16th IEEE International Symposium on Embedded Multicore/Many-core Systems-on-Chip (MCSoC-2023)

Revealing Secret Key from Low Success Rate Deep Learning-Based Side Channel Attacks

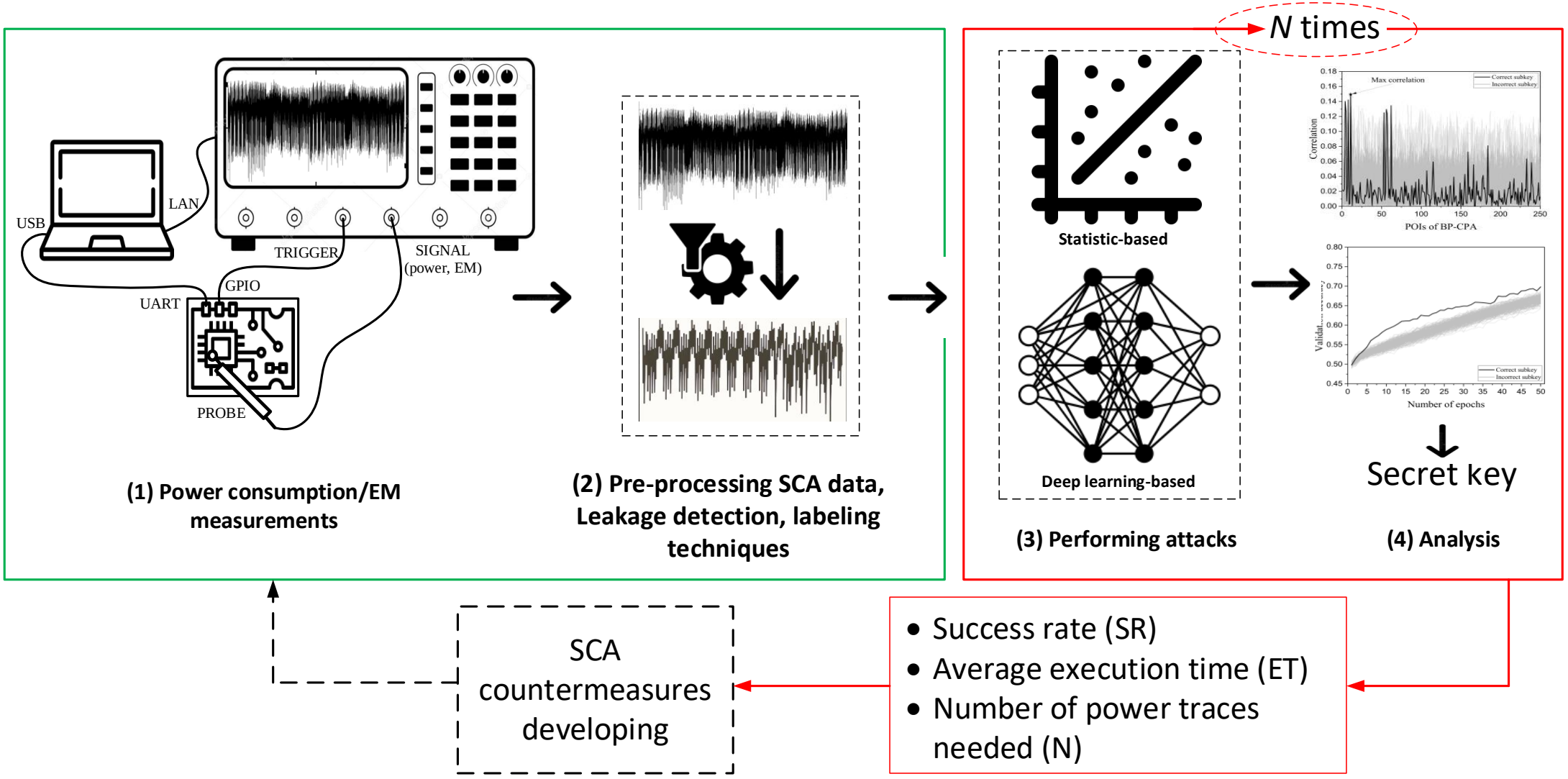
Van-Phuc Hoang[1], Ngoc-Tuan Do[1], Trong-Thuc Hoang[2], and Cong-Kha Pham[2]
[1]Le Quy Don Technical University, Hanoi, Vietnam
[2]University of Electro-Communications, Tokyo, Japan

Singapore, December 2023

Outline

1. Introducing side-channel attack
2. Non-profiled deep learning-based side channel attack
3. Propose a new SCA metric for non-profiled DLSCA
4. Validation experiments
5. Conclusion

Introduction to side-channel attack

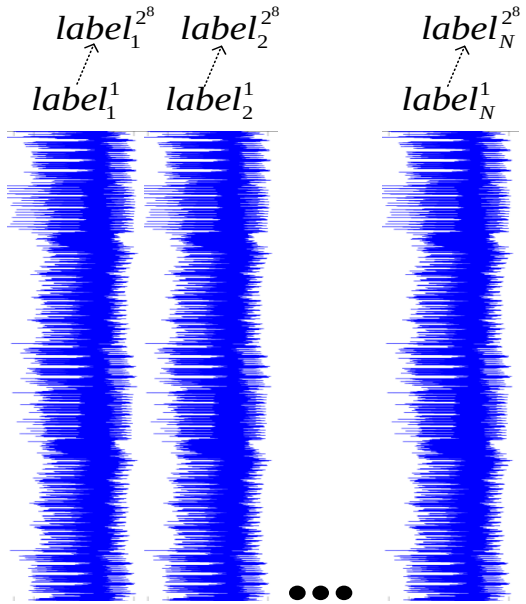


✓ Side-channel attacks can leverage statistical or deep learning techniques to reveal the secret key.

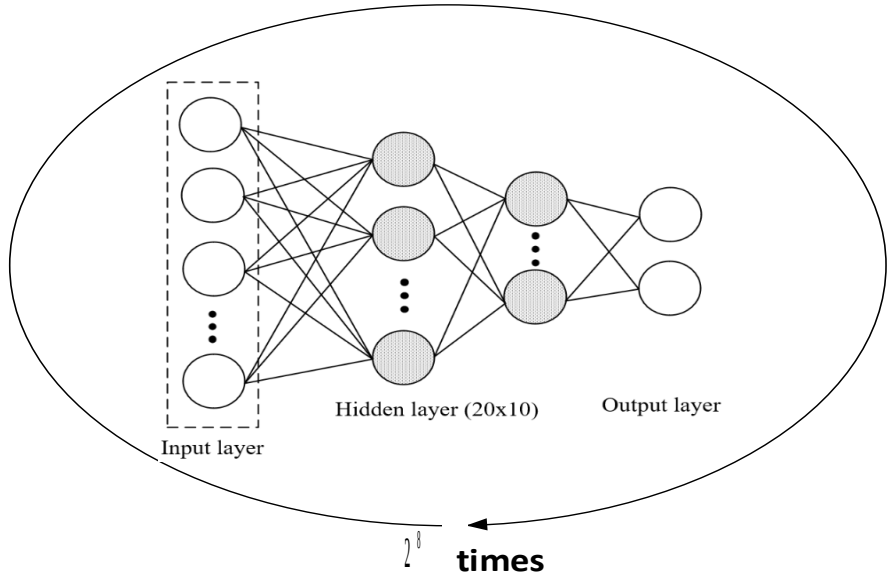
✓ They are commonly categorized into two approaches: **profiled attack and non-profiled attack.**

Non-profiled deep learning based side channel attack

➢ *Differential deep learning analysis (DDLA) [*]*



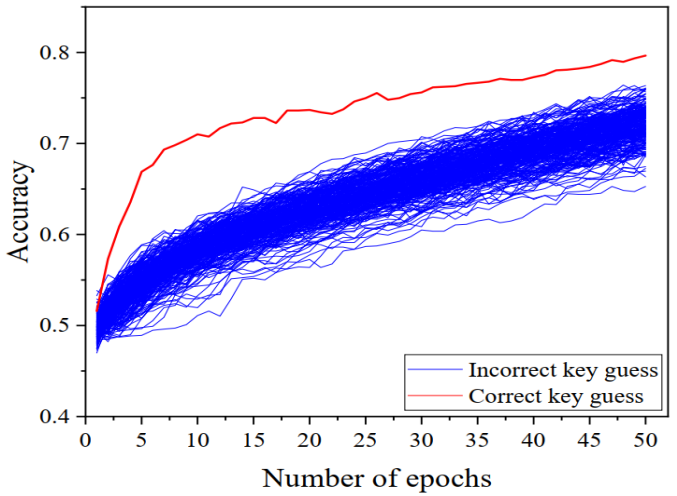
N labeled power traces



DDLA attack

Training metric

2^8



DDLA attack's procedure

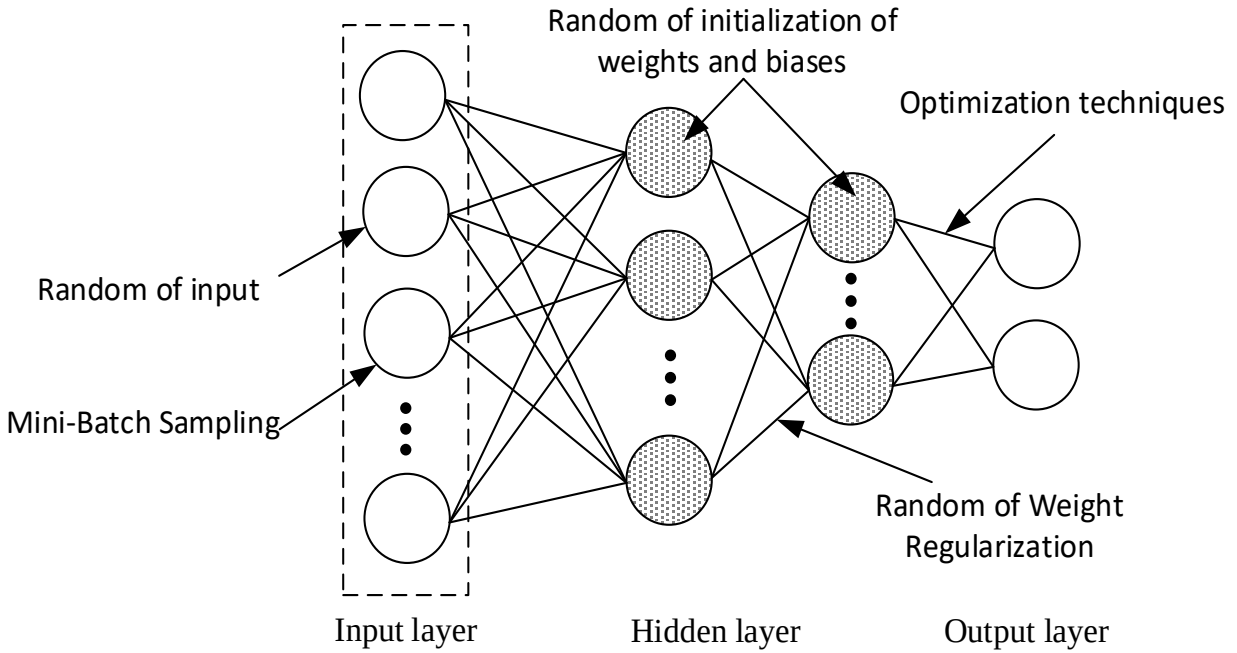
- Using only one model for attack
- Require repeating training process for each key hypothesis
- The secret key is determined by training metrics, such as loss or accuracy

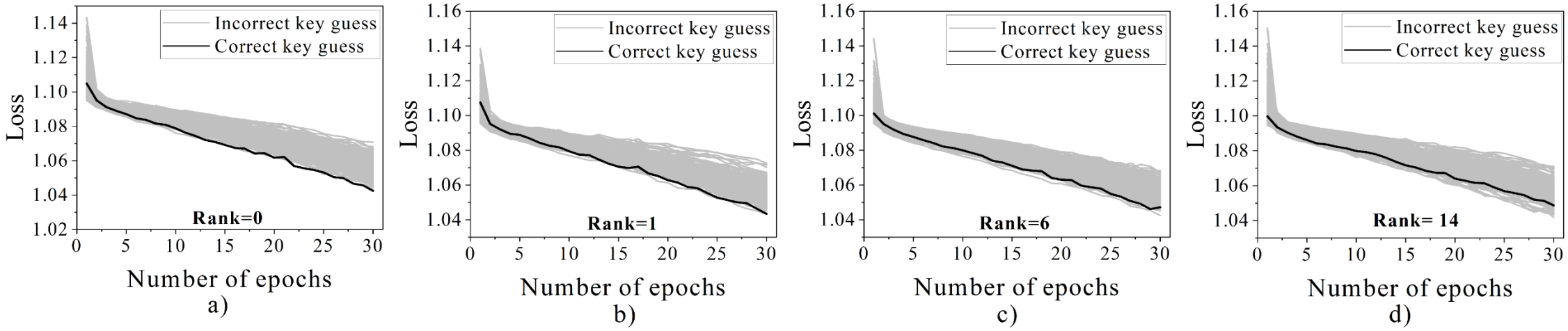
[*] B. Timon, "Non-profiled deep learning-based side-channel attacks with sensitivity analysis," *IACR Transactions on Cryptographic Hardware and Embedded Systems*, vol. 2019, no. 2, pp. 107 131, Feb. 2019. [Online]. Available: https://tches.iacr.org/index.php/TCHES/article/view/7387

Sources of randomness in non-profiled deep learning based SCA

- ✓ Different sources of randomness in DL training:
  - + The randomness of inputs
  - + The initialization of weights and biases
  - + Regularization techniques
  - + Optimization techniques
- ✓ Non-profiled based on deep learning usually using training metrics, such as loss and accuracy to determine secret key.

=> **Unstable achieved results**





The secret key ranks differently in various training instances using the same dataset and model.

SCA metrics for non-profiled DLSCA

- Score and Rank:

SCA attack on 8-bit Sbox produces 256 scores $[score_0, score_1, \ldots, score_{255}]$ where $score_i$ is attack score of the key candidate i. Then we have vector: $[rank_0, rank_1, \ldots, rank_{255},]$ where $rank_k$ is rank of key candidate k and the best possible rank equals 1.

For example, if the best score came from k = 17: $rank_{17} = 1$

- Success rate (SR):

The success rate of order o is the average empirical probability that the correct key is located within the first o elements of the key guessing vector g

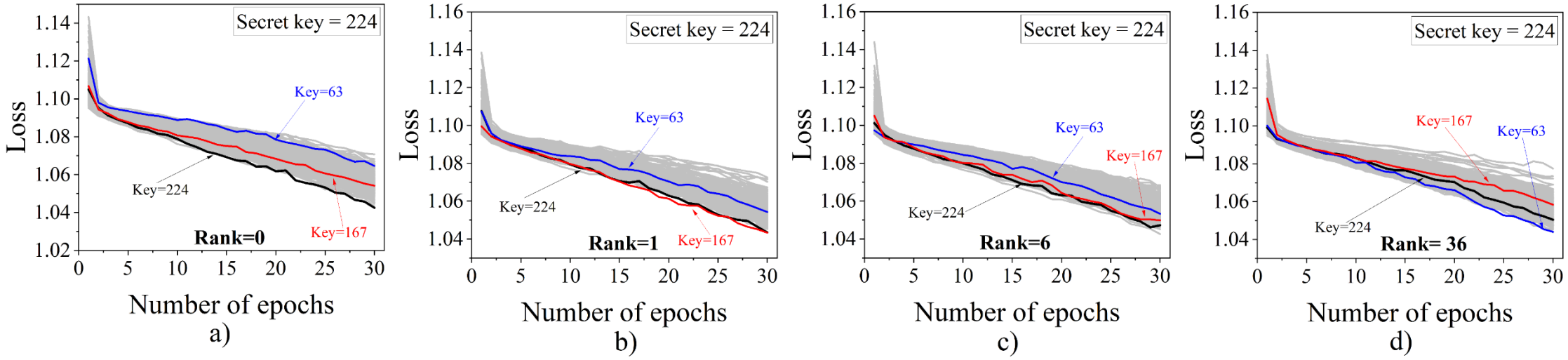
$$R_o^i = \begin{cases} 1, \text{if } rank_{k_c} \le o \\ 0, \text{otherwise} \end{cases} \qquad SR_o = \frac{1}{n}\sum_{i=1}^{n} SR_o^i.$$

- Guessing Entropy (GE):

$$GE = \frac{1}{n}\sum_{i=1}^{n} GE^i.$$

Output of training processes using the correct key guess

- ✓ Due to the randomness of model, the correct key is not always determined over all attacks, especially in the case of using un-optimized model.
- ✓ For the existence of the correlation between power traces and the power model (using correct key), the model always has stable output when trained with the correct key compared to others.



The secret key yields a stable loss over different training processes (attacks)

- Key 224 usually has lower rank compared to others.
- The ranks of other keys, such as Key 63 and Key 167 are inconsistent.

Proposed Inversion of Exponential Rank (IER) metric

- SR, Rank or GE metrics are primarily employed for known-key analysis, indicating level of difficulty for an attacker to extract the secret key from a given set of measured traces.
- Due to the sources of randomness in the DL training process, conventional SCA metrics cannot provide reliable results.
- DLSCA attack requires repetition for reliable outcomes, making hyperparameter tuning in DL a time-consuming and costly process.

=> It is necessary for a new metric to evaluate consistency of DL based non-profiled attacks.

$$IER_j = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\alpha^{KR_{i,j}}}, (\alpha > 1) \quad \text{where } 1 \le i \le n \text{ and } 0 \le j \le 255$$

- ✓ IER of the correct key will reach 1 when $KR_{i,ck}$ equals zero for all i.
- ✓ A significantly small IER value indicates higher rank.
- ✓ Key guesses with higher ranks have a negligible impact on IER.
- ✓ The more consistently non-profiled DLSCA attacks yield low-ranked keys, the higher the IER value.

Rank (KR)	IER by Equation 4						
	α=1.01	α=1.05	α=1.1	α=1.3	α=1.5	α=1.7	α=1.9
0	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
1	**0.99**	**0.95**	**0.91**	**0.77**	**0.67**	**0.59**	**0.53**
2	**0.98**	**0.91**	**0.83**	**0.59**	**0.44**	**0.35**	**0.28**
3	**0.97**	**0.86**	**0.75**	**0.46**	**0.30**	**0.20**	**0.15**
4	**0.96**	**0.82**	**0.68**	**0.35**	**0.20**	**0.12**	0.08
5	**0.95**	**0.78**	**0.62**	**0.27**	**0.13**	0.07	0.04
10	**0.91**	**0.61**	**0.39**	0.07	0.02	0.00	0.00
20	**0.82**	**0.38**	**0.15**	0.01	0.00	0.00	0.00
40	**0.67**	**0.14**	0.02	0.00	0.00	0.00	0.00
80	**0.45**	0.02	0.00	0.00	0.00	0.00	0.00
160	**0.20**	0.00	0.00	0.00	0.00	0.00	0.00

IER metric helps detecting a key that has a stable KR within a first "k" elements of key ranks. "k" depends on the value of alpha.

Proposed distinguisher based on IER metric

- Proposed metric shows that it is capable of revealing the secret key without requiring prior knowledge of the correct key.
- The secret key could be detected by the following steps:
  - The attack is performed and repeated N times on the same dataset. The ranks KR of all hypothesis keys are calculated on each attack.
  - IER_j of the key guess number j is determined following Equation 4.
  - The hypothesis key corresponding to the highest IER is specified as the correct key.

Algorithm 2 Proposed non-profiled DLSCA using IER metric

Input: D traces $(\boldsymbol{t}_i)_{1\leq i\leq D}$, corresponding plaintexts $(d_i)_{1\leq i\leq D}$, and K key hypotheses. A network Net and number of epochs n_e

Output: $k_{cr} \in \boldsymbol{k}$

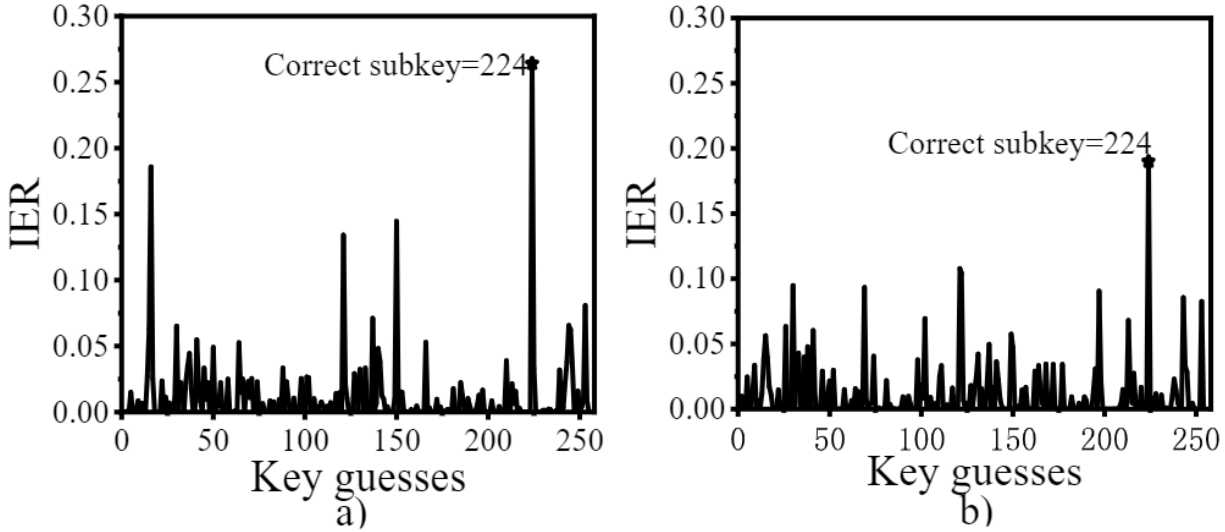
Set training data as $X = (\boldsymbol{t}_i)_{1\leq i\leq D}$.
for $i \in interation$ **do**
  for $k_j \in \boldsymbol{k}$ **do**
    Re-initialize trainable parameters of Net
    Compute the series of hypothetical values $\left(h_{k_j,i}\right)_{1\leq i\leq D}$
    Set training labels as $y_{k_j,i} = \left(h_{k_j,i}\right)_{1\leq i\leq D}$
    $DL\left(Net, X, y_{k_j,i}, n_e\right)$
    Calculate the key rank $KR_{i,j}$
  end for
end for
Calculate the IER for all key guesses using (4)
return key k_{cr} which leads to the highest IER

Validation experiments with DDLA-SHW attack

Attack	No. of epochs	Results	Byte
			3
DDLA-SHW [7]	30	SR (%)	23.33
DDLA-SHW+IER ($\alpha = 1.3$)			✓
DDLA-SHW [7]	20	SR (%)	10
DDLA-SHW+IER ($\alpha = 1.3$)			✓

✓: Successful revealing secret key

Comparison between DDLA-SHW and DDLA-SHW combined IER on the ASCAD (fixed key) dataset.



Attack results using DDLA-SHW [*] and IER.
a) 30 epochs; b) 20 epochs.

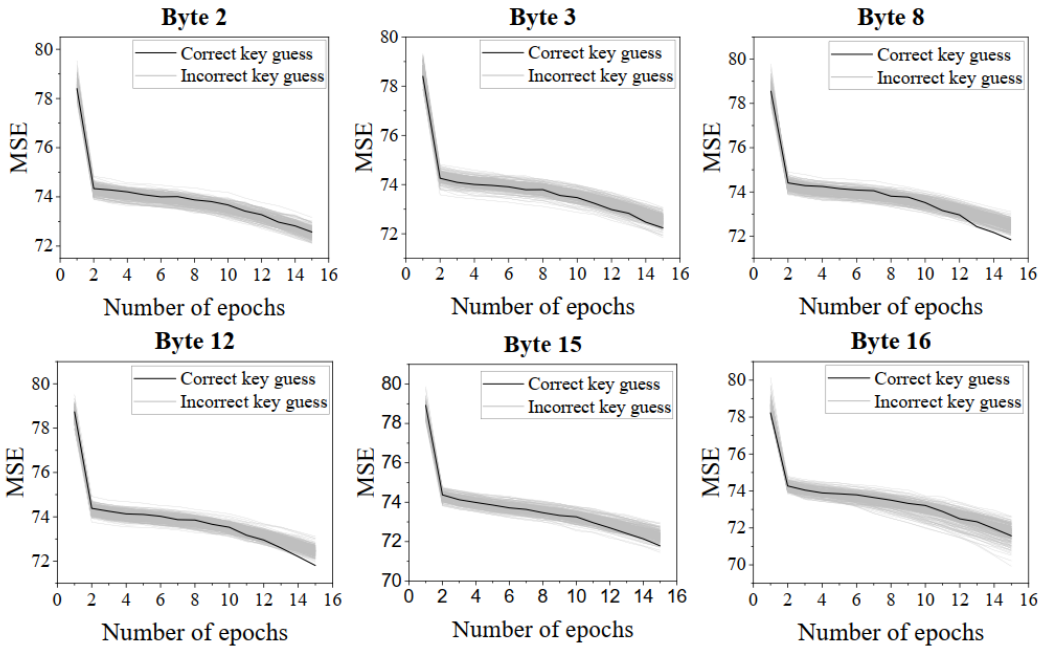
- DDLA-SHW yields poor and unstable results, resulting in low success rate (SR), especially when trained with small number of epochs.
- Difficult to determine the correct key.
- By using IER metric, the best candidate can be clearly seen.
- The higher the IER value, the more stable the results achieved.

[*] N.-T. Do, V.-P. Hoang, V. S. Doan, and C.-K. Pham, "On the performance of non-profiled side channel attacks based on deep learning techniques," *IET Information Security,* vol. 17, no. 3, pp. 377–393, dec 2022

Validation experiments with MOR attack

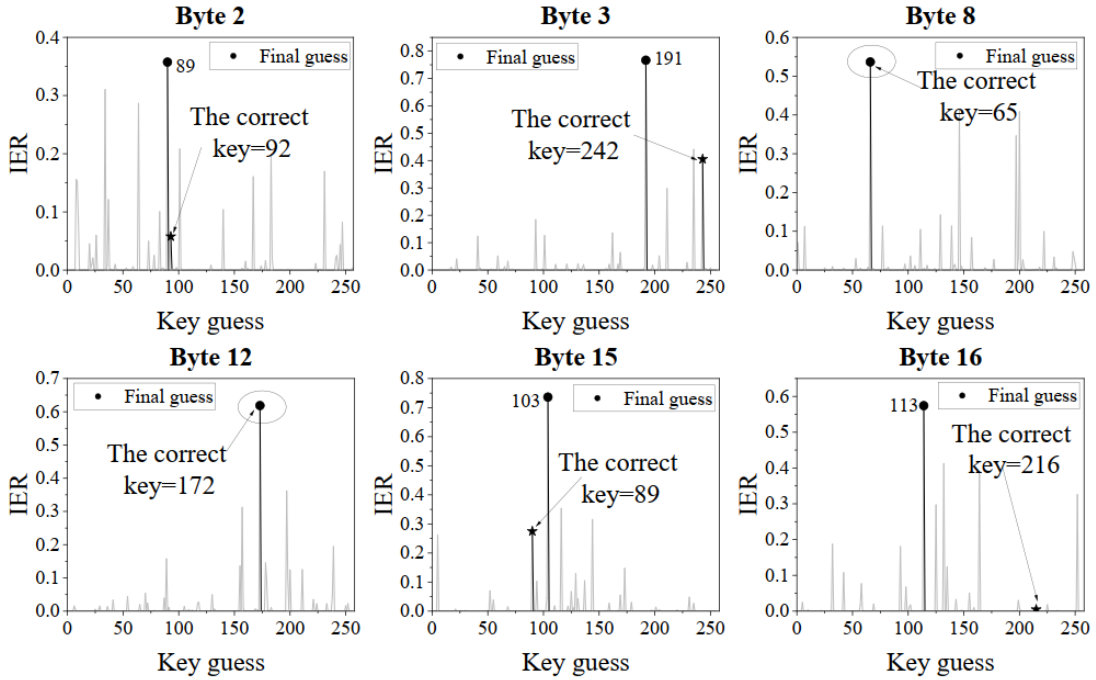
Byte	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Value	23	92	242	153	122	133	131	65	60	119	223	172	126	108	89	216

The values of all secret key bytes on CHES2018-CTF dataset



Attack results using MOR architecture [*]

IER over 30 attacks



Attack results using MOR architecture [*] combined IER metrics

- The loss metric primarily indicates the most promising candidate for the secret key, while the success rate (SR) is subsequently computed based on numerous repeated attacks.
- IER metric offers additional insights by revealing which candidate is consistently detected more frequently than others across repeated attacks.

[*] N.-T. Do, V.-P. Hoang, and V. S. Doan, "A novel non-profiled side channel attack based on multi-output regression neural network," *Journal of Cryptographic Engineering*, mar 2023

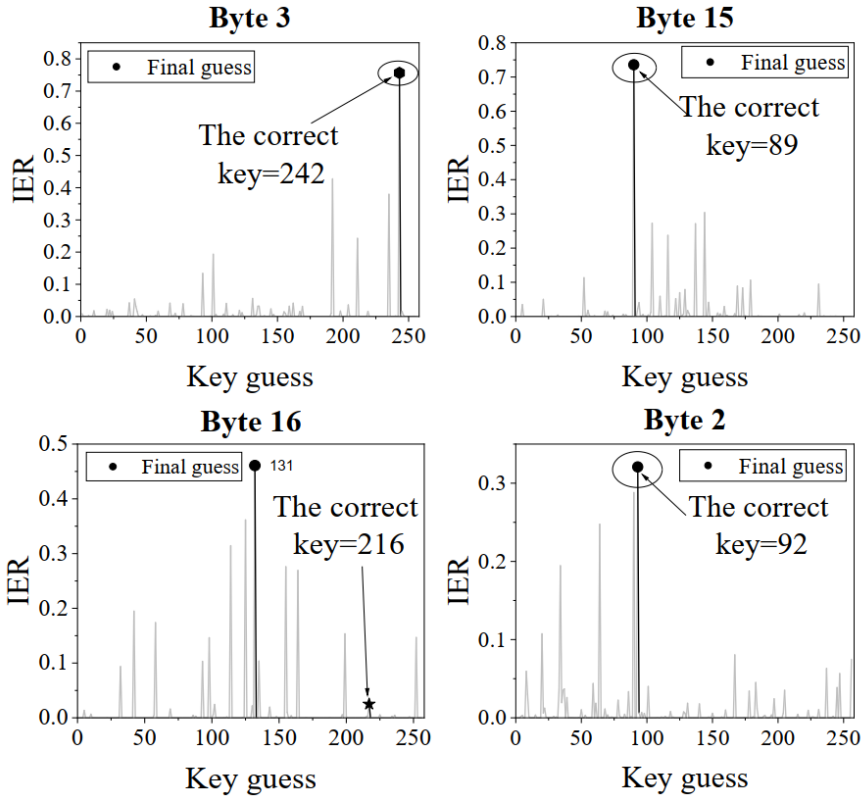
Validation experiments with MOR attack (cont.)

Attack	No. of epochs	Results	Byte															
			1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
MOR [6]	15	SR (%)	96.67	**3.33**	**26.67**	93.33	100	60	86.67	**36.67**	73.33	60	70	**36.67**	70	73.33	**10**	**0**
MOR+IER ($\alpha = 1.3$)			✓	✗	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗
MOR [6]	20	SR (%)	-	**20**	**53.33**	-	-	-	-	**90**	-	-	**-**	**60**	-	-	**60**	**0**
MOR+IER ($\alpha = 1.3$)			✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗

✓: Successful revealing secret key

Attack results of all bytes on CHES-CTF2018 dataset (increasing number of epochs)

- When employing the same dataset, varying attack outcomes emerge with different bytes, posing a challenge in assessing all bytes consistently with a single model.
- Fine-tuning the model for each byte causes substantial costs and time consumption.
- Utilizing IER enhances the accuracy of detecting the correct key byte, notably increasing from 10 bytes to 12 bytes.
- The IER metric provides a clear distinction in identifying the correct key.



Conclusions

- Non-profiled DLSCA encounters challenges when the metric of the correct key is not distinguishable from incorrect ones.
- The proposed metric was applied to improve the performance of non-profiled DL-based attacks, particularly in cases with low SR outcomes.
- The IER metric can be utilized to assess the stability of attack results across different models.
- The IER metric can be combined with other techniques to improve the performance of SCA attacks and enable the comparison of effectiveness among various DLSCA techniques.

THANK YOU FOR YOUR ATTENTION!

Acknowledgement

This publication is the output of the ASEAN IVO project, "Artificial Intelligence Powered Comprehensive Cyber-Security for Smart Healthcare Systems (AIPOSH)", and financially supported by NICT, Japan.