

From Trigger to Impact: Knowledge-Graph Reasoning and Risk-Aware Classification for Hardware Trojan Detection

Yang Zhang¹, Xing Hu^{2*}, Xiaowen Chen¹, Huan Guo¹, Zhenyu Zhao¹, Sheng Liu¹

¹College of Computer Science and Technology, National University of Defense Technology, China

²School of Physics and Electronic Science, Changsha University of Science and Technology, China

Abstract—Hardware Trojans (HTs) in modern ICs pose severe threats to system security. Existing methods often treat HT detection as a binary classification task, overlooking functional behavior and impact. This work introduces an impact-aware framework that models triggers, payloads, and attack targets, forming localized subgraphs that reflect activation dependencies and interactions. These are embedded into a novel knowledge-graph representation—enabling explainable reasoning based on structural, functional, and criticality semantics. A risk-aware classifier then ranks HT severity, helping engineers prioritize responses. Unlike prior approaches, our method not only detects HTs but explains their intent and impact. Experiments on Trust-Hub benchmarks show a 14.05% accuracy gain over state-of-the-art methods, with enhanced interpretability bridging low-level analysis and high-level security.

Index Terms—Hardware Trojan detection, subgraph expansion, security alarm levels, knowledge graph reasoning

I. INTRODUCTION

The rise of stealthy hardware Trojans (HTs) in complex, globalized IC supply chains highlights the urgent need to move beyond traditional binary classification approaches that simply label circuits as Trojan-infected or clean, toward more robust and systematic detection methods. Existing HT detection methods—ranging from signal-activity heuristics to machine learning and information-flow tracking—suffer from high false positives, limited interpretability, and narrow threat coverage, highlighting the need for a unified, risk-aware approach grounded in semantic analysis. Trigger-based heuristics often misclassify benign modules like counters due to their naturally low controllability, highlighting the insufficiency of rarity as a detection criterion. This work addresses the issue by introducing a multi-dimensional framework that jointly analyzes triggers, payloads, and attack targets for accurate, explainable, and risk-aware HT diagnosis. We present a multi-dimensional characterization of HTs that jointly incorporates trigger mechanisms, payload logic, and attack targets, moving beyond single-dimensional rarity-based heuristics. We are the first to introduce knowledge-graph reasoning into HT detection, enriching structural subgraphs with controllability imbalance, functional roles, and target criticality, thereby enabling semantic and intent-aware diagnosis. We propose a graded

alarm classification framework that not only detects the presence of HTs but also quantifies their severity, providing risk-aware diagnostics that distinguish between benign anomalies and high-impact malicious modifications.

II. METHODOLOGY

The proposed methodology establishes a multi-dimensional diagnostic paradigm for gate-level HTs, centered on the interplay of triggers, payloads, and attack targets, as illustrated in Fig. 1. Suspicious structures are not treated in isolation, but abstracted into localized subgraphs that capture their functional and security dependencies. These subgraphs are embedded into a knowledge-graph representation, where structural motifs are enriched with semantic context to support inference about malicious intent. A graded alarm mechanism further maps each candidate subgraph onto a spectrum of risk levels, reflecting the completeness of the trigger–payload–target chain and the criticality of the affected nodes.

To identify stealthy triggers, we introduce a controllability imbalance metric that captures asymmetric state reachability and highlights hard-to-activate nodes. Payload-driven analysis traces the functional impact of HTs by propagating from triggers to identify logic patterns—like overrides or redirections—that can alter, disable, or degrade circuit behavior. Attack target analysis differentiates between critical and non-critical nodes to assess the true impact of HTs, linking structural anomalies to their security significance. By tracing how modified signals propagate to sensitive sinks, the framework assigns graded risk levels, enabling principled and interpretable threat classification.

Subgraph expansion aggregates triggers, payloads, and targets into cohesive graph regions that capture their structural dependencies. We model suspicious subgraphs as knowledge graphs enriched with controllability, functionality, and criticality attributes, capturing both structural and semantic relationships. By combining rule-based and graph-neural reasoning, the framework identifies malicious patterns with improved accuracy and interpretability. We propose a graded alarm classification system that evaluates suspicious subgraphs based on the presence of triggers, functional payloads, and targeted attack points. This risk-based stratification translates detection results into actionable diagnostics, enabling engineers to prioritize responses by severity.

* Corresponding author: Xing Hu, email: 009377@csust.edu.cn.

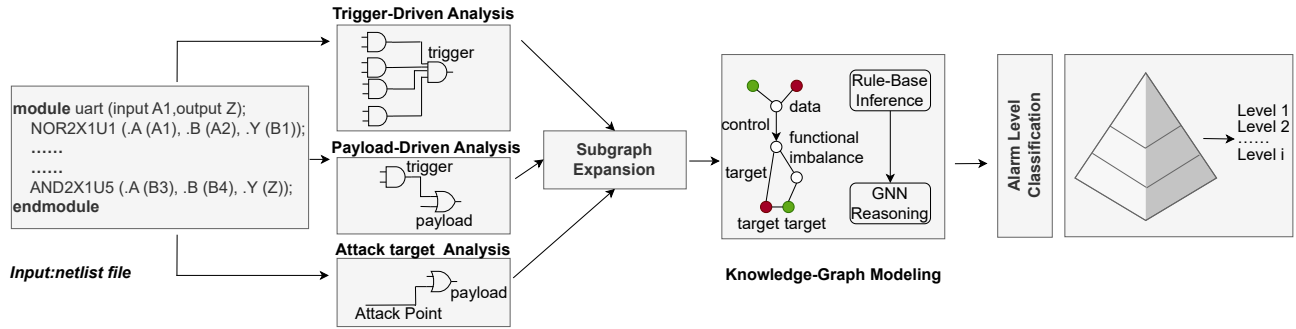


Fig. 1: Overall framework of the proposed methodology.

III. EXPERIMENTS

Table I demonstrates that our framework accurately identifies HT types, functionality, and structural components across Trust-Hub benchmarks, offering both detection and semantically grounded, risk-aware diagnostics through alarm-level classification. The results show that our framework accurately captures diverse HT types and their functional impacts, achieving 95% classification accuracy. It effectively distinguishes high-risk cases like control-targeted HTs from medium-risk or ambiguous ones, while also revealing challenges in borderline scenarios such as DoS vs. change-function misclassifications.

Compared to prior ML/DL-based methods [1]–[3] and GNN4HT [4], our framework advances beyond binary detection by performing multi-dimensional decomposition and knowledge-graph reasoning, enabling structural interpretation, impact assessment, and risk stratification. This holistic approach boosts functionality classification accuracy to 95%, outperforming GNN4HT’s 80.95% while offering deeper, more actionable diagnostics.

IV. CONCLUSION

This paper proposes a knowledge-graph-based framework for HT detection that analyzes triggers, payloads, and attack

targets to enable semantic reasoning and risk-aware diagnostics. Unlike traditional heuristics or binary classifiers, our approach explains both how HTs operate and how severe their impact is. Experiments on Trust-Hub benchmarks show a 14.05% improvement in functionality classification accuracy over state-of-the-art methods, with enhanced interpretability.

ACKNOWLEDGMENT

This work was supported by the Natural Science Foundation of Hunan Province, China, under Grant No. 2026JJ50512.

REFERENCES

- [1] K. Hasegawa, K. Yamashita, S. Hidano, K. Fukushima, K. Hashimoto, and N. Togawa, “Node-wise hardware trojan detection based on graph learning,” *IEEE Transactions on Computers*, vol. 74, no. 3, pp. 749–761, 2025.
- [2] W. Chen, Z. Bai, G. Pan, and J. Wang, “A fast modularity hardware trojan detection technique for large scale gate-level netlists,” *Computers & Security*, vol. 148, no. 000, 2025.
- [3] R. Yasaei, L. Chen, S. Y. Yu, and M. A. A. Faruque, “Hardware trojan detection using graph neural networks,” *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, vol. 44, no. 1, pp. 25–38, 2025.
- [4] L. Chen, C. Dong, Q. Wu, X. Liu, X. Guo, Z. Chen, H. Zhang, and Y. Yang, “Gnn4ht: A two-stage gnn-based approach for hardware trojan multifunctional classification,” *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 44, no. 1, pp. 172–185, 2025.

TABLE I: Detection Results on Representative Trust-Hub Benchmarks

Benchmark	HT Type	Trigger	Payload	Attack Target	HT Functionality	Alarm Level
RS232-T1000	CF	Yes	OR → AND	Non-Critical	CF (✓)	Level 2
RS232-T1100	CF	Yes	OR → AND	Non-Critical	CF (✓)	Level 2
RS232-T1200	CF	Yes	OR → AND	Non-Critical	CF (✓)	Level 2
RS232-T1300	CF	Yes	OR → AND	Non-Critical	CF (✓)	Level 2
RS232-T1400	CF	Yes	OR → AND	Non-Critical	CF (✓)	Level 2
RS232-T1500	CF	Yes	OR → AND	Non-Critical	CF (✓)	Level 2
RS232-T1600	CF	Yes	OR → AND	Non-Critical	CF (✓)	Level 2
S15850-T100	CF, DoS	Yes	AND → MUX	Non-Critical	CF (✓)	Level 2
S35932-T100	CF, LI	Yes	AND → MUX	Non-Critical	CF (✓)	Level 2
S35932-T200	DoS	Yes	AND → NOR	Non-Critical	CF (×)	Level 2
S35932-T300	DoS, DP	Yes	AND → INV	Non-Critical	DoS (✓)	Level 2
S38417-T100	CF, DoS	Yes	AND → OR	Non-Critical	CF (✓)	Level 2
S38417-T200	CF, DoS	Yes	AND → OR	Non-Critical	CF (✓)	Level 2
S38417-T300	CF, DoS	Yes	AND → DFF(CLK)	Non-Critical	CF (✓)	Level 2
vga_lcd-T100	CF, DoS	Yes	AND → OR	Non-Critical	CF (✓)	Level 2
AES-T100	LI	No	XOR	Critical	LI (✓)	Level 3
AES-T200	LI	No	XOR	Critical	LI (✓)	Level 3
AES-T600	LI	Yes	AND → XOR	Critical	LI (✓)	Level 1
AES-T700	LI	Yes	AND → XOR	Critical	LI (✓)	Level 1
AES-T800	LI	Yes	AND → XOR	Critical	LI (✓)	Level 1