal{L}_{\mathsf{GAN}}(G_A, D_B, A, B) \\ &+ \mathcal{L}_{\mathsf{GAN}}(G_B, D_A, B, A) \\ &+ \lambda \mathcal{L}_{\mathsf{cyc}}(G_A, G_B)\end{aligned} \tag{3}$$

with consistency loss:

$$\begin{aligned}\mathcal{L}_{\mathsf{cyc}}(G_A, G_B) &= \mathbb{E}_{a\sim A}||G_B(G_A(a)) - a||_1 \\ &+ \mathbb{E}_{b\sim B}||G_A(G_B(b)) - b||_1\end{aligned} \tag{4}$$

Aims at finding:

$$G_A^*, G_B^* = \operatorname*{argmin}_{G_A, G_B} \operatorname*{argmax}_{D_A, D_B} \mathcal{L}(G_A, G_B, D_A, D_B) \tag{5}$$

Multiplexer CycleGAN self-supervised mitigation

1. Build the anomalies matrix M such that:

$$m_{i,j} = \begin{cases} 1, & \text{if } x_{i,j} \in \xi(8) \vee x_{i,j} \notin R(x_{:,j}) \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

Multiplexer CycleGAN self-supervised mitigation

1. Build the anomalies matrix M such that:

$$m_{i,j} = \begin{cases} 1, & \text{if } x_{i,j} \in \xi(8) \vee x_{i,j} \notin R(x_{:,j}) \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

2. Split into A, B sets based on maximum Hamming distance:

$$\underset{i,j \in \{1,\ldots,n\}}{\operatorname{argmax}} \ \mathrm{HW}(m_{i,:} \oplus m_{j,:}), \quad i \neq j \tag{7}$$

SECURE YOUR FUTURE

Multiplexer CycleGAN self-supervised mitigation

1. Build the anomalies matrix M such that:

$$m_{i,j} = \begin{cases} 1, & \text{if } x_{i,j} \in \xi(8) \vee x_{i,j} \notin R(x_{:,j}) \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

2. Split into A, B sets based on maximum Hamming distance:

$$\underset{i,j \in \{1,\dots,n\}}{\operatorname{argmax}} \ \mathrm{HW}(m_{i,:} \oplus m_{j,:}), \quad i \neq j \tag{7}$$

3. Train the model with gradient descent (previous slide).

Multiplexer CycleGAN self-supervised mitigation

1. Build the anomalies matrix M such that:

$$m_{i,j} = \begin{cases} 1, & \text{if } x_{i,j} \in \xi(8) \vee x_{i,j} \notin R(x_{:,j}) \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

2. Split into A, B sets based on maximum Hamming distance:

$$\underset{i,j \in \{1,\dots,n\}}{\operatorname{argmax}} \ \mathrm{HW}(m_{i,:} \oplus m_{j,:}), \quad i \neq j \tag{7}$$

3. Train the model with gradient descent (previous slide).
4. Replace by generated values through multiplexers:

$$\begin{aligned} A'' &= \mathrm{mux}(A, G_B(B), M_A) = (M_A \wedge G_B(B)) \vee (\neg M_A \wedge A) \\ B'' &= \mathrm{mux}(B, G_A(A), M_B) = (M_B \wedge G_A(A)) \vee (\neg M_B \wedge B) \end{aligned} \tag{8}$$

SECURE YOUR FUTURE

Benefits of proposed architecture

SECURE YOUR FUTURE

- Selective correction that include the anomaly models.
- Only values marked as **anomalies** are generated, others are untouched.
- Multiplexers add training stability for GANs, reduce complexity.
- Sets matching on Hamming distance allows optimal correction.

Results

Mitigation results

	Cswap Pointer			Cswap Arith		
	Before	RAE	GAN	Before	RAE	GAN
Outliers (%)	4.32	5.36	**1.67**	5.15	4.73	**1.35**
Saturation (%)	30.27	**1.19**	10.22	12.88	**0.01**	5.19
Total (%)	33.39	**6.55**	11.85	16.54	**4.75**	6.49

Table: Percentage of outliers and extremes obtained on original patterns, after applying the RAE and the CycleGAN. Best results are highlighted in bold.

Impact on distributions

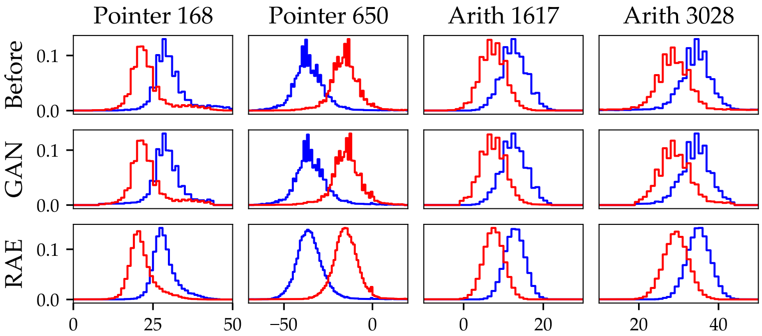
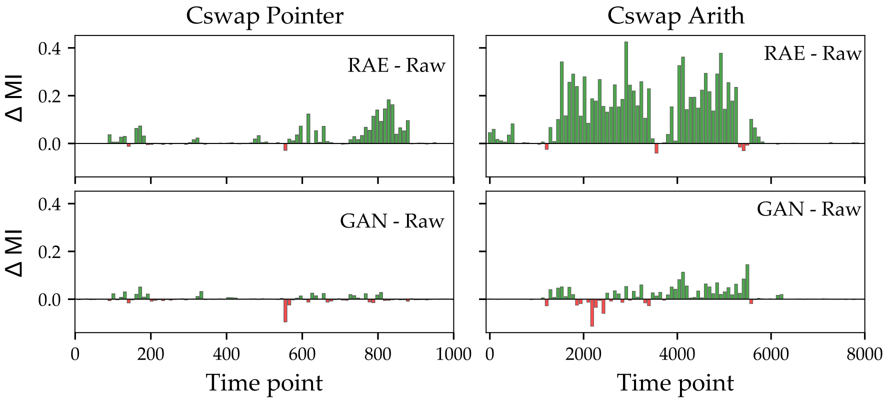


Figure: Empirical p.d.f of four samples before and after application of the RAE and CycleGAN to mitigate abnormal values. Blue p.d.f corresponds to class $c = 0$ (resp. red $c = 1$).

Information conservation

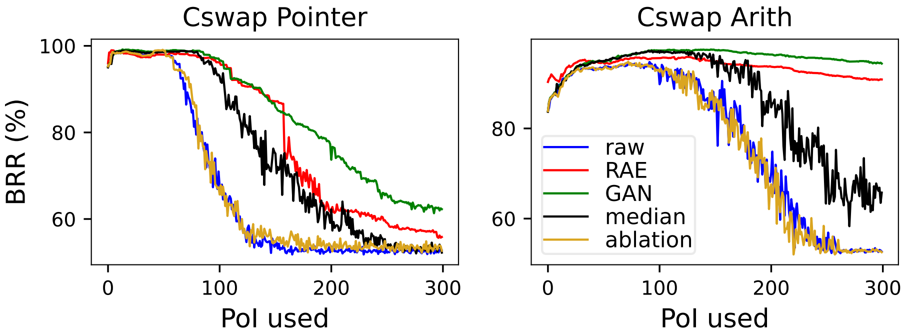
No change in the global MI. [1]



[1]Estimated with MINE.

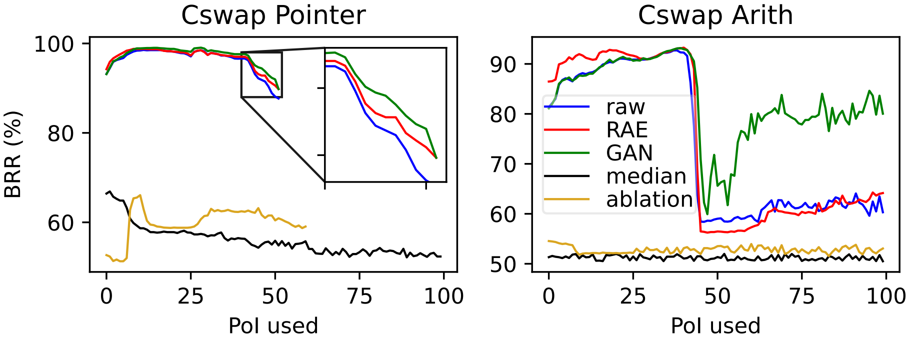
Supervised selection - upper bound

Select k PoI with highest t-values and apply multidimensional clustering.



Unsupervised selection

Multidimensional clustering on the best k PoI from Cler *et al.* 2023 unsupervised selection.



Conclusion

Conclusion

Benefits

- Anomalies mitigation **improves leakage exploitability**
- Methods are applicable in a completely unsupervised context

Conclusion

Benefits

- Anomalies mitigation **improves leakage exploitability**
- Methods are applicable in a completely unsupervised context

Limitations

- Architecture choice and parameters tuning can be hard in practice
- Attack success **still** depends on the exploitation method

Conclusion

Benefits

- Anomalies mitigation **improves leakage exploitability**
- Methods are applicable in a completely unsupervised context

Limitations

- Architecture choice and parameters tuning can be hard in practice
- Attack success **still** depends on the exploitation method

Future work

- Consider additional anomalies models
- Generalize on other targets/algorithms

Thank you for your attention

Do you have any questions ?

- Read the thesis: **hal.science/tel-04730413v1**
- Paper: **CASCADE 2025**, soon to be published (Springer)
- Contact: **g.cler@serma.com**

SECURE YOUR FUTURE