114學年度大學部專題海報展

國立清華大學 資訊工程學系
National Tsing Hua University Department of Computer Science

ICCAD 2025 CAD Contest Problem A: Hardware Trojan Detection on Gate Level Netlist

組員：熊恩伶、李彥呈、楊力衡、黃梓齊、陳啟綸、謝采晏

Abstract

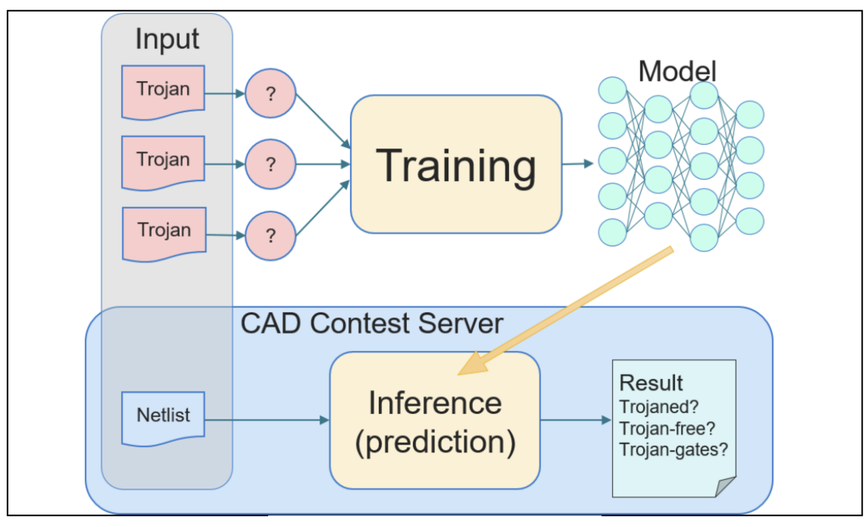
Traditionally, hardware security has been addressed through functional verification and side-channel analysis. However, as integrated circuits grow in complexity and third-party IP usage increases, hardware Trojans (HTs) have become a major threat to the integrity of modern designs. Conventional detection techniques often fail to generalize across unseen designs due to their reliance on handcrafted features or simulation-based methods.

In this work, we propose a graph-based detection framework that leverages deep learning on gate-level netlists to identify Trojan-inserted gates. Each circuit is modeled as a graph, where nodes represent logic gates and edges capture signal connectivity. We adopt and compare multiple architectures, including Graph Convolutional Networks (GCN), GraphSAGE, and a hybrid GCN–GraphSAGE model that combines global structure learning with inductive neighborhood aggregation. The proposed approach effectively captures both topological and functional characteristics of circuits, enabling accurate and generalizable Trojan detection.

All members of this project participated in the CAD Contest 2025, and has won one 特優 award and four 佳作 awards in this contest.

Problem Formulation

The problem is to detect hardware Trojans in a flattened gate-level Verilog netlist. Each input netlist contains only primitive gates, wires, constants, and flip-flops. The task is to develop a program that determines whether the circuit is Trojan-free or Trojaned and, if Trojaned, identifies the specific Trojan gates. The output must list the detection result and Trojan gate names, with correctness and F1 score used for evaluation.



Approach

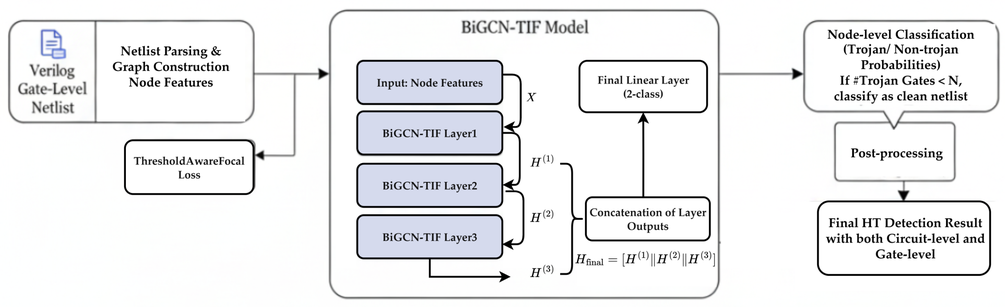
Our method employs the BiGCN-TIF (Bidirectional GCN with Topology and Information Fusion) dual-branch architecture for precise hardware Trojan detection and localization in gate-level netlists.

1) Multi-Level Feature Extraction

We extract node features and subgraph features to capture both local structure and regional context.

- Node Features (35 dimensions): Encode gate type, DFF roles, connectivity, local fan-in/fan-out, combinational depth, and basic dynamic activity (transitions and logic duration).

- Subgraph Features (23 dimensions): Represent regional statistics for each weakly connected component, including node/edge counts, gate type distribution, and averaged node-level metrics.



2) Dual-Branch GNN Architecture

The BiGCN-TIF model fuses structural and contextual representations for node-level classification.

- Node Branch: A three-layer Bidirectional GCN processes node features through both forward (driver-to-load) and backward (load-to-driver) message passing, producing a 192-dimensional local representation.
- Subgraph Branch: A two-layer MLP encodes subgraph-level features into a 64-dimensional context vector, which is assigned to all nodes in the same WCC.
- Fusion and Classification: The two vectors are concatenated into a 256-dimensional representation and passed through a classification head to predict Trojan or Free labels for each gate.

3) Topology-Aware Post-Processing

To enhance robustness and reduce false positives, predictions are refined through structural heuristics:

- Neighbor Consistency: Isolated Trojan predictions without Trojan neighbors are reverted to Free.
- Small-Group Pruning: Predicted Trojan clusters smaller than five nodes are removed.

Best Results

The two tables below demonstate the best results achieved in the CAD Contest@ICCAD 2025, including both public and hidden test cases. The results are summarized in following figures. These outstanding results earned us one 特優 award and four 佳作 awards in the contest.