Cracking the Chip: AI-Powered Security for Semiconductor Threats

Shreyas Kumar, Shruti Oruganti and Isha Virk
Department of Computer Science and Engineering, Texas A&M University, College Station, USA

shreyas.kumar@tamu.edu
shruti_oruganti@tamu.edu
Ishavirk@tamu.edu

Abstract: Semiconductor supply chains have become increasingly vulnerable to sophisticated, low-level threats that originate in the early phases and propagate undetected across various stages of semiconductor device production. As semiconductor systems grow more complex and globally interconnected, these low-level design threats present significant risks, including data breaches, system failures, and long-term erosion of reliability. This paper presents a comprehensive AI-driven framework to detect, model, and mitigate hardware security threats across the semiconductor supply chain, from design and fabrication to assembly. We begin with the design phase, illustrating how vulnerabilities like hardware Trojans in third-party IP blocks, compromised EDA scripts, and speculative execution side-channels can be exploited. AI techniques, such as anomaly detection for logic integrity, dynamic hashing for secure script flows, and entropy-based instruction shuffling, are shown to proactively block or obfuscate these attacks. These models serve as templates for following stages, including fabrication (tampered masks or altered process flows), assembly and packaging (hardware fingerprinting), and post-silicon validation (malicious test routines or data exfiltration). Our contributions include a stage-wise breakdown of threat surfaces across the supply chain and the design of threat models with corresponding AI-driven defenses that analyze patterns, enforce trust boundaries, and obfuscate system behavior. Additionally, to assess the viability of these defenses, we outline a validation framework involving simulated and prototyped defenses, which include instruction shuffling, JTAG interface monitoring, and machine learning-based fault pattern analysis. Proposed evaluation metrics include detection accuracy, computational overhead, entropy of runtime traces, and classification accuracy. By addressing persistent security threats early and continuously through the chip lifecycle, we aim to leverage AI to shift hardware security from reactive patching to proactive risk management. Our work emphasizes the importance of securing semiconductor systems at their root, offering a path toward proactive hardware security and highlights the need for scalable, interdisciplinary solutions at the intersection of AI, hardware design, and supply chain security.

Keywords: AI-Driven security, Hardware threat modeling, Adaptive defense, Semiconductor supply chain security

1. Introduction

Semiconductors, also commonly known as integrated circuits in today's age, drive nearly all devices around the world. From mobile devices to life-saving medical devices, these small yet powerful chips are the foundation of the digital age and would not exist without the critical industry behind it. The semiconductor supply chain is an intricate system that is composed of many interconnected stages. Its sprawl lies on a global scale, which is crucial in upholding its infrastructure. Due to the number and complexity of its stages, an effective global framework is critical as "no singular country", or company, can perform all roles in the supply chain in a reliable manner to produce the numerous types of semiconductors needed in the international market (Thadani and Allen 2023). Considering the vital role that the semiconductor supply chain plays and its reliance on global interconnection, it is increasingly important that vulnerabilities are rooted out along each stage to preserve its function. The existence of security threats at any stage of the supply chain serves not only to weaken its infrastructure but will also cause detrimental repercussions for the consumers that depend on semiconductors. With the rapid advancement of artificial intelligence, it is critical for this high-value industry to seek how they can leverage techniques powered by AI to reinforce its supply chain, an infrastructure that is crucial to global economic function.

1.1 Motivation

Hardware security vulnerabilities often originate during the early stages of chip design and persist throughout the entire lifecycle of a semiconductor. These low-level threats often serve as the root cause of cascading security issues, ultimately manifesting as critical software-level security breaches. As modern systems become increasingly complex and interconnected, the consequences of such vulnerabilities grow more severe, ranging from compromised data integrity to large-scale cyber-attacks. Such weaknesses create rippling consequences not only in the semiconductor industry, but for the various economic sectors that depend on semiconductor technology as well. This paper is motivated by the urgent need to proactively identify and mitigate hardware-level threats at their point of origin. By leveraging AI-driven techniques, we aim to enhance the resilience of semiconductor systems and disrupt the attack chain before it escalates to higher system layers.

1.2 Contributions

This paper seeks to advance existing research concerning the usage of artificial intelligence to target hardware vulnerabilities in the semiconductor supply chain and demonstrate how key figures within the industry can reinforce their sector to mitigate current and theoretical threats. The key contributions of this paper include:

- Comprehensive Threat Categorization Across Supply Chain Stages: We present a stage-wise categorization of hardware threat vectors, such as design-time Trojans, test access misuse (e.g., JTAG abuse), and counterfeit components, and map them to realistic attack scenarios and propagation paths.
- AI-Powered Detection and Mitigation Techniques: We propose a suite of machine learning-based techniques tailored to specific vulnerabilities, including FSM anomaly detection, optical/SEM image classification of chip defects, and instruction-level runtime obfuscation guided by AI policies.
- Framework for Adaptive Defense: We introduce a modular defense framework where AI agents can adaptively respond to threats depending on the supply chain phase, enabling detection as well as real-time mitigation (e.g., dynamic instruction shuffling or scan chain monitoring).
- Validation Methodology: We propose a structured evaluation foundation using detection time, classification accuracy, and reverse engineering resistance as metrics, and describe prototype implementations to support future empirical validation.

2. Background

The semiconductor supply chain can be divided into three broad stages: research and design, manufacturing, and assembly and test packaging. Each stage involves specialized sectors distributed across key geographic regions including the United States, South Korea, Japan, China, Taiwan, and Europe (Varas et al 2021). Understanding the roles these regions play is essential for identifying vulnerable phases, as disruptions at any point can cause cascading consequences throughout the chain. These risks are not unique to semiconductors. Critical infrastructure industries like energy have faced similar supply chain cyberattacks, underscoring the urgent need for AI-powered defenses and zero-trust frameworks (Kumar, Tripathi, and Das 2025). Securing the semiconductor supply chain, the foundation of digital systems, can also improve resilience in sectors like energy, healthcare, and defense.

Manufacturing begins with pre-competitive research, which advances semiconductor fabrication and composition to meet evolving needs. Design translates this research into real-world architectures under complex parameters. Reusable intellectual property (IP) blocks streamline development by eliminating the need to recreate common elements, allowing engineers to focus on innovation. This is enabled by electronic design automation (EDA) tools, which include simulation, design, and verification. Simulation tests circuit behavior under various conditions. The design stage includes logic synthesis, schematic capture, and layout construction. Verification ensures the design meets specifications and is manufacturable, using physical and functional checks. Undetected flaws here can lead to costly delays. As noted by Varas et al (2021), chip design and EDA contribute roughly 54% of the semiconductor supply chain's total value added, highlighting their strategic importance.

The next major phase is manufacturing, where chips are fabricated through complex, nanoscale processes. Semiconductor fabs manufacture integrated circuits on silicon wafers, each containing hundreds of identical chips. Extreme precision and cleanliness are required—any contamination or tampering can compromise the batch, leading to significant losses. Depending on the application, fabrication can span 400 to 1,400 steps over 14 to 20 weeks (Varas et al 2021). Fabs rely heavily on specialized semiconductor manufacturing equipment (SME) to carry out these intricate procedures.

Finally, wafers enter the assembly, testing, and packaging stage. Known as back-end manufacturing, this phase prepares chips to meet functional and mechanical requirements. Primarily concentrated in Asia, it involves dicing wafers into individual dies, placing them onto substrates, and encapsulating them to protect against environmental and mechanical damage (Varas et al 2021). While this summary simplifies a highly technical process, the step is vital to delivering functional, defect-free chips to global markets. Major U.S. firms like Apple, Intel, and Texas Instruments depend heavily on this stage to commercialize their designs (Tomoshige 2022).

3. Methodology

Our methodology follows a multi-step approach involving threat modeling, AI framework design, and validation design, based on a comprehensive literature review and analysis of documented hardware attack cases.

First, we conducted an extensive review of academic literature, vulnerability databases (e.g., NIST's National Vulnerability Database), and industry case studies to identify and classify hardware threats across key stages of the semiconductor supply chain. Threats such as malicious IP insertion, JTAG exploitation, and counterfeit component introduction were examined in terms of their entry points, propagation behavior, and observable surfaces for detection.

Building on this analysis, we designed an AI-driven mitigation framework in which each stage of the supply chain is assigned targeted defense strategies. For example, the design phase employs anomaly detection and instruction obfuscation techniques, while the manufacturing phase uses image classification models for physical inspection. Each AI approach was chosen based on the nature of the threat's signal surface and the operational constraints of that stage.

To evaluate the feasibility and potential effectiveness of these defenses, we propose a validation framework that includes: (1) prototype simulation; (2) metric-driven comparison of proposed AI techniques with existing verification tools; and (3) threat injection scenarios to evaluate detection accuracy and response time. Although full implementation is left to future work, we outline practical experimental setups and tooling pathways, such as leveraging open-source RTL designs, SEM image datasets, and accessible FPGA development platforms.

4. Threat Modelling

To systematically address supply-chain security, we formalized a stage-wise threat model that specifies adversary capabilities, trust boundaries, attacker goals, and operational constraints in Table 2. This structured model provides the foundation for mapping AI-driven defences to stage-wise risk scenarios.

- **Adversary Capabilities:** Potential attackers include malicious insiders at IP vendors, EDA providers, fabrication facilities, with varying levels of access to RTL/netlists, process manufacturing parameters, packaging steps, or test/debug interfaces.
- **Trust Boundaries:** We assume that inhouse designs and OEMs retain partial trust, while third-party IP vendors, EDA toolchains, and offshore/decentralized fabs are considered potentially compromised.
- **Attacker Goals:** Depending on stage, goals include violating integrity (e.g., Trojan insertion), confidentiality (e.g., side-channel extraction of keys), authenticity (e.g., counterfeit parts), and availability (e.g., parameter drift leading to failure).
- **Operational Constraints:** Practical deployment requires low false-positive rates, tight resource budgets and testability compliance for real-time systems.

Table 1: Formalized Threat Model

Stage	Adversary Capabilities	Trust Boundaries	Attackers Goals	Constraints
Design	Insert malicious logic into IP/netlists; inject compromised EDA scripts; exploit open test/debug ports.	IP vendors are untrusted; EDA providers are partially trusted.	Integrity/Trojan-free design; confidentiality (protect cryptographic keys).	≤1% false positives; monitor overhead <5% area/power.
Manufacturing	Modify process parameters (doping, etching); introduce dopant-level tampering; conceal physical defects.	Foundries are partially trusted; Semiconductor Manufacturing Equipment vendors are untrusted.	Authenticity: genuine wafers only; reliability and reproducibility.	Inline inspection latency < 5 sec per wafer; minimal throughput loss.
Assembly	Swap with counterfeit ICs; alter packaging to enable probing; exploit environmental stress (humidity, delamination).	Outsourced semiconductor assembly and test vendors are untrusted; Original Equipment Manufacturers are trusted.	Authenticity (prevent counterfeits); confidentiality (block invasive probing).	Fingerprinting stable across voltage, current and temperature; provenance auditable and traceable.

4.1 Design Phase Threats

4.1.1 Hardware Trojans in IP blocks

Intellectual property (IP) blocks are reusable components of logic that support reusable designs, which allows for efficient implementation of logical components and Integrated Circuits. However, they also introduce significant security threats. For example, hardware Trojans embedded into these third-party IP-blocks can be extremely difficult to identify, especially when they are delivered as black-box gate-level netlists. Artificial Intelligence offers a promising solution to this conundrum by enhancing and deepening the reach of traditional testing methods. AI can be trained not only on traditional inputs and outputs, but also on side-channel indicators such as timing glitches, power consumption, and electromagnetic emissions. By monitoring these additional parameters, AI improves the likelihood of identifying subtle anomalies in chip behavior, indicating the presence of unwanted logic or circuits. Thus, AI can significantly strengthen the verification process of IP blocks, as its speed and adaptability allow for broader and deeper testing against potential threat vectors.

4.1.2 Exploitation of infected EDA tools

Another point of entry for critical vulnerabilities lies in EDA tools. Bugs in these EDA tools present a unique level of complexity as they are often embedded in automated steps of the EDA processes. Malicious actors can take advantage of this and inject compromised scripts in between the different stages without altering the source HDL code, making them hard to detect in source code audits or even in post-silicon validation once the chip is fabricated. In their 2023 study, Mavarapu et al highlight that attacks such as HeisenTrojan specifically exploit vulnerabilities in EDA tools, “establishing a permanent presence and providing a beachhead for intrusion into that system.” Figure 1 illustrates how such attacks can propagate. The script shown compiles a CPU core using the synth_design command. The flatten_hierarchy flag removes module boundaries, hindering traceability and making previously inserted malicious logic harder to detect. While it does not insert Trojans directly, it complicates verification workflows. By flattening the design hierarchy, any malicious logic can become deeply embedded in the netlist and no longer traceable to its original HDL modules, significantly deterring detection. Artificial Intelligence can support analysis of the provenance of the chips by verifying signed and reproducible EDA workflows. Anomaly detection is implemented on authenticated metadata, such as versioned commits and digital signatures, to identify suspicious changes rather than modifying hash semantics. These hashes will be flexible to accommodate valid and trusted edits but trigger alerts when any suspicious modifications are made. Furthermore, AI models can be used to secure the script pipelines by classifying changes and verifying their authenticity and enforcing credential-based editing leveraging the previously discussed hashes. For example, when the compromised script's timing and utilization reports show any unexpected spikes in logic usage or delays, the trained AI models can quickly assess whether the flattened netlist matches with the original high-level structure. This allows authorized users to make safe updates while effectively blocking tampering attempts.

```
# Read in all RTL files
read_verilog ALU.v SCC.v IM.v DM.v SCP.v

# Set the top-level module
set_top SCP

#Elaborate and compile the design
compile_ultra

# Write out synthesized netlist
write -format verilog -hierarchy -output synthesizedSingleCycle.v

#Setting parameters
set_clock_uncertainty 0.25
set_max_delay -from [get_ports A] -to [get_ports B] 2.0

#Malicious modification
set_dont_touch [get_cells *ALU.v*] # prevent ALU from being optimized or inspected
set_app_var suppress_messages {TIM-123} #hide specific timing warnings from logs

# Synthesize CPU design
synth_design -top SCP -part xc7a35tcsg324-1 -flatten_hierarchy rebuilt
#flattern hierarchy flag obscures the original module structure
#making detection of malicous code harder
```

Figure 1: Infected TCL script

4.1.3 Side-Channel leakages

Information leakage through side channels poses a serious security risk by exploiting a system's architectural behavior rather than its logic. These attacks extract sensitive data by monitoring power consumption, network activity, storage, and cache usage—revealing internal chip operations. A notable example is the exploitation of branch prediction and speculative execution. In branch prediction, the CPU guesses whether a branch will be taken; if it guesses wrong, the wrongly predicted instructions are rolled back. However, these discarded instructions can leave behind observable traces such as cached entries or memory timing variations. Attackers can exploit this by training the branch predictor to speculatively access sensitive memory, leading to data leakage of confidential information, though speculative execution does not modify program state. AI offers a promising defense against such threats. As Demme et al (2012) noted, side-channel attacks rely on detecting patterns in a program's behavior. AI can disrupt these patterns through dynamic instruction shuffling, rearranging non-critical instructions, similar to how operating systems schedule code blocks across threads, to eliminate predictability. Furthermore, AI can detect emerging trace patterns and inject obfuscation blocks, such as padded memory accesses or dummy instructions, maintaining functional correctness while increasing system unpredictability.

4.1.4 Insecure/open access test interfaces

Another significant threat during the design phase is the improper implementation of design-for-test features, particularly the exploitation of the JTAG (Joint Test Action Group) interface. JTAG is an industry-standard protocol widely used for testing and debugging integrated circuits (ICs), SoCs, and PCBs, allowing external access to internal logic and registers without physical intervention. However, if JTAG access is not properly fused or locked during IP block transfer or at any point in the chain of custody, it becomes a powerful attack vector. Malicious actors can gain low-level control of the device, including CPU, memory, and I/O, to inject code or modify program behavior. As Dong et al (2017) note, hardware Trojans have successfully exploited unsecured JTAG ports to manipulate device memory via the data bus. Artificial intelligence can play a key role in mitigating such threats both at design time and during runtime. At design time, AI-powered static analysis can identify unsecured or improperly fused JTAG/TAP ports in RTL/netlists. At runtime, anomaly detection of scan activity and power traces can detect and respond to abnormal JTAG usage events. Unlike traditional rule-based tools, AI models trained on diverse design patterns offer deeper vulnerability recognition and proactive detection. At runtime, embedded AI logic can monitor activation events and respond adaptively shutting down access to sensitive chip regions in the event of unexpected JTAG activity.

4.2 Manufacturing-Phase Threats

4.2.1 Physical tampering

During wafer fabrication, malicious actors or insider threats can introduce or mask physical defects such as dopant impurities, contaminants, or open circuit faults. Due to their microscopic and localized nature, these flaws often evade detection through traditional methods like electrical testing or visual inspection. Left unaddressed, they can degrade chip functionality and security over time, potentially compromising entire product lines before deployment. Though seemingly minor, such defects may slip past quality assurance and create systemic vulnerabilities. As Nam and Kim (2023) note, chip-level defect analysis is becoming critical due to the unpredictability and cost of customer-discovered flaws that wafer-level testing often misses. AI offers a scalable solution to mitigate these risks. Modern testing systems generate massive datasets, and convolutional neural networks (CNNs) trained on SEM images and micrographs can efficiently detect and classify anomalies. This reduces testing overhead, making robust defect detection more feasible and incentivizing early intervention. Furthermore, AI can be used to develop degradation models that monitor drift patterns over time. This can prove to be useful in the early identification of unusual drift patterns indicative of those infinitesimal defects. Such predictive capabilities can help mitigate the large-scale impact of these faults by enabling proactive correction before devices reach critical deployment stages.

4.2.2 Deviation in fabrication parameters

During wafer manufacturing, the integrity of physical parameters such as doping concentration, etching precision, and dielectric quality is crucial to chip security. Deviations, accidental or malicious, can degrade performance or functionality, compromising chips even if the digital design is errorproof. For example, metal layers that are thinner than specified can reduce component isolation, increasing vulnerability to side-channel leaks. Keith, Castillo-Villar and Bhuiyan (2023) highlight the growing exposure of fabrication facilities to cyber-

physical threats and demonstrate through stochastic modeling how even small variations can be part of larger adversarial campaigns. To counter these risks, AI provides powerful tools for real-time adaptation and specification verification. AI systems can dynamically adjust manufacturing parameters based on live sensor data, using reinforcement learning and feedback control to maintain consistency across wafers and prevent drift in metrics like etch duration or deposition thickness. To reinforce this process, chips produced at trusted facilities can serve as "model keys" with detailed profiles of power consumption, leakage, delay, and timing jitter. By modeling these features with Gaussian Mixture Models (GMM), AI can continuously compare ongoing output to the baseline. Any deviation beyond accepted tolerances triggers alerts or corrective actions, enabling proactive protection against fabrication-layer tampering.

4.3 Assembly and Packaging Threats

4.3.1 Counterfeit insertion and lack of traceability

The back-end stages of the semiconductor supply chain, particularly assembly, packaging, and testing, present a critical point of vulnerability for counterfeit chip production and insertion. The geographic concentration of foundries in East and Southeast Asia, combined with the decentralized nature of packaging operations, increases the risk of authentic chips being replaced with visually identical counterfeit ones. These substitutes may lack essential cryptographic modules or tamper detection circuits, making them highly exploitable. Moreover, if counterfeit chips carry cloned serial numbers or markings, they undermine traceability and accountability within the global supply chain. To address this, a hybrid approach combining artificial intelligence and blockchain-based provenance tracking offers a compelling solution. AI models can create unique chip-level fingerprints by analyzing subtle variations in timing, thermal response, and power-up behavior, which result from minor manufacturing variations. These fingerprints provide an identifier that allows physically identical clones to be differentiated from legitimate chips. Blockchain technology has shown promise in enhancing transparency, security, and automation in cyber risk management through decentralized ledgers and smart contracts. This approach can improve risk assessment and streamline processes, offering valuable lessons for securing complex supply chains (Kumar, Kocian and Loo 2024). When coupled with blockchain technology, each legitimate chip's lifecycle events can be recorded securely in a decentralized ledger (Neisse, Steri and Nai-Fovino 2017). AI can enhance this system by continuously scanning ledger entries, calculating risk scores, and identifying flows that may have unauthorized substitutions or modifications. This combination of AI-based fingerprinting and blockchain accountability enhances both hardware authentication and the integrity of provenance records, thereby mitigating the risks of counterfeit infiltration in semiconductor supply chains.

4.3.2 Environmental factors -based attacks

Moisture penetration during the packaging stage of semiconductor manufacturing presents an often overlooked but critical security risk. Moisture can degrade encapsulation materials, weaken structural barriers, and facilitate invasive physical attacks such as signal probing or micro-wiring modification. In plastic-encapsulated microelectronics, moisture diffuses over time—particularly through cracks or voids—leading to dielectric breakdown and interfacial delamination (Hadi, Ahmed and Sayyid 2011). These pathways can be exploited by adversaries to manipulate internal circuit elements. Moreover, certain hardware Trojans may remain dormant under standard conditions but activate when triggered by environmental factors like humidity or temperature. These sleeper mechanisms pose a significant threat in long-lifecycle devices, as their malicious behavior emerges only during field operation. To address these vulnerabilities, artificial intelligence can be integrated into non-destructive testing (NDT) workflows, offering a proactive and scalable defense. AI-enhanced NDT methods such as infrared thermography, X-ray imaging, and acoustic microscopy can improve defect detection, automate anomaly classification, and enable predictive diagnostics (Safhi, Keserle and Blanchard 2024). When applied to semiconductor packaging, these techniques can detect early signs of delamination, moisture-induced stress, or material degradation by comparing current chips with reference profiles. Additionally, time-series AI models can analyze environmental and electrical data to identify degradation trends, allowing manufacturers to detect and address issues before they evolve into security threats. By combining predictive moisture behavior with AI-driven inspection, manufacturers can ensure environmental conditions do not undermine device security. Table 1 summarizes these threats and corresponding mitigations across each stage of the supply chain.

Table 2: Threat Model Summary Table

Stage	Threat	Proposed AI defense
Design	Compromised IP Blocks	Comprehensive analysis of side channel indicators beyond traditional inputs
Design	Infected EDA Tool/Scripts	AI-based contextual script hashing, behavior verification, credential-based editing
Design	Side-Channel Leak	Instruction shuffling, AI-derived memory padding and obfuscation blocks
Design	Insecure Test Interface	Detection of open TAPs using AI trained on diverse patterns, runtime monitoring and response for unexpected access
Manufacturing	Physical Tampering	Using CNNs for SEM image classification, drift tracking for early anomaly detection
Manufacturing	Parameter Deviations	Modelling parameters using GMM to identify deviation from base behavior
Assembly	Counterfeit Chips	AI-based chip fingerprinting and blockchain, creation and comparison to key models
Assembly	Moisture/Environmental factors	NDT enhanced by AI, Time series AI models of environmental degradation

5. Framework Validation

To assess the effectiveness and overhead of the proposed AI-driven threat detection and mitigation framework, we outline a multi-pronged validation blueprint centered on simulation, comparative benchmarking, modeling of real-world vulnerability scenarios and aligned with established practices in hardware security and industrial inspection. While full-scale experimentation is reserved for future work, this plan lays out concrete evaluation paths using open-source tools, widely available datasets, and standardized test protocols to ensure reproducibility and practical relevance.

5.1 Trojan and Logic-Based Threat Detection

We simulate hardware Trojans and architectural vulnerabilities in RTL designs using open-source toolchains like Verilator and OpenROAD. Trust-Hub benchmarks will serve as the primary dataset, providing ground-truth examples of Trojan-inserted netlists. AI-based FSM anomaly detectors and graph-based classifiers will be compared against rule-based verification approaches and SAT-based deobfuscation attacks (Subramanyan et al., 2015). Evaluation metrices will be detection accuracy, false positive rate, false negative rate, runtime latency, and resource overhead (area, power, LUT utilization).

5.2 Runtime Obfuscation and Side Channel Defense

AI-enhanced instruction shuffling and AI- powered obfuscation mechanisms on FPGA SoCs built with LiteX, synthesized via Vivado (for FPGAs) or Design Compiler (for ASICs), and simulated using ModelSim or Verilator. Metrics include entropy of observable memory/cache access patterns, IPC (instructions per cycle), and runtime delays. Reverse engineering resistance will be evaluated using SAT-based deobfuscation attacks (Subramanyan et al., 2015) and formal equivalence checking, replacing software reverse engineering tools to ensure hardware-appropriate evaluation. Side channel leakage will be tested using standardized techniques such as Test Vector Leakage Assessment (TVLA) and mutual information analysis (Gierlichs et al., 2008).

5.3 Tamper Detection Using SEM Imagery

To identify packaging and physical tampering, we train lightweight convolutional or transformer models on scanning electron microscope (SEM) images. State-of-the-art industrial anomaly detection baselines such as PatchCore (Roth et al., 2022), RD++ (Tien et al., 2023), and PNI (Bae et al., 2023) will be included for comparison.

Models will be built using TensorFlow or PyTorch and evaluated on their ability to detect surface abrasion, lithographic drift, classification accuracy, localization precision, robustness and other subtle anomalies. Public SEM imagery and held-out defect sets will be used to benchmark classification accuracy and robustness.

5.4 Counterfeit Detection via Behavior Fingerprinting

We simulate authentic and counterfeit chips on FPGAs and analyze power and timing behaviors. We explicitly scope these experiments to logical-level fingerprints and do not generalize to dopant- or process-level attacks without silicon datasets. RNNs or sequence classifiers will be trained to fingerprint genuine behavior patterns and detect deviations. Modified variants—such as those simulating dopant shifts or corner process changes—will test generalizability. Metrics include classification accuracy and false positive rate.

5.5 Process Drift Detection in Fabrication

Process deviations such as bridging faults, voids, and misalignments will be simulated using TCAD tools (e.g., Silvaco). A time-series model (e.g., RNN) will be trained to detect and localize these variations. Baselines will include frameworks such as M2D2 (Lin et al., 2025). Performance evaluation metrics will focus on detection precision and robustness under varying process conditions.

5.6 Reproducibility and Deployment Constraints

All experiments will specify dataset splits, baseline implementations, parameters, and tool versions to ensure reproducibility. Hardware prototypes will report overhead in terms of area, power, and timing. Practical deployment requirements, such as maintaining area overhead under 5% and ensuring acceptable false positive rates, will be explicitly analyzed.

6. Discussion

Our framework is designed for seamless integration into existing semiconductor workflows with minimal disruption. The detection models are lightweight and can be deployed either on an isolated security-dedicated chip region or a separate system, consuming no more than 5% computational overhead. Training data is largely available from public sources, such as Trojan benchmarks like TrustHub and archival fabrication data collected for quality control. While the framework primarily targets new chip design and production, legacy systems can still benefit through offloaded processes like cloud-based system monitoring.

At this stage, our work remains a theoretical design study. Prototyping and precise overhead measurements, including power and performance impact, are planned for future work. Although our evaluation plan outlines how the framework could be validated, implementation and empirical testing will follow due to current resource limitations. Some strategies, such as runtime obfuscation, introduce overhead and may complicate worst-case execution-time (WCET) certification in real-time systems. These trade-offs must be carefully considered when defining security budgets and QoS (Quality of Service) goals. Nevertheless, once realized, the proposed measures could enable chipmakers to issue machine-verifiable integrity certificates per wafer lot, providing downstream integrators with traceable, auditable assurance—an increasingly vital requirement in regulated sectors such as automotive and medical devices.

Much prior work has addressed isolated stages of the supply chain, offered targeted solutions but lacked a unified approach. For example, Wilson et al (2024) introduced RAPTOR, a deep-learning discriminator for detecting tampering in post-fabrication packaging, yet it does not account for earlier threats such as malicious IP or test access insertion. Lin, Tsend and Tsai (2025) developed M2D2, a multi-modal drift detection framework for monitoring equipment and process health in wide-bandgap fabrication, focusing on operational anomalies rather than adversarial behavior. Sun et al (2024) proposed a CNN-based method for lithographic hotspot detection during the CAD stage, offering layout-level accuracy but limited to physical design vulnerabilities. In contrast, our framework spans the entire semiconductor supply chain, integrating CNN-based SEM classification, AI-powered provenance ledgers, and anomaly detection across both logical and physical domains. In contrast, our framework spans the entire semiconductor supply chain, combining CNN based SEMs classification, AI powered ledgers, and anomaly detection across both logical and physical domains. It integrates insights from these works but extends them into a modular, stage-specific, and adversary-aware architecture designed for proactive security response.

7. Conclusion

As semiconductor systems continue to evolve in complexity, they also grow increasingly susceptible to security vulnerabilities introduced across the design, manufacturing, and assembly pipeline. This paper presents a comprehensive, AI-driven framework for detecting, modeling, and mitigating hardware-level threats at each critical stage of the semiconductor supply chain. By integrating machine learning with side-channel analysis, image-based defect detection, provenance tracking, and adaptive runtime obfuscation, we demonstrate how AI can go beyond passive threat identification and serve as an active, scalable defense mechanism.

Through detailed threat modeling, our work underscores the potential of artificial intelligence as a cornerstone for secure and resilient hardware ecosystems. Future directions include expanding these defenses into edge-deployed AI chips and collaborating with industry partners to operationalize these solutions at a scale. Our findings lay the groundwork for trustable semiconductor manufacturing in an era where both performance and security are paramount.

Ethics Declaration: This research did not require any ethical clearance.

AI Declaration: OpenAI's ChatGPT was used to check and correct any in-text citation formatting and reference list formatting issues. The tool was not used to generate original content or conduct analysis.

References

Bae, J., Lee, J.-H., & Kim, S. (2023). PNI: Industrial anomaly detection using position and neighborhood information. In Proceedings of the IEEE/CVF International Conference on Computer Vision. https://arxiv.org/abs/2211.12634

Demme, J., Martin, R., Waksman, A. and Sethumadhavan, S. (2012) 'Side-channel vulnerability factor: A metric for measuring information leakage', *Proceedings of the 39th Annual International Symposium on Computer Architecture (ISCA)*, pp. 106–117. Available at: https://doi.org/10.1145/2366231.2337172.

Dong, C., Xu, Y., Liu, X., Zhang, F., He, G. and Chen, Y. (2020) 'Hardware Trojans in chips: A survey for detection and prevention', *Sensors*, Vol. 20, No. 18, p. 5165. Available at: https://doi.org/10.3390/s20185165.

Hadi, M.A., Ahmed, A. and Sayyid, M.I. (2011) 'Moisture induced failure mechanisms in plastic encapsulated microelectronics', *Journal of Physics D: Applied Physics*, Vol. 44, No. 3, p. 034007. Available at: https://doi.org/10.1088/0022-3727/44/3/034007.

Gierlichs, B., Batina, L., Tuyls, P., & Preneel, B. (2008). Mutual information analysis. In E. Oswald & P. Rohatgi (Eds.), Cryptographic Hardware and Embedded Systems – CHES 2008 (pp. 426–442). Springer. https://doi.org/10.1007/978-3-540-85053-3_27

Keith, K., Castillo-Villar, K.K. and Bhuiyan, T.H. (2023) 'Attack graph-based stochastic modeling approach for enabling cybersecure semiconductor wafer fabrication', *Journal of Manufacturing Systems*, Vol. 69, pp. 508–521. Available at: https://doi.org/10.1016/j.cie.2024.109912.

Khan, S.M., Peterson, D. and Mann, A. (2021) *The semiconductor supply chain*, Center for Security and Emerging Technology, Vol. 8, No. 8.

Kumar, S., Kocian, L. and Loo, L. (2024) 'Blockchain applications for cyber liability insurance', *International Journal on Cybernetics & Informatics (IJCI)*, Vol. 13, No. 5, pp. 119–139.

Kumar, S., Tripathi, K. and Das, S. (2025) 'Cybersecurity of energy systems', in *Method of Process Systems in Energy Systems: Emerging Energy Systems Part II*, Volume 9, Elsevier, Amsterdam.

Lin, C.-Y., Tseng, T.-L. (B.) and Tsai, T.-H. (2025) 'A multi-machine and multi-modal drift detection (M2D2) framework for semiconductor manufacturing', *Applied Sciences*, Vol. 15, No. 12, p. 6500. Available at: https://doi.org/10.3390/app15126500.

Mavurapu, A., Shan, H., Guo, X., Arias, O. and Sullivan, D. (2023) 'HeisenTrojans: They are not there until they are triggered', *Proceedings of the 2023 Asian Hardware Oriented Security and Trust Symposium (AsianHOST)*, pp. 1–7. Available at: https://doi.org/10.1109/AsianHOST59942.2023.10409305.

Nam, S. and Kim, J. (2023) 'Chip-level defect correlation and data handling strategies in semiconductor testing', *Electronics*, Vol. 12, No. 4, p. 116. Available at: https://www.mdpi.com/2673-8392/4/4/116.

Neisse, R., Steri, G. and Nai-Fovino, I. (2017) 'A blockchain-based approach for data accountability and provenance tracking', *Proceedings of the 12th International Conference on Availability, Reliability and Security (ARES)*. Available at: https://doi.org/10.1145/3098954.3098958.

Roth, K., Pemula, L., Zepeda, J., Schölkopf, B., Brox, T., & Gehler, P. (2022). Towards total recall in industrial anomaly detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. https://arxiv.org/abs/2106.08265

Safhi, A.E.M., Keserle, G.C. and Blanchard, S.C. (2024) 'AI-driven non-destructive testing insights', *Encyclopedia*, Vol. 4, No. 4, pp. 1760–1769. Available at: https://doi.org/10.3390/encyclopedia4040116.

Subramanyan, P., Ray, S., & Malik, S. (2015). Evaluating the security of logic encryption algorithms. 2015 IEEE International Symposium on Hardware Oriented Security and Trust (HOST), 137–143. https://doi.org/10.1109/HST.2015.7140252

Sun, H., Liu, C., Zhao, J., Wang, Y. and Pan, D. (2025) 'Interpretable CNN-based lithographic hotspot detection through error marker learning', *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, Vol. 44, No. 3, pp. 1031–1044. Available at: https://doi.org/10.1109/TCAD.2024.3468016.

Thadani, A. and Allen, J. (2023) *Mapping the semiconductor supply chain: The critical role of the Indo-Pacific region*, Center for Strategic and International Studies (CSIS). Available at: https://www.csis.org/analysis/mapping-semiconductor-supply-chain-critical-role-indo-pacific-region.

Tien, T. D., Nguyen, A. T., Tran, N. H., Huy, T. D., Duong, S. T. M., Nguyen, C. D. T., & Truong, S. Q. H. (2023). Revisiting reverse distillation for anomaly detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 24511–24520)

Thompson, K. (1984) 'Reflections on trusting trust', *Communications of the ACM*, Vol. 27, No. 8, pp. 761–763. Available at: https://doi.org/10.1145/358198.358210.

Tomoshige, H. (2022) *CHIPS+ and semiconductor packaging*, Center for Strategic and International Studies (CSIS). Available at: https://www.csis.org/blogs/perspectives-innovation/chips-and-semiconductor-packaging.

Varas, A., Varadarajan, R., Goodrich, J. and Yinug, F. (2021) *Strengthening the global semiconductor supply chain in an uncertain era*, Boston Consulting Group (BCG) and Semiconductor Industry Association (SIA). Available at: https://www.semiconductors.org/strengthening-the-global-semiconductor-supply-chain-in-an-uncertain-era/.

Wilson, B., Chen, Y., Singh, D.K., Ojha, R., Pottle, J., Bezick, M., Boltasseva, A., Shalaev, V.M. and Kildishev, A.V. (2024) 'Authentication through residual attention-based processing of tampered optical responses', *Advanced Photonics*, Vol. 6, No. 5, p. 056002. Available at: https://doi.org/10.1117/1.AP.6.5.056002.