World Journal of Advanced Engineering Technology and Sciences

eISSN: 2582-8266

Cross Ref DOI: 10.30574/wjaets

Journal homepage: https://wjaets.com/

(REVIEW ARTICLE)

Check for updates

Leveraging Artificial Intelligence in In-System Test: A New Paradigm for Predictive and Adaptive Chip Validation

Jayesh Kumar Pandey *

Independent Researcher, USA.

World Journal of Advanced Engineering Technology and Sciences, 2025, 15(03), 2234-2245

Publication history: Received on 12 May 2025; revised on 21 June 2025; accepted on 24 June 2025

Article DOI: https://doi.org/10.30574/wjaets.2025.15.3.1155

Abstract

The exponential growth in complexity of modern system-on-chip (SoC) designs, characterized by heterogeneous integration of billions of transistors across diverse functional blocks, has fundamentally transformed semiconductor validation requirements. As these sophisticated chips increasingly power mission-critical applications—from autonomous transportation systems to medical devices and critical infrastructure—traditional test methodologies focused primarily on manufacturing defects have proven inadequate for ensuring sustained reliability throughout operational lifetimes. In-System Test (IST) mechanisms embedded within semiconductor devices offer a promising solution by extending validation capabilities beyond manufacturing into deployment environments, enabling continuous monitoring and diagnostics throughout the product lifecycle. However, conventional IST implementations remain predominantly static and pattern-based, executing predetermined test sequences that cannot adapt to the dynamic operational conditions, workload variations, and evolving stress patterns encountered in real-world environments. This architectural limitation creates a critical gap between test coverage and actual reliability requirements, particularly for advanced process nodes where subtle degradation mechanisms may manifest uniquely based on specific usage conditions.

This article explores the transformative potential of Artificial Intelligence (AI) in addressing these limitations by augmenting traditional IST frameworks with adaptive, learning-enabled capabilities. Machine learning techniques can enable predictive fault detection by identifying subtle precursors to potential failures before they manifest as functional errors, dynamic test scheduling that optimizes validation coverage based on operational conditions, and intelligent analytics that accelerate root cause identification for complex failure modes. The integration of these capabilities creates a fundamental shift from reactive to proactive reliability management, enabling semiconductor devices to continuously adapt their validation strategies based on actual operational experience. The article provides a framework for implementing AI-augmented IST, detailing the architectural requirements, data collection infrastructure, edge processing considerations, and secure update mechanisms necessary for practical deployment. The discussion examines potential AI models and implementation strategies across diverse application domains, from autonomous vehicles to data center processors, medical devices, and consumer electronics, highlighting domain-specific considerations and optimization techniques. Finally, the article examines the significant challenges that must be addressed to realize the full potential of AI-augmented IST, including model efficiency requirements for resource-constrained environments, training data limitations for reliability applications, security and privacy considerations for distributed learning systems, and standardization needs for cross-platform interoperability. By mapping both the opportunities and obstacles in this emerging field, the article provides a roadmap for developing intelligent test frameworks for next-generation semiconductor systems.

Keywords: In-System Test; AI in Testing; Predictive Diagnostics; Adaptive Test Scheduling; Runtime Fault Detection; IST Architecture; Machine Learning; SoC Reliability

* Corresponding author: Jayesh Kumar Pandey

1. Introduction

Modern semiconductor devices have evolved into intricate system-on-chip (SoC) architectures incorporating multiple processor cores, specialized accelerators, and complex interconnect networks that power mission-critical applications across various domains. The verification challenge for these designs has grown exponentially with each technology node, creating fundamental gaps in traditional testing methodologies that focus primarily on functional correctness through simulation-based approaches [1]. These highly integrated chips serve as the computational backbone for autonomous vehicles navigating complex urban environments, hyperscale data centers processing massive volumes of information, and edge AI systems making real-time decisions where human lives may be at stake. The combination of increasing design complexity, tight time-to-market constraints, and the prohibitive computational resources required for exhaustive verification has necessitated paradigm shifts in how reliability assurance is approached for modern SoCs.

In-System Test (IST) architectures have emerged as a critical solution by embedding self-diagnostic capabilities directly within silicon, enabling runtime monitoring and validation mechanisms that operate throughout the product lifecycle. These approaches complement conventional verification by focusing on temporal properties and assertions that can be monitored during actual system operation, addressing the fundamental incompleteness of pre-silicon verification [2]. The formal specification of these properties has proven essential for capturing the intended behavior of complex hardware modules and detecting violations during operation. However, despite their utility, current IST implementations remain predominantly rule-based and deterministic, operating according to fixed parameters that cannot adapt to the dynamic conditions encountered in real-world deployment scenarios.

The integration of Artificial Intelligence represents a revolutionary opportunity to transform IST from a static diagnostic tool into a dynamic, learning-enabled system. By leveraging machine learning algorithms trained on operational telemetry, AI-augmented IST can adapt test behavior based on actual usage patterns, environmental conditions, and historical failure trends. This paradigm shift enables predictive rather than merely reactive testing strategies, where potential points of failure can be identified and addressed before they manifest as functional errors, bridging the gap between the inherent limitations of formal verification methods and the practical needs of runtime assurance for complex SoCs [1].

Furthermore, AI-driven IST can intelligently allocate test resources by prioritizing critical components based on runtime stress analysis and operational context. This context-aware approach enhances reliability while optimizing power consumption and performance overhead associated with continuous monitoring—a crucial consideration for energy-constrained devices. The ability to specify and verify temporal assertions at runtime provides a foundation upon which these adaptive AI methodologies can build, combining the rigor of formal methods with the adaptability of machine learning [2].

As semiconductor technology continues to advance toward increasingly specialized designs for AI acceleration, the synergistic relationship between AI and IST creates a positive feedback loop: more powerful AI capabilities enable more sophisticated test strategies, while more effective testing ensures the reliability of the very systems implementing these AI functionalities. The evolution of runtime verification from static temporal logic assertions to dynamic, learning-based systems represents the next frontier in addressing the verification gap for complex SoCs in safety-critical applications.

2. Limitations of Traditional IST

Traditional In-System Test methodologies, while providing essential validation capabilities, exhibit fundamental limitations that hinder their effectiveness in modern semiconductor applications. These constraints manifest across multiple dimensions of test architecture and execution paradigms, creating significant challenges for ensuring reliability in complex integrated circuits deployed in mission-critical systems.

The fixed test sequence approach prevalent in conventional IST implementations represents a significant constraint for evolving system requirements. Current IST architectures typically employ pre-determined test patterns that remain static throughout the product lifecycle, creating a critical disconnect between testing strategies and the dynamic nature of operational degradation mechanisms. Software quality assurance research has demonstrated that such static testing approaches fail to address the evolutionary nature of system behavior in field conditions, particularly when systems operate in environments not anticipated during initial test development [3]. This limitation mirrors challenges identified in software testing, where fixed test suites gradually lose effectiveness as the system evolves during operation. The inability to adapt test coverage to emergent failure modes results in diminishing test effectiveness over

time, particularly for subtle parametric shifts that manifest uniquely based on usage conditions and environmental factors that were not modeled during test development.

Context awareness limitations further undermine the effectiveness of traditional IST frameworks, which typically operate without sufficient integration with the operational environment of the device. Modern verification methodologies, including Universal Verification Methodology (UVM), have established that effective test strategies must incorporate contextual factors such as power states, thermal conditions, and application workloads to achieve meaningful coverage [4]. However, current IST implementations rarely leverage these contextual factors when scheduling or configuring test operations. The disconnection between functional execution context and test operations leads to suboptimal resource utilization and coverage gaps, particularly in heterogeneous systems where computational workloads shift dynamically across processing elements. This isolated approach to testing also complicates fault diagnosis, as the relationship between operational conditions and observed failures remains largely opaque, limiting the actionability of test results for root cause analysis and system optimization.

The reactive nature of traditional IST represents perhaps its most significant limitation for reliability-critical applications. Conventional test approaches focus primarily on fault detection rather than prediction, activating only after performance degradation or functional errors have already manifested [3]. This reactive paradigm fundamentally limits the preventative potential of IST, particularly in safety-critical domains where even momentary malfunctions can have severe consequences. Software reliability engineering research has demonstrated that proactive detection approaches that leverage operational data to anticipate failures significantly outperform reactive methodologies in both detection effectiveness and resource efficiency. The temporal gap between fault occurrence and detection in reactive frameworks further complicates remediation efforts, as system state information crucial for diagnosis may be lost by the time testing is initiated, creating persistent challenges for intermittent fault scenarios that may not reproduce consistently during subsequent test cycles.

Table 1 Comparison of Traditional IST vs. AI-Augmented IST. [3, 4]

Characteristic	**Traditional IST**	**AI-Augmented IST**
Test Sequence Adaptation	Static, predefined patterns	Dynamic, context-aware patterns
Operational Context	Limited or no awareness	Comprehensive integration with workload and environmental data
Failure Detection Approach	Reactive, post-occurrence	Predictive, pre-emptive
Resource Utilization	Fixed test schedules regardless of conditions	Adaptive scheduling based on operational risk
Learning Capability	None, fixed implementation	Continuous improvement from operational data

3. AI Opportunities in IST

The integration of Artificial Intelligence with In-System Test frameworks offers transformative opportunities that address the fundamental limitations of conventional testing methodologies. These AI-enabled capabilities span the entire testing lifecycle, creating new possibilities for semiconductor validation and reliability management in complex systems-on-chip.

3.1. Predictive Fault Detection

The application of machine learning for prognostics and health management represents a paradigm shift in semiconductor testing strategy. By analyzing real-time telemetry streams from operational chips, predictive models can identify subtle degradation patterns well before conventional threshold-based methods detect anomalies. Recent advances in convolutional neural networks have demonstrated remarkable capability in processing multi-parameter sensor data to predict remaining useful life in electronic components under various operational conditions [5]. These models learn to recognize the complex relationships between electrical parameters, thermal profiles, and performance metrics that precede specific failure modes [13]. The time series analysis capabilities of recurrent architectures prove particularly valuable for capturing the temporal progression of degradation phenomena that evolve over extended operational periods. This predictive capability transforms reliability management from reactive response to proactive intervention, enabling targeted testing or workload adjustment before functional failures manifest. The ability to establish correlations between specific operational conditions and subsequent failure probabilities creates

opportunities for design feedback that can enhance inherent reliability in future generations, addressing issues at their source rather than merely detecting their manifestations.

Table 2 Key AI Models for IST Applications. [3, 4]

Application	Recommended AI Approach	Key Advantages
Predictive Fault Detection	Convolutional Neural Networks	Effective for pattern recognition in time-series sensor data
Adaptive Test Scheduling	Reinforcement Learning	Optimizes test resource allocation under dynamic conditions
Root Cause Analysis	Transfer Learning	Leverages limited failure data effectively for diagnosis
Anomaly Detection	Self-Supervised Learning	Identifies unusual patterns without extensive labeled examples

3.2. Adaptive Test Scheduling

Traditional fixed-interval testing approaches fail to account for the dynamic nature of modern semiconductor operation, where stress distribution varies dramatically based on workload characteristics and environmental conditions. AI techniques enable intelligent allocation of test resources through dynamic scheduling algorithms that prioritize validation based on real-time assessment of reliability risk. Edge computing research has demonstrated the effectiveness of reinforcement learning approaches in optimizing resource-constrained decision processes where multiple competing objectives must be balanced [6]. Applied to IST, these techniques can determine optimal test timing by weighing factors including execution history, current workload intensity, thermal conditions, and criticality of active functions. The continuous adaptation of scheduling policies based on observed outcomes creates self-improving test strategies that progressively optimize coverage while minimizing performance impact. This approach proves particularly valuable in heterogeneous systems where computational workloads shift dynamically across diverse processing elements with different reliability characteristics and failure modes. The context-awareness inherent in adaptive scheduling ensures that test resources target the most vulnerable system components based on actual operational stress rather than static assumptions established during design, significantly enhancing the efficiency and effectiveness of reliability management.

3.3. Intelligent Diagnosis and Root Cause Analysis

The complexity of modern semiconductor failure modes often creates ambiguous relationships between observed symptoms and underlying causes, complicating traditional diagnostic approaches. Machine learning classification algorithms excel at mapping complex, non-linear relationships between multi-parameter test signatures and specific defect mechanisms. Transfer learning techniques have proven especially effective for building robust fault classification models by leveraging knowledge from simulation and accelerated life testing to improve diagnostic accuracy for field failures with limited examples [5]. Graph-based neural networks can model fault propagation through interconnected circuit elements, distinguishing between root causes and secondary effects to isolate the fundamental failure mechanism. This diagnostic precision dramatically accelerates debug processes during system-level testing and failure analysis, reducing time-to-resolution for complex reliability issues. The ability to capture and formalize the relationship between test signatures and underlying defect mechanisms creates a continuously improving knowledge base that enhances both test coverage and design robustness through systematic identification of reliability vulnerabilities that might otherwise remain undetected until field deployment.

Machine learning approaches have demonstrated significant potential for fault detection and diagnosis across various domains, from industrial equipment to semiconductor systems. Previous work has established effective frameworks for applying these techniques to complex mechanical systems with multiple potential failure modes [13], providing methodological foundations that can be adapted for semiconductor test applications

3.4. Anomaly Detection in Test Logs

The sheer volume and dimensionality of data generated by comprehensive IST implementations often overwhelm traditional analysis methods, obscuring subtle patterns that might indicate emerging reliability issues. Unsupervised and self-supervised learning techniques provide powerful mechanisms for identifying anomalous behaviors without requiring predefined fault models. Recent work in representation learning for time series data has demonstrated that

contrastive learning approaches can effectively distinguish between normal variations and potentially significant anomalies in multi-parameter operational telemetry [6]. These techniques establish normative behavioral baselines across complex parameter spaces and identify contextually unusual patterns that warrant further investigation. Semi-supervised approaches have proven particularly effective for reliability monitoring by leveraging the abundant normal-operation data available while requiring minimal examples of fault conditions. The ability to automatically distinguish meaningful deviations from benign variations significantly improves the signal-to-noise ratio in reliability monitoring, enabling earlier intervention for emerging issues while reducing false alarms that might otherwise undermine confidence in the testing infrastructure. This enhanced anomaly detection capability proves especially valuable for identifying novel failure modes that might escape detection by supervised models trained only on previously observed fault categories.

4. System Architecture for AI-Augmented IST

The implementation of AI capabilities within In-System Test frameworks necessitates a carefully structured architectural approach that addresses computational efficiency, resource limitations, and security requirements. The following sections outline the essential components of an AI-augmented IST architecture that enables intelligent, adaptive testing while maintaining system integrity.

4.1. Data Collection Layer

The effectiveness of any AI-based testing system fundamentally depends on a comprehensive data acquisition infrastructure capable of capturing diverse operational parameters with sufficient temporal resolution to support meaningful analytics. The data collection layer must integrate multiple telemetry streams, including traditional test results, performance metrics, thermal readings, power consumption patterns, and signal characteristics across critical interfaces. Modern edge computing frameworks in manufacturing environments have demonstrated that multi-tier data architectures with local preprocessing capabilities significantly reduce bandwidth requirements while preserving essential information content for analytics purposes [7]. These architectures implement selective sampling strategies that adjust data resolution based on detected anomalies, preserving detailed information around potential failure events while implementing compression for normal operation periods. The implementation of time-series databases with specialized indexing schemes has proven particularly effective for semiconductor telemetry, enabling efficient retrieval of historical patterns that correspond to specific operational conditions or failure precursors. Manufacturing systems research has established that edge-based data preprocessing directly on the production equipment dramatically improves the responsiveness of analytical systems while reducing the communication overhead between factory systems and cloud platforms. The synchronization of heterogeneous data streams with precise timestamps represents a particular challenge in distributed sensing environments, requiring specialized protocols that account for communication latencies and clock drift between subsystems while maintaining the temporal relationships critical for correlation analysis.

4.2. Edge Processing Unit

The latency-sensitive nature of many semiconductor testing scenarios necessitates localized AI inference capabilities that can analyze telemetry and detect anomalies without dependence on external processing resources. The edge processing element typically comprises specialized hardware accelerators optimized for the matrix operations predominant in machine learning workloads, often implemented as extensions to existing microcontroller architectures or as dedicated subsystems within the test infrastructure. Research in IoT-based manufacturing has demonstrated that heterogeneous computing architectures combining conventional processors with specialized AI accelerators achieve an optimal balance between flexibility and performance efficiency for industrial analytics applications [7]. Memory architecture considerations prove particularly critical in these implementations, as model parameters must be efficiently accessible while minimizing energy impact. The computational workload distribution between edge and cloud resources must be carefully optimized based on the specific requirements of different analytical models, with time-sensitive anomaly detection typically implemented at the edge while more complex predictive analytics may leverage cloud resources when immediate response is not required. The increasing availability of specialized neural processing hardware has dramatically improved the feasibility of sophisticated analytics within power-constrained environments, enabling implementations that were previously impractical due to computational limitations. Virtualization techniques that provide logical isolation between inference workloads and critical system functions have proven essential for maintaining deterministic performance in real-time applications where testing operations must not interfere with primary system functionality.

Table 3 System Architecture Components for AI-Augmented IST. [7, 8]

Component	Primary Function	Implementation Considerations
Data Collection Layer	Aggregates multi-source telemetry	Edge preprocessing to reduce bandwidth requirements
Edge Processing Unit	Executes inference models	Heterogeneous computing architecture for efficiency
Model Update Mechanism	Securely deploys updated models	Differential updates with cryptographic verification
Telemetry Integration	Communicates insights to host systems	Standardized interfaces with security protocols

4.3. Model Update Mechanism

The dynamic nature of semiconductor failure mechanisms necessitates regular refinement of analytical models based on operational experience and emerging reliability patterns. A secure model update infrastructure provides the foundation for deploying revised AI models without compromising system integrity or introducing security vulnerabilities. Research on firmware update mechanisms for constrained devices has established that open standards-based approaches combining manifest-based authentication with differential updates significantly reduce bandwidth requirements while maintaining robust security properties [8]. These systems implement cryptographic verification at multiple stages in the update pipeline to prevent unauthorized modifications that could potentially compromise system reliability or intellectual property protection. A critical consideration in model update architectures involves the verification of updated models against established performance baselines to ensure that new versions maintain or improve analytical accuracy without introducing regression in critical detection capabilities. Research has demonstrated that challenges in resource-constrained environments include limited cryptographic capabilities, unreliable network connectivity, and power constraints that complicate secure update procedures. Implementation strategies must carefully balance security requirements against practical limitations, with tiered approaches that apply different security measures based on the criticality of specific update components. Resilience considerations dictate that update mechanisms must maintain fallback capabilities to restore previous configurations if performance degradation is detected after deployment, ensuring continuous system functionality even when update processes encounter unexpected conditions.

4.4. Telemetry Integration

The analytical insights generated by on-device AI systems achieve maximum value when effectively incorporated into broader system management and diagnostic frameworks. The telemetry integration layer establishes standardized interfaces for communicating analytical results to host systems, maintenance platforms, and enterprise analytics environments. Manufacturing systems research has demonstrated that standardized data models and communication protocols significantly improve interoperability across diverse analytical platforms while reducing integration complexity for heterogeneous device ecosystems [7]. These integration frameworks implement appropriate authentication and encryption to protect sensitive diagnostic information while ensuring that critical alerts propagate with minimal latency to response systems. The implementation of edge analytics capabilities that perform initial data reduction before transmission has proven particularly valuable in bandwidth-constrained environments, enabling effective remote diagnostics without requiring continuous transmission of raw telemetry. Security research has established that telecommunications between edge devices and cloud platforms represent a potential attack vector that must be protected through layered security measures, including mutual authentication, message integrity verification, and encrypted communication channels [8]. The telemetry architecture must support both synchronous query operations for interactive diagnostics and asynchronous notification mechanisms for critical alerts, providing flexibility for diverse operational scenarios while maintaining communication efficiency. Integration with standardized enterprise systems enables coordination between device-level analytics and broader organizational processes, including maintenance scheduling, spare parts management, and continuous improvement initiatives that leverage insights from field operations to enhance future designs.

5. Use Cases and Applications

The integration of AI with In-System Test methodologies creates transformative opportunities across diverse application domains. This section explores implementation scenarios where AI-augmented IST delivers significant value across various critical systems.

5.1. Autonomous Vehicles

The safety-critical nature of autonomous driving systems demands robust reliability assurance for the complex SoCs that enable perception, planning, and control functions. AI-augmented IST provides crucial capabilities for proactive validation in these challenging environments where traditional testing approaches prove insufficient. Research in verification and validation methodologies for autonomous systems has identified that conventional testing frameworks struggle with the combinatorial explosion of operational scenarios that must be validated, particularly when environmental conditions and sensor uncertainties are considered [9]. AI-based testing can address this challenge through intelligent scenario prioritization that focuses validation resources on the most safety-critical operational modes based on current driving conditions. Before transitioning to higher autonomy levels or complex driving maneuvers, the system can dynamically execute targeted test sequences focusing on perception modules under current lighting and weather conditions, or decision-making components about to navigate complex traffic scenarios. This context-aware approach significantly enhances safety assurance compared to static testing regimes. Deep learning techniques for anomaly detection have demonstrated particular effectiveness in identifying edge cases and corner conditions that traditional rule-based testing might miss, enabling more comprehensive validation coverage for autonomous systems where unanticipated scenarios present the greatest safety risks. The continuous learning capabilities of AI-based test frameworks enable adaptation to emerging failure patterns unique to autonomous driving deployments, such as perception uncertainties or decision-making inconsistencies that conventional validation processes might not anticipate. The fusion of operational telemetry with simulation-based testing creates powerful hybrid validation approaches that can predict system behavior across a far broader range of scenarios than physical testing alone could practically cover.

5.2. Data Center SoCs

Modern data center processors operate in highly dynamic environments where computational demands fluctuate continuously based on application requirements and infrastructure conditions. Research on energy efficiency in computing systems has established that workload characteristics directly influence both performance requirements and reliability stress patterns, creating opportunities for adaptive management approaches that optimize resources based on actual operational conditions [10]. AI-augmented IST enables thermal and workload-aware test scheduling that dynamically adjusts validation coverage based on observed stress patterns rather than predetermined intervals. During periods of elevated computational intensity or thermal load, the system can increase test frequency and coverage for vulnerable components while reducing testing overhead during lower-risk operational phases. This adaptive approach significantly improves reliability assurance while minimizing performance impact on production workloads. Machine learning models trained on operational telemetry can establish correlations between specific application characteristics and reliability risk profiles, allowing predictive scheduling of test operations before potential failure conditions develop. Energy efficiency research has demonstrated that workload phase detection techniques can identify transitions between computational patterns with distinct resource requirements, providing natural scheduling points for test operations that minimize performance impact. The scale of data center deployments creates unique opportunities for fleet-wide learning, where reliability insights from multiple systems can be aggregated to refine predictive models while preserving the confidentiality of specific workload characteristics. Thermal management challenges in high-density computing environments create particular reliability concerns that AI-augmented testing can address through continuous monitoring of thermal gradients and hotspot formation patterns that might indicate emerging cooling inefficiencies before they reach critical thresholds that could trigger failure events.

5.3. Medical Devices

The increasing integration of sophisticated semiconductor components in medical devices presents unique reliability challenges due to the critical nature of healthcare applications and the extended operational lifetimes required in many therapeutic scenarios. Research in verification methodologies for autonomous systems has established that safety-critical applications require multi-layered validation approaches that combine formal verification, simulation-based testing, and runtime monitoring to achieve adequate assurance levels [9]. AI-augmented IST enables continuous health monitoring capabilities that can detect subtle parametric shifts indicative of potential reliability issues before they affect therapeutic functions. For implantable devices, where accessibility for maintenance is severely limited and failure consequences can be life-threatening, the predictive capabilities of AI-based testing provide particular value by potentially identifying reliability concerns early enough to schedule intervention before critical failures occur. Machine learning techniques can recognize the specific degradation signatures associated with environmental factors unique to medical applications, such as tissue interaction effects or chemical exposure that may affect electrical parameters and system performance over time. The ability to correlate observed parametric shifts with specific failure mechanisms enables more precise diagnostic information for healthcare providers, potentially reducing unnecessary interventions while ensuring a timely response to genuine reliability concerns. The continuous learning capabilities of AI-augmented

IST frameworks allow testing strategies to adapt to the unique physiological environment of each patient, accounting for individual variations that may influence device reliability over extended deployment periods. Federated learning approaches can enable cross-device knowledge sharing while maintaining strict privacy protection for sensitive medical data, creating increasingly refined predictive models without centralizing the underlying patient information.

5.4. Consumer Electronics

The consumer electronics sector presents distinct reliability challenges due to highly variable usage patterns and diverse operational environments that traditional testing methodologies struggle to address comprehensively. Energy efficiency research has established that consumer devices experience widely varying utilization patterns that directly impact component stress and aging characteristics, creating opportunities for usage-aware reliability management that optimizes both performance and longevity [10]. AI-augmented IST enables adaptive reliability strategies that tailor testing approaches to observed usage behavior rather than generic assumptions, improving reliability outcomes while optimizing resource utilization. Machine learning techniques can identify correlations between specific application patterns and reliability stress factors, enabling targeted testing of the components most vulnerable under observed conditions. Energy management research has demonstrated that usage pattern classification can effectively identify distinct operational profiles with different reliability implications, enabling differentiated testing strategies that focus resources where they provide maximum value for each specific device. The ability to adapt testing based on environmental factors detected through onboard sensors further enhances reliability assurance by focusing validation on the specific failure mechanisms most relevant to current conditions. Over-the-air update capabilities in modern consumer devices create a natural infrastructure for deploying refined AI models and test sequences based on fleet-wide learning, enabling continuous improvement of reliability management throughout the product lifecycle. Energy-aware testing approaches can schedule validation operations during charging periods or low-utilization intervals, minimizing impact on battery life and user experience while maintaining comprehensive reliability coverage. The integration of user behavior modeling with reliability management creates opportunities for personalized longevity optimization that balances performance capabilities against device lifespan based on individual usage patterns and preferences.

6. Challenges and Research Opportunities

While AI-augmented In-System Test offers transformative capabilities for semiconductor validation, several significant challenges must be addressed to realize its full potential. This section explores key obstacles and emerging research directions that will shape the evolution of intelligent testing methodologies.

6.1. Model Size and Efficiency

The deployment of AI capabilities within resource-constrained semiconductor environments presents fundamental challenges regarding computational efficiency and implementation footprint. Recent systematic reviews of on-device machine learning have identified that memory consumption and computational overhead remain primary obstacles for edge AI deployment, particularly in systems where power budgets and silicon area are tightly constrained [11]. Unlike cloud implementations where computational resources are abundant, on-chip AI for IST must operate within strict power and performance envelopes while maintaining sufficient analytical capabilities to deliver meaningful insights. Traditional deep learning architectures often require substantial parameter storage and significant matrix operations for inference, making them impractical for direct implementation within test infrastructures that must minimize impact on primary system functions. Research in neural network compression has identified several promising approaches for reducing implementation requirements while preserving analytical accuracy, including quantization techniques that reduce numerical precision, structured pruning methods that systematically remove redundant connections while preserving critical network pathways, and knowledge distillation approaches that transfer insights from larger models to compact networks. The effectiveness of these techniques varies considerably across different model architectures and application domains, requiring careful optimization for specific reliability monitoring scenarios. Hardware-software co-design methodologies have demonstrated significant efficiency improvements by tailoring both model architectures and execution hardware to specific application constraints, a particularly relevant approach for test infrastructure, where customized hardware acceleration may be feasible. The unique temporal characteristics of reliability monitoring present additional optimization opportunities, as many degradation phenomena evolve gradually over time, potentially enabling specialized architectures that trade instantaneous computation capability for improved energy efficiency through intermittent processing approaches that capitalize on the relatively slow evolution of the underlying phenomena being monitored.

6.2. Training Data Availability

The effectiveness of machine learning models fundamentally depends on the quality and representativeness of training data, creating significant challenges for IST applications where relevant failure data is inherently scarce and difficult to acquire. Research in data-centric AI has identified that data quality issues often present greater obstacles to effective model deployment than algorithmic limitations, particularly in specialized domains where labeled examples are limited [12]. Semiconductor reliability failures—particularly those associated with aging mechanisms and wear-out phenomena—may require extended operational periods to manifest under normal conditions, creating a fundamental temporal mismatch between development timelines and data availability. This scarcity is exacerbated by the rapidly evolving nature of semiconductor technology, where each process node and design generation introduces new reliability mechanisms that may not be adequately represented in historical data. Manufacturing testing generates substantial data volumes but focuses predominantly on initial quality rather than the in-field reliability challenges that IST aims to address. Research in few-shot learning and semi-supervised approaches has demonstrated promising results for domains with limited labeled examples, leveraging auxiliary tasks and data augmentation to improve model generalization despite training constraints. Self-supervised learning techniques that extract supervisory signals from unlabeled data offer particularly promising approaches for reliability monitoring, where abundant normal operation data can be leveraged to establish baseline behavioral models against which anomalies can be detected. Simulation-based approaches can supplement limited empirical data through physics-informed modeling of failure mechanisms, though the fidelity of these synthetic representations depends heavily on the accuracy of the underlying physical models and their ability to capture the complex interactions between multiple degradation factors. The inherent class imbalance in reliability data, where normal operation vastly outnumbers failure events, creates additional methodological challenges requiring specialized training approaches to improve model performance for rare event detection without being overwhelmed by the dominant class distribution.

6.3. Security and Privacy

The integration of AI capabilities within semiconductor test infrastructures introduces significant security and privacy considerations that extend beyond traditional test methodologies. Recent research in machine learning security has identified numerous attack vectors specific to AI systems, including adversarial examples that can induce misclassification through imperceptible input perturbations, model inversion attacks that can potentially extract training data from deployed models, and poisoning attacks that compromise model behavior through manipulated training examples [11]. These vulnerabilities present particular concerns for reliability monitoring, where compromised test systems could potentially be exploited to introduce reliability vulnerabilities, create side-channel opportunities, or enable denial-of-service conditions by triggering unnecessary test operations. The machine learning models themselves may become targets for adversarial manipulation designed to induce false positives or negatives in failure detection, potentially undermining reliability assurance or causing unnecessary interventions. Securing model update mechanisms presents unique challenges in semiconductor environments, where update processes must maintain integrity despite potential resource constraints and limited cryptographic capabilities. Beyond malicious manipulation, the telemetry data collected for AI-based testing often contains sensitive information about system operation and potential vulnerabilities that require careful protection. This data might inadvertently reveal proprietary details about chip design, manufacturing processes, or performance characteristics that could have competitive implications if improperly disclosed. In consumer contexts, telemetry might capture usage patterns that raise significant privacy concerns, particularly when aggregated across multiple devices for fleet learning applications. Emerging privacy-preserving machine learning techniques, including federated learning approaches that enable collaborative model development without centralizing sensitive data and differential privacy methods that systematically introduce calibrated noise to protect individual data points, offer promising directions for addressing these concerns while maintaining analytical capabilities. The evolving regulatory landscape surrounding data privacy creates additional complexity, requiring flexible implementation approaches that can adapt to diverse compliance requirements across global markets.

6.4. Standardization and Interoperability

The fragmented landscape of semiconductor test methodologies presents significant challenges for integrating AI capabilities across diverse platforms and implementation environments. Research in industrial analytics has identified that interoperability limitations often present greater obstacles to widespread adoption than technological constraints, particularly in domains with established legacy systems and heterogeneous technology stacks [12]. The absence of standardized formats for telemetry data, diagnostic information, and reliability metrics complicates the development of reusable analytical models and cross-platform learning capabilities. Current test infrastructures frequently employ proprietary data formats and communication protocols that inhibit interoperability, creating artificial barriers to collaborative development and knowledge sharing across the industry. The diversity of semiconductor applications

further complicates standardization efforts, as the critical parameters and failure mechanisms relevant for automotive applications may differ substantially from those in data center or consumer contexts. Research in industrial data standardization has demonstrated that semantic models and ontologies can bridge heterogeneous data sources by providing formal frameworks for relating diverse measurements to common conceptual models, enabling more sophisticated cross-domain analytics. Open data exchange formats designed specifically for time-series telemetry have shown promise in adjacent industries, potentially offering templates for semiconductor-specific adaptations that address the unique characteristics of test data. Beyond data formats, standardized interfaces between test infrastructure and AI processing elements would enable more flexible deployment options, allowing implementations to evolve independently while maintaining compatibility. Metadata standards for model provenance and validation metrics would further enhance interoperability by providing consistent frameworks for documenting model characteristics and expected performance, essential for safety-critical applications where reliability assurance must meet rigorous verification requirements. The establishment of benchmark datasets and standardized evaluation methodologies would similarly accelerate progress by enabling direct comparison between different analytical approaches and implementation strategies while providing common targets for improvement.

Table 4 Challenges and Research Directions for AI-Augmented IST. [11, 12]

Challenge	Key Research Direction	Potential Approach
Model Size and Efficiency	Neural network compression	Hardware-software co-design methodologies
Training Data Availability	Learning from limited examples	Self-supervised and few-shot learning techniques
Security and Privacy	Protecting model integrity	Federated learning with differential privacy
Standardization	Cross-platform interoperability	Semantic models and data exchange formats

7. Conclusion

The convergence of Artificial Intelligence and In-System Test methodologies represents a transformative shift in the semiconductor validation paradigm, fundamentally reimagining how reliability assurance is structured for complex systems-on-chip. By embedding intelligence directly into test infrastructure, the industry can transition from static, predetermined validation practices to dynamic, context-aware frameworks that continuously adapt to operational conditions and evolving reliability challenges. This integration creates semiconductor devices that are essentially self-aware, capable of monitoring their operational health, predicting potential failure mechanisms before they manifest, and adapting their behavior to maintain functional integrity under diverse stress conditions.

The architectural framework presented in this article—comprising comprehensive data collection mechanisms, efficient edge processing capabilities, secure model update infrastructure, and standardized telemetry integration—provides a blueprint for implementing AI-augmented IST across diverse application domains. The layered structure enables tailored implementations that balance analytical sophistication against resource constraints, creating practical deployment pathways for both resource-limited edge devices and performance-critical systems. By leveraging machine learning techniques, including convolutional neural networks for pattern recognition, reinforcement learning for adaptive scheduling, transfer learning for diagnostic efficiency, and self-supervised approaches for anomaly detection, these frameworks create multiple layers of reliability protection that complement traditional test methodologies while addressing their fundamental limitations.

The application-specific implementations discussed demonstrate how AI-augmented IST can be optimized for diverse operational contexts, from safety-critical autonomous systems where predictive capabilities directly impact human welfare, to data center environments where operational efficiency and service continuity drive implementation priorities, to medical devices where extended reliability horizons and minimal intervention opportunities necessitate sophisticated prognostic capabilities. These tailored examples highlight the versatility of the underlying architectural concepts while illustrating how domain-specific considerations shape practical deployment strategies.

Significant challenges remain on the pathway to widespread adoption, including the development of more efficient model architectures suitable for resource-constrained environments, innovative solutions for generating representative training data despite the scarcity of failure examples, robust security frameworks that protect both model integrity and data privacy, and standardization efforts that enable cross-platform interoperability and knowledge sharing. These challenges present fertile ground for future innovation, with promising directions emerging at the

intersection of hardware-software co-design, physics-informed machine learning, privacy-preserving analytics, and semantic data modeling.

Despite these obstacles, the fundamental value proposition of AI-augmented IST—enabling predictive rather than merely reactive reliability management—creates compelling incentives for continued investment and development. As semiconductor technology continues advancing toward increasingly specialized designs for artificial intelligence acceleration, the synergistic relationship between AI and IST creates a positive feedback loop: more sophisticated AI capabilities enable more effective test strategies, while more reliable semiconductor platforms enable more ambitious AI deployments in critical applications. This virtuous cycle promises to accelerate progress toward increasingly resilient electronic systems that can maintain functional integrity despite the growing complexity of both silicon technology and deployment environments.

By embedding intelligence into the foundational validation infrastructure of next-generation semiconductor devices, AI-augmented IST offers a promising pathway toward electronic systems that are not merely functional, but genuinely trustworthy—capable of maintaining operational integrity across extended lifespans, adapting to unforeseen operational challenges, and providing meaningful assurance of their reliability state. This evolution from static testing to intelligent, adaptive validation represents a crucial step toward electronic systems that can meet the reliability demands of increasingly autonomous, connected, and safety-critical applications that will define the next generation of computing.

References

[1] Niranjana Gurushankar, "Challenges in Verifying Complex SOC Designs," Journal of Artificial Intelligence, Machine Learning and Data Science, 2022. https://www.researchgate.net/publication/388769128_Challenges_in_Verifying_Complex_SOC_Designs

[2] Thomas Reinbacher et al., "Runtime verification of embedded real-time systems," Formal Methods in System Design, Springer, 2013. https://link.springer.com/article/10.1007/s10703-013-0199-z

[3] Katharina Juhnke et al., “Challenges concerning test case specifications in automotive software testing: assessment of frequency and criticality," Software Quality Journal, Springer, 2020. https://link.springer.com/article/10.1007/s11219-020-09523-0

[4] Oscar Amelia, "AI-Driven Testing and Validation Techniques for Low-Power Semiconductor Design Verification Using UVM," ResearchGate Publication, 2024. https://www.researchgate.net/publication/384896673_AI-Driven_Testing_and_Validation_Techniques_for_Low-Power_Semiconductor_Design_Verification_Using_UVM

[5] Rachita Ghoshhajra et al., "A Review on Machine Learning Approaches for Predicting the Effect of Device Parameters on Performance of Nanoscale MOSFETs," 2021 Devices for Integrated Circuit (DevIC), 2021. https://ieeexplore.ieee.org/document/9455840

[6] Ana Pereira, Carsten Thomas, "Challenges of Machine Learning Applied to Safety-Critical Cyber-Physical Systems," Mach Learn Knowl Extr, 2020. https://www.mdpi.com/2504-4990/2/4/31

[7] Baotong Chen et al., "Edge Computing in IoT-Based Manufacturing," IEEE Communications Magazine, 2018. https://www.researchgate.net/publication/327705750_Edge_Computing_in_IoT-Based_Manufacturing

[8] Koen Zandberg et al., “Secure Firmware Updates for Constrained IoT Devices Using Open Standards: A Reality Check," IEEE Access, 2019. https://www.researchgate.net/publication/333472928_Secure_Firmware_Updates_for_Constrained_IoT_Devices_Using_Open_Standards_A_Reality_Check

[9] Francesco Concas et al., "VALIDATION FRAMEWORKS FOR SELF-DRIVING VEHICLES: A SURVEY," arXiv:2007.11347v1 [cs.SE] 2020. https://arxiv.org/pdf/2007.11347

[10] Abbas Akbari et al., "Thermal-Aware Virtual Machine Allocation for Heterogeneous Cloud Data Centers," Energies, 2020. https://www.mdpi.com/1996-1073/13/11/2880

[11] Xiangzhong Luo et al., "Efficient Deep Learning Infrastructures for Embedded Computing Systems: A Comprehensive Survey and Future Envision," ACM Transactions on Embedded Computing Systems, 2024. https://dl.acm.org/doi/10.1145/3701728

[12] Duan-Yang Liu et al., "Machine learning for semiconductors," Chip, 2022. https://www.sciencedirect.com/science/article/pii/S2709472322000314

[13] Pavan Kumar Datla Jagannadha et al., “Special Session: In-System-Test (IST) Architecture for NVIDIA Drive-AGX Platforms,” IEEE Xplore, 2019. https://ieeexplore.ieee.org/document/8758636