Transformers for Secure Hardware Systems: Applications, Challenges, and Outlook

Banafsheh Saber Latibari
University of Arizona
Tucson, AZ, USA
banafsheh@arizona.edu

Najmeh Nazari
University of California, Davis
Davis, CA, USA
nnazari@ucdavis.edu

Avesta Sasan
University of California, Davis
Davis, CA, USA
asasan@ucdavis.edu

Houman Homayoun
University of California, Davis
Davis, CA, USA
hhomayoun@ucdavis.edu

Pratik Satam
University of Arizona
Tucson, AZ, USA
pratiksatam@arizona.edu

Soheil Salehi
University of Arizona
Tucson, AZ, USA
ssalehi@arizona.edu

Hossein Sayadi
California State University, Long Beach
Long Beach, CA, USA
hossein.sayadi@csulb.edu

Abstract

The rise of hardware-level security threats, such as side-channel attacks, hardware Trojans, and firmware vulnerabilities, demands advanced detection mechanisms that are more intelligent and adaptive. Traditional methods often fall short in addressing the complexity and evasiveness of modern attacks, driving increased interest in machine learning-based solutions. Among these, Transformer models, widely recognized for their success in natural language processing and computer vision, have gained traction in the security domain due to their ability to model complex dependencies, offering enhanced capabilities in identifying vulnerabilities, detecting anomalies, and reinforcing system integrity. This survey provides a comprehensive review of recent advancements on the use of Transformers in hardware security, examining their application across key areas such as side-channel analysis, hardware Trojan detection, vulnerability classification, device fingerprinting, and firmware security. Furthermore, we discuss the practical challenges of applying Transformers to secure hardware systems, and highlight opportunities and future research directions that position them as a foundation for next-generation hardware-assisted security. These insights pave the way for deeper integration of AI-driven techniques into hardware security frameworks, enabling more resilient and intelligent defenses.

CCS Concepts

• **Security and privacy → Security in hardware**.

Keywords

Hardware Systems, Transformer, Security, Threat Detection.

ACM Reference Format:
Banafsheh Saber Latibari, Najmeh Nazari, Avesta Sasan, Houman Homayoun, Pratik Satam, Soheil Salehi, and Hossein Sayadi. 2025. Transformers for Secure Hardware Systems: Applications, Challenges, and Outlook. In *Great Lakes Symposium on VLSI 2025 (GLSVLSI '25), June 30-July 2, 2025, New Orleans, LA, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3716368.3735281

1 Introduction

Hardware security has emerged as a critical pillar in the protection of modern computing systems, which are increasingly targeted by a diverse and evolving array of sophisticated threats [38]. Attacks such as side-channel exploits, hardware Trojans, and firmware-level vulnerabilities pose significant risks across a wide spectrum of platforms, from resource-constrained embedded devices to large-scale high-performance computing systems. These threats are particularly difficult to detect due to the complexity of modern hardware architectures and the ability of attackers to operate below the software stack [37, 56]. Traditional security mechanisms, such as rule-based heuristics, signature-based detection, and static analysis, often fail to generalize across evolving threat vectors and adapt to the dynamic nature of contemporary attacks.

As adversaries continue to develop more advanced and adaptive techniques, there is a pressing need for more intelligent, data-driven, and robust detection mechanisms. In recent years, to address these challenges, machine learning (ML)-based approaches have gained attention in the hardware security community. ML models can learn complex patterns from data, enabling anomaly detection, behavioral analysis, and predictive security measures [46]. However, conventional ML models often face limitations in this domain: they may struggle to capture long-range dependencies in hardware data, require extensive manual feature engineering, and frequently fail to scale effectively with the increasing heterogeneity and complexity of hardware systems data.

Transformer models have recently emerged as a powerful solution to these limitations. Originally designed for natural language

GLSVLSI '25, New Orleans, LA, USA

ACM ISBN 979-8-4007-1496-2/2025/06
https://doi.org/10.1145/3716368.3735281

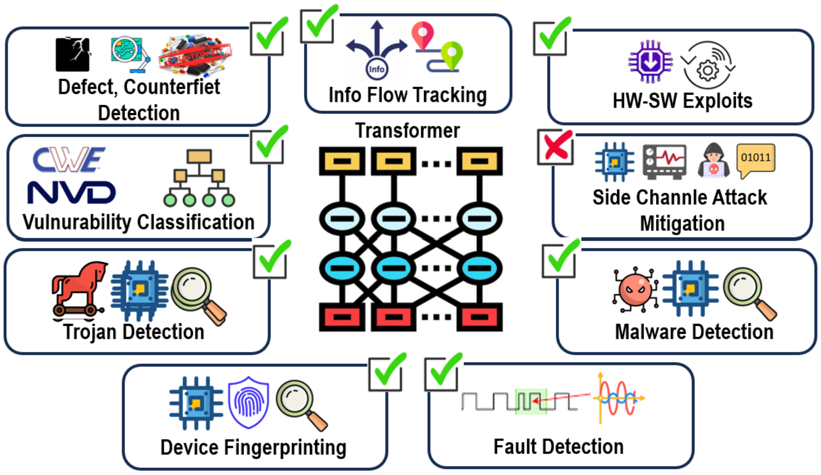

Figure 1: Scenarios where Transformers detect, classify, and mitigate vulnerabilities.

processing, Transformers leverage self-attention mechanisms to capture intricate relationships within sequential data [22, 23, 45]. This makes them especially well-suited for fast and accurate modeling of temporal and multi-modal patterns of the low-level data in hardware systems, without the need for high-overhead feature analysis. As depicted in Figure 1, Transformers offer promising capabilities across a range of hardware security applications, from hardware Trojan detection and vulnerability assessment to malware detection and device fingerprinting. Their versatility and modeling power position them as a strong candidate for enabling next-generation, intelligent threat detection systems.

This paper presents a comprehensive survey of the application of Transformers in the context of hardware security. We review the current state of research, highlight key challenges and limitations, and outline future opportunities for advancing secure and intelligent hardware systems through this emerging paradigm. Section 2 provides background on deep learning models relevant to this domain. Section 3 reviews recent research applying Transformers to secure hardware systems. In Section 4, we discuss the current challenges and opportunities, including practical constraints, architectural considerations, and potential directions. Lastly, Section 5 concludes this study.

2 Background on Transformers

Vaswani et al. [55] introduced the Transformer, which leverages the attention mechanism to model dependencies between input and output in machine translation. This approach eliminates the sequential nature of RNNs, enabling more efficient parallel processing and capturing long-range dependencies more effectively. Figure 2 illustrates the detailed computations within a Transformer architecture. This encoder-decoder-based architecture consists of multiple Transformer blocks, each containing a multi-head attention (MHA) module and a feed-forward network (FFN) module. Each block includes Layer Normalization (LayerNorm) and a residual connection. The MHA module first projects the input sequence using weight matrices W_Q, W_K, and W_V, generating query, key, and value representations. These representations are then split into h heads, where each head has a hidden dimension of d/h, and processed as follows:

$$\begin{aligned} Q^i &= QW_Q^i \\ K^i &= KW_K^i, \qquad i \in heads \\ V^i &= VW_V^i \end{aligned} \tag{1}$$

The scaled dot-product attention function computes the attention scores and output:

$$O = SoftMax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2}$$

The attention outputs from all heads are concatenated along the hidden dimension d and projected using W_{out}. The result is processed with LayerNorm and a residual connection to form the MHA output. The FFN module consists of two linear layers: the first projects the input from d to a higher dimension, and the second projects it back to d.

3 Transformers in Hardware Security

3.1 Hardware Trojan Detection

Trojan is a malicious modification of a hardware design that can compromise functionality, integrity, or confidentiality. Since HTs are stealthy, ML-based models have been proposed to help with identification [14]. For pre-silicon Hardware Trojan (HT) detection and localization, a Transformer-based method called HTrans is proposed. It utilizes a Graph Convolutional Network (GCN) in the preprocessing stage to ensure scalability across various design sizes. The model achieves 96.7% F1 score for HT detection and 91.7% accuracy for HT localization on the Trusthub benchmark, with detection completed in under a second at the Register Transfer Level (RTL) [25]. Authors in this paper propose a non-destructive, golden-chip-free transformer-based framework for Hardware Trojan Detection (HTD), utilizing Power Side-Channel (PSC) data. They apply Generative AI techniques such as GPT, BERT, and transformers to classify

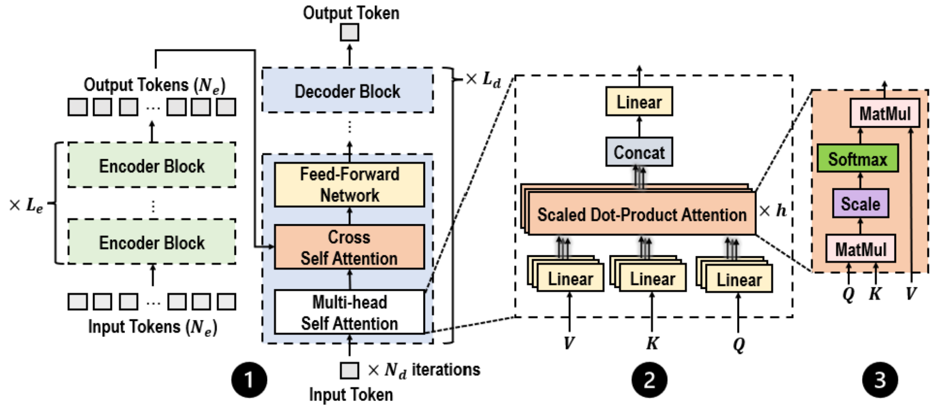

Figure 2: Architecture of the Vanilla Transformer, featuring (1) an Encoder-Decoder structure, (2) Multi-Head Self-Attention for capturing diverse contextual relationships, and (3) Scaled Dot-Product Attention for efficient information weighting.

hardware trojans into Enabled, Disabled, and Triggered categories. The framework processes side-channel data through different transformer networks and achieves 87.74% accuracy in HT detection, demonstrating superior performance compared to existing methods in identifying abnormal IC behaviors [34]. TrojanFormer incorporates a unique message-passing scheme within a graph transformer network to enhance detection performance while reducing computational complexity. It achieves an average F1 score of 97.66% on medium and small-scale datasets, surpassing other graph learning baseline models. On large-scale circuit datasets, TrojanFormer shows a 4% performance improvement and an 18% reduction in computational overhead, highlighting its effectiveness and efficiency in large-scale integrated circuit HT detection scenarios [4].

TrojanWhisper [10] explores the potential of general-purpose LLMs for detecting HTs in Register Transfer Level (RTL) designs, including modules like SRAM, AES, and UART. The tool systematically evaluates state-of-the-art LLMs (GPT-4o, Gemini 1.5 pro, and Llama 3.1) in HT detection without prior fine-tuning. To address training data bias, perturbation techniques such as variable name obfuscation and design restructuring are implemented. The experimental results show perfect detection rates (100% precision/recall) for GPT-4o and Gemini 1.5 pro, with performance degradation under code perturbation for all models, particularly in payload localization. This work demonstrates the potential of LLMs for hardware security applications [10].

NtNDet leverages large-scale pre-trained NLP models. It introduces a method called Netlist-to-Natural-Language (NtN) to convert gate-level netlists into a format suitable for Natural Language Processing (NLP) models, applying the self-attention mechanism of Transformers to capture complex dependencies. Experiments on the Trust-Hub, TRIT-TC, and TRIT-TS benchmarks demonstrate that NtNDet outperforms existing methods, achieving improvements of 5.27% in precision, 3.06% in True Positive Rate (TPR), 0.01% in True Negative Rate (TNR), and a 3.17% increase in F1 score, setting a new state-of-the-art in HT detection [21]. The work in [5] propose the TA-MobileViT lightweight model, which integrates the triplet attention (TA) mechanism with the MobileViT network to address the challenges in Hardware Trojan detection. This model enhances cross-channel interaction without increasing the parameter count, improving both classification accuracy and generalization ability. By combining convolutional and transformer blocks, TA-MobileViT effectively extracts both local and global features, achieving 100% recognition accuracy for single Trojan types and 72.2% accuracy for AES-600 detection. In multi-Trojan detection, the model reaches an impressive 97.04% mean accuracy. Compared to other deep learning methods, TA-MobileViT offers better performance with fewer parameters, making it a highly competitive model for hardware Trojan detection.

3.2 Packaging Defect and Counterfeit Detection

In the field of integrated circuit (IC) packaging and PCB design surface defect detection, ensuring high precision in identifying defects such as cracks, scratches, and contamination is crucial for maintaining product quality and manufacturing efficiency [6, 27, 32, 33]. The work in [57] addresses challenges in industrial IC surface defect detection, particularly the imbalance in information density due to data collection difficulties. The proposed hybrid model, SDDM, combines ResNet and Vision Transformer (ViT) to enhance defect detection by leveraging multi-channel image segmentation and convolution operations within patches. This approach improves the identification of high-information-density areas while optimizing computational efficiency for industrial applications. Experimental results show that SDDM achieves 98.6% accuracy on imbalanced datasets, improving productivity in IC packaging and testing.

Bhure et al. [3] discuss the challenges posed by counterfeit components in the global semiconductor supply chain, which threaten product quality and reliability. The distributed nature of manufacturing and distribution increases the risk of fraud. To address this, they propose a Vision Transformer model for counterfeit IC detection, capable of classifying authentic samples and 11 types of counterfeit defects. Their model achieves 88% accuracy on the test set, with attention map visualizations providing insights into its predictions [3].

3.3 Side-Channel Attacks

Recent advancements in leveraging Transformer models for the mitigation of side channel attacks (SCA) have significantly improved

Table 1: System-Level Hardware Vulnerability Detection using Attention-Based Methods

Category	Subsection	Description	Attention Usage
Design-Time Security	Vulnerability Classification [12, 26]	Categorizing flaws in CWE and NVD datasets.	LLMs for processing datasets through a storytelling framework to suggest mitigations.
	Trojan Detection [10, 25]	Detect hardware trojans in 3P IPs or RTL.	Graph Attention Networks (GATs) to detect unusual paths.
	Information Flow Tracking (IFT) [30]	Trace propagation of sensitive data.	LLM-driven hierarchical dependency analysis across modules.
	Defect & Packaging Analysis [3, 57]	Detect packaging-related defects & fake components.	Transformer models trained on SEM/X-ray images or electrical response sequences.
Runtime Security	Malware Detection [8, 44, 51]	Detect malicious behavior during operation.	Binaries converted into images, and ViTs are used instead.
	Side-channel Attack Detection [15, 16]	Identify information leakage via timing, EM, power, etc.	attention-modal on power traces, timing, etc.
	Fault Injection / Reliability-Aware Security [53]	Detect when injected faults cause vulnerabilities.	Multi-modal attention models combining error signals and system state.
	HW-SW Interface Exploits [1, 59, 60]	Firmware and IoT systems face security risks.	Transformers model firmware, system calls, and network traffic as sequences to detect complex attacks.
Post-Deployment Assurance	Device Fingerprinting [43, 52]	Identify unique behavior profiles of devices.	Attention over power, Radio Frequency (RF) characteristics, or EM patterns.
	Remote Attestation [11]	Verify system integrity over a network.	Transformers for cyber threat detection in IoT networks.
	Insider Threat Detection [41]	Legitimate user abusing access privileges.	Attention models over system access logs and patterns.

the robustness and effectiveness of cryptographic systems [39]. Berreby and Sauvage's work [2] employed the ANSSI Side-Channel Attack Database (ASCAD) and introduced a JAX-based framework specifically tailored for exploiting side-channel vulnerabilities in AES implementations.

The SCAR framework [54] introduces a pre-silicon mitigation strategy utilizing Graph Neural Networks (GNNs) integrated with LLMs for early detection and automatic mitigation of power side-channel (PSC) leakage vulnerabilities at the Register-Transfer Level (RTL). SCAR converts RTL designs into control-data flow graphs (CDFGs), where nodes represent RTL basic blocks, and edges represent control flow. A GNN-based classification approach is applied for identifying modules susceptible to leakage, supported by LLM-generated mitigation code to automatically fortify identified vulnerable RTL code segments. This technique significantly enhances the early detection and mitigation of side-channel vulnerabilities, embedding security deeply within the hardware design cycle. TransNet [16] offers a significant advance by incorporating shift invariance into transformer networks, a critical feature to effectively manage desynchronized power traces, and a common side channel countermeasure. TransNet leverages a modified Transformer architecture incorporating relative positional encoding and self-attention mechanisms that effectively capture dependencies among distant points of interest (PoIs).

Building on TransNet, EstraNet [15] further optimizes the Transformer network specifically for SCA by achieving linear time and memory complexity. EstraNet introduces the GaussiP self-attention layer, an approach featuring relative positional encoding for enhanced shift-invariance and computational efficiency. Additionally, EstraNet employs a new layer-centering normalization method instead of traditional batch or layer normalization techniques, overcoming typical normalization challenges encountered in SCA applications. EstraNet demonstrated substantial robustness against masking, random delays, and clock jitter. Its scalability to very long traces (exceeding several thousand points) highlights its practical applicability in real-world cryptographic security scenarios.

3.4 Malware Detection

Transformer models, with their powerful ability to capture intricate dependencies through self-attention mechanisms, have demonstrated significant potential in malware detection. A notable study by Seneviratne et al. introduces SHERLOCK [51], a self-supervised deep learning framework that leverages ViT to detect Android malware. SHERLOCK adopts a self-supervised approach to learn robust feature representations from unlabeled binary samples. The binaries are converted into grayscale images, allowing the model to recognize malware patterns visually. To address the challenge of deploying effective malware detection on resource-constrained edge devices, Ravi et al. proposed ViT4Mal [44]. ViT4Mal also converts executable byte-code into images and employs a customized, lightweight ViT architecture designed explicitly for limited-resource hardware.

Another innovative approach, TransMalDE, developed by Deng et al. [8], targets the detection of IoT malware through a hierarchical Transformer-based framework. The TransMalDE framework migrates computationally intensive malware detection tasks from IoT devices to edge computing nodes, thus significantly reducing detection latency. TransMalDE captures the latent behavior patterns characteristic of evolving IoT malware by analyzing the textual semantic patterns that traditional methods might overlook.

Further extending Transformer models to utilize process resource utilization metrics, Natsos and Symeonidis present a dynamic malware detection technique [35]. The authors encode input data as sequences of processes, each represented by its resource metrics (CPU, memory, disk usage). Additionally, this research introduces dynamic malware signatures derived from resource metrics, revealing indirect malware activity indicators through cascading effects on system-wide processes. These studies underscores the Transformer model's versatility and robustness in diverse malware detection contexts, from resource-rich cloud environments [28] to resource-constrained edge devices, outperforming traditional ML techniques.

3.5 Device Fingerprinting

Similar to machine learning fingerprinting [36], device fingerprinting is also crucial for securing IoT communications against spoofing and cloning attacks by extracting unique radio frequency (RF) characteristics. Recent advancements in this domain extensively use deep learning methods for automatic feature extraction from transmitted signals, significantly improving the accuracy and robustness of device identification. Wu et al. [58] leveraged an LSTM recurrent neural network for RF fingerprinting. Their model automatically captures intrinsic hardware-specific features, such as frequency drift and transient behaviors, achieving high accuracy even in environments with significant noise interference. Lee et al. [24] further explored deep-learning-aided RF fingerprinting in Near Field Communication (NFC) systems and utilized neural networks, including CNN and RNN.

Focusing on IoT scenarios, Jafari et al. [19] used CNN and RNN to identify individual ZigBee devices. Their experiments across varied signal-to-noise ratio (SNR) conditions showed robust accuracy in distinguishing identical devices.Shen et al. [52] introduced Transformer-based models specifically for LoRa device fingerprinting, effectively managing variable-length signals and improving accuracy significantly through data augmentation and multi-packet inference, especially in low SNR conditions. Building upon Transformer architectures, Parpart et al. [43] proposed transformer masked autoencoders pre-trained on large-scale unlabeled RF data. Their method significantly enhanced classification accuracy, surpassing traditional CNN-based models, and demonstrated efficient handling of extensive datasets with thousands of devices.

3.6 Remote Attestation

The rapid growth of IoT devices demands efficient mechanisms for detecting network-based attacks. SecurityBERT is a lightweight BERT-based model for cyber threat detection in IoT networks, using a novel Privacy-Preserving Fixed-Length Encoding (PPFLE) combined with a Byte-level tokenizer. Trained on the Edge-IIoTset dataset, it achieves 98.2% accuracy across 14 attack types, outperforming traditional ML/DL methods while maintaining low inference time and small model size, making it ideal for resource-constrained IoT devices [11].

3.7 Insider Threat Detection

Insider threats pose a serious challenge for hardware systems, as they can exploit behavioral drift and data imbalance to stay hidden within normal activities. Detecting such threats is difficult due to the rarity of malicious behavior compared to regular user actions. [41] proposes an insider threat detection approach using an ensemble of stacked-LSTM and stacked-GRU attention models trained on sequential activity logs. A new equally-weighted random sampling technique balances different threat categories, improving model fairness and performance.

3.8 Hardware-Software Inference Exploits

The hardware-software interface is a possible source of security vulnerabilities caused by misconfiguration. Transformers can be applied to detect the vulnerability patterns through analyzing firmware routines, system calls, and instruction-level execution traces as sequences. The rise of IoT devices has further exposed firmware over-the-air (OTA) updates to threats like distributed denial of service (DDoS) attacks. Notably, the Mirai botnet exploited IoT vulnerabilities to launch massive attacks. Recent works, such as DDoSViT, leverage Vision Transformers (ViTs) by converting attack flows into images and training on datasets like CICIoT2023 and CICIoMT2024. These models achieved detection rates as high as 99.50%, demonstrating the effectiveness of attention-based methods in identifying complex multi-vector attacks across the hardware-software boundary [1]. FirmVulSeeker is a firmware vulnerability search tool that leverages BERT pretraining and a Siamese network to match semantically similar functions [60]. SLFHunter is a LLM-based framework for detecting command injection vulnerabilities in embedded Linux firmware. By identifying sensitive dynamically linked library functions (DLLFs) with ChatGPT, it marks new sinks for static analysis tools like EmTaint [59].

3.9 Information Flow Tracking

Information flow tracking (IFT) is a method for monitoring how data moves through a system. It aims to detect unauthorized or suspicious flows that may lead to vulnerabilities such as data leakage or privilege escalation. LLM-IFT leverages LLM-driven hierarchical dependency analysis across intra- and inter-module levels to overcome the scalability and adaptability limitations of traditional IFT methods. The approach achieves 100% success in confidentiality and integrity checks on Trust-Hub benchmarks, shows the effectiveness of LLMs for security analysis in integrated circuits [30].

3.10 Fault Detection

Fault attacks, such as electromagnetic (EM) injection, voltage glitching, and clock manipulation, are powerful techniques for inducing errors in hardware and bypassing security mechanisms. Smart Monitor is a hardware framework that uses on-chip sensors and an AI core to detect and classify fault attacks like electromagnetic (EM) and clock-glitch (CG) injections. It achieves 92% detection accuracy and 78% classification accuracy with zero false positives [53].

3.11 Vulnerability Classification

With the increasing number of reported vulnerabilities, datasets like the National Vulnerability Database (NVD), Common Weakness Enumeration (CWE), and Common Vulnerabilities and Exposures (CVE) have become essential for developing effective security solutions. However, extracting meaningful insights from these vast datasets is not easy, making it necessary to apply AI techniques for effective vulnerability classification. VulExplainer is a deep learning approach for classifying and explaining vulnerabilities using a Transformer-based hierarchical distillation framework. It improves classification accuracy by 5%–29% and is compatible with models like CodeBERT, GraphCodeBERT, and CodeGPT without architectural modifications [12]. The HW-V2W-Map Framework is a machine learning-based approach for mapping hardware vulnerabilities to weaknesses, with a focus on IoT security. It incorporates an Ontology-driven Storytelling framework to track vulnerability trends. Additionally, it leverages GPT-based LLMs to generate mitigation strategies, helping to predict and prevent future vulnerabilities [26]. To enhance Industrial Control System (ICS) security, this study proposes deep learning-based automated vulnerability categorization. Given the limitations of national NVDs, the authors demonstrate that LSTM-tuned BERT models achieve superior precision, F1 score, accuracy, and recall. This approach strengthens cyber threat intelligence (CTI) and improves attack mitigation in ICS environments [29].

4 Challenges and Opportunities

4.1 Computational Cost and Overhead

Transformer models offer strong performance for hardware security tasks like side-channel analysis, Trojan detection, and malware classification, but they come with high computational costs. Their self-attention mechanism has quadratic complexity, making it difficult to deploy them directly on resource-constrained systems like embedded devices or FPGAs. To address this, some models like EstraNet and TA-MobileViT use efficient attention mechanisms or lightweight architectures that reduce memory and power usage while maintaining detection accuracy. Various optimization techniques help reduce overhead. These include pruning unnecessary weights, quantizing models to use lower precision numbers, and using efficient attention mechanisms like sparse or linear attention. Such changes can make models faster and more suitable for low-power environments with minimal impact on performance. Designs often combine convolution layers with attention or reduce input size early to ease processing.

In real-world deployments, heavy computation is often offloaded to edge servers to lighten the burden on local devices, as seen in frameworks like TransMalDE. While this reduces latency and energy use on-device, it can introduce new trade-offs like network dependency. Overall, by applying the right balance of architectural changes and deployment strategies, Transformers can be effectively used for secure, real-time applications in hardware systems.

4.2 Lack of Explainability

One major challenge in using transformers for hardware vulnerability detection and mitigation is the lack of explainability and interpretability. These models function as black boxes, making it difficult to understand their decision-making process, which is crucial for security-critical applications. Without clear insights into how vulnerabilities are identified, debugging and trust in the system become significant concerns. Additionally, the absence of interpretability hinders the ability to validate predictions and ensure robustness against adversarial attacks. Solutions to address these limitations include attention visualization, feature attribution methods, saliency-guided training to enhance model focus on relevant features, and hybrid approaches that integrate symbolic reasoning for improved transparency [20].

4.3 Vulnerability Against Adversarial Attacks

Despite their effectiveness, deep learning models remain highly susceptible to adversarial attacks, and attention-based architectures, including Transformers, are no exception. Adversarial attacks involve crafting subtly perturbed inputs that mislead the model into making incorrect predictions or failing to detect malicious behavior [31]. This vulnerability poses a significant challenge for security-critical applications such as malware detection, where adversaries can potentially exploit model weaknesses to evade detection [9].

To address these risks, researchers have explored defense mechanisms such as adversarial training, which improves robustness by incorporating adversarial examples during model training. He et al. [17] has demonstrated that ML-based malware detection systems are susceptible to adversarial attacks, especially those operating on structured tabular data like processor performance counters. To address this, a multi-phased defense framework based on Deep Reinforcement Learning (DRL) was introduced, combining adversarial training with real-time attack pattern prediction and dynamic defense assignment via a UCB-guided controller, This approach significantly improved detection robustness, achieving up to an 86% increase in F1-score.

In addition, recent studies have also highlighted the vulnerability of Transformer models to Bit-Flip Attacks (BFAs), where adversaries manipulate a small number of model parameters at the binary level to degrade performance. A novel defense strategy, Forget and Rewire (FaR) [40], introduces targeted rewiring in Transformer Linear layers to obscure critical neurons and redistribute computation, significantly improving model robustness with minimal accuracy loss. Additionally, hybrid approaches that combine deep learning with rule-based or statistical methods may offer enhanced reliability by introducing complementary detection layers. As Transformer-based models are increasingly adopted in hardware and system security, developing robust, interpretable, and attack-resilient architectures remains a key research priority.

4.4 Availability of Dataset

One of the main challenges in applying attention-based models is their reliance on large-scale, high-quality datasets for effective training. However, in this research domain, such datasets are often scarce or inaccessible. In many cases, access to design files or detailed system information needed to construct meaningful datasets is restricted, as hardware companies are unwilling to share proprietary or sensitive data. This lack of transparency significantly hinders dataset creation and model development. To address this limitation, researchers must explore alternative solutions such as synthetic data generation, simulation-based environments, or data augmentation techniques to produce diverse and representative training samples.

4.5 Real-Time Hardware-Level Threat Detection

Hardware-level threat detection has gained momentum as a robust complement to traditional software-based defenses, offering deeper visibility into runtime behavior through low-level microarchitectural signals. Hardware-Assisted Malware Detection (HMD) techniques leverage sources such as Hardware Performance Counters (HPCs) and other on-chip telemetry to collect fine-grained execution traces, enabling the identification of anomalous and potentially malicious activities in real-time [7, 13, 49, 50]. Transformers present compelling opportunities for advancing HMDs. Their self-attention mechanism enables context-aware modeling of execution behavior, allowing the capture of subtle correlations in tabular hardware data that may be overlooked by traditional models. This capability is particularly valuable for detecting complex and stealthy attack patterns, including zero-day malware that manifests across temporal sequences or diverse microarchitectural events [18, 47]. Moreover, Transformers' ability to handle multi-modal inputs makes them well-suited for fusing heterogeneous side-channel signals (e.g., HPCs, power telemetry) into a unified representation, enabling a more in-depth analysis of system behavior under attack.

However, there exist several challenges in employing Transformers for hardware-assisted malware detection. Transformers are computationally demanding, limiting their deployment in resource-constrained environments where HMD is most needed. Their use with structured tabular data, such as HPC traces, also requires careful architectural tuning and feature encoding. Additionally, interpretability and robustness remain critical concerns, especially in security-sensitive applications where transparency is essential [42, 48]. Despite these challenges, the continued evolution of Tiny Transformers, efficient attention mechanisms, and self-supervised pretraining on hardware execution traces presents a path forward. Future work could explore hybrid architectures that balance accuracy and efficiency, domain-specific pretraining to improve generalization, and adversarial training techniques to strengthen resilience. As malware grows more sophisticated, the integration of Transformers into HMD systems, and more broadly across hardware-assisted security domains, holds the potential to unlock next-generation, adaptive, and intelligent threat detection frameworks.

5 Concluding Remarks

This paper presents a comprehensive review of recent developments in the application of Transformer-based models for hardware security. We examine their application across a wide range of critical domains, including side-channel analysis, hardware Trojan detection, device fingerprinting, vulnerability classification, malware detection, and firmware security. With their ability to capture complex dependencies and process multi-modal inputs, Transformers represent a significant shift in the design of intelligent, context-aware threat detection mechanisms at the hardware level. Despite their considerable promise, the application of Transformers to hardware-based security remains an emerging area with several open challenges, each of which also presents opportunities for future exploration. Key issues include computational overhead, susceptibility to adversarial attacks, limited interpretability, limited publicly available datasets, and the need for enhanced adaptability to structured hardware telemetry. Addressing these limitations will be critical to enabling scalable, trustworthy, and widely deployable Transformer-based security solutions, and offers a rich landscape for advancing research in secure and intelligent hardware systems.

Acknowledgments

This work is supported in part by the National Science Foundation under Award No. 2139034.

References

[1] Muhammad Ali, Yasir Saleem, Sadaf Hina, and Ghalib A Shah. 2025. DDoSViT: IoT DDoS attack detection for fortifying firmware Over-The-Air (OTA) updates using vision transformer. *Internet of Things* (2025), 101527.

[2] Yohaï-Eliel Berreby and Laurent Sauvage. 2023. Investigating efficient deep learning architectures for side-channel attacks on AES. *arXiv preprint arXiv:2309.13170* (2023).

[3] Chaitanya Bhure, Dhruvakumar Aklekar, Wenjie Che, and Fareena Saqib. 2024. Vision Transformers for Counterfeit IC Detection. In *2024 IEEE Physical Assurance and Inspection of Electronics (PAINE)*. 1–7. doi:10.1109/PAINE62042.2024.10792793

[4] Menghui Chen, Xiaoyong Kou, and Gongxuan Zhang. 2024. TrojanFormer: Resource-Efficient Hardware Trojan Detection Using Graph Transformer Network. In *2024 7th International Conference on Electronics Technology (ICET)*. IEEE, 165–170.

[5] Shouhong Chen, Guanxiang Qin, Ying Lu, Tao Wang, and Xingna Hou. 2025. A lightweight Hardware Trojan detection approach in the waveform diagram based on MobileViT and attention mechanism. *The Journal of Supercomputing* 81 (2025), 580.

[6] Xing Chen, Yonglei Wu, Xingyou He, and Wuyi Ming. 2023. A comprehensive review of deep learning-based PCB defect detection. *IEEE Access* 11 (2023), 139017–139038.

[7] John Demme, Matthew Maycock, Jared Schmitz, Adrian Tang, Adam Waksman, Simha Sethumadhavan, and Salvatore Stolfo. 2013. On the feasibility of online malware detection with performance counters. *ACM SIGARCH computer architecture news* 41, 3 (2013), 559–570.

[8] Xiaoheng Deng, Zhe Wang, Xinjun Pei, and Kaiping Xue. 2023. TransMalDE: an effective transformer based hierarchical framework for IoT malware detection. *IEEE Transactions on Network Science and Engineering* 11, 1 (2023), 140–151.

[9] Sai Manoj Pudukotai Dinakarrao et al. 2019. Adversarial attack on microarchitectural events based malware detectors. In *Proceedings of the 56th Annual Design Automation Conference 2019*. 1–6.

[10] Md Omar Faruque, Peter Jamieson, Ahmad Patooghy, and Abdel-Hameed A Badawy. 2024. TrojanWhisper: Evaluating Pre-trained LLMs to Detect and Localize Hardware Trojans. *arXiv preprint arXiv:2412.07636* (2024).

[11] Mohamed Amine Ferrag et al. 2024. Revolutionizing Cyber Threat Detection With Large Language Models: A Privacy-Preserving BERT-Based Lightweight Model for IoT/IIoT Devices. *IEEE Access* 12 (2024), 23733–23750. doi:10.1109/ACCESS.2024.3363469

[12] Michael Fu, Van Nguyen, Chakkrit Kla Tantithamthavorn, Trung Le, and Dinh Phung. 2023. Vulexplainer: A transformer-based hierarchical distillation for explaining vulnerability types. *IEEE Transactions on Software Engineering* 49, 10 (2023), 4550–4565.

[13] Yifeng Gao, Hosein Mohammadi Makrani, Mehrdad Aliasgari, Amin Rezaei, Jessica Lin, Houman Homayoun, and Hossein Sayadi. 2021. Adaptive-hmd: Accurate and cost-efficient machine learning-driven malware detection using microarchitectural events. In *2021 IEEE 27th International Symposium on On-Line Testing and Robust System Design (IOLTS)*. IEEE, 1–7.

[14] Kevin Immanuel Gubbi, Banafsheh Saber Latibari, Anirudh Srikanth, Tyler Sheaves, Sayed Arash Beheshti-Shirazi, Sai Manoj PD, Satareh Rafatirad, Avesta Sasan, Houman Homayoun, and Soheil Salehi. 2023. Hardware trojan detection using machine learning: A tutorial. *ACM Transactions on Embedded Computing Systems* 22, 3 (2023), 1–26.

[15] Suvadeep Hajra et al. 2024. Estranet: An efficient shift-invariant transformer network for side-channel analysis. *IACR Transactions on Cryptographic Hardware and Embedded Systems* 2024, 1 (2024), 336–374.

[16] Suvadeep Hajra, Sayandeep Saha, Manaar Alam, and Debdeep Mukhopadhyay. 2022. Transnet: Shift invariant transformer network for side channel analysis. In *International Conference on Cryptology in Africa*. Springer, 371–396.

[17] Zhangying He, Houman Homayoun, and Hossein Sayadi. 2024. Beyond conventional defenses: Proactive and adversarial-resilient hardware malware detection using deep reinforcement learning. In *Proceedings of the 61st ACM/IEEE Design Automation Conference*. 1–6.

[18] Zhangying He, Houman Homayoun, and Hossein Sayadi. 2024. Guarding against the unknown: Deep transfer learning for hardware image-based malware detection. *Journal of Hardware and Systems Security* 8, 2 (2024), 61–78.

[19] Hossein Jafari, Oluwaseyi Omotere, Damilola Adesina, Hsiang-Huang Wu, and Lijun Qian. 2018. IoT Devices Fingerprinting Using Deep Learning. In *MILCOM 2018 - 2018 IEEE Military Communications Conference (MILCOM)*. 1–9. doi:10.1109/MILCOM.2018.8599826

[20] Ali Karkehabadi, Banafsheh Saber Latibari, Houman Homayoun, and Avesta Sasan. 2024. HLGM: A novel methodology for improving model accuracy using saliency-guided high and low gradient masking. In *2024 14th International Conference on Information Science and Technology (ICIST)*. IEEE, 909–917.

[21] Shijie Kuang, Zhe Quan, Guoqi Xie, Xiaomin Cai, Xiaoqian Chen, and Keqin Li. 2025. NtNDet: Hardware Trojan detection based on pre-trained language models. *Expert Systems with Applications* (2025), 126666.

[22] Banafsheh Saber Latibari, Houman Homayoun, and Avesta Sasan. 2025. Optimizing Vision Transformers: Unveiling'Focus and Forget'for Enhanced Computational Efficiency. *IEEE Access* (2025).

[23] Banafsheh Saber Latibari, Najmeh Nazari, Muhtasim Alam Chowdhury, Kevin Immanuel Gubbi, Chongzhou Fang, Sujan Ghimire, Elahe Hosseini, Hossein Sayadi, Houman Homayoun, Soheil Salehi, et al. 2024. Transformers: A Security Perspective. *IEEE Access* (2024).

[24] Woongsup Lee, Seon Yeob Baek, and Seong Hwan Kim. 2021. Deep-Learning-Aided RF Fingerprinting for NFC Security. *IEEE Communications Magazine* 59, 5 (2021), 96–101. doi:10.1109/MCOM.001.2000912

[25] Yilin Li, Shan Li, and Haihua Shen. 2023. Htrans: Transformer-based method for hardware trojan detection and localization. In *2023 IEEE 32nd Asian Test Symposium (ATS)*. IEEE, 1–6.

[26] Yu-Zheng Lin, Muntasir Mamun, Muhtasim Alam Chowdhury, Shuyu Cai, Mingyu Zhu, Banafsheh Saber Latibari, Kevin Immanuel Gubbi, Najmeh Nazari Bavarsad, Arjun Caputo, Avesta Sasan, et al. 2023. Hw-v2w-map: Hardware vulnerability to weakness mapping framework for root cause analysis with gpt-assisted mitigation suggestion. *arXiv preprint arXiv:2312.13530* (2023).

[27] Fei Liu, Heng Wang, Pingfa Feng, and Long Zeng. 2024. Integrated Circuit Packaging Defect Analysis and Deep Learning Detection Method. *IEEE Transactions on Components, Packaging and Manufacturing Technology* 14, 9 (2024), 1707–1719. doi:10.1109/TCPMT.2024.3447040

[28] Hosein Mohammadi Makrani et al. 2021. Security threats in cloud rooted from machine learning-based resource provisioning systems. In *Silicon Valley Cybersecurity Conference*. Springer, 22–32.

[29] Mounesh Marali et al. 2024. A hybrid transformer-based BERT and LSTM approach for vulnerability classification problems. *International Journal of Mathematics in Operational Research* 28, 3 (2024), 275–295.

[30] Nowfel Mashnoor, Mohammad Akyash, Hadi Kamali, and Kimia Azar. 2025. LLM-IFT: LLM-Powered Information Flow Tracking for Secure Hardware. *arXiv preprint arXiv:2504.07015* (2025).

[31] Ali Mirzaeian, Zhi Tian, Sai Manoj PD, Banafsheh S Latibari, Ioannis Savidis, Houman Homayoun, and Avesta Sasan. 2022. Adaptive-Gravity: A Defense Against Adversarial Samples. In *2022 23rd International Symposium on Quality Electronic Design (ISQED)*. IEEE, 96–101.

[32] Mohamad Mohamadsalehi. 2022. *Expansion of Conforming to Interface Structured Adaptive Mesh Refinement Algorithm to Higher Order Elements and Crack Propagation*. The Ohio State University.

[33] Mohamad Mohmadsalehi and Soheil Soghrati. 2022. An automated mesh generation algorithm for simulating complex crack growth problems. *Computer Methods in Applied Mechanics and Engineering* 398 (2022), 115015.

[34] Abdurrahman Nasr, Khalil Mohamed, Mohamed Zaki, et al. 2024. Improving Hardware Trojan Detection with Transformer-Based Power Analysis. (2024).

[35] Dimosthenis Natsos and Andreas L Symeonidis. 2025. Transformer-based malware detection using process resource utilization metrics. *Results in Engineering* (2025), 104250.

[36] Najmeh Nazari et al. 2024. LLM-FIN: Large Language Models Fingerprinting Attack on Edge Devices. In *2024 25th International Symposium on Quality Electronic Design (ISQED)*. IEEE, 1–6.

[37] Najmeh Nazari et al. 2024. Securing On-Chip Learning: Navigating Vulnerabilities and Potential Safeguards in Spiking Neural Network Architectures. In *2024 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 1–5.

[38] Najmeh Nazari et al. 2024. SpecScope: Automating Discovery of Exploitable Spectre Gadgets on Black-Box Microarchitectures. In *2024 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, 1–6.

[39] Najmeh Nazari, Chongzhou Fang, Hosein Mohammadi Makrani, Behnam Omidi, Mahdi Eslamimehr, Setareh Rafatirad, Avesta Sasan, Hossein Sayadi, Khaled N Khasawneh, and Houman Homayoun. 2024. Architectural Whispers: Robust Machine Learning Models Fingerprinting via Frequency Throttling Side-Channels. In *Proceedings of the 61st ACM/IEEE Design Automation Conference*. 1–6.

[40] Najmeh Nazari, Hosein Mohammadi Makrani, Chongzhou Fang, Hossein Sayadi, Setareh Rafatirad, Khaled N Khasawneh, and Houman Homayoun. 2024. Forget and Rewire: Enhancing the Resilience of Transformer-based Models against Bit-Flip Attacks. In *33rd USENIX Security Symposium*. 1349–1366.

[41] Preetam Pal, Pratik Chattopadhyay, and Mayank Swarnkar. 2023. Temporal feature aggregation with attention for insider threat detection from activity logs. *Expert Systems with Applications* 224 (2023), 119925.

[42] Zhixin Pan, Jennifer Sheldon, and Prabhat Mishra. 2022. Hardware-assisted malware detection and localization using explainable machine learning. *IEEE Trans. Comput.* 71, 12 (2022), 3308–3321.

[43] Gavin Parpart, Jonathan H Tu, Bradley Clymer, Jung Lee, and Jasen Babcock. 2024. Transformer Masked Autoencoders for RF Device Fingerprinting. In *MILCOM 2024-2024 IEEE Military Communications Conference (MILCOM)*. IEEE, 859–862.

[44] Akshara Ravi, Vivek Chaturvedi, and Muhammad Shafique. 2023. Vit4mal: Lightweight vision transformer for malware detection on edge devices. *ACM Transactions on Embedded Computing Systems* 22, 5s (2023), 1–26.

[45] Banafsheh Saber Latibari, Soheil Salehi, Houman Homayoun, and Avesta Sasan. 2024. IRET: Incremental Resolution Enhancing Transformer. In *Proceedings of the Great Lakes Symposium on VLSI 2024*. 620–625.

[46] Hossein Sayadi, Mehrdad Aliasgari, Furkan Aydin, Seetal Potluri, Aydin Aysu, Jack Edmonds, and Sara Tehranipoor. 2022. Towards ai-enabled hardware security: Challenges and opportunities. In *2022 IEEE 28th International Symposium on On-Line Testing and Robust System Design (IOLTS)*. IEEE, 1–10.

[47] Hossein Sayadi, Yifeng Gao, Hosein Mohammadi Makrani, Tinoosh Mohsenin, Avesta Sasan, Setareh Rafatirad, Jessica Lin, and Houman Homayoun. 2020. Stealthminer: Specialized time series machine learning for run-time stealthy malware detection based on microarchitectural features. In *Proceedings of the 2020 on Great Lakes Symposium on VLSI*. 175–180.

[48] Hossein Sayadi, Zhangying He, Hosein Mohammadi Makrani, and Houman Homayoun. 2024. Intelligent malware detection based on hardware performance counters: A comprehensive survey. In *2024 25th International Symposium on Quality Electronic Design (ISQED)*. IEEE, 1–10.

[49] Hossein Sayadi, Hosein Mohammadi Makrani, Sai Manoj Pudukotai Dinakarrao, Tinoosh Mohsenin, Avesta Sasan, Setareh Rafatirad, and Houman Homayoun. 2019. 2smart: A two-stage machine learning-based approach for run-time specialized hardware-assisted malware detection. In *2019 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, 728–733.

[50] Hossein Sayadi, Nisarg Patel, Avesta Sasan, Setareh Rafatirad, and Houman Homayoun. 2018. Ensemble learning for effective run-time hardware-based malware detection: A comprehensive analysis and classification. In *Proceedings of the 55th Annual Design Automation Conference*. 1–6.

[51] Sachith Seneviratne, Ridwan Shariffdeen, Sanka Rasnayaka, and Nuran Kasthuriarachchi. 2022. Self-supervised vision transformers for malware detection. *IEEE Access* 10 (2022), 103121–103135.

[52] Guanxiong Shen, Junqing Zhang, Alan Marshall, Mikko Valkama, and Joseph Cavallaro. 2021. Radio frequency fingerprint identification for security in low-cost IoT devices. In *2021 55th Asilomar conference on signals, systems, and computers*. IEEE, 309–313.

[53] Ritu-Ranjan Shrivastwa, Sylvain Guilley, and Jean-Luc Danger. 2021. Multi-source fault injection detection using machine learning and sensor fusion. In *International Conference on Security and Privacy*. Springer, 93–107.

[54] Amisha Srivastava et al. 2024. SCAR: Power Side-Channel Analysis at RTL Level. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems* (2024).

[55] Ashish Vaswani et al. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[56] Han Wang, Hossein Sayadi, Sai Manoj Pudukotai Dinakarrao, Avesta Sasan, Setareh Rafatirad, and Houman Homayoun. 2021. Enabling micro AI for securing edge devices at hardware level. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems* 11, 4 (2021), 803–815.

[57] Xiaobin Wang, Shuang Gao, Jianlan Guo, Chu Wang, Liping Xiong, and Yuntao Zou. 2024. Deep learning-based integrated circuit surface defect detection: Addressing information density imbalance for industrial application. *International Journal of Computational Intelligence Systems* 17, 1 (2024), 29.

[58] Qingyang Wu, Carlos Feres, Daniel Kuzmenko, Ding Zhi, Zhou Yu, Xin Liu, and Xiaoguang 'Leo'Liu. 2018. Deep learning based RF fingerprinting for device identification and wireless security. *Electronics Letters* 54, 24 (2018), 1405–1407.

[59] Junjian Ye et al. 2024. Detecting command injection vulnerabilities in Linux-based embedded firmware with LLM-based taint analysis of library functions. *Computers & Security* 144 (2024), 103971.

[60] Yingchao Yu, Shuitao Gan, and Xiaojun Qin. 2021. firm VulSeeker: BERT and Siamese based Vulnerability for Embedded Device Firmware Images. In *2021 IEEE Symposium on Computers and Communications (ISCC)*. 1–7.