I. Approach: BiGCN-TIF Dual-Branch Fusion Model

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

A. Multi-Level Feature Extraction

We extract **35-dimensional node features** ($\mathbf{x}$) and **23-dimensional subgraph features** ($\mathbf{Z}_s$) from the gate-level netlist graph to comprehensively capture both local structure and regional context.

1) Node Features ($\mathbf{x}$) – 35 Dimensions (Input to GNN): This feature vector captures the functional, structural, and dynamical properties of each gate.

Gate Type and Pin Roles (14 features): Nine one-hot gate type indicators (AND, OR, XOR, DFF, etc.) describe the node's logical function. Five binary flags indicate whether the node acts as a consumer of critical DFF pins such as Clock, Data, or Reset.

Topology and Structure (16 features): Two I/O flags denote whether the node connects directly to a primary input (PI) or primary output (PO). Six local structural counts represent sink status, two-hop fan-in and fan-out counts, two-hop DFF count, and combinational level. Four neighborhood descriptors measure the number of nodes and distinct gate types within a four-hop combinational range. Four additional features record the shortest and longest topological distances to PI and PO, as well as reachability and distance to power (VDD) and ground (GND) sources.

Simulation Statistics (5 features): These features quantify dynamic signal activity, including transition counts ($1\rightarrow0$ and $0\rightarrow1$), longest consecutive logic levels, and total logic-1 duration.

2) Subgraph Features ($\mathbf{Z}_s$) – 23 Dimensions (Regional Context): Subgraph features describe the regional context within each weakly connected component (WCC), which is defined by cutting DFF connections to isolate combinational regions.

Topology Statistics (5 features): These include node count, number of inner edges, subgraph density, number of cut edges to other WCCs, and cut ratio.

Gate Type Distribution (9 features): A normalized histogram describes the proportions of the nine gate types within each WCC, capturing regional logic composition.

Node Feature Averages (7 features): Each WCC also contains averaged values of selected node-level metrics, such as average two-hop fan-in, average distance to PO, and average combinational depth, providing contextual aggregation.

B. Dual-Branch GNN Architecture

The BiGCN-TIF architecture fuses local structural representations ($\mathbf{h}_{\text{node}}$) learned from graph convolution and global contextual representations ($\mathbf{z}_{\text{node}}$) obtained from subgraph analysis.

1) Node Branch ($\mathbf{h}_{node}$): This branch adopts a **three-layer Bidirectional GCN (BiGCN)**. It processes the $\mathbf{x}$ features by performing bidirectional message passing—both forward (driver-to-load) and backward (load-to-driver). The outputs from both directions are concatenated to form a 192-dimensional structural vector $\mathbf{h}_{\text{node}}$ for each gate, effectively encoding the local structural dependencies.

2) Subgraph Branch ($\mathbf{z}_{node}$): The subgraph branch employs a **two-layer feed-forward network (MLP)** that encodes the 23-dimensional $\mathbf{Z}_s$ vector into a 64-dimensional latent contextual representation. This encoded vector is then propagated back to all nodes belonging to the same WCC, forming the contextual embedding $\mathbf{z}_{\text{node}}$ for each node.

3) Fusion and Classification: The two representations are fused by concatenating $\mathbf{h}_{\text{node}}$ (192 dimensions) and $\mathbf{z}_{\text{node}}$ (64 dimensions), producing a 256-dimensional fused vector for each node. This fused vector is then passed through a final MLP classification head to output per-gate predictions of `Trojan` or `Free` status.

C. Topology-Aware Post-Processing

To improve robustness and suppress false positives, the model's outputs undergo a topology-aware refinement stage that enforces spatial consistency across the predicted Trojan nodes.

Neighbor Consistency Filtering: A predicted Trojan node is reverted to Free if none of its 1-hop neighbors are also predicted as Trojan, preventing isolated false alarms.

Small-Group Pruning: Clusters of predicted Trojan nodes with fewer than five members are discarded entirely, as small disconnected groups are less likely to represent actual Trojan structures.

Trojan Count Enforcement: If the total number of predicted Trojan nodes in a netlist is below ten, the entire design is classified as `Trojan-Free`, overriding individual gate-level predictions.