AI-Driven Chip Testing on the Cloud

Elena Petrova
Asst Professor

Abstract
The semiconductor testing process is a key bottleneck in the chip manufacturing flow. To address this inefficiency, data engineering pipelines have been built in the cloud that unify the ingest of test vectors and rich parametric telemetry for real-time testing and AI-powered diagnosis. New chips are tested using algorithms that can decide whether to release or scrap a chip based on the test results and preceding telemetry. Because clearance decisions are made during test, failed chips can quickly be relabeled or logged without being held in test. Data science pipelines examine failed chips and use the conclusions from these analysis to detect issues in other chips in near real-time. Such AI-powered diagnostics also have the potential to help prevent future recurrences. Acting on these insights can ultimately improve yield and shorten test time for future designs.

Data pipelines are about two orders of magnitude less expensive than the test process. They enable continuous decision-making for testing, clearance, and failure analysis without manual steps and facilitate timely action based on telemetry and failed part diagnostics. Making the testing process more efficient can represent significant cost savings for high-volume products. Ultimately, delivering better and cheaper parts to customers faster improves competitiveness and profitability.

Keywords: Semiconductor testing, semiconductor diagnostics, data engineering, cloudscale development, cloud-scale data engineering .

1. Introduction

Semiconductor generation and testing remain costly and time-consuming processes that benefit from cloud-scale data engineering. High-fidelity semiconductor test vectors are computed in a cloud environment and the test operations of chips are distributed to multiple test systems. The semiconductor testing becomes a cloud-scale data generation system, with the ATE data ingested in real time into cloud-based storage and serving systems. In parallel, an AI-based visual diagnostics pipeline is enabled. High-quality visual diagnosis labels are created to train AI-based visual diagnostics models, thereby providing a scalable solution. High-speed cloud ingestion and storage systems handle the large amounts of diagnostic data generated by the physical testing operations. Real-time AI-based visual diagnosis of the chips and non-intrusive test vector noise analysis enable root-cause analysis on the tested chips. The presented cloud-scale data engineering approach and test-vector-management system enable high-fidelity and real-time semiconductor data generation for AI-powered diagnostics. The AI-powered visual-chip-diagnostic system and the cloud-scale semiconductor ATE ingestion are capable of visual analysis and scaling with only automatic test equipment data.

The successful deployment and operation of high-speed data pipelines focused on the real-time generation of high-fidelity cloud-based data for AI applications are showcased. The methodology enabled is relevant to all data-science and AI disciplines for which high-quality, wide-area data-labeling pipelines and infrastructure are critical. Deep learning is an advanced data-mining technology that continuously pushes the limits of SI systems. At the same time, rubber-stamp manufacturing processes, decreases in the size of semiconductor technologies and constant pressure to find faults at each level, increase the yield and reduce the cost of production are forcing data generation at as high speed as the manufacturing process permits. The ultimate prize is finding a solution to the problem of previous black-box development methodologies of artificial intelligence based chips.

Table 1. Overview of the Proposed Semiconductor Testing Framework

Component	Description	Purpose	Technologies/Approaches
Cloud-Scale Data Engineering	Distributed cloud infrastructure for semiconductor testing data	Real-time ingestion, storage, and analytics	Open-source cloud technologies, streaming systems
Semiconductor Test Systems	Automated Test Equipment (ATE) executing test vectors	Validate chip functionality and quality	Real-time chip testing
Data Lake	Centralized storage repository	Store telemetry, diagnostics, and test measurements	Streaming ingestion, scalable storage
AI-Powered Diagnostics	Machine learning models for failure analysis	Detect defects and improve yield	CNNs, deep learning, predictive diagnostics
Test Vector Management	Organizes and distributes test vectors	Improve test coverage and reduce redundancy	CCTV and TVD strategies
Streaming Serving Layer	Real-time data dissemination layer	Continuous monitoring and diagnostics	Cloud streaming architecture

2. Background and Motivation

With a more connected world, there is a growing need for more interconnected devices, appliances, and systems, and semiconductors are the fundamental building blocks of these

ISSN: 3067-4190
JOURNAL OF ARTIFICIAL INTELLIGENCE AND BIG DATA DISCIPLINES
Volume: 04 Issue: 01
RECEIVED: JANUARY 18 REVISED: FEBRUARY 02 ACCEPTED: FEBRUARY 20 PUBLISHED: MARCH 04

devices, which has attracted investments of hundreds of billion dollars in research facilities and product manufacturing. After an overall picture of this booming market, this section presents the two focal areas of this work with a more academic perspective: semiconductor testing for validating and certifying products, and Artificial Intelligence, drawing bridges between those two areas. Real-time data engineering processes, transported from the cloud to the edge, seize test signals and guarantee a method to generate the best test samples for advanced kayseq technologies.

An increasing number of products and systems are based on Artificial Intelligence techniques, such as information classification, image processing, anomaly detection, speech recognition, information generation, and many others. Nevertheless, many products are restricted to specific applications, and the potential empowered by these Artificial Intelligence techniques will only be accessible for all products when the development of high-quality information labeling tools and processes finally breaks the existing barriers. One of those barriers is the lack of high-quality labeled information, especially for training models that classify information in abnormal behavior and during testing for edge inference. One of the AI models in this work aims to help build high-quality labeled information repositories for semi-conductors, boosting test coverage for these products that necessarily play a central role in enabling the integration of AI techniques in all products and systems.

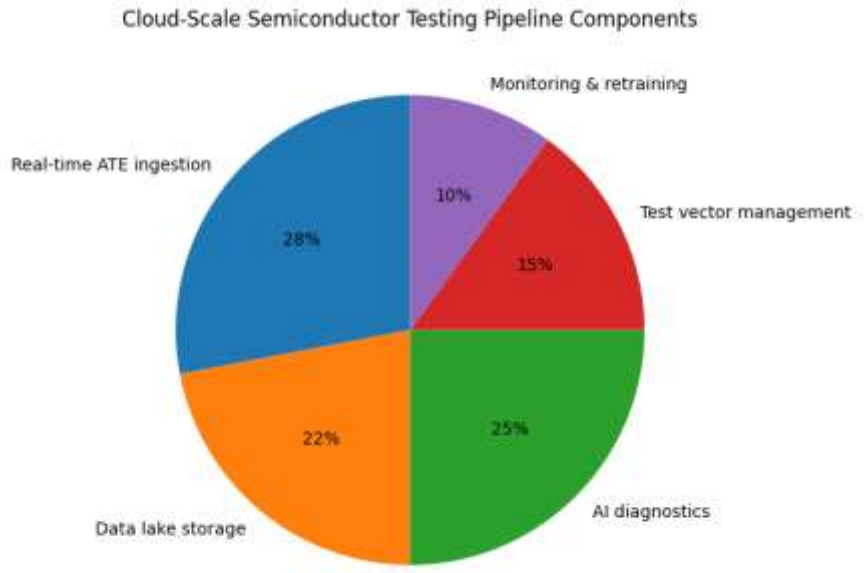


Table 2. Real-Time Data Engineering Pipeline

ISSN: 3067-4190
JOURNAL OF ARTIFICIAL INTELLIGENCE AND BIG DATA DISCIPLINES
Volume: 04 Issue: 01
RECEIVED: JANUARY 18 REVISED: FEBRUARY 02 ACCEPTED: FEBRUARY 20 PUBLISHED: MARCH 04

Pipeline Stage	Input Data	Processing Activity	Output
Data Generation	Semiconductor test vectors and telemetry	Functional testing of chips	Raw test data
Data Ingestion	ATE measurement data	Continuous upload to cloud	Streamed semiconductor data
Data Storage	Streaming data	Storage in cloud data lake	Structured datasets
AI Diagnostics	Historical and live test data	Predictive analysis and classification	Failure diagnostics
Monitoring & Analytics	Diagnostic outputs	Visualization and root-cause analysis	Yield optimization insights
Feedback Loop	AI-generated insights	Test vector improvement	Enhanced future testing

3. System Architecture for Cloud-Scale Data Engineering

A distributed architecture based on open-source cloud-scale data engineering principles establishes a cloud-scale solution stack that ingests, processes, and distributes semiconductor test data in real time. Cloud-scale computing enables data engineering systems that perform the requisite data processing and data movement at something approaching the required rates for industrial applications. Relatively small volumes of data are created and made available within seconds to hopefully help engineers narrow the root cause analysis as fast as possible. For semiconductor devices previously subjected to real-time diagnostic tests, the data is used to power a suite of AI-based diagnostics running in urban legend.

The first component of the architecture is a streaming serving layer designed to facilitate the dissemination of the data up and down the testing process from various levels of testing (wafer, package, board, system) to the diagnostic model developers and back. This layer is complemented by a test vector management system that organizes and processes the test vectors delivered to the tester during production testing. Together with the previous component, they provide the requisite data flow for real-time monitoring of semiconductor testing.

Mathematical Formulas:

1. Signal-to-Noise-and-Distortion Ratio (SINAD)

Used in semiconductor diagnostics and test-quality evaluation.

$$SINAD = 10\log_{10}\left(\frac{P_{signal} + P_{noise} + P_{distortion}}{P_{noise} + P_{distortion}}\right)$$

$$SINAD = 10\log_{10}\left(\frac{P_{signal} + P_{noise} + P_{distortion}}{P_{noise} + P_{distortion}}\right)$$

2. Data Throughput Equation

Represents streaming data ingestion performance in cloud pipelines.

$$Throughput = \frac{Data\ Volume}{Processing\ Time}$$

$$Throughput = \frac{Data\ Volume}{Processing\ Time}$$

3. Chip Yield Formula

Useful for semiconductor manufacturing efficiency.

$$Yield = \frac{Number\ of\ Good\ Chips}{Total\ Chips\ Tested}$$

$$Yield = \frac{Number\ of\ Good\ Chips}{Total\ Chips\ Tested}$$

4. Classification Accuracy

Applied in AI-powered chip diagnostics.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

5. Precision of Failure Detection

Measures diagnostic correctness.

$$Precision = \frac{TP}{TP + FP}$$

$$Precision = \frac{TP}{TP + FP}$$

6. Recall for Fault Detection

Useful in semiconductor anomaly detection.

$$Recall = \frac{TP}{TP + FN}$$

$$Recall = \frac{TP}{TP + FN}$$

7. Latency in Real-Time Streaming

$$Latency = Processing\ Time - Arrival\ Time$$

8. CNN Convolution Operation

Related to AI-based chip image diagnostics.

$$F(i,j) = \sum_{m} \sum_{n} I(i-m, j-n) K(m,n)$$

9. Mean Squared Error (MSE)

Used during AI model training.

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

10. Data Compression Ratio

Useful for cloud-scale storage optimization.

$$Compression\ Ratio = \frac{Original\ Size}{Compressed\ Size}$$

11. Failure Probability

$$P(Failure) = \frac{Failed\ Chips}{Total\ Chips}$$

12. Test Coverage Equation

$$Coverage = \frac{Detected\ Faults}{Total\ Faults}$$

$$Coverage = \frac{Detected\ Faults}{Total\ Faults}$$

13. AI Model Loss Function

Binary classification for pass/fail chip prediction.

$$Loss = -[y\log(\hat{y}) + (1-y)\log(1-\hat{y})]$$

14. Streaming Queue Utilization

$$Utilization = \frac{Arrival\ Rate}{Service\ Rate}$$

15. Cloud Storage Scaling

$$Storage_{total} = N \times Storage_{per\ chip}$$

3.1.DataIngestionandStreaming

The first layer of the cloud-scale data architecture addresses the use case of real-time semiconductor testing. A key requirement is the ingestion of measurement data that is generated during semiconductor tests at a 24-by-7 scale. Each test of a semiconductor device requires the execution of test vectors that are applied to the device terminals, and the test results are returned to the testing system and stored in a SQL storage system. An

ISSN: 3067-4190
JOURNAL OF ARTIFICIAL INTELLIGENCE AND BIG DATA DISCIPLINES
Volume: 04 Issue: 01
RECEIVED: JANUARY 18 REVISED: FEBRUARY 02 ACCEPTED: FEBRUARY 20 PUBLISHED: MARCH 04

efficient data upload system is provided to continuously gather test measurement data from these SQL systems at the end of every test into streaming storage of a data lake.

The data-lake serving these test measurements also supports the use case for AI-generated chip test diagnostic data and leverages the streaming ingestion of chip test measurement data from the real-time chip test system.

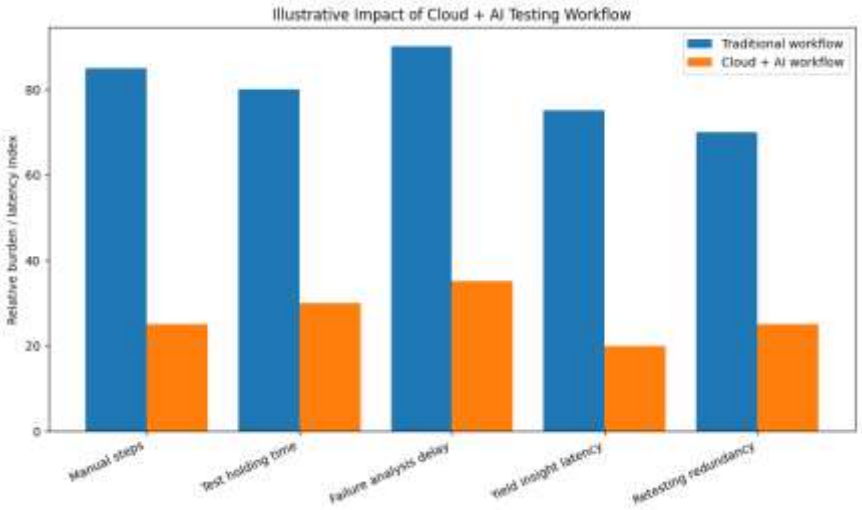


Table 3. Semiconductor Testing Workflow

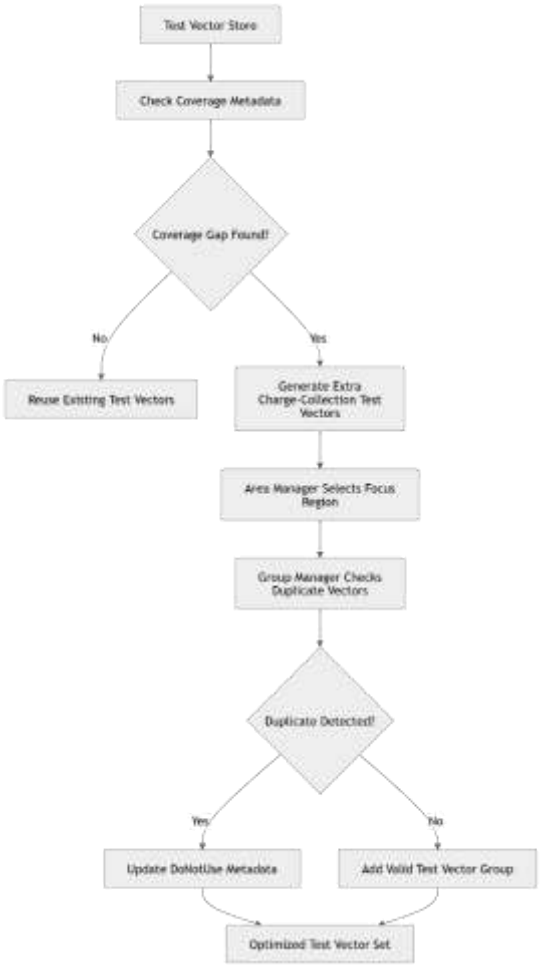
Stage	Activity	Objective	Result
Test Vector Generation	Create functional test patterns	Stimulate chip functions	Test matrix
Functional Testing	Execute vectors on semiconductor devices	Validate chip behavior	Response measurements
Signal Analysis	Compare outputs with expected values	Detect deviations	SINAD assessment
Data Logging	Store simulation and test logs	Trace chip behavior	Diagnostic datasets
Failure Analysis	Analyze abnormal responses	Identify root causes	Failure classification
AI-Based Recommendations	Predict corrective actions	Improve production quality	Yield enhancement

4. Real-Time Testing Pipelines for Semiconductors

The core of semiconductor testing consists in the assessment of chip responsiveness to thousands of functional patterns, called test vectors. Test vectors characterize the chip functions of interest and are explicitly defined by the chip designer. To disseminate these vectors, a dedicated Test Pattern Generator (TPG) is implemented. The test matrix produced by the TPG serves as input for various semiconductor testers or channels. Each test vector stimulus activates the chip, and the chip response is measured and compared to the expected output.

In parallel, an industry-standard register transfer-level simulation set-up provides logging coverage of all stimulated chip functions, offering a baseline to determine the quality of chip test results as well as the reasons for any eventual failure. The test matrix from the TPG, combined with the simulation log coverage and the results from all testing channels, enables the generation of a Signal-to-Noise-and-Distortion Ratio (SINAD) assessment for the chip functions, which is essential for diagnosis.

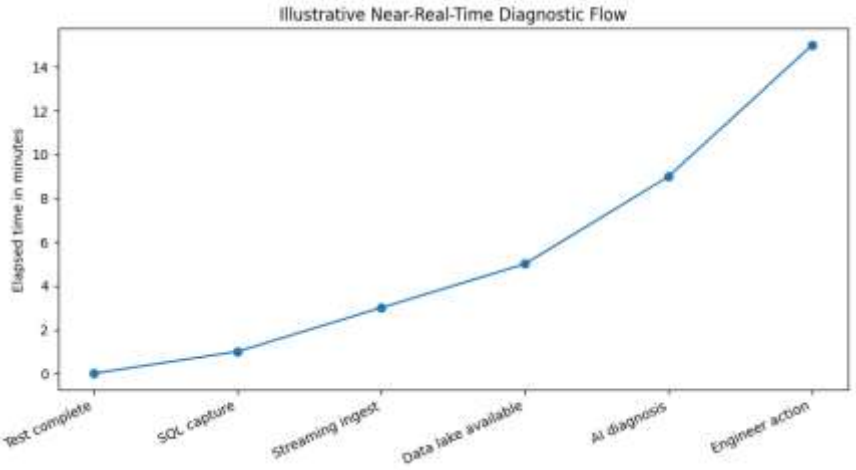
ISSN: 3067-4190
JOURNAL OF ARTIFICIAL INTELLIGENCE AND BIG DATA DISCIPLINES
Volume: 04 Issue: 01
RECEIVED: JANUARY 18 REVISED: FEBRUARY 02 ACCEPTED: FEBRUARY 20 PUBLISHED: MARCH 04



4.1.TestVectorManagement

Test vectors are the input patterns used to stimulate a silicon chip when testing it after fabrication and are cached and reused in final test. Charge-Collection Test Vectors (CCTV) are specifically for detecting charge-collection defects caused by radioactive particles. They check if charge injected into the chip is trapped in the oxide near the MOSFET channel and can leak through the oxide. These defects tend to only trigger a handful of test vectors, so a careful selection and categorization process is needed for future tests. Other test vectors are for detecting gaps in the testing space. Typically, for CTVs, if a small part of the chip has already been tested with a CCTV, the DoNotUse metadata is updated to prevent redundancy.

ISSN: 3067-4190
JOURNAL OF ARTIFICIAL INTELLIGENCE AND BIG DATA DISCIPLINES
Volume: 04 Issue: 01
RECEIVED: JANUARY 18 REVISED: FEBRUARY 02 ACCEPTED: FEBRUARY 20 PUBLISHED: MARCH 04



The extra TVD management uses groups of CTVs and is triggered when the TX store is populated enough. Four stages help decide whether to generate extra CTVs, determine which area to focus on, how many areas to cover, and the distribution of the test vectors based on missing coverage. This management is adapted for the extra TVD generator. When a duplicate is detected, the group manager update the valid DoNotUse area and moves to the group with the same area received from the area manager.

Table 4. Test Vector Management Strategies

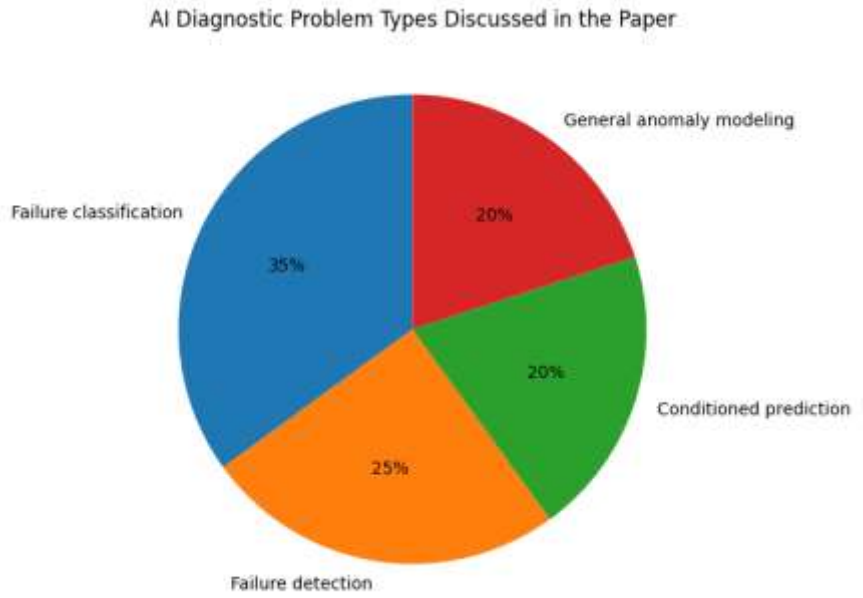
Strategy	Description	Benefit
Charge-Collection Test Vectors (CCTV)	Detect charge-collection defects caused by radioactive particles	Improve defect identification
DoNotUse Metadata	Prevent repeated testing in already-covered areas	Reduce redundancy
Extra TVD Generation	Generate additional vectors for uncovered regions	Improve test coverage
Group-Based Vector Management	Organize vectors into coverage groups	Efficient vector allocation
Duplicate Detection Mechanism	Detect overlapping vector coverage	Optimize resource utilization

5. AI-Powered Chip Diagnostics

ISSN: 3067-4190
JOURNAL OF ARTIFICIAL INTELLIGENCE AND BIG DATA DISCIPLINES
Volume: 04 Issue: 01
RECEIVED: JANUARY 18 REVISED: FEBRUARY 02 ACCEPTED: FEBRUARY 20 PUBLISHED: MARCH 04

As semiconductor manufacturing is becoming increasingly complex, so are the test-based failures. Traditionally, the analysis of these failure modes was carried out manually by engineers and specialists. However, the growing number of chips to test—with the associated volume of potential test-based failures—has prompted the industry to look into solutions that improve the cadence, automation, and cost efficiency of the analysis. To address this challenge, AI has been adopted as a solution to fill the gap of human expertise. One example from the automotive industry is a Deep Learning-based failure classification model provided with redesign recommendations. Such applications necessitate the consumption of not only test-side information but also traditional design-for-test information. These relationships, however, are not always straightforward.

Four different AI problems can be identified based on the designer's goal concerning the classification and the provided input. Failures can either be classified or detected, while the model can be conditioned either with an input feature that strongly helps to predict the failure or in the most general form. There are two main reasons that explain why traditional supervised classification is not sufficient. First, a chip can be affected by too many failure modes. Second, semiconductor manufactures follow a test-based strategy, which often leads to unbalanced data—presence of a failure in a limited number of tested samples compared to its total amount. The proposed AI-powered chip diagnostics accounts for these issues and consists of a combination of multiple AI models.



5.1.DiagnosticModelsandTrainingRegimes

Reliable chip diagnostics not only help manufacturers identify the chip failure reasons but also aid design engineers to figure out possible design flaws. Unfortunately, there are no

ISSN: 3067-4190
JOURNAL OF ARTIFICIAL INTELLIGENCE AND BIG DATA DISCIPLINES
Volume: 04 Issue: 01
RECEIVED: JANUARY 18 REVISED: FEBRUARY 02 ACCEPTED: FEBRUARY 20 PUBLISHED: MARCH 04

adequate solutions publicly available to facilitate scalable chip diagnostics modeling. Given the increasing complexity of modern chips, traditional diagnostic modeling approaches such as manually designed rules or expert systems may become impractical. Consequently, various state-of-the-art deep learning-based techniques for failing chip diagnostic modeling have been developed but generally offer poor performance. These techniques share several limitations. First, they use a multi-layer perception as a classifier, which is commonly believed to be less effective than convolutional neural networks (CNNs) when modeling classification tasks. Second, they are trained either in a single-stage or in a two-stage manner during the whole chip-level classification process. However, the two-stage training configuration introduces layer-wise category imbalance in the first-stage model, which may further introduce suboptimal first-stage models. Last but not least, these techniques lack a diagnostic failure-path analysis which helps designers to identify potential circuit design errors.

To address these limitations, researchers propose a novel diagnostic modeling framework for failing chip classification named ChipDiags, which comprises two stages: a diagnostic model and a following diagnostic sub-category analysis. A CNN-based deep learning configuration is adopted in the diagnostic modeling. The first stage investigates the chip diagnostic model training via an end-to-end level-failing-chip classification scheme, and the second stage develops a sub-category analysis system that is capable of shining a light along the failure paths of malfunctioned chips and identifying the abnormal circuit areas contributing most to the chip failure.

Table 5. AI-Powered Chip Diagnostics Models

AI Model Type	Function	Input Data	Expected Outcome
Failure Classification Model	Categorize chip failures	Test telemetry and logs	Failure categories
Failure Detection Model	Detect abnormal chip behavior	Real-time test data	Fault alerts
CNN-Based Diagnostic Model	Deep learning-based defect analysis	Chip diagnostic datasets	Improved classification accuracy
Predictive Diagnostics Model	Predict future chip failures	Historical testing patterns	Preventive maintenance
Sub-Category Analysis System	Analyze failure paths	Diagnostic outputs	Root-cause identification

6. Conclusion

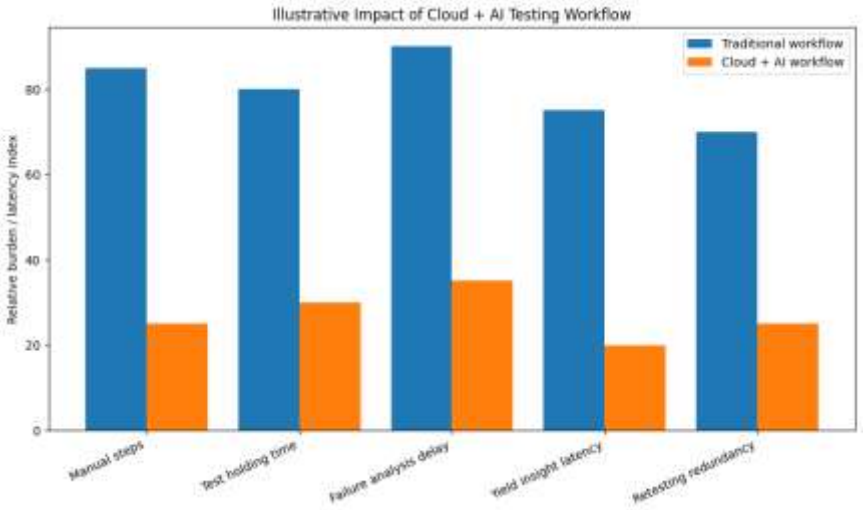
The presented work illustrates a robust, scalable system architecture for the cloud-scale data engineering of semiconductor testing and diagnostics. Real-time consistency and fault tolerance requirements are met through a combination of open-source tools for data ingestion, quality control, storage, and analytics. Data is streamed in real time from semiconductor test stations during wafer manufacturing. Test vector quality is monitored and outliers highlighted for analyst review. The diagnostics subsystem deploys an advanced set of AI-powered machine learning models that reduce lab testing time in production by several orders of magnitude. These models are designed to withstand concept drift, with automatic retraining triggered by data quality monitoring.

Data engineering pipelines that serve as the foundations of semiconductor artificial intelligence are also described. New test data are ingested into the data lake on a per-test-case basis as they are generated during manufacturing. The test-condition space is monitored for gaps and the quality of surrounding test data is shown to affect AI model accuracy.

Table 6. Comparison Between Traditional and AI-Based Diagnostics

Feature	Traditional Diagnostics	AI-Powered Diagnostics
Analysis Method	Manual engineering analysis	Automated machine learning
Scalability	Limited	Highly scalable
Processing Speed	Slow	Real-time or near real-time
Failure Detection	Rule-based	Data-driven and adaptive
Handling Large Data Volumes	Difficult	Efficient
Predictive Capability	Limited	Strong predictive analytics
Cost Efficiency	High operational cost	Reduced operational cost

ISSN: 3067-4190
JOURNAL OF ARTIFICIAL INTELLIGENCE AND BIG DATA DISCIPLINES
Volume: 04 Issue: 01
RECEIVED: JANUARY 18 REVISED: FEBRUARY 02 ACCEPTED: FEBRUARY 20 PUBLISHED: MARCH 04



References

1. Amrouch, H., Chowdhury, A. B., Jin, W., Karri, R., Khorrami, F., Krishnamurthy, P., Polian, I., Van Santen, V. M., Tan, B., & Tan, S. X. D. (2021). Special session: Machine learning for semiconductor test and reliability. In Proceedings of the 2021 IEEE 39th VLSI Test Symposium. IEEE.
2. Kareem, P., & Shin, Y. (2021). Synthesis of lithography test patterns using machine learning model. IEEE Transactions on Semiconductor Manufacturing, 34(1), 49–57.
3. Davuluri, P. S. L. (2023). AI-Augmented Sanctions Screening: Enhancing Accuracy and Latency in Real Time Compliance Systems. AI-Augmented Sanctions Screening: Enhancing Accuracy and Latency in Real Time Compliance Systems (December 15, 2023).
4. Cheng, K. C. C., Chen, L. L. Y., Li, J. W., Li, K. S. M., Tsai, N. C. Y., Wang, S. J., Huang, A. Y. A., Chou, L., Lee, C. S., Chen, J. E., Liang, H. C., & Hsu, C. L. (2021). Machine learning-based detection method for wafer test induced defects. IEEE Transactions on Semiconductor Manufacturing, 34(2), 161–167.
5. Jiang, D., Lin, W., & Raghavan, N. (2021). A Gaussian mixture model clustering ensemble regressor for semiconductor manufacturing final test yield prediction. IEEE Access, 9, 22253–22263.
6. Chowdhury, A. B., Tan, B., Garg, S., & Karri, R. (2022). Robust deep learning for IC test problems. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 41(1), 183–195.
7. Li, K. S. M., Chen, L. L. Y., Cheng, K. C. C., Liao, P. Y. Y., Wang, S. J., Huang, A. Y. A., Chou, L., Tsai, N. C. Y., & Lee, C. S. (2022). TestDNA-E: Wafer defect signature for pattern recognition by ensemble learning. IEEE Transactions on Semiconductor Manufacturing, 35(2), 373–380.
8. Li, K. S. M., Jiang, X. H., Chen, L. L. Y., Wang, S. J., Huang, A. Y. A., Chen, J. E., Liang, H. C., & Hsu, C. L. (2022). Wafer defect pattern labeling and recognition using

semi-supervised learning. IEEE Transactions on Semiconductor Manufacturing, 35(2), 291–299.

9. Kolla, S. K., & Reddy, V. A. R. (2024). Evaluating Cloud-Native vs. Hybrid Architectures for Health Benefit Administration Systems. International Journal of Medical Toxicology and Legal Medicine, 27(5), 1042-1053.
10. Xu, H. W., Qin, W., & Sun, Y. N. (2022). An improved XGBoost prediction model for multi-batch wafer yield in semiconductor manufacturing. IFAC-PapersOnLine, 55(10), 2162–2166.
11. Xu, Q., Xu, C., & Wang, J. (2022). Forecasting the yield of wafer by using improved genetic algorithm, high dimensional alternating feature selection and SVM with uneven distribution and high-dimensional data. Autonomous Intelligent Systems, 2, Article 12.
12. Shi, Z., Li, M., Khan, S., Wang, L., Wang, N., Huang, Y., & Xu, Q. (2022). DeepTPI: Test point insertion with deep reinforcement learning. In Proceedings of the 2022 IEEE International Test Conference (pp. 194–203). IEEE.
13. Liao, Y., Najafi-Haghi, Z. P., Wunderlich, H.-J., & Yang, B. (2022). Efficient and robust resistive open defect detection based on unsupervised deep learning. In Proceedings of the 2022 IEEE International Test Conference. IEEE.
14. Loganathan, R., & Kolla, S. H. (2024). HARNESSING GENERATIVE AI FOR ADAPTIVE RISK POLICY GENERATION IN MULTI-REGULATORY DATA CENTER ENVIRONMENTS. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 15(3), 505–515. https://doi.org/10.61841/vakvfh91
15. Liu, J., Qiao, F., Zou, M., Zinn, J., Ma, Y., & Vogel-Heuser, B. (2022). Dynamic scheduling for semiconductor manufacturing systems with uncertainties using convolutional neural networks and reinforcement learning. Complex & Intelligent Systems, 8, 4641–4662.
16. Lee, Y. H., & Lee, S. (2022). Deep reinforcement learning based scheduling within production plan in semiconductor fabrication. Expert Systems with Applications, 191, 116222.
17. Parmar, T. (2022). Overcoming challenges in cloud adoption for semiconductor manufacturing data engineering. Journal of Artificial Intelligence, Machine Learning and Data Science, 1(1), 2201–2204.
18. Lim, H. G., Jang, J., Ju, B. K., Ko, J. W., & Kim, C. O. (2023). A test vector selection method based on machine learning for efficient presilicon verification. Expert Systems with Applications, 220, 120056.
19. Lee, Y., & Roh, Y. (2023). An expandable yield prediction framework using explainable artificial intelligence for semiconductor manufacturing. Applied Sciences, 13(4), 2660.
20. Bandi, V. D. V. K. (2024). Intelligent Data Platforms For Personalized Retail Analytics At Scale. Metallurgical and Materials Engineering, 30 (4), 1011–1027.
21. Gallagher, J. C., Mastro, M. A., Ebrish, M. A., Jacobs, A. G., Gunning, B. P., Kaplar, R. J., Hobart, K. D., & Anderson, T. J. (2023). Using machine learning with optical profilometry for GaN wafer screening. Scientific Reports, 13, 3352.
22. Batool, U., Shapiai, M. I., Mostafa, S. A., & Ibrahim, M. Z. (2023). An attention-augmented convolutional neural network with focal loss for mixed-type wafer defect classification. IEEE Access, 11, 108891–108905.

23. Li, R., & Kang, Z. (2023). Deep learning for wafer map defect detection: A review. In Proceedings of the 2023 Global Reliability and Prognostics and Health Management Conference (PHM-Hangzhou). IEEE.
24. Prasad, A., Dey, B., Blanco, V., & Halder, S. (2024). An evaluation of continual learning for advanced node semiconductor defect inspection. In Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD 2024).
25. Gallagher, J. C., Mastro, M. A., Jacobs, A. G., Kaplar, R. J., Hobart, K. D., & Anderson, T. J. (2024). Detecting defects that reduce breakdown voltage using machine learning and optical profilometry. Scientific Reports, 14, 7440.
26. Zhai, W., Shi, X., Wong, Y. D., Han, Q., & Chen, L. (2024). Explainable AutoML (xAutoML) with adaptive modeling for yield enhancement in semiconductor smart manufacturing. In Proceedings of the 2024 2nd International Conference on Artificial Intelligence and Automation Control. IEEE.
27. Mangalampalli, B. M. Intelligent Data Profiling for Healthcare Data Lakes Using AI-Enhanced Analytics.
28. Jo, U. K., & Kim, S. B. (2025). Semi-supervised learning with wafer-specific augmentations for wafer defect classification. IEEE Access, 13, 56–66.
29. Tsai, T. H., & Wang, C. Y. (2025). Wafer map defect classification using deep learning framework with data augmentation on imbalance datasets. EURASIP Journal on Image and Video Processing, 2025, Article 6.
30. Ingle, R., Shahade, A. K., Gaikwad, M., & Patil, S. (2025). Deep learning driven silicon wafer defect segmentation and classification. MethodsX, 14, 103158.
31. Qiao, Y., Chen, Y., Liu, F., Mei, Z., Luo, Y., Chen, Y., Liao, Y. Y., Wu, B., & Deng, Y. (2025). RA-UNet: A new deep learning segmentation method for semiconductor wafer defect analysis on fine-grained scanning electron microscope (SEM) images. IEEE Transactions on Semiconductor Manufacturing, 38(2), 185–193.
32. Kim, J., Hwang, S., Jeong, J., Jeong, J., Chang, K. B., Choi, H., Shon, S., & Kim, S. B. (2025). Range-aware deep learning framework for multi-parameter electrical test prediction in semiconductor manufacturing. In Proceedings of the 2025 IEEE 21st International Conference on Automation Science and Engineering (CASE). IEEE.
33. Zhao, D., Chen, Y., Guo, P., Chen, F., Jing, R., Xu, X., Chen, Y., Liu, F., & Wu, B. (2026). A survey on semiconductor wafer yield prediction by artificial intelligence. Microelectronics Journal, 167, 106958.
34. Awais, M., Postolache, O. A., & Oliveira, S. M. (2026). Graph-based contrastive learning for self-supervised semiconductor wafer defect detection. Journal of Intelligent Manufacturing.
35. Yue, K., Li, J., Deng, S., Chen, Z., Kwoh, C. K., & Li, W. (2026). Toward wafer defect detection: An incremental industrial large model with prompt-aware memory replay. Applied Soft Computing, 194, 114815.