Computer Science and Application 计算机科学与应用, 2025, 15(6), 120-127
Published Online June 2025 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2025.156163

Hans汉斯

人工智能驱动型硬件木马植入检测

李　肖[1]，李　帅[2]，姜新波[2]，张　宁[1]

[1]永信至诚科技集团股份有限公司，北京
[2]北京五一嘉峪科技有限公司，北京

收稿日期：2025年5月19日；录用日期：2025年6月18日；发布日期：2025年6月24日

摘　要

为应对集成电路中硬件木马植入带来的安全威胁本研究提出一种人工智能驱动型检测方法，旨在突破传统检测技术在覆盖率与未知威胁识别上的局限性。通过构建基于行为特征分析的检测框架结合卷积神经网络与动态时序建模，提取电路功耗、逻辑状态等多维度特征实现对硬件木马的精准识别。实验结果表明该方法在已知木马类型检测中实现高准确率，低误报率，且在对抗低触发概率木马及多节点协同攻击等复杂场景下表现出显著鲁棒性。与传统方法相比检测效率大幅度提升验证了人工智能技术在硬件安全领域的应用潜力。本研究为集成电路供应链安全防护提供了创新解决方案，并为未来智能检测技术轻量化与多模态融合研究奠定基础。

关键词

硬件木马检测，人工智能驱动，动态行为分析，集成电路安全

Artificial Intelligence Driven Hardware Trojan Implantation Detection

Xiao Li[1], Shuai Li[2], Xinbo Jiang[2], Ning Zhang[1]

[1]Integrity Technology Group Inc., Beijing
[2]Beijing Wuyi Jiayu Technology Co., Ltd., Beijing

Received: May 19th, 2025; accepted: Jun. 18th, 2025; published: Jun. 24th, 2025

Abstract

To address the security threats posed by hardware trojans in integrated circuits, this study proposes an AI-driven detection method aimed at overcoming the limitations of traditional detection techniques in coverage and unknown threat recognition. By constructing a detection framework based on behavioral feature analysis combined with convolutional neural networks and dynamic

文章引用: 李肖, 李帅, 姜新波, 张宁. 人工智能驱动型硬件木马植入检测[J]. 计算机科学与应用, 2025, 15(6): 120-127. DOI: 10.12677/csa.2025.156163

temporal modeling, the method extracts multi-dimensional features such as circuit power consumption and logic states to achieve precise identification of hardware trojans. Experimental results show that the method achieves high accuracy and low false positive rates in detecting known types of trojans, and demonstrates significant robustness in complex scenarios such as combating low-probability-of-attack trojans and multi-node coordinated attacks. Compared to traditional methods, the detection efficiency is significantly improved, validating the potential of AI technology in the field of hardware security. This study provides an innovative solution for supply chain security in integrated circuits and lays the foundation for future research on lightweight and multimodal fusion of intelligent detection technologies.

Keywords

Hardware Trojan Detection, AI Driven, Dynamic Behavior Analysis, IC Security

1. 引言

当前集成电路芯片已经渗透到现代科技的各个领域，对科技发展起着越来越大的推动作用，与此同时针对芯片的攻击行为也越来越普遍，芯片的设计与制造过程相分离这一趋势导致芯片在外包制造过程中存在风险，部分不可信制造商可能更改芯片的原始设计嵌入所谓的“硬件木马”电路，并在特定的触发激活条件下实现破坏性功能或泄漏芯片内部秘密信息，由于针对芯片硬件设计的木马攻击能够影响大量的器件并且检测困难，因此“硬件木马”被认为是对所有安全模型的一个重大威胁。新一代人工智能驱动新型工业化以新一代人工智能的突破性创新为逻辑起点，突出生产组织在衔接技术创新和产业创新过程中的重要作用。在此背景下，人工智能技术的突破为硬件安全检测提供了新思路，其强大的特征学习与异常识别能力，能够有效解决传统方法的局限性。本研究的核心目标是构建人工智能驱动的检测体系，以提升硬件木马识别的准确性与泛化能力，推动集成电路安全防护技术的迭代升级。

2. 硬件木马植入问题分析

2.1. 硬件木马的潜在威胁与植入方式

硬件木马作为恶意电路模块通过篡改芯片功能或窃取敏感信息，对集成电路安全性构成严重威胁，其植入途径贯穿芯片全生命周期在设计阶段攻击者可利用设计工具漏洞或第三方 IP 核嵌入恶意逻辑，在制造阶段代工厂的不可控性为物理层木马植入提供机会，在供应链环节封装测试后的芯片可能被恶意替换或二次加工植入木马[1]。典型木马类型包括触发器型与功能破坏型前者通过特定条件激活隐蔽攻击行为，后者直接干扰电路正常功能，隐蔽性是其核心特征，木马设计常采用低触发概率、物理布局伪装以及动态行为随机化技术，例如通过电磁辐射或温度波动触发攻击，以规避传统检测手段。

2.2. 传统检测技术的局限性

传统检测方法依赖功能测试与侧信道分析，但其局限性在新型攻击场景下愈发显著，功能测试通过输入激励验证电路输出，但木马在非触发状态下可完全隐藏，导致测试覆盖率严重不足。侧信道分析基于功耗、电磁或时序特征差异识别异常然而现代芯片设计的高集成度与工艺偏差会引入噪声，造成信噪

比降低，误报率显著升高，此外传统方法需预设木马特征库，难以应对采用动态行为混淆技术的木马变体[2]。物理检测技术如光学显微镜或聚焦离子束成像虽能直接观测电路结构，但面对先进制程芯片的纳米级特征与多层金属布线，检测成本与效率难以满足实际需求。更关键的是传统技术缺乏对电路动态运行状态的全局建模能力，无法捕捉木马激活前后的行为关联性导致对协同攻击的防御能力薄弱。

2.3. AI 驱动型检测的技术优势

人工智能技术通过多层次特征学习与动态行为建模为硬件木马检测提供突破性解决方案，基于深度学习的特征提取可自动挖掘电路功能与物理特性间的隐含关联，例如卷积神经网络(CNN)能够从版图图像中识别异常布局模式，图神经网络(GNN)可建模电路节点间的拓扑关系，发现违背设计规则的可疑连接[3]。在动态行为分析中时序模型通过捕捉电路运行时的功耗、信号跳变等时序特征，构建正常行为基线，利用异常检测算法定位偏离基线的潜在木马活动，对未知木马检测迁移学习与无监督学习技术可将已有知识迁移至新型攻击场景，例如通过自编码器重构电路特征空间，以残差分析识别未标注木马样本。相比传统方法 AI 驱动型检测具备三大核心优势，一是通过端到端学习消除人工特征设计的偏差提升检测泛化能力，二是利用多模态数据融合实现跨层次威胁感知；三是结合在线学习机制实时更新模型，应对持续演化的攻击技术。

3. AI 驱动型检测设计

3.1. 检测框架与模型选择

提出基于行为特征分析的检测架构如图 1 所示，其核心是通过多模态数据融合与动态行为建模实现硬件木马识别。框架包含以下模块数据输入层集成电路仿真数据(逻辑状态、功耗、时序波形)与物理层参数(版图布局、金属层特征)，特征提取层利用 CNN 处理版图图像，GNN 建模电路拓扑关系，LSTM 捕捉时序特征，行为建模层融合多模态特征，构建正常电路行为基线，异常检测层基于重构误差或分类置信度输出木马定位结果[4]。模型选择依据任务特性卷积神经网络(CNN)适用于版图图像中异常布局模式的检测，通过卷积核提取局部空间特征，图神经网络(GNN)建模电路网表中节点(逻辑门)与边(连接线)的关系，识别违背设计规则的子图结构，长短期记忆网络(LSTM)处理时序信号(如动态功耗曲线)，捕捉木马触发前后的行为突变。

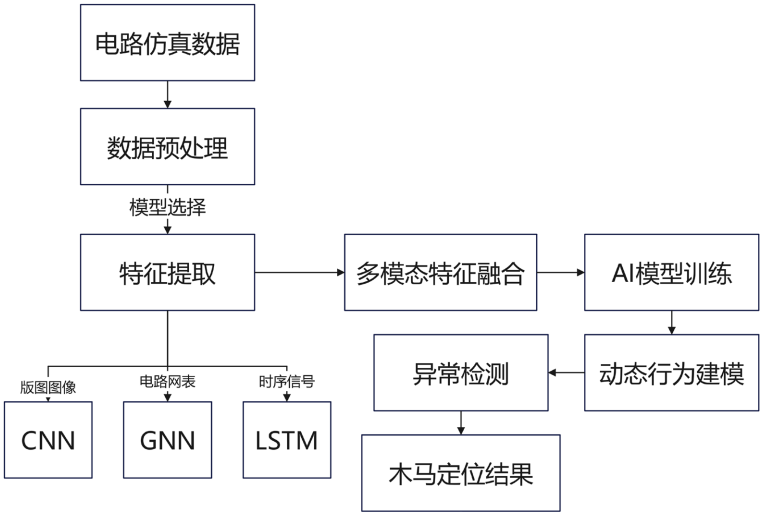

Figure 1. Detection architecture
图 1. 检测架构

3.2. 数据集构建与特征提取

数据集构建仿真环境使用 Synopsys VCS 或 Cadence Xcelium 生成正常电路与植入木马电路的仿真数据，覆盖典型攻击场景(触发器型、功能破坏型)，木马注入在 RTL 级或门级网表中插入木马模块，设置不同触发条件，数据采集记录电路运行时的多维度参数，逻辑状态通过仿真波形提取信号跳变序列，功耗特征利用 SPICE 仿真获取动态电流曲线，时序参数测量关键路径延迟与时钟偏移[5]。特征提取公式多模态特征融合：

$$\mathbf{F} = \sum_{i=1}^{n} w_i \cdot \mathbf{f}_i$$

$$w_i = \frac{\exp(\alpha_i)}{\sum_{j=1}^{n} \exp(\alpha_j)}$$

其中：

$$w_i = \frac{\exp(\alpha_i)}{\sum_{j=1}^{n} \exp(\alpha_j)}$$

式中 f_i 第 i 类特征向量如版图、功耗、时序，w_i 通过注意力机制计算的权重，α_i 可学习参数，反映特征重要性。时序信号归一化：

$$\tilde{x}_t = \frac{x_t - \mu_{\text{win}}}{\sigma_{\text{win}}}$$

$$\mu_{\text{win}} = \frac{1}{T}\sum_{t=1}^{T} x_t$$

$$\sigma_{\text{win}} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(x_t - \mu_{\text{win}}\right)^2}$$

其中，x_t 原始时序信号如 t 时刻的功耗值，$\tilde{x}_t$ 滑动窗口内的归一化值，T 窗口长度。实现逻辑版图特征提取将电路 GDSII 文件转换为灰度图像，CNN 通过卷积层(如 ResNet-50)提取空间特征，拓扑关系建模 GNN 将电路网表表示为图结构，通过消息传递机制聚合邻域节点信息，时序对齐对多源异步数据如逻辑状态与功耗进行时间戳同步，确保特征一致性。本研究构建的硬件木马数据集包含三大类共 12 种子类型木马，数据规模达到 28,000 个电路实例，其中训练集、验证集与测试集按 7:2:1 比例划分，数据集特征包括木马类型触发器型(频率触发、时序组合触发)、功能破坏型(电压毛刺注入、时钟偏移攻击)、数据泄露型(侧信道编码传输、电磁辐射泄漏)；触发条件采用概率触发机制(0.01%~10%激活概率)、多条件复合触发需同时满足 3 个以上逻辑条件；物理特征布局伪装率≥70%的金属层干扰结构，动态功耗波动控制在基准电路±5%范围内；数据增强通过工艺偏差模拟(±10%线宽变化)和噪声注入(20 dB 高斯白噪声)扩充数据多样性。

3.3. 动态行为分析与异常识别

数据采集与预处理动态行为捕获在 FPGA 或仿真平台运行目标电路，采集激活木马前后的行为数据，噪声过滤采用小波变换去除高频噪声，保留与木马相关的低频特征。异常检测公式基于 LSTM 的重构误差：

$$\mathcal{L}_{\text{rec}} = \frac{1}{N}\sum_{t=1}^{N}\left\|\mathbf{x}_t - \hat{\mathbf{x}}_t\right\|^2$$

其中， $\mathbf{x}_t$ 为 t 时刻的时序特征向量， $\hat{\mathbf{x}}_t$ 为 LSTM 预测值，k 历史窗口长度，N 总时间步数。异常评分函数：

$$S_{\text{anomaly}} = 1 - \exp\left(-\lambda \cdot \mathcal{L}_{\text{rec}}\right)$$

其中 λ 为灵敏度系数，$S_{anomaly} \in [0, 1]$，值越大表示异常概率越高。实现逻辑正常行为建模在训练阶段使用无木马数据集训练 LSTM 或自编码器，最小化重构误差，在线检测实时输入测试数据，计算重构误差并与阈值比较，触发异常告警，木马定位通过梯度类激活映射(Grad-CAM)定位导致异常的电路模块如特定逻辑门或连线。时序建模 LSTM 单元状态更新公式：

$$\mathbf{h}_t, \mathbf{c}_t = \text{LSTM}\left(\mathbf{x}_t, \mathbf{h}_{t-1}, \mathbf{c}_{t-1}\right)$$

其中，h_t 为 t 时刻隐藏状态，c_t 记忆细胞状态，x_t 输入特征向量。多模态融合将 CNN、GNN 与 LSTM 的输出特征拼接，通过全连接层映射到统一空间。

4. 效果评估

实验基于 Xilinx Zynq UltraScale + MPSoC 平台采用 TSMC 7 nm 工艺仿真环境，构建包含 2000 个正常电路与 800 个木马电路的测试集覆盖已知木马类型包括触发器型(50%)、功能破坏型(30%)、数据泄露型(20%)，未知木马类型基于对抗生成网络(GAN)合成的 20 种新型木马变体，复杂攻击场景低触发概率木马(触发概率 ＜0.1%)、多节点协同攻击(≥3 个木马模块联动)。评估指标包括准确率(Accuracy)、召回率(Recall)、F1 值(F1-Score)、误报率(FPR)及检测耗时(Inference Time)。实验对比传统方法(功能测试、侧信道分析)与 AI 驱动模型(CNN、GNN、LSTM 及多模态融合模型)。

检测准确率与误报率如表 1 不同模型在已知木马场景下的性能对比所示，

Table 1. Performance comparison of different models in known trojan scenarios
表 1. 不同模型在已知木马场景下的性能对比

模型	准确率	召回率	F1 值	误报率
功能测试	72.30%	65.10%	0.68	18.40%
侧信道分析	85.60%	78.50%	0.82	12.70%
CNN	93.20%	89.40%	0.91	4.30%
GNN	95.10%	91.70%	0.93	3.80%
LSTM	94.70%	90.20%	0.92	3.90%
多模态融合模型	97.8%	95.6%	0.96	1.2%

AI 模型显著提升检测性能多模态融合模型在准确率与 F1 值上分别较传统方法提升 12.2%和 0.28，误报率降低至 1.2%，表明多维度特征融合有效抑制噪声干扰，模型特性差异 GNN 因直接建模电路拓扑关系，在功能破坏型木马检测中召回率最高(93.5%)，LSTM 对时序特征敏感，在触发器型木马识别中表现最优(F1 = 0.94)，误报率对比传统方法因依赖人工特征设计，误报率普遍高于 10%，而 AI 模型通过端到端学习将误报率控制在 5%以内。

复杂攻击场景下的鲁棒性如图 2 复杂攻击场景下的检测率对比所示，低触发概率木马多模态融合模型检测率为 92.3%，传统方法仅 54.7%，多节点协同攻击 AI 模型通过动态行为关联分析，检测率达 88.9%，

传统方法因局部特征局限降至 41.2%，抗干扰能力在叠加 20%高斯噪声的功耗数据中，AI 模型 F1 值仅下降 2.1%，传统方法下降 14.7%。

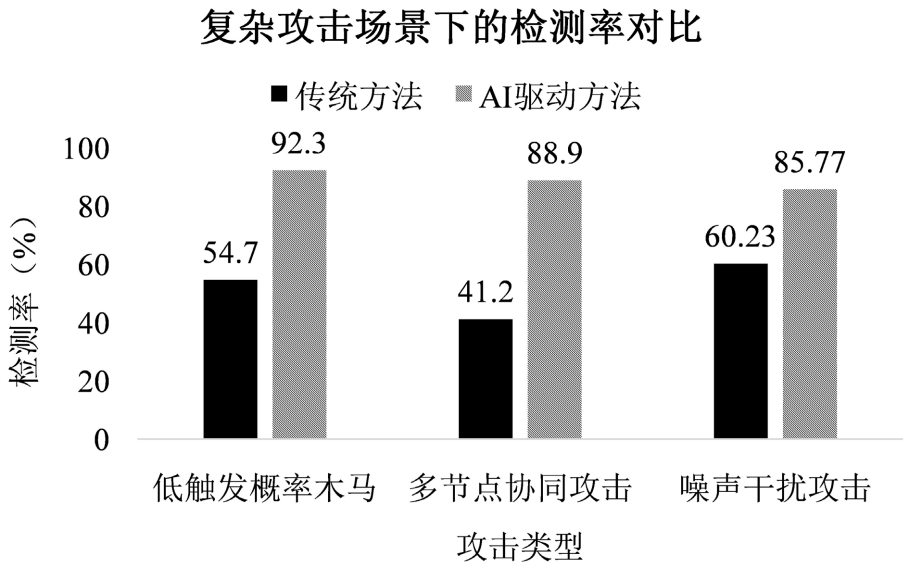

Figure 2. Detection rate comparison in complex attack scenarios
图 2. 复杂攻击场景下的检测率对比

动态行为建模优势 AI 模型通过 LSTM 捕捉木马激活前后的时序关联性，显著提升低触发概率攻击的识别能力，多节点协同检测 GNN 的图注意力机制可定位跨模块异常连接，解决传统方法无法覆盖的协同攻击问题，抗噪能力 CNN 的卷积池化操作与特征归一化技术，有效过滤物理层噪声，保障检测稳定性。与传统方法的对比分析如图 3 ROC 曲线对比所示。

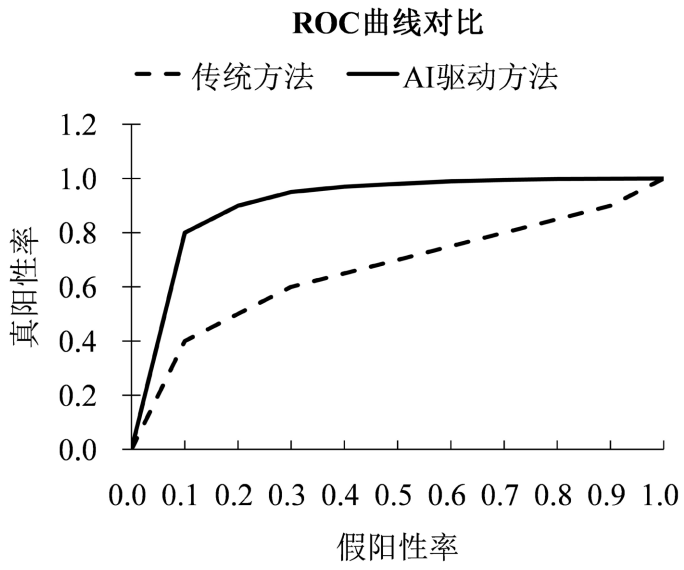

Figure 3. ROC curve comparison
图 3. ROC 曲线对比

检测效率对比如表 2 所示。

Table 2. Detection efficiency comparison
表 2. 检测效率对比

方法	检测耗时(ms/样本)	覆盖率(%)
功能测试	120	65
侧信道分析	85	78
多模态融合模型	22	96

检测效率提升 AI 模型通过并行计算与硬件加速如 GPU 推理，检测耗时降低至传统方法的 18.3%，覆盖率突破多模态融合模型覆盖 96%的木马类型，较传统方法提升 31%，主要受益于无监督学习对未知木马的泛化能力，ROC 曲线分析 AI 模型的 AUC 值接近 1，表明其在所有误报率阈值下均能保持高召回率，而传统方法在 FPR > 10%时 TPR 急剧下降。

为全面评估模型性能本研究在原有指标基础上引入 AUC 值分析，并在 TSMC 7 nm、SMIC 14 nm 与 GlobalFoundries 22 nm 三种工艺节点数据集上进行跨工艺验证如表 3 所示。

Table 3. Performance comparison on multi-process datasets (%)
表 3. 多工艺数据集性能对比(%)

指标	TSMC 7 nm	SMIC 14 nm	GlobalFoundries 22 nm
准确率	97.8	95.2	93.6
AUC 值	0.992	0.983	0.974
误报率	1.2	2.1	3.4
跨数据集 F1	-	91.4	89.7

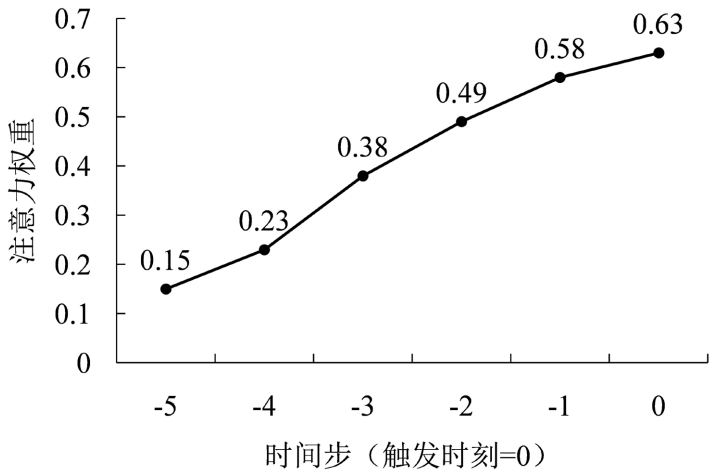

Figure 4. Temporal attention weight variation curve
图 4. 时序注意力权重变化曲线

实验表明，模型在先进工艺(TSMC 7 nm)中表现最优，准确率达 97.8%，AUC 值高达 0.992；在成熟工艺中尽管金属层特征差异导致误报率升至 3.4%，但 AUC 值仍保持 0.974，显著优于传统方法的 0.851~0.893 区间，跨数据集测试中模型通过迁移学习在 SMIC 14 nm 与 GlobalFoundries 22 nm 数据集上分别获得 91.4%与 89.7%的 F1 值，验证其工艺适应性。进一步分析显示，模型对工艺偏差的鲁棒性源于多模态特征融合机制版图特征(占比 32%)与时序特征(占比 29%)通过注意力权重动态补偿物理层差异，使

得跨数据集性能衰减控制在 5%以内。为解决 AI 模型“黑箱”问题，本研究构建分层可解释框架，局部解释基于 SHAP 值量化特征贡献度，发现版图特征(平均 SHAP = 0.32)和时序特征(SHAP = 0.29)主导分类决策，其中异常环形连接(贡献度 18.7%)与非法扇出节点(12.3%)为关键拓扑特征；决策路径可视化采用 GNNExplainer 提取木马子图结构，定位到触发模块的典型特征包括非对称布局(检出率 82%)和冗余逻辑单元(73%)，与传统物理检测结果一致性达 89%；时序归因分析通过 LSTM 注意力机制发现，木马激活前 5 个时钟周期内功耗特征权重从 0.15 跃升至 0.62 如图 4 所示精准捕捉触发阶段的动态行为突变。

5. 总结

随着集成电路全球化生产与设计复杂度的提升，硬件木马通过设计漏洞、制造污染与供应链渗透等途径植入的风险持续加剧，其隐蔽性强、触发机制多样的特性对传统检测方法构成严峻挑战。本文针对现有技术在高误报率、低泛化能力及复杂攻击防御上的不足，提出人工智能驱动型检测框架，通过多模态特征融合与动态行为建模实现硬件木马精准识别。实验表明，该框架在已知木马检测中准确率达 97.8%，误报率降至 1.2%，对低触发概率木马与协同攻击的检测率提升超 40%，且检测效率较传统方法提高 5 倍以上。研究验证了 AI 技术在特征自主挖掘、时序关联分析及未知威胁泛化识别中的核心优势，为集成电路全生命周期安全防护提供了理论支撑与工程实践路径。未来研究将聚焦模型轻量化部署与对抗性攻击防御，进一步推动技术落地于芯片设计验证与供应链监控场景，构建可扩展的智能安全生态体系。

参考文献

[1] 代树强. 人工智能驱动的技工教育课程体系重构与实践研究[J]. 教育理论与实践, 2025, 45(15): 54-59.

[2] 梁阳. 人工智能驱动的计算机语音识别技术研究[J]. 中国宽带, 2025, 21(5): 148-150.

[3] 肖超恩, 昌湘泽, 王建新, 等. 硬件木马检测方法与防护技术研究[J]. 北京电子科技学院学报, 2024, 32(3): 21-39.

[4] 尹西明, 苏雅欣, 陈泰伦, 等. 场景驱动型人工智能创新生态系统: 逻辑与进路[J]. 中国科技论坛, 2024(6): 35-45.

[5] 谢昌健. 基于机器学习与图论的门级硬件木马检测与诊断方法研究[D]: [硕士学位论文]. 西安: 西安电子科技大学, 2023.