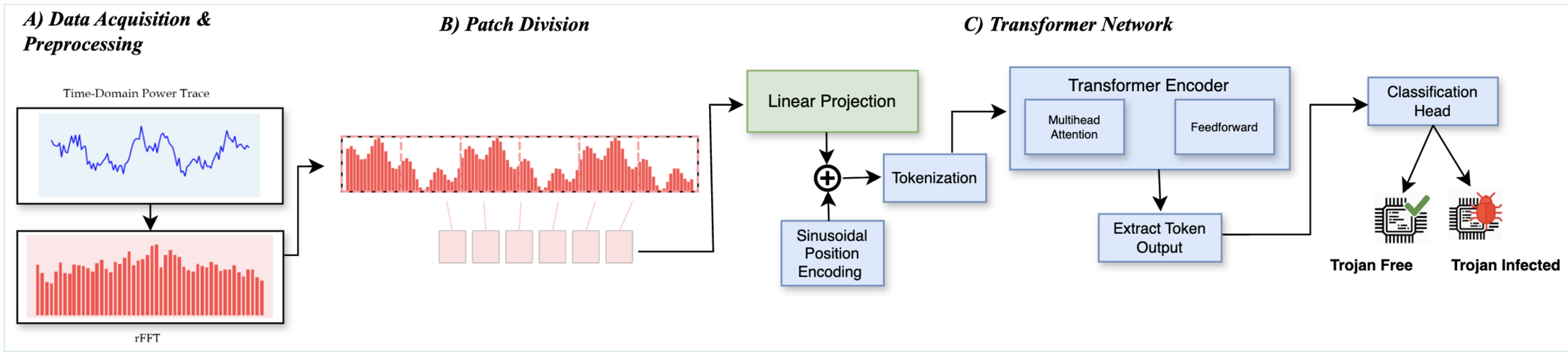
PATCH-FFT: Unmasking Dormant Hardware Trojans with Patch-Based Frequency-Domain Transformers

Hasala Senevirathne and Amin Rezaei



PATCH-FFT Pipeline. rFFT log-magnitude → 1D patch embedding (+ CLS, sinusoidal position) → Transformer encoder → binary classification head.

Introduction

Hardware Trojans (HTs). Malicious modifications inserted into untrusted IC supply chains, can leak cryptographic keys through power side-channel emissions. For information-leaking HTs, detection after the trigger fires is too late because the secret has already been exfiltrated; therefore, the only useful defense window is while the Trojan remains **dormant**.

The Gap. ML detectors on time-domain dynamic power (LSTM, HTM, CNN) reliably flag only active Trojans; a dormant Trojan perturbs the trace minimally and slips through. Static-power / IDDQ methods can target dormant logic but need multi-pad probing, on-die sensors, or population calibration.

Contributions

- Patch-based Transformer on frequency-domain power traces — local spectral features via patches, global harmonic context via self-attention.
- First evidence that rFFT representations expose **dormant** Trojan signatures invisible to time-domain dynamic-power methods.
- Interpretable layer-wise attention maps showing which frequency bands drive each decision, aligned with known leakage mechanisms.
- Validated on 8 information-leaking variants: **90.94% average accuracy**, stable under synthetic process variation.

Threat Model & Assumptions

Scope. Key-leakage HTs inserted during design or fabrication, evaluated on the public IEEE/TrustHub dataset reproduced on a commercial FPGA.

Deployment. Golden chips train the model; at inference it takes a **single externally measured power trace** (on-board shunt + oscilloscope) and returns a binary verdict — no multi-pad probing, on-die sensors, IDDQ, or invasive access.

Frequency-Domain Transformation

$$\tilde{X}_{\log}(k) = \frac{\log(|X_{\mathrm{rFFT}}(k)| + \varepsilon) - \mu_x}{\sigma_x + \varepsilon}$$

Each power trace x[n] (T = 2500) is mapped to its ⌊T/2⌋ + 1 = **1251** non-redundant rFFT coefficients (Hermitian symmetry). We take log-magnitude and apply per-trace z-score normalization, removing absolute scale and offset so the model learns relative spectral shape rather than raw power.

Patch Embedding & Transformer

$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V$$

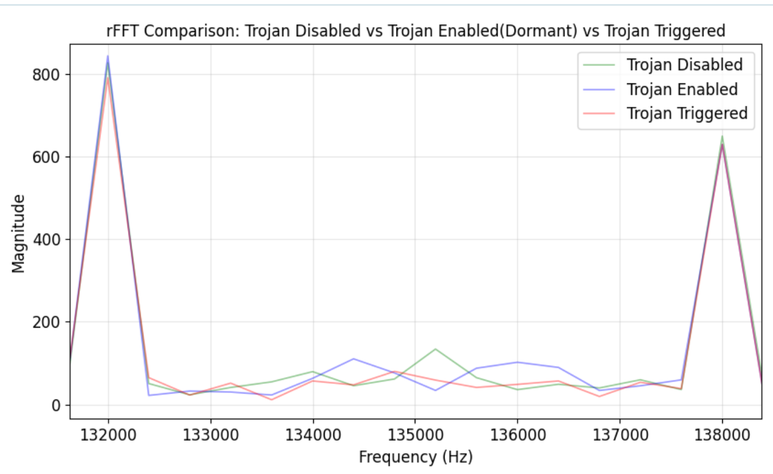
The 1251-bin spectrum is split into **N = 79 patches** (p = 16), each linearly projected to a d = 128 token; a learnable CLS token and sinusoidal positional encodings give an 80-token sequence. **L = 4** encoder blocks with **h = 8**-head self-attention and a 256-d feed-forward fuse local bands with global harmonic structure; the CLS embedding feeds a linear two-class head trained with label-smoothed cross-entropy.

Dataset & Training

IEEE HT Power & EM Side-Channel set [1]: AES-128 on a Xilinx Spartan-6 (45 nm) SAKURA-G FPGA, supply current sensed via an on-board shunt on the $V_{cc\text{-}int}$ rail and digitized at 25 °C. ~10,000 traces per configuration, each T = 2500 samples, over 8 key-leakage variants.

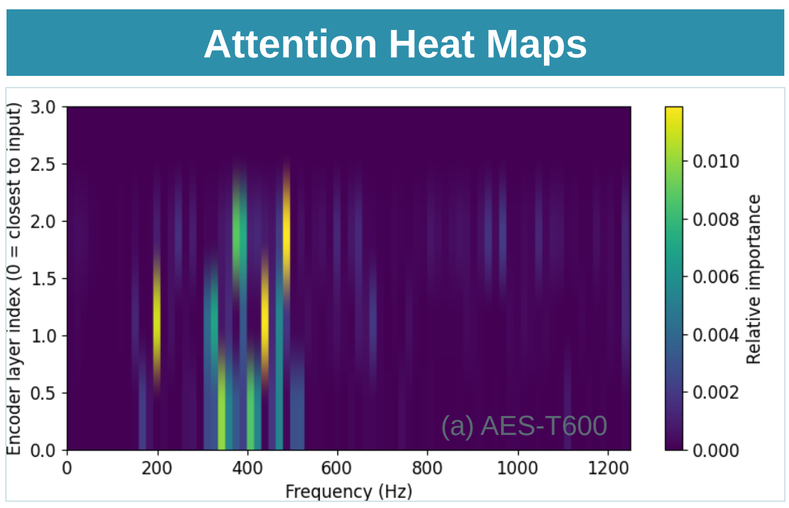
Protocol. Per benchmark, traces are pooled into TrojanDisabled / TrojanEnabled (or Triggered) and split 70/10/20, stratified and trace-disjoint. AdamW, lr 1e-3 (cosine, 2-epoch warm-up), label smoothing 0.02, batch 128, early stopping (patience 5).

Frequency-Domain Signatures of HTs



rFFT magnitude of AES-T400 (baseline vs dormant vs triggered). Zoomed to the band where modes differ most. Per-bin gaps are tiny; aggregated across patches under global self-attention they become discriminative.

Attention Heat Maps



Encoder layer × frequency attention. (a) AES-T600 : early layers respond broadly; deeper layers concentrate on narrow bands where dormant traces deviate from baseline. (b) AES-T800: attention collapses onto harmonics of the CDMA spreading sequence — the model learns physically meaningful spectral features, not spurious correlations.

Experimental Results — 90.94% average · 89.56% dormant · 92.31% triggered

Benchmark	Key-Leakage Mechanism	Native Channel	Dormant %	Active %
AES-T600	Leakage current modulated by key bits	Leakage current	77.3	83.6
AES-T2000	Increased leakage current for key bits equal to 0	Leakage current	88.4	92.1
AES-T700	CDMA-based modulation of dynamic power after trigger	Dynamic power	96.7	99.2
AES-T800	CDMA-based modulation of dynamic power after trigger	Dynamic power	100.0	100.0
AES-T1000	Increased dynamic power after trigger sequence	Dynamic power	95.5	97.8
AES-T400	Key bits via modulated RF signal on unused pin	RF / EM emission	89.2	92.4
AES-T1600	Key bits via RF signal after trigger sequence	RF / EM emission	82.8	84.3
AES-T1700	Key bits via RF signal driven by counter	RF / EM emission	86.6	89.1

Architecture & Training

Hyperparameter	Value
Input length F	1251
Patch size p	16
Model dim d	128
Attention heads h	8
Encoder layers L	4
FFN dim d_ff	256
Dropout	0.1
Label smoothing α	0.02
Optimizer	AdamW
Learning rate	1e-3 cos
Batch size	128

Comparison with State-of-the-Art Works

Method	Accuracy %
HTM [2]	92.2 (trig.)
Siamese LSTM [3]	86.78 (trig.)
PATCH-FFT (dormant)	89.56
PATCH-FFT (triggered)	92.31
PATCH-FFT (average)	90.94

Process Variation

σ (% RMS)	Dormant %	Δ clean
0	89.56	–
5	88.71	−0.85
10	86.94	−2.62
15	84.10	−5.46
20	80.42	−9.14

Inference

```
DetectTrojan(x, model, pos_class):
  X ← rFFT(x)
  X_log ← log(|X| + ε)
  X̃ ← z-score(X_log)
  h ← Encodear(Patchify(X̃))
  ŷ ← argmax softmax(W·h + b)
 if ŷ=1
     return Enabled
else
     return Disabled
```

Conclusion & Future Work

PATCH-FFT is the first frequency-domain detector to reliably flag dormant information-leaking Trojans while matching prior art on active ones and stays stable under process variation. Modulation-based HTs with sharp spectral peaks are caught near-perfectly; subtle leakage-current variants remain hardest.

Future Work. Time + frequency feature fusion for leakage-current Trojans, cross-benchmark transfer, EM-trace validation, and silicon validation on fabricated ASICs.

Selected References

[1] R. Yasaei et al, “Hardware Trojan Power & EM Side-Channel dataset,” In IEEE Dataport, 2021. [Online]. Available: https://dx.doi.org/10.21227/9fwb-8978
[2] S. Faezi,et al., “Brain-inspired golden chip free hardware Trojan detection,” In IEEE Transactions on Information Forensics and Security, vol. 16, pp. 2697-2708, 2021.
[3] A. Nasr, et al., “A Siamese deep learning framework for efficient hardware Trojan detection using power side-channel data,” In Scientific Reports, vol. 14, no. 13013, pp. 1-13, 2024.
[4] A. Vaswani, et al., “Attention is all you need,” In International Conference on Neural Information Processing Systems (NIPS), pp. 6000-6010, 2017.
[5] A. Dosovitskiy et al., “An image is worth 16×16 words: Transformers for image recognition at scale,” In arXiv:2010.11929, 2020.
[6] Y. Nie et al., “A time series is worth 64 words: Long-term forecasting with transformers,” In International Conference on Learning Representations (ICLR), 2023.

Computer Engineering & Computer Science • Hung Family College of Engineering • California State University, Long Beach
Contact: Hasala.Senevirathne01@student.csulb.edu • Amin.Rezaei@csulb.edu