WPI

Too Hot To Be True:

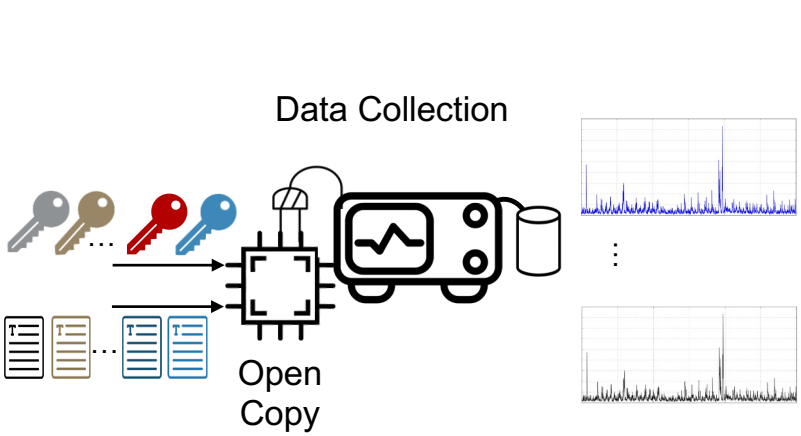
Temperature Calibration for Higher Confidence in NN-assisted SCA

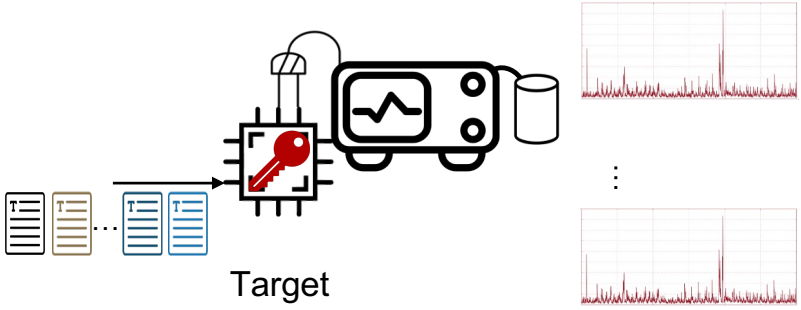
Seyedmohammad Nouraniboosjin and Fatemeh Ganji

Date: 3/30/2026

CASCADE'26

Profiled SCA: in search for probability distributions!





Prob. Density Function

Profiling

g : (,) ↦ Pr (| ,)

g : (,) ↦ Pr (| ,)

Attack

s = g ( )

s = g ( )

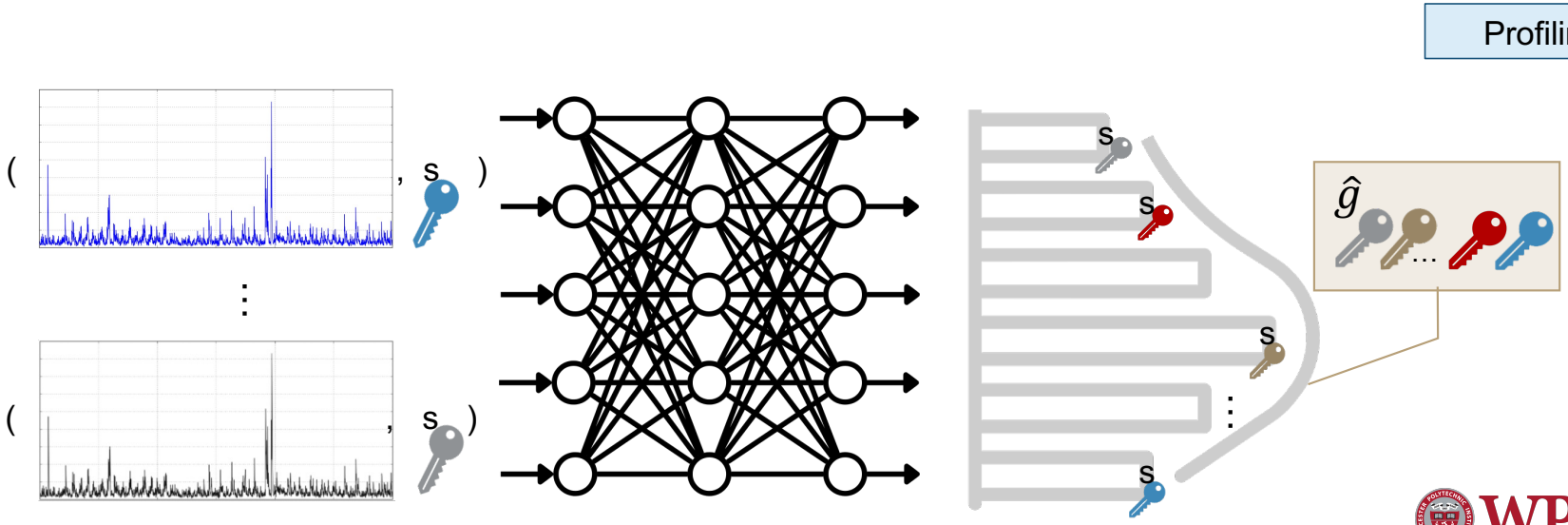
s > s > s > …>

Rank ( )=2

WPI

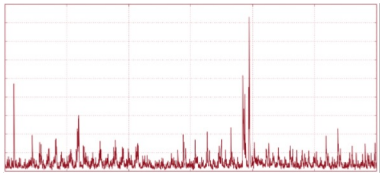
Let's bring in NNs

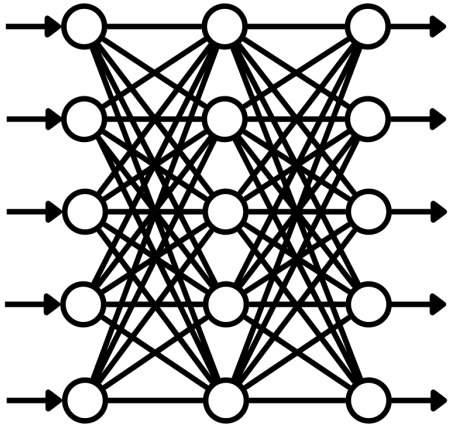
- Characterizing the leakage precisely through statistical techniques: costly in terms of the number of traces needed
- NN-assisted profiled SCA: effective against un-/protected cryptographic implementations, as well as noisy and shuffled traces

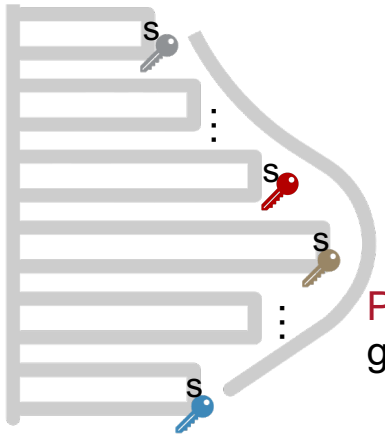


WPI

NN-assisted SCA







Attack

Rank ( )=10

Problem: a high prediction probability given to an incorrect key

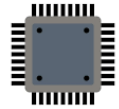
NN model is too confident in its predictions

Confidence = Pr (is the correct key)

WPI

How does it look like in practice?

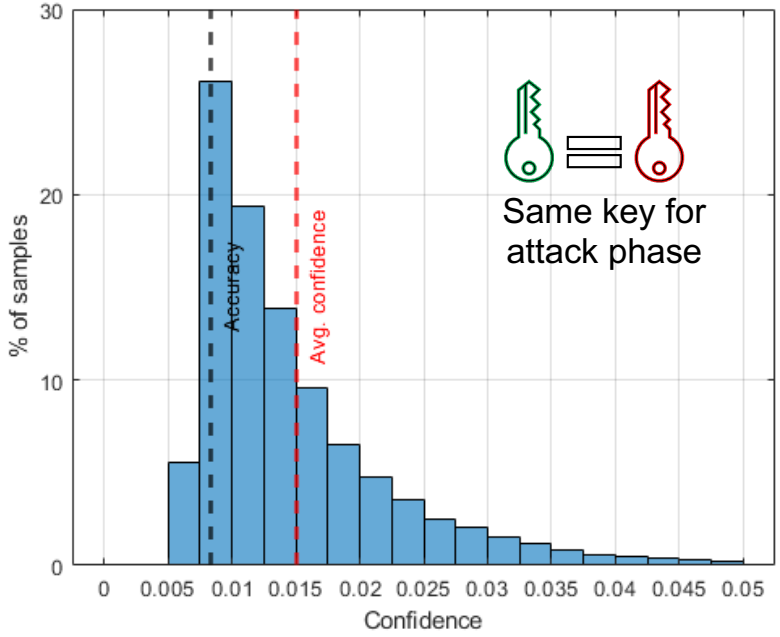
- ASCAD dataset [1]
- Models proposed in [2,3]

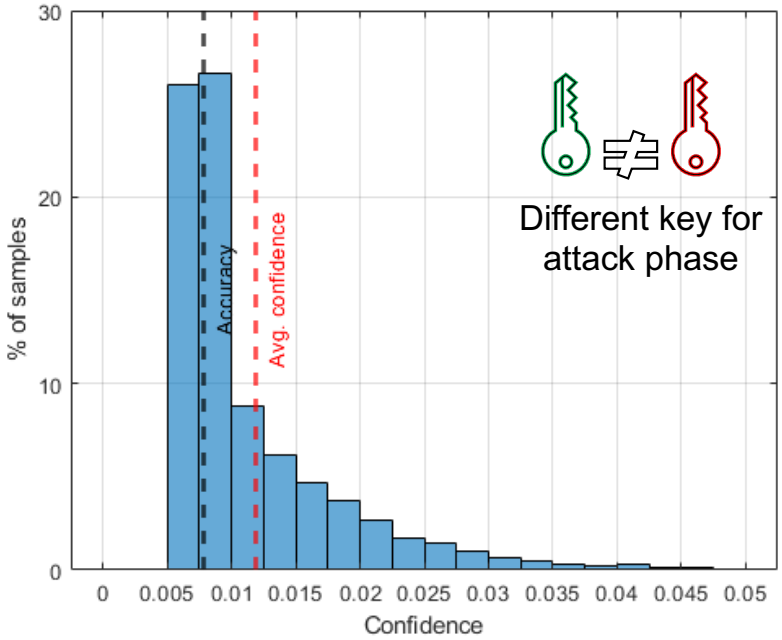


8-bit AVR

AES-128

Masked

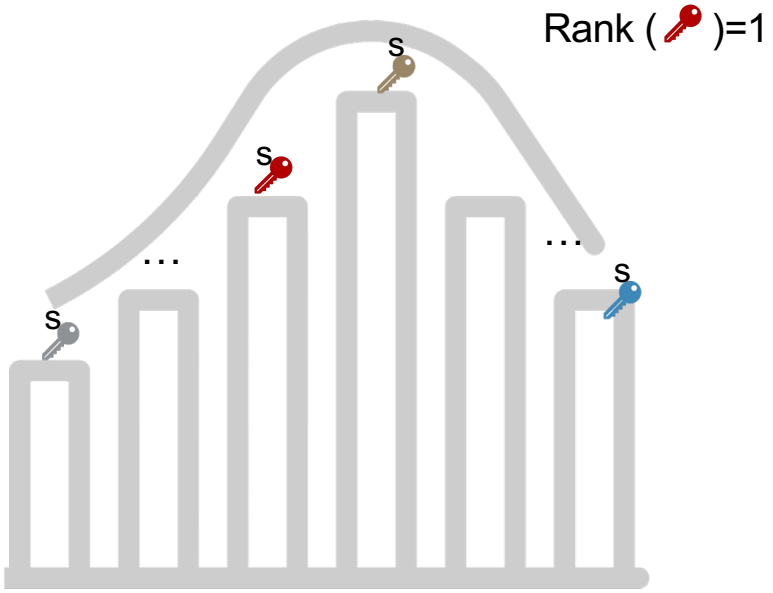




[1] Benadjila, Journal of Cryptographic Engineering, 2020. [2] Wouters, TCHES, 2020. [3] Wu, TETC, 2020.

WPI

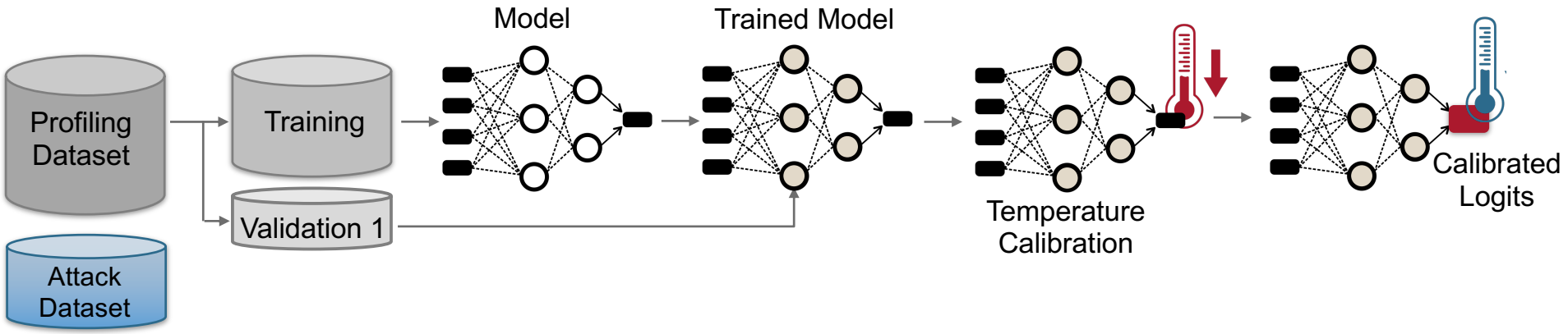
Temperature scaling



- Temperature calibration outputs a calibrated probability
  - Take the same raw NN scores, so-called logits
  - Divide them by a temperature T
  - Converting calibrated logits to probabilities
- This makes the output either sharper or softer.
  - T> 1: output becomes softer / less overconfident
  - T < 1: output becomes sharper / more confident.
- How to compute T?!

WPI

Finding T



Training a logistic regression model on the *validation set* to learn a scalar $T > 0$

WPI

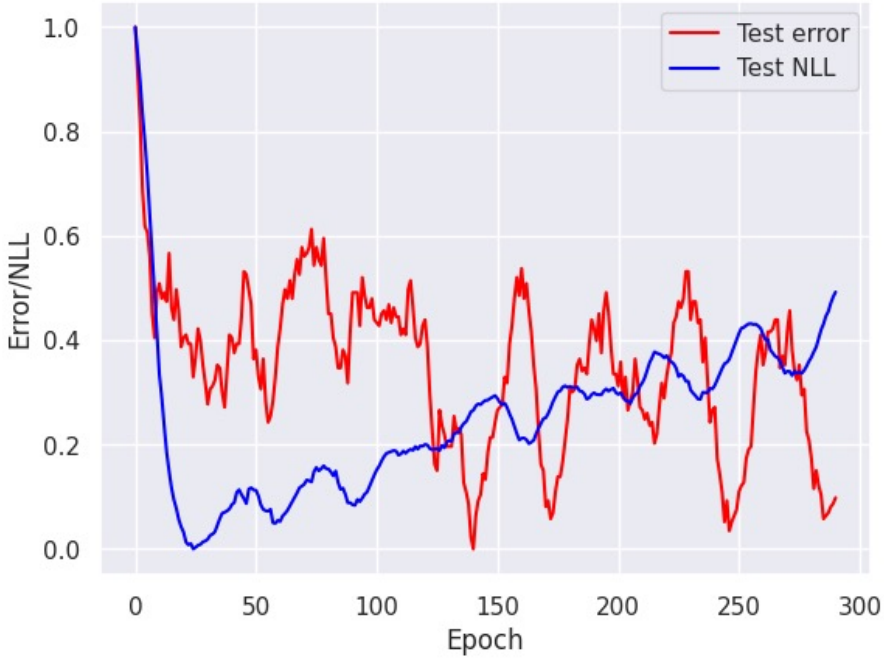
Some results…

Dataset	Model	Leakage	T (uncal.)	TGE_o (uncal.)	T (cal.)	TGE_o (cal.)	Std(T)
ASCADf	MLP	HW [1]	6.470	3368	0.950	2736	0.09
	CNN	ID [2]	3.677	508	0.910	442	0.24
ASCADr	MLP	HW [1]	7.976	3443	0.977	2686	0.73
	CNN	HW [1]	1.670	1851	0.983	1615	0.097
CHES-CTF	MLP	HW [1]	15.194	4403	1.028	3381	1.70

[1] Wu, TETC, 2020. [2] Wouters, TCHES, 2020.

Why SCA should not push for higher accuracy levels

- SCA goal: key recovery, not top-1 classification
  - Accuracy: predicted class only
  - GE: ranking quality across key hypotheses
- Problem: high accuracy ≠ well-calibrated probabilities
- Overfitting to NLL: better classification, worse probability distribution
  - Consequence: degraded GE despite acceptable accuracy
- Takeaway: monitor calibration, not accuracy alone



Test error and NLL of the MLP model trained on ASCAD-f datasets

WPI

Conclusion

- Temperature scaling: simple post hoc fix
- Design guideline: smaller, better-calibrated networks over larger, overconfident ones
  - Large capacity: lower classification error
    - But also: higher overconfidence risk
    - More epochs: possible NLL overfitting
    - Effect: weaker probability quality for SCA
- Link between temperature scaling and effective perceived information (PI)/cross-entropy (CE) [1,2]
  - "Raw" CE or PI can be misleading as attack-performance predictors
  - Temperature: the source of CE/PI ambiguity
  - Effective CE/PI by optimizing CE (or PI) over $\beta = \frac{1}{T}$
  - A practical way to pick an inverse temperature using the validation dataset, rather than sweeping β abstractly as in [1,2]

[1] Ito, TCHES, 2022. [2] Ito, TCHES, 2025.

Thank you

FGANJI@WPI.EDU