Abstract:

Hardware Trojans (HTs) pose a significant threat to the trustworthiness of integrated circuits, particularly in offshore manufacturing environments where untrusted facilities may access critical design stages. To prevent the production of compromised or costly defective devices, it is crucial to ensure circuit integrity before fabrication. In this work, we focus on netlist-level detection of HTs introduced through both combinational and sequential insertions. While such Trojans often leave little to no visual or structural evidence, they introduce subtle variations in synthesis-level features—differences that can be exploited for detection. Using benchmark netlists, we trained and evaluated several machine learning models, including K-Nearest Neighbors (KNN), Random Forest, and One-Class SVM (OCSVM). Among these, KNN achieved the highest accuracy of 99.26% in distinguishing clean circuits from Trojan-infected ones. These results demonstrate the effectiveness of pre-silicon testing as a proactive measure against hardware-level attacks and emphasize the potential of machine learning as a scalable and robust approach to enhancing hardware security in modern IC design and manufacturing flows. Future work will expand the dataset with diverse benchmark circuits, explore additional synthesis parameters, test model robustness with broken/empty netlists, and evaluate new unsupervised learning approaches for Trojan detection.

Table 1: Confusion matrix results for HT detection models.

Model	True Clean → Pred. Clean	True Clean → Pred. Trojan	True Trojan → Pred. Clean	True Trojan → Pred. Trojan
KNN	42	0	1	92
Random Forest	42	0	2	91
OCSVM	39	3	0	93

Figure 1: Model Accuracy