Uncertainty-Aware Hardware Trojan Detection Using Multimodal Deep Learning

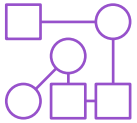
Rahul Vishwakarma
Graduate Research Assistant
California State University Long Beach

Amin Rezaei
Assistant Professor
California State University Long Beach

Agenda

Introduction

Problem statement and preliminaries



NOODLE

Uncertainty aware multimodal approach

Experiments

Evaluation of performance metrics

Hardware Trojans

Applied machine learning to detect hardware Trojans

Hardware Trojans

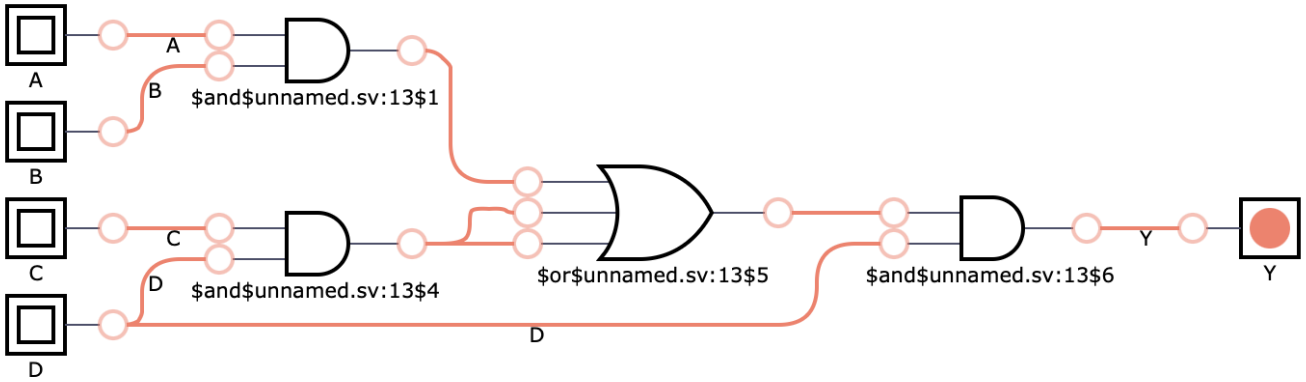


Figure 1: Trojan free circuit

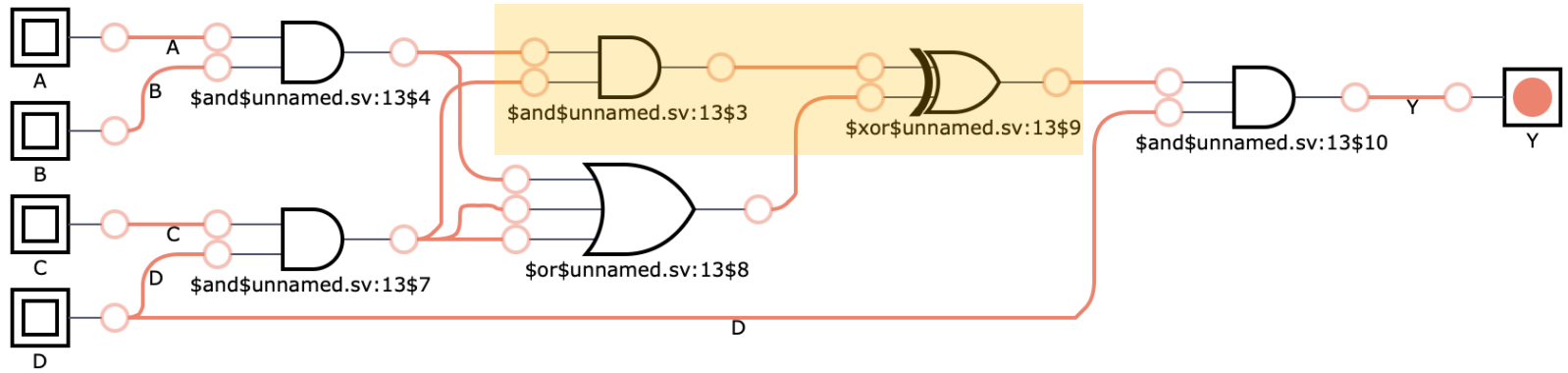
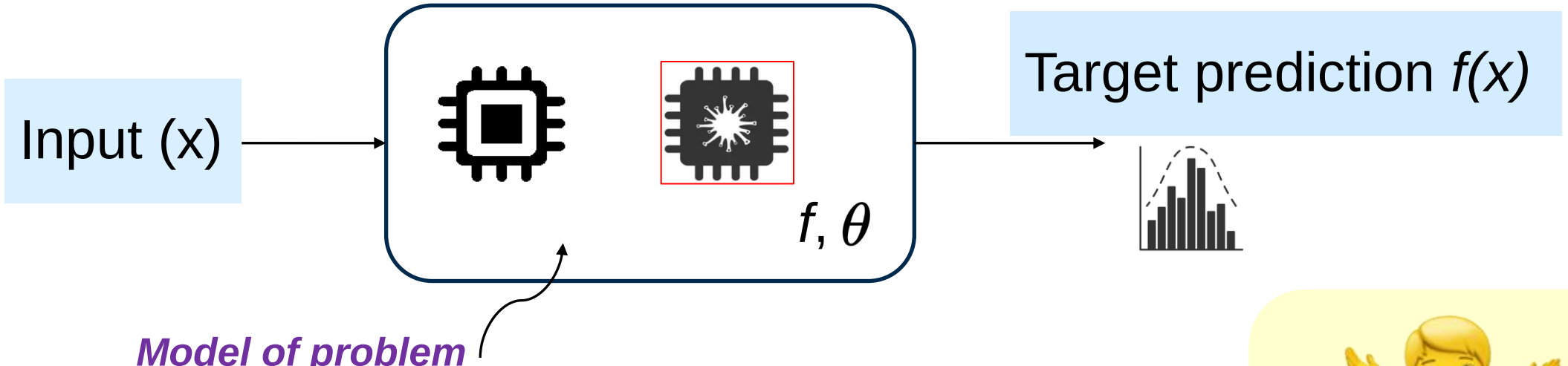


Figure 1: Trojan inserted circuit

Role of machine learning



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

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

All models are wrong, but some are useful!

Related Works

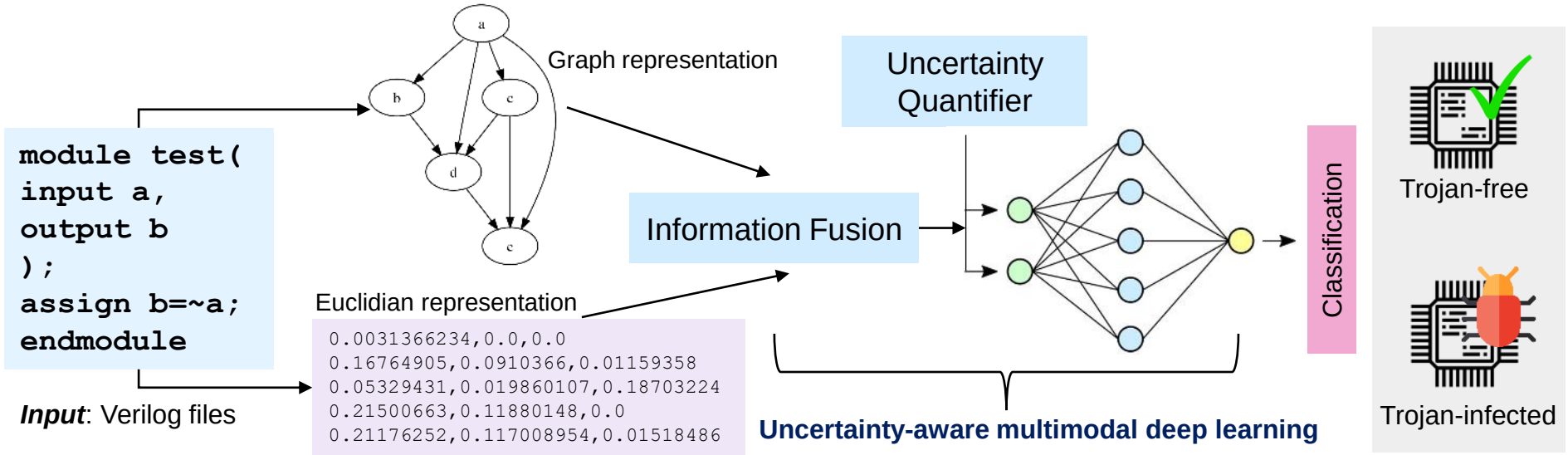
Prior work and challenges

- ML approaches for hardware Trojan detection.
  - RTL based – Random Forest, Support Vector Machines.
  - Deep learning approach (image classification).
- Uncertainty aware multimodal learning in healthcare.
- Challenges in hardware security domain
  - Learning with less data.
  - Highly imbalanced data.

NOODLE

Uncertainty-aware Hardware Trojan Detection using Multimodal deep Learning

Proposed solution



NOODLE Framework

Pseudocode

Algorithm 1: Uncertainty-aware information fusion

Input : Number of data sources N;
Training sets for each data source $T_1 = \{(x_1^{(1)}, y_1), \ldots, (x_n^{(1)}, y_n)\}, \ldots, T_N = \{(x_1^{(N)}, y_1), \ldots, (x_n^{(N)}, y_n)\}$, where $x_i^{(j)}$ is the ith data point belonging to the jth data source and y_i is the class label of the ith data point;
Number of classes M;
Class labels $y^{(i)} \in Y = \{y^{(1)}, y^{(2)}, \ldots, y^{(M)}\}$;
Classifiers $S_1, \ldots, S_N$ for each data source;
Confidence level E.

Output: Conformal prediction regions $r_E = \{y^{(j)} : \hat{p}_j > 1 - E, y^{(j)} \in Y\}$.

1 Get the new unlabeled example w.r.t each data source $x_{n+1}^{(1)}, \ldots, x_{n+1}^{(N)}$.
2 Evaluate conformal predictors and classifiers $S_1, \ldots, S_N$ corresponding to each data source, compute p-values $p_j^{(i)}$, where $i = 1, \ldots, N$ corresponds to the ith data source and $j = 1, \ldots, M$ corresponds to the jth class label.
3 **for** *each class label* $y^{(j)}$, $j = 1, \ldots, M$ **do**
4 Compute p-value, $\hat{p}_j$, of combined hypothesis from N modalities
5 **return** r_E.

Algorithm 2: Multimodal deep learning

Input : RTL-level files (Verliog) of circuits
Output: Decision (D) = Trojan-free or Trojan-infected

1 **for** *each circuit* C **do**
2 Convert C to Graph data **G** and Euclidean data **T**.
if ∃ *missing modalities* **then**
3 perform GAN to impute the missing modality.
4 Feed the modalities to *CNN*-based classifier.
for *each modalities* M **do**
5 Use Algorithm 1 for uncertainty-aware information fusion.
6 Perform early fusion.
7 Perform late fusion.
8 Choosing the winning fusion method.
9 **return** D.

Novelty

- Multimodal learning approach in hardware security.
  - Euclidian and graph.
- Information fusion with uncertainty quantification.
  - Model fusion strategy employs p-values.
- Mitigating missing modalities.
  - GANs to augment data and overcome small datasets issue.

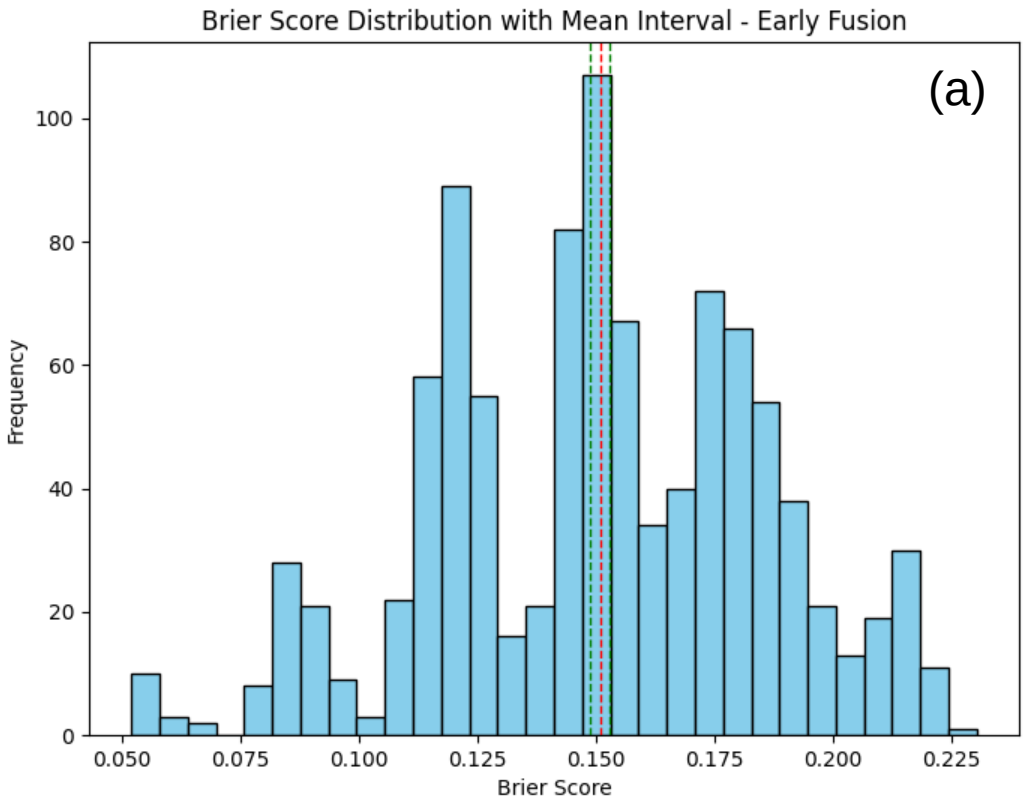
Results

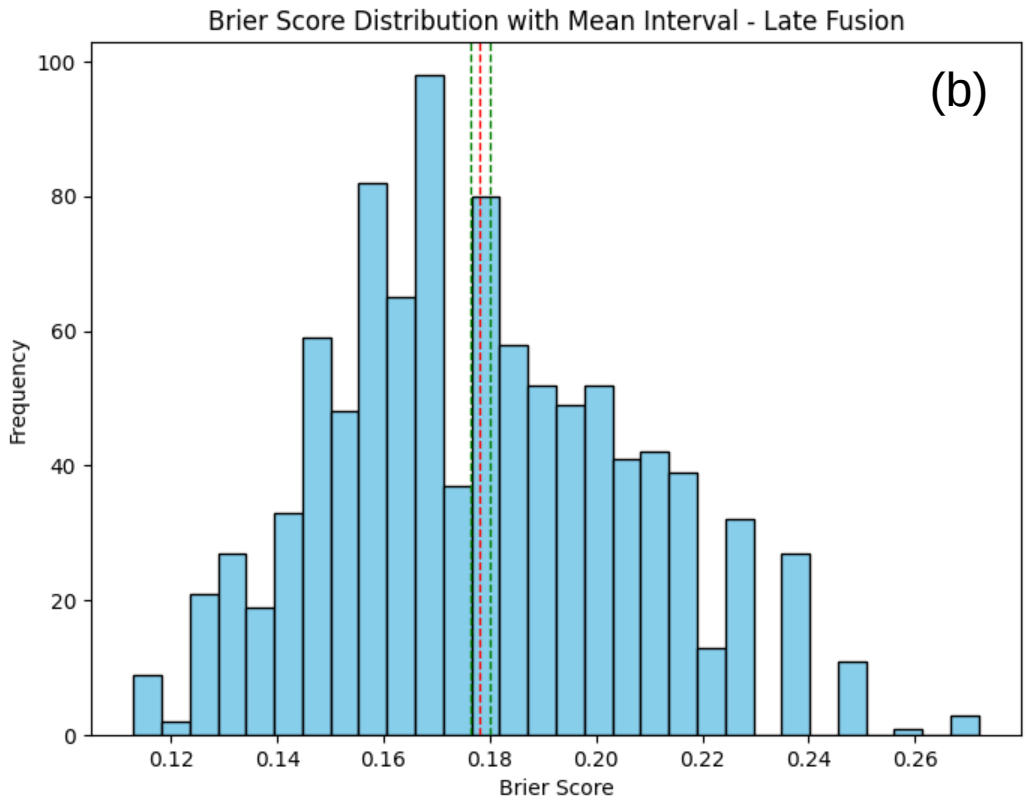
Dataset

- Chip-level Trojan: Trust-Hub[1]
  - Modal 1: Represent circuit using AST to Euclidian data.
  - Modal 2: Graph representation of circuit.
    - Convert each graph to “Tensor”.

[1] H. Salmani, M. Tehranipoor, S. Sutikno, and F. Wijitrisnanto, “Trust-hub Trojan benchmark for hardware trojan detection model creation using machine learning,” 2022. [Online]. Available: https://dx.doi.org/10.21227/ px6s-sm21

Performance metrics





NOODLE's Brier score (a) Early fusion (b) Late fusion

Brier score

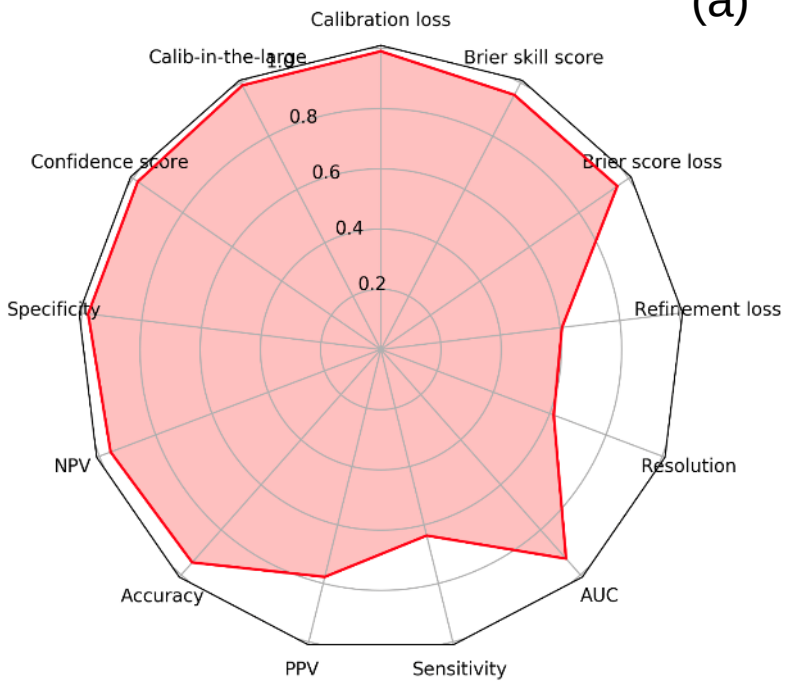
Brier score comparison for different modalities

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

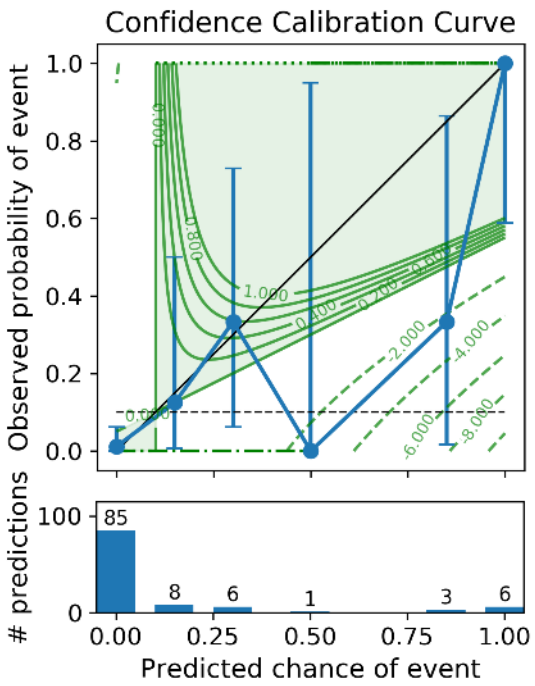
https://github.com/cars-lab-repo/NOODLE[1]

Experimental results

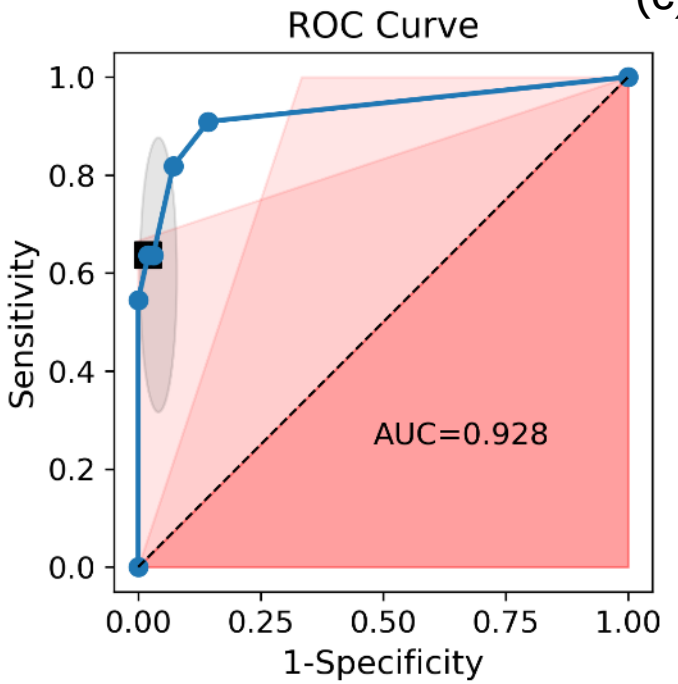
(a)



(b)



(c)



NOODLE's (a) consolidated metrics (b) confidence calibration curve, (b) ROC-AUC curve

DATE 24

Conclusion

Key takeaways

- Missing modalities – GAN.
- Uncertainty aware multi-modal fusion.
- Ensemble of fusion approach.

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

Rahul Vishwakarma

Amin Rezaei

Computer Architecture, Reliability, and Security Laboratory (CARS-Lab)

California State University Long Beach