Risk-Aware and Explainable Framework for Ensuring Guaranteed Coverage in Evolving Hardware Trojan Detection

Rahul Vishwakarma and Amin Rezaei

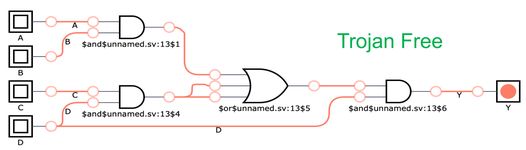
IEEE/ACM ICCAD 2023 INTERNATIONAL CONFERENCE ON COMPUTER-AIDED DESIGN 42nd Edition

Introduction & Contributions

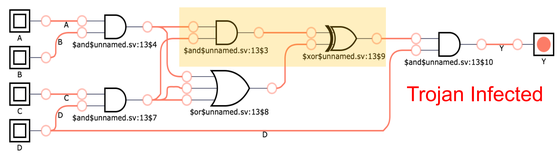
As the semiconductor industry has shifted to a fabless paradigm, the risk of hardware Trojans being inserted at various stages of production [1] has also increased. Recently, there has been a growing trend toward the use of machine learning solutions [2] to detect hardware Trojans more effectively, with a focus on the accuracy of the model as an evaluation metric. However, in a high-risk and sensitive domain, we cannot accept even a small misclassification. Additionally, it is unrealistic to expect an ideal model, especially when Trojans evolve over time. Therefore, we need metrics to assess the trustworthiness of detected Trojans and a mechanism to simulate unseen ones.

We generate evolving hardware Trojans using conformalized generative adversarial networks and offer an efficient approach to detecting them based on a non-invasive algorithm-agnostic [3] statistical inference framework that leverages the Mondrian conformal predictor. The method acts like a wrapper over any of the machine learning models and produces set predictions along with uncertainty quantification for each new detected Trojan for more robust decision-making. In the case of a NULL set, a novel method to reject the decision by providing a calibrated explainability.

Hardware Trojan Insertion



When A through D are all switched to 1, the output of Y is 1 as expected.



When we test out the same inputs with the Trojan inserted code, the output of Y is 0 when all pins are activated. So, the functionality of both RTL designs behave almost identical.

Conventional Approach & Challenges

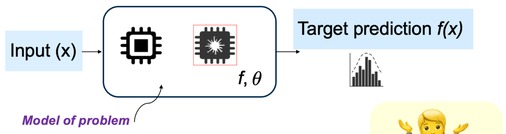
Conventional methodologies like Side-channel power analysis exploits power consumption patterns to identify anomalous behavior indicative of a Trojan. Temperature analysis monitors localized differentials, wireless transmission power analysis assesses electromagnetic emissions, focusing on signal anomalies as Trojan markers. As hardware Trojans continue to evolve, they may employ sophisticated techniques to bypass the scrutiny of traditional detection methods.

Notion of evolution:

$$E_{HT} \rightarrow HT' \blacksquare HT_{structural\_changes}$$

Modern integrated circuits are incredibly complex, making it challenging to spot tiny hardware Trojans. Detecting these threats early in the design process is crucial for effective security. However, doing so in a cost-efficient manner presents yet another obstacle in the realm of hardware Trojan detection.

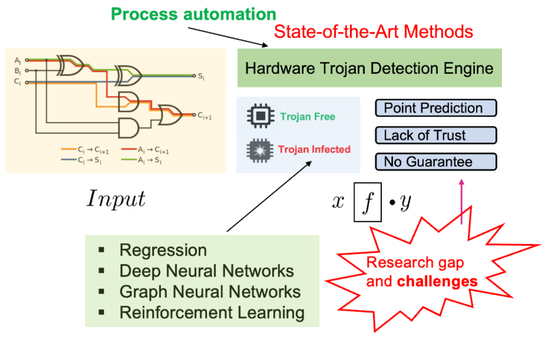
Machine Learning



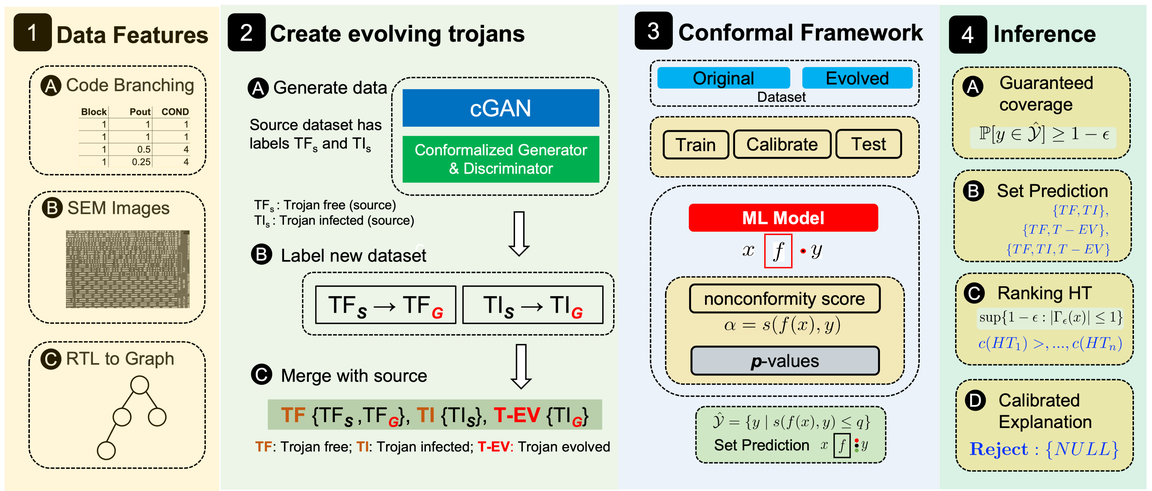
- Given a dataset $\{(x_1, y_1), \ldots, (x_N, y_N)\}$
- Pick θ that minimizes the Loss

$$\mathcal{L}(\theta) = \frac{1}{N} \sum_i L(f_\theta(x_i), y_i)$$

Machine Learning based solution



Proposed Solution to Identify Evolving Hardware Trojan



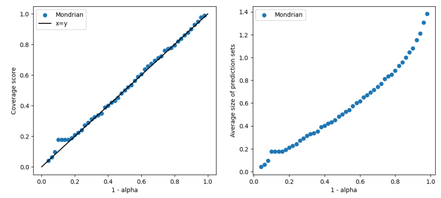
Experimental Results

Dataset: GAINESIS (binary classification) and Chip-Level Trust-Hub Trojans (including evolving hardware trojans)

Performance metrics of conformal inference

sig	mean_err	avg_c	n_correct	mean_T-EV
0.05	0.049	1.040	589	0.012
0.1	0.102	0.941	556	0.045
0.2	0.204	0.812	493	0.133
0.3	0.303	0.701	431	0.220
0.4	0.406	0.596	367	0.319
0.5	0.504	0.497	307	0.423
0.6	0.604	0.397	245	0.536
0.7	0.702	0.298	184	0.650
0.8	0.798	0.202	125	0.764
0.9	0.900	0.100	61	0.884

Coverage and Efficiency of Predictions



Ranking Mechanism

Adoption of confidence for risk-aware ranking on trust-hub chip-level trojan dataset.

$$\text{Confidence}(x) = \sup\{1 - \epsilon : |\Gamma_\epsilon(x)| \leq 1\}$$

$$\alpha_{0.05}(\text{ circuit } 12) = \{T - EV\}_{C=0.88}$$
$$\alpha_{0.05}(\text{ circuit } 13) = \{T - EV\}_{C=0.81}$$
$$\alpha_{0.05}(\text{ circuit } 14) = \{T - EV\}_{C=0.61}$$

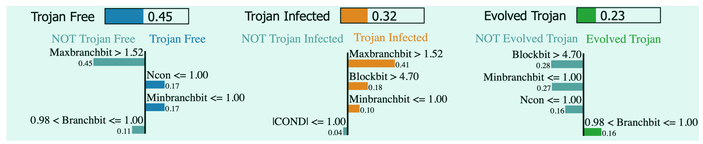
Model says – "*I don't know*"

Algorithm 1: Prediction with Reject Option

Input: model, instance
Output: prediction

```
confidence_scores = model.predict(instance);
if max(confidence_scores) < threshold then
    return "I don't know";
return class_with_highest_confidence(confidence_scores);
```

Calibrated Explanation For Rejecting A Prediction Made By The Model



How It's Going...

Takeaway 1: The need for algorithm-agnostic and explainability-aware rejection of predictions along with guaranteed coverage of decisions.

Takeaway 2: While there's no silver bullet for zero-day attacks, adopting a proactive risk-aware defense strategy significantly reduces the attack.

References

[1] J. Francq and F. Frick, "Introduction to hardware trojan detection methods," in 2015 Design, Automation & Test in Europe Conference & Exhibition (DATE). IEEE, 2015, pp. 770–775.

[2] T. C. . KÁNoylÁNu, C. R. W. Reinbrecht, A. Gebregiorgis, S. Hamdioui, and M. Taouil, "A survey on machine learning in hardware security," ACM 2023.

[3] G. Shafer and V. Vovk, "A tutorial on conformal prediction." Journal of Machine Learning Research, vol. 9, no. 3, 2008.