# Uncertainty-Aware Hardware Trojan Detection Using Multimodal Deep Learning

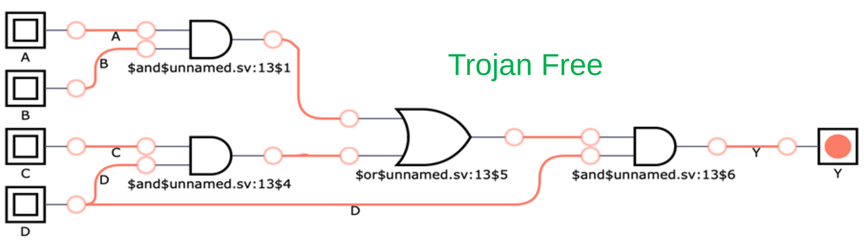
Rahul Vishwakarma and Amin Rezaei

## Introduction & Contributions

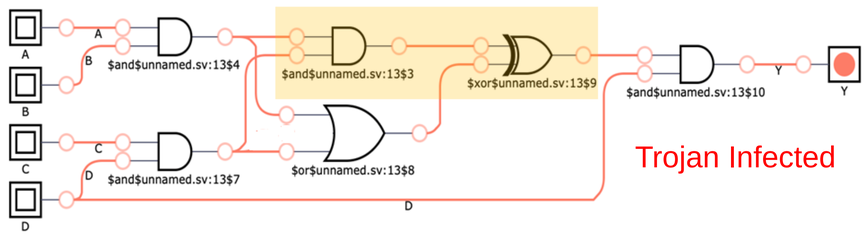
The risk of hardware Trojans being inserted at various stages of chip production has increased in a zero-trust fabless era. To counter this, various machine learning solutions have been developed for the detection of hardware Trojans. While most of the focus has been on either a statistical or deep learning approach, the limited number of Trojan-infected benchmarks affects the detection accuracy and restricts the possibility of detecting zero-day Trojans. To close the gap, we employ Generative Adversarial Networks (GANs) to amplify our data in two alternate representations of modalities: a graph and a tabular, which ensures a representative distribution of the dataset.

We propose a multimodal deep learning approach called **NOODLE** to detect hardware Trojans and evaluate the results from both early fusion and late fusion strategies. We also estimate the uncertainty quantification metrics of each prediction for risk-aware decision-making. The results validate the effectiveness of the proposed hardware Trojan detection technique and pave the way for future studies utilizing multimodality and uncertainty quantification to tackle other hardware security problems.

## Hardware Trojans




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

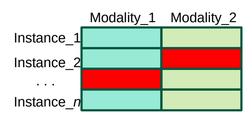



When we test out the circuit with the Trojan inserted component, the output of Y is 0 when all pins are 1.

**Note:** Stealth is advantageous to attackers inserting hardware Trojans, so in a high-level design, we expect a Trojan to be inserted in rare nets that can skip the testing phase.

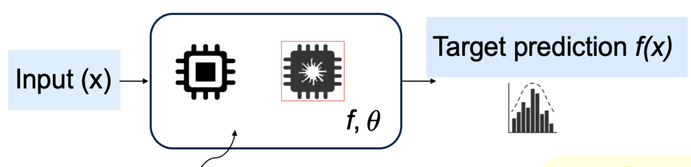
## Conventional Approach & Challenges

Conventional methodologies like Register-Transfer Level (RTL) based dataset for hardware Trojan detection (i.e., Random Forest [1], Support Vector Machines [2], etc.) and the use of Deep Learning techniques, specifically leveraging image classification have limitations in a risk sensitive domain like hardware Trojan classification. One of the challenge of concern is **missing modalities** while working with multimodal data.




The missing modalities ca be handled by data imputation; however, we choose to use GAN complemented with conformal prediction to generate quality data. In real-time, we also have challenges like working with less datapoints and the cases of highly imbalance dataset.

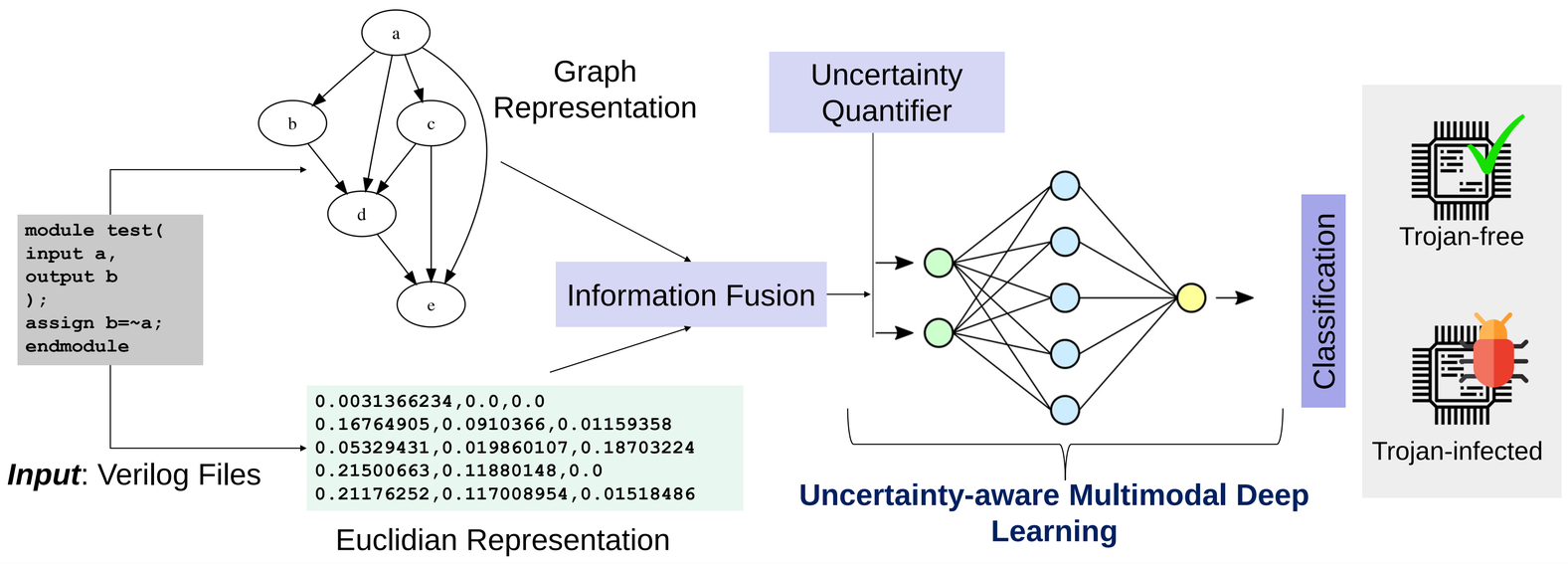
## Machine Learning




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

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


## Proposed Multimodal Approach for Hardware Trojan Detection




## Uncertainty Aware Learning

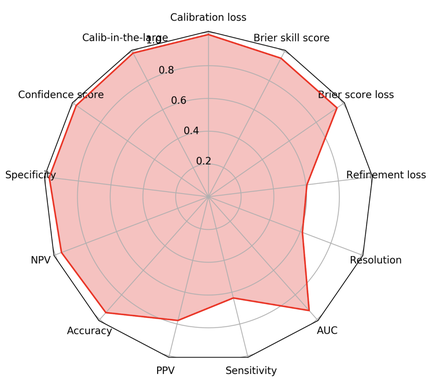
**Algorithm 1** Uncertainty aware multimodal approach

**Input:** Graph and euclidean dataset
**Output:**
$D$ (Trojan detection), $r_E$ (Conformal prediction regions).
**Procedure:**

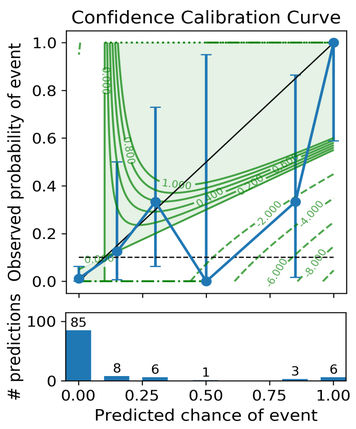
**for** each circuit $C$ **do**
  Convert $C$ to $G(graph), T$ (euclidean).
  **if** missing modalities **then**
    Perform conformalized GAN for imputation.
  **end if**
**end for**
Feed modalities to Deep Learning classifier.
**for** each modality $M$ **do**
  Get new unlabeled examples.
  Evaluate conformal predictors, compute $p$-values.
  **for** each class label $y^{(j)}$ **do**
    Compute $\hat{p}_j$ for combined hypothesis.
  **end for**
  Use uncertainty-aware fusion, perform early and late fusion.
**end for**
Choose winning fusion method.
Return $D, r_E$.

## Experimental Results

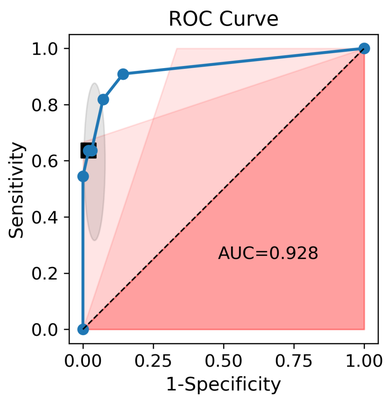
**Dataset:** Chip-level Trust-Hub Trojans based on code branching [3] and graph representation [4]




(a)




(b)




(c)

a) The **radar plot** offers a comprehensive evaluation of predictor performance across diverse dimensions, including discrimination metrics (AUC, resolution, and refinement loss) and combined metrics (Brier score and Brier skill score)
b) The **confidence calibration curve** evaluates alignment between a classification model's predicted and observed probabilities, revealing a deviation due to dataset imbalance
c) The **ROC-AUC curve** evaluates the balance between sensitivity and specificity, with a value of 0.928 indicating effective discrimination between Trojan-free and Trojan-infected cases

| Dataset | Brier Score |
|---|---|
| Graph-based Data | 0.1798 |
| Tabular-based Data | 0.1913 |
| NOODLE - Early Fusion (Graph + Tabular) | 0.1685 |
| NOODLE - Late Fusion (Graph + Tabular) | 0.1589 |

Brier score comparison for different modalities

## Conclusion

We addressed the gaps in the current machine learning approaches for the identification of hardware Trojans by proposing an uncertainty-aware hardware Trojan detection framework using multimodal deep learning. The utilization of multimodality and uncertainty quantification shows great potential for addressing other critical challenges in hardware security too.

## References

[1] K. Hasegawa et al., “Trojan-feature extraction at gatelevel netlists and its application to hardware-trojan detection using random forest classifier,” in ISCAS, 2017.
[2] C. Bao et al., “On application of one-class svm to reverse engineering-based hardware trojan detection,” in ISQED, pp. 47–54, 2024.
[3] H. Salmani et al., “Trust-hub trojan benchmark for hardware trojan detection model creation using machine learning,” 2022 [Online].
[4] S. -Y. Yu et al., "Hw2vec: A graph learning tool for automating hardware security,"2021 [Online].