ISSN: 3067-4166

AMERICAN DATA SCIENCE JOURNAL FOR ADVANCED COMPUTATIONS

VOLUME: 02 ISSUE: 02

RECEIVED: APRIL 22 REVISED: MAY 11 ACCEPTED: MAY 27 PUBLISHED: JUNE 15

Smart Predictive Testing for Semiconductors

Niklas Andersson
Asst Professor

***Abstract*—Next-generation semiconductor chips require predictive testing to assure the quality, reliability, and time-to-market of increasing complex systems-on-chip with multiple processor cores and dedicated hardware accelerators. AI-enabled methodologies are needed such that the test coverage and resource commitments are commensurate with the risk of undetected latent defects.**

Against this background, this work presents the architecture of a predictive testing framework that deploys AI technologies to address all main elements crucial for predictive testing—component fault prediction, test conditions that prioritize areas with elevated risk, and test vectors that consider fault interactions. Core capabilities are outlined, the integration approach specified, and the expected outcomes discussed. The proposed predictive testing framework contributes to the body of knowledge by providing AI-enabled methodologies that go beyond simple fault detection and create a predictable testing roadmap for semiconductor chips.

Keywords—Predictive Semiconductor Testing, AI-Enabled Test Engineering, System-on-Chip Reliability, Component Fault Prediction, Risk-Aware Test Coverage, Intelligent Test Vector Generation, Latent Defect Detection, Hardware Accelerator Validation, Multi-Core SoC Testing, Test Resource Optimization, Fault Interaction Modeling, Predictive Quality Assurance, Silicon Yield Improvement, Time-to-Market Optimization, AI-Driven Verification Frameworks, Test Prioritization Strategies, Integrated Test Architectures, Reliability Analytics, Semiconductor Manufacturing Intelligence.

I. INTRODUCTION (*HEADING 1*)

Semiconductor chips have reached advanced levels of performance and complexity. New manufacturing technologies allow for miniaturization, but they also result in lowered tolerances and correspondingly increased risk, especially with regard to hardware faults. Consequently, predictive testing is gaining popularity in the community—not only as a means of increasing yield but also as a means of improving reliability and speeding up time-to-market. However, managing and coordinating predictive testing effectively is also a considerable challenge. Different components of the predictive testing process are typically implemented as isolated modules, with minimal integration and little or no built-in provision for scalability or reproducibility.

An integrated framework based on statistical and machine-learning (ML) techniques is proposed for managing predictive testing for next-generation semiconductor chips. The framework covers support for data pipelines and feature engineering; predictive-model libraries with associated training protocols; AI-augmented generation and selection of predictive-test scenarios; and validation and verification of the prediction process, along with proposals for using data from the prediction process to generate benchmarks for predictive testing.

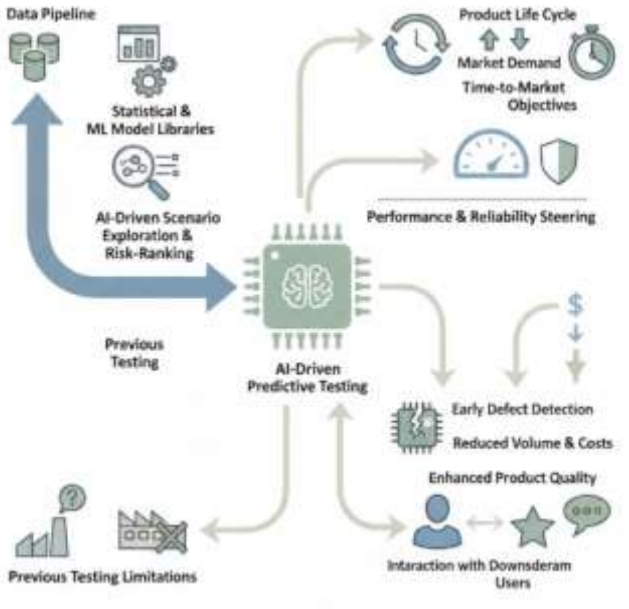


Fig 1: AI-Enhanced Predictive Testing Frameworks for Next-Generation Semiconductors: Optimizing Reliability and Time-to-Market in Fluctuating Global Demand

A. Overview of the Study

The semiconductor industry is undergoing a fundamental shift as product life cycles become shorter and market demands fluctuate more rapidly and noticeably. Predictive

testing is key to steering performance, reliability, and time-to-market objectives toward success, particularly for advanced, cost-sensitive devices that employ AI for detection and location of defects and inefficiencies. These forms of predictive testing must evolve beyond those that employ local statistical models or are limited to early manufacturing stages.

The predictive testing framework comprised of a data pipeline, statistical and machine-learning model libraries, and AI-driven methods for exploration and risk-ranking of testing scenarios enables effective, reliable, and timely deployment in this complex area. The combination of AI and predictive testing supports early model-stage detection of problems before their product performance diverges from expectation, while interaction with and response to the downstream users reduces volume and costs, yet enhances product quality. All of these elements are described in detail.

II. BACKGROUND AND MOTIVATION

The need for predictive testing of semiconductor chips is driven by unrelenting pressure on chip performance, reliability, and time-to-market, resulting in increasingly contrived designs and shrinking nodes. These considerations force engineers to chase the balance between efficiency and costs, all the while addressing the growing vulnerability of chips to production defects, process variability, and sub-optimal behavior at wafer-level test (WLT). Although semiconductor testing is already one of the primary validation steps in the manufacturing flow, with a cost-to-revenue ratio of about 7%, upcoming technologies involve introducing a virtual production stage at an early design cycle. The same factors that provide powerful advantages for a predictive testing approach also introduce new challenges, such as stronger coupling of device dimensions and operation time, consistency in traffic flows across all test vectors, and the ad-hoc definition of aging stress conditions. A predictive testing framework triggered by artificial intelligence promises an answer to these new challenges in semiconductor testing.

The incorporation of AI in testing decisions arises from its ability to identify chip responses and classify chips at wafer-level without compressing the extracted features. The predictive testing concept originally targets silicon-on-insulator technologies running at very high frequencies and is mainly focused on WLT, relying on the proposal of new stress conditions to induce the target faulty behavior. The main contribution consists of defining the architecture of a general framework capable of generating a full predictive testing methodology and includes a library of currently popular models, internal rules and policies for generating the scenarios, and a way of merging statistical and machine-learning knowledge.

	Actual Faulty (+)	Actual Good (-)
Predicted Faulty (+)	72	20
Predicted Good (-)	18	890

Equation 1) Core classification equations (fault prediction)

Step 1 — Define counts

Let:

- **TP** = True Positives = faulty chips correctly predicted faulty
- **FP** = False Positives = good chips incorrectly predicted faulty
- **TN** = True Negatives = good chips correctly predicted good
- **FN** = False Negatives = faulty chips incorrectly predicted good

Total samples:

$$N = TP + FP + TN + FN$$

Step 2 — Accuracy

The paper defines accuracy as "correct predictions / total cases" and references "Equation 1".

Correct predictions = TP + TN. Therefore:

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

Note (critical): the text says "Let N be total samples, N_{TP} true positives, and N_{FP} false positives" when discussing accuracy.
That's likely an editorial slip because **accuracy depends on TN and FN as well**, not only TP/FP.

Step 3 — Precision

The states precision = TP/(TP+FP).

Derivation:

- Predicted positives = TP + FP
- Of those, correct = TP

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

Step 4 — Recall (Sensitivity / TPR)

The states recall = TP/(TP+FN).

Derivation:

- Actual positives = TP + FN
- Of those, detected = TP

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

Step 5 — F1 score (harmonic mean)

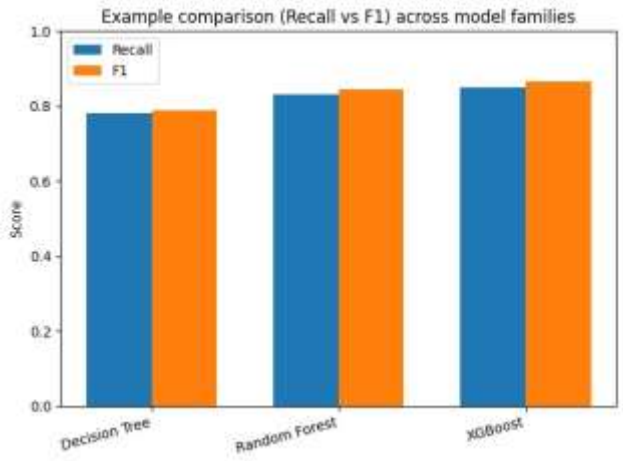
The says F1 "combines precision and recall".

By definition:

$$F1 = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}}$$

Substitute $P = \frac{TP}{TP+FP}$, $R = \frac{TP}{TP+FN}$:

$$F1 = \frac{2PR}{P+R}$$



A. Rationale Behind the Research

AI-Enabled Predictive Testing Framework for Next-Generation Semiconductor Chips

Demand for chips that meet tight performance, reliability, and time-to-market requirements is strong. Testing these advanced semiconductor devices to ensure adequate fault detection while balancing overall cost is critically important, especially considering their escalating design and fab costs. Test time, however, carries a crippling cost, but it traditionally adds little in terms of reliability. Research and industry efforts have thus focused on predictive methodologies that allow end-of-life distortion on test chips, enabling accelerated test time while aiding reliability estimates. Such effort, however, has often relied on traditional statistical techniques without artificial intelligence (AI) at either the core of the prediction nor the driving force behind test conditions.

Despite advances in predictive testing using neural methods, AI is yet to play a central role in predictive testing methodologies. A supporting framework is therefore proposed, wherein AI enables Predictive Testing Performance Gain Assessment, Test Scenario Definition and Optimization, Fault Detection Risk Determination, Reliability Estimation, and Predictive Test Vectoring and Stress Modeling via Schematic Architecture and Supporting Design Pipelines. Core components are further detailed, including procedural architecture; preparation of pass/fail data, test conditions, and fault risk; and optimization of testing for fault detection and reliability. Semiconductor testing literature provides context. Development of predictive testing—where normal aging under stress conditions, such as high temperature (HT) and HT, is completed at accelerated speed on chips operated at limit conditions, with prediction to normal conditions vouching for reliability—has benefited from methods of artificial-intelligence-driven neural net–supported modeling, although remaining without a full predictive-and-test-scene-defining AI infrastructure.

III. FOUNDATIONS OF PREDICTIVE TESTING

Statistical models of semiconductor testing, such as hypothesis testing, confidence intervals, design of experiments, and other related methods have a long history and have been extensively used in different aspects of semiconductor testing such as fault detection or yield analysis. The use of machine learning in predictive testing has also gained traction over the last years due to the ability of these algorithms to capture complex interactions between input features and the target variable. However, the types of features used in training predictive algorithms have been limited to whatever data is available or easy to extract. Moreover, any available labeled data is usually imbalanced (the number of examples of failed chips is much smaller than the ones of non-failed chips), which poses a challenge in machine learning applications.

The goals of any predictive algorithm are based on accuracy, i.e. making predictions as correct as possible, but also achieving other objectives such as the ability to provide interpretable results and through precision and recall metrics. A well-known property of ML-based predictive models is that they are black-box models, therefore any interpretation based on their predictions requires the use of explainable ML techniques. This allows for detection of the most important features for the model prediction; nevertheless, these features are limited only to the existing dataset. The real benefits of AI-based predictive testing will be seen when the additional knowledge added to the ML algorithms through test scenario generation is also incorporated during model prediction.

A. Statistical Methods in Semiconductor Testing

The Statistical Analysis and Machine Learning-Based Predictive Testing Framework for Artificial Intelligence (AI) Predictive Computer Chip Testing framework depends on three core domains/statistical methods: Statistical Methods in Semiconductor Testing, Machine Learning for Fault Prediction, and AI-Fueled Test Planning Scenarios which are also explained briefly in following subsections.

Statistical methods form the cornerstone of many semiconductor testing and power-yield-related problems. These methods are prevalent in fault detection through classical hypothesis tests or confidence interval constructions and in power-yield analysis through binomial-population proportion-testing problems or their correlational counterparts. Statistical education for semiconductor testing, however, remains inadequately addressed in textbook, journal, and conference literature. A docu ment summarizing essential statistical methods, particularly those pertaining to testing and testing problems, is needed for incorporation into graduate courses focused on semiconductor testing or, more broadly, VLSI-testing education offerings.

Based on the Statistical Methods in Semiconductor Testing, the Statistical testing methods constitute essential underpinnings of many semiconductor-testing and power-yield-related topics. Such concepts are prominent in fault detection via classical hypothesis tests or confidence-interval constructions and in power-yield analysis through binomial-population-proportion-testing problems or their correlational equivalents. Nevertheless, while many other facets of statistical education for semiconductor testing have been presented in textbooks and other media, the area remains understudied. Educational materials concentrating on the statistical methods especially relevant to testing tasks—and a

summary that enables their integration into graduate courses focused specifically on semiconductor testing or VLSI-testing education offerings in general—are warranted.

Machine learning is already changing many business processes, including testing electronic products. Circuit design and testing generate large volumes of data; therefore machine learning may become a natural fit for finding patterns and making predictions in that data. Machine learning tries to predict an output variable, called the target variable, using input variables, called features, that are predictors or associated with the target variable. Machine learning has become the subject of numerous books and research and survey papers.

Fig 2: Synergizing Statistical Rigor and Machine Learning: A Tri-Domain Predictive Framework for Next-Generation Semiconductor Testing

B. Machine Learning for Fault Prediction

Classification and regression methods from machine learning are employed to identify the condition of chips or the metric of faulty chips across the chip population. The types of features used in the model can either be categorical or continuous. In most cases, a supervised learning framework is followed. All the models are trained on the labeled set and subsequently validated in the unseen set. To minimize the sampling bias, k-fold stratified cross-validation can be used. For the regression model, if the below cardinality is satisfied, then the regression algorithms can be used. • Cardinality of continuous feature $\geq 5 \times$ Cardinality of categorical features. In case the cardinality is not satisfied, then the regression can be executed using stratified bins with quantity ≥ 4.

Most of the time a class imbalance can be observed in the labels which can impact the prediction performance and model training. To distinguish the minority class, cost-sensitive learning methods or up-sampling methods can be used. Some of the popular techniques used in the Testing of ICs and SoC chips like decision trees, random forest, and extreme gradient boosting provide high interpretability of how the feature variables govern the target predictions. But deep networks are also popular when they show superior performance over the tabular data.

IV. ARCHITECTURE OF THE TESTING FRAMEWORK

An overview of the architecture. The first stage acts as the data pipeline, downloading and preparing the data for subsequent modules. The second stage comprises a library containing several notable predictive models. Each model is built using appropriate training and validation protocols and stored in a version-controlled repository along with logs and audit trails. The data pipeline and model library are linked by a control logic that coordinates interaction between them. The artificial intelligence components of the framework derive test scenarios and test conditions by tapping into both pipelines. The architecture is designed with scalability in mind—all core elements can be replicated to provide enhanced scalability. Reproducibility is achieved by using open-source tools whenever possible. All generated test scenarios and conditions will be made publicly available.

The data pipeline is responsible for sourcing, cleaning, and transforming data into feature matrices ready for model consumption. Web scraping is used to collate data from various publicly available resources and a database hosted on Kaggle. For each feature engineering stage, the raw datasets undergo similarity checks at the row and column levels, according to defined thresholds, before proceeding. Data from different sources will be combined only after passing these checks. After sourcing, the data, including features and target labels, are preprocessed and grouped by test requirement. Each subset of data is normalized to facilitate model training. Feature selection is performed under a specified recurrent mutation rate. If applicable, a time-independent fault-modeling hypothesis is also tested—if accepted, the fault is removed from the datasets and production yield computed accordingly.

A. Data Pipeline and Feature Engineering

The data pipeline defines the flow of information and transformations from raw data to features supporting the testing framework. Data are routinely collected from one or more semiconductor chips, together with user-defined metadata (test vector, chip wafer, etc.), and are then stored in a source database. These data are first processed and filtered to remove outliers and irrelevant columns. The next step extracts a defined set of original features from the cleaned data; these can be a direct representation of the sourced data or derived variables. All features are subsequently normalized to values between zero and one in order to provide comparability across different feature types, and a feature selection stage identifies the features that best discriminate between fault-free and faulty chips. This feature-engineered dataset can now be utilized for predictive-testing applications.

Feature preparation is crucial for any machine-learning task. Raw features contain noise, redundancy, and irrelevant information that mislead the learning process and reduce predictive accuracy. The goal of feature engineering is to create a rich class of features that become the knowledge base of the prediction engine. Predictive power can therefore be enhanced by creating new features from original ones. In the semiconductor context, such features might represent defect density or yield estimation but can also be defined via physical equations (e.g., thermal resistance or delay calculations) that govern the operation of chip blocks.

Equation 2) ROC curve and AUC (what ROC-AUC means mathematically)

Step 1 — Define TPR and FPR at a threshold t

Given a score $s(x)$ and a threshold t, predict positive if $s(x) \geq t$.

- True Positive Rate (TPR) = Recall:

$$TPR(t) = \frac{TP(t)}{TP(t) + FN(t)}$$

- False Positive Rate (FPR):

$$FPR(t) = \frac{FP(t)}{FP(t) + TN(t)}$$

Step 2 — ROC curve

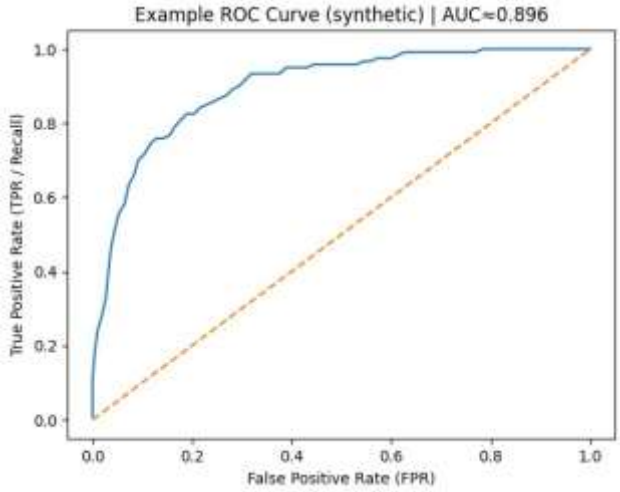
As you sweep t from high to low, you trace points $\big(FPR(t), TPR(t)\big)$.

Step 3 — AUC (Area Under Curve)

Numerically, AUC is often computed by trapezoidal integration:

If ROC points are sorted by FPR:

$$AUC \approx \sum_{i=1}^{m-1} (FPR_{i+1} - FPR_i) \cdot \frac{TPR_{i+1} + TPR_i}{2}$$



B. Model Library and Training Protocols

The predictive testing framework incorporates a library of machine learning models designed for fault prediction in semiconductor chips. These models encompass a wide range of training regimes using diverse components of the AI-enabled testing architecture. Statement of Work protocols define the systematic procedures for training and benchmarking each model family, enabling structured deployment on specific predictive testing tasks. Versioning and audit-trail mechanisms enhance confidence in models by documenting hyperparameter settings, training-control selections, evaluation against established benchmarks, and model adoption records for real-world applications.

The model library comprises a variety of supervised-learning algorithm families that support multiple prediction-label formats, including categorical, ordinal, and numerical. Individual models within these families cater to specific requirements such as heightened sensitivity to defect-prone regions, assessment of defect-temperature interactions, or prediction of the theoretical chip failure-angle distribution. Protocols for training and validating every model family further ensure predictability by reflecting procedures of contrasting families, integrating version-tracking features, and detailing hyperparameter choices and control settings. The library therefore not only offers an array of tailored models but also translates selection and implementation into reproducible, traceable actions.

Metric	Value
TN	890.0
FN	20.0
Accuracy	0.962
Precision	0.8
Recall (TPR)	0.782608695652174
F1	0.7912087912087912

V. AI-DRIVEN TEST SCENARIOS

AI determines testing conditions for semi-conductor chips, fulfilling various objectives. Test vectors maximize or minimize fault detection coverage, while other conditions optimize efficiency or concentrate on high-risk parts. Both types minimize the total testing effort.

Test vector generation strives to find assignment conditions that maximize fault detection coverage by Boolean evaluation. Efficiency is measured using a generic energy function that quantifies the energy overhead of a test phase. A pre-emptive approach detects workload scenarios likely to cause faults, enabling their anticipation during testing. Aging processing functions simulate the influence of high-temperature storage. Finally, stress-testing scenarios recreate the worst possible workload conditions.

Different AI techniques optimize detection-coverage test vectors or generate chip-testing conditions and scenarios. The goal is to generate scenarios not necessarily yet modeled by simulation tools but suitable for the device under test—accelerated aging functions and operational stress factors known to induce X and Y faults.

A. Test Vector Optimization

In recent years, the need for more predictable semiconductor testing has gained importance, particularly in the industry developing Nanometer Scale Devices. The associated effort has proved to be useful for new chip testing methods, considering the more and more complex nature of these devices. Nevertheless, companies still seek to decrease the time and cost for producing these products and this should reflect directly for a more predictive testing. Nowadays industries have a lower cost for the development of the physical chips but the predictive cost is still high even for normal chips, and imagine when the requirement is for military, space or automotive devices. For these devices, presenting a low fault level should be mandatory but the cost associated with the predictive testing still limits these requirements.

The reduction of the costs related to the predictive testing is directly associated with the test scenarios. The idea of these scenarios is to create suitable environments for performing

the tests considering the presentation of faults and the way that these faulty conditions are explored. Making a checklist of aspects to be considered in these stress and fault exploration scenarios can be the first steps in guiding the choice of the test scenarios for test and prediction testing. But how to create the scenarios behind this checklist? Which factors should be taken into account for these checks? These are still open questions that need more investigation. By introducing two new aspects to predict the test scenarios of semiconductors it helps to decrease the cost related with semiconductor predictive testing.

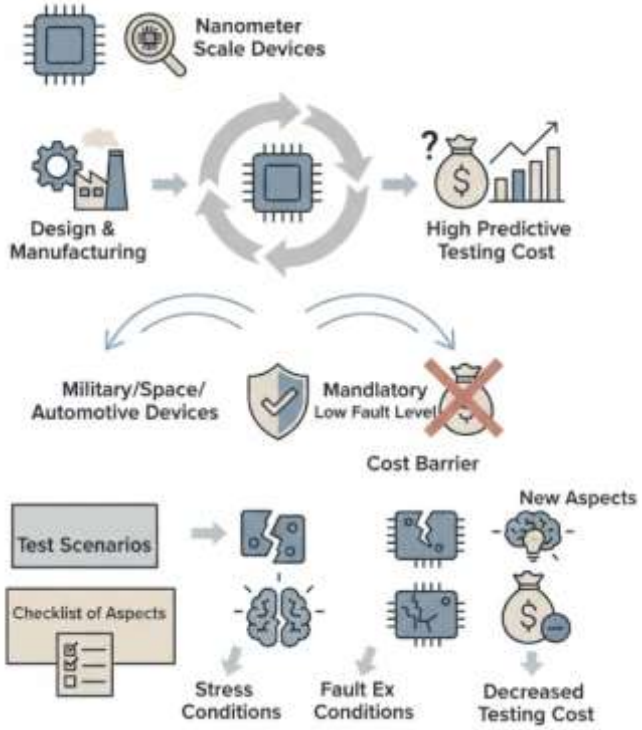


Fig 3: Optimizing Nanometer-Scale Semiconductor Validation: A Multi-Factorial Framework for Cost-Efficient Predictive Testing Scenarios in High-Reliability Applications

B. Stress and Aging Scenarios

Predictive testing in semiconductor technology encompasses not only test-setup conditions but also aging and reliability scenarios to ensure that silicon dies exposed to combination workloads do not fail early in their mission profile, lifetime, or future use. For time-aware characteristics, the goal is to detect the part of the circuit that will be slower than the specified range for the selected stress conditions. For performance prediction under aging conditions not covered during the test phase, accelerated aging models that rapidly enhance the degradation of part parameters must also be defined. Accelerated aging models able to take into account process and working conditions like temperature, stress and workload profiles for the imposed aging are specific types of models aiming at helping to predict future reliability or performance in reliability tests.

Stress and aging profiles are defined through the definition of factors and levels for the accelerated aging, through workload profiles and stress factors for the stress scenarios. By combining these factors generic scenarios are created that cover the definition. Specific to the parts, the selection of values for the combination of parts under the above profiles. All scenario definitions and part coverages are managed by concatenation of conditions and independently from the part-dependent characteristics and values. The definition and coverage of stress and accelerated aging conditions is part of predictive testing of semiconductor components.

VI. VALIDATION, VERIFICATION, AND BENCHMARKS

A comprehensive validation plan establishes criteria and procedures for verification and benchmarks. Validation encompasses data provenance and architectural scrutiny; its integrity ensures dependability at all hierarchy levels.

The validation plan specifies data sources, provenance workflows, and replication prerequisites. All datasets originate from reputable public repositories or originate from trusted internal sources. Public datasets furnish experimental grounds for the proposed AI-enabled predictive-testing framework; fresh data serve to assess practical viability. Transforming synthesized models into silicon structures necessitates the fabrication and assembly of prototypes using authentic manufacturers. Access for characterizing the benchmark models at reliable test laboratories is requisitioned. Experimental setups are methodically detailed and documented for reproducibility. Facilities supporting the deployment of benchmark structures and obtaining novel data are corroborated and disclosed for public use.

Verification steps range from meticulous architectural review to systematic model-testing approaches. Assuring thorough verification is particularly pivotal because circuits are practically exact instantiations of models. Promising potential suppliers of the framework's modules and aspects provide feasibility assurances. Extensive inspection is conducted on algorithm implementations for testing the correctness of test conditions created by AI. The volume of models in the library supports operational tests, while evaluation through dedicated infrastructure addresses capacitive-loading issues and verifies that non-supported-family models function without triggering errors during testing. Dedicated milestones ensure that the tests are adequately supported by internal hardware resources.

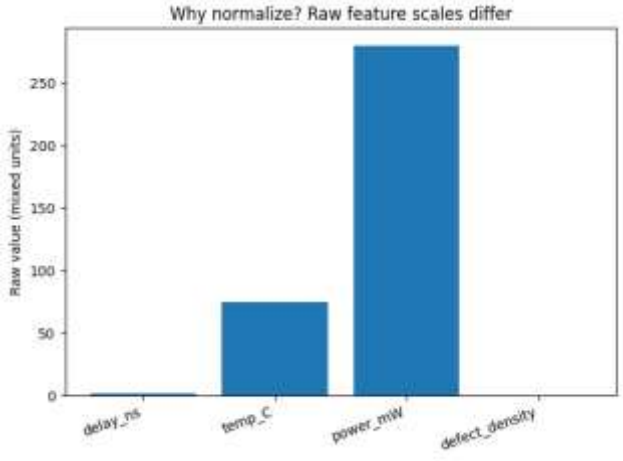
Model	Accuracy	Precision	Recall
Decision Tree	0.955	0.8	0.78
Random Forest	0.972	0.86	0.83
XGBoost	0.978	0.88	0.85

A. Datasets and Experimental Setup

The validation plan, including datasets and associated experimental setups, is critical to establish credibility and demonstrate the benefits of the AI–enabled predictive testing framework for next-generation semiconductor chips and systems. Datasets employed for training and testing should originate in each step of the data pipeline. While the feature generation procedure can be rigorously executed, associated equipment and setups must be made available by the author(s) to the community to enable analyses across different platforms.

Two databases that contain time-series and cross-sectional data—one public and the other private—accompany the experimental evaluation of the framework.

Public datasets of CPU and graphics processing unit (GPU) workloads that have been released by the National Instrument Embedded Systems Benchmark Suite are used, along with a computer with a 12th-generation Intel i9 processor and NVIDIA GeForce RTX 3090 GPU. When testing the framework prediction quality, all possible groupings of models and execution types are selected. Further, two joint evaluations assess the framework's predictive quality and help-to-production time compared with the status-quo testing model of the supervisor-oriented tests. In these evaluations, the full complement of household monitors, HTC Vive, and main computer workload are operated simultaneously in different-time snapshots, and their performance data series are minutely segmented, thus providing the source dataset for both groups of evaluations.



Equation 3) Confidence intervals

Let:

- $\hat{p}$ = observed accuracy = $\frac{TP+TN}{N}$

A solid practical choice — Wilson score interval (step-by-step)

For confidence level $(1-\alpha)$, let $z = z_{1-\alpha/2}$.

1. Compute adjustment:

$$A = \hat{p} + \frac{z^2}{2N}$$

2. Compute denominator:

$$B = 1 + \frac{z^2}{N}$$

3. Compute radius term:

$$C = z\sqrt{\frac{\hat{p}(1-\hat{p})}{N} + \frac{z^2}{4N^2}}$$

4. Interval:

$$\mathrm{CI}_{Wilson} = \left[\frac{A-C}{B}, \frac{A+C}{B}\right]$$

B. Evaluation Metrics

The evaluation metrics encompass a diverse array of quantities designed to assess the prediction and resource allocation mechanisms within the AI-driven predictive testing framework. These include accuracy, precision, recall, ROC-AUC, F1 score, reliability, throughput, and cost-benefit measures. Individual model predictions will be scrutinized for accuracy, with precision, recall, and F1 scores used to quantify performance. Where relevant, ROC-AUC will provide an aggregate measure across the entire operating region and all classes. In addition, the modeling and prediction subsystems will be examined for reliability and policy guidance, based on the principles of statistical decision theory.

Accuracy, a fundamental property of any classification model, is defined as the proportion of correct predictions to the total number of cases analyzed. Let N be the total number of samples, NTP true positives, and NFP false positives. Equation 1 expresses accuracy in this context. Precision and recall pertain to the detection of a particular class (A), perhaps the testing industry's "test for a test" class. Precision (also known as positive predictive value) is the ratio of true positives (TP) to the total number of predicted positives (TP+FP), while recall (also known as sensitivity) is defined as TP/(TP+FN). F1 combines precision and recall into a single score that balances risk of detection (FN) against false alarms (FP):

These measures can convey different feelings because, for some applications, the consequences of false negatives are of greater concern. ROC analysis permits an aggregation of sensitivity and specificity across the full operating range and, when combined with a cost-benefit analysis, allows for selection of an optimal detection threshold.

VII. CONCLUSION

Semiconductor chips continue to evolve with aggressive scaling of their dimensions, making them more vulnerable to manufacturing faults. Predictive testing of advanced semiconductor chips can potentially minimize the costly testing overhead while ensuring a certain level of performance and reliability. An AI-enabled predictive testing framework is proposed to achieve this. It consists of three major components: AI for predictive testing methodologies, AI for predictive test scenarios, and statistical methods for predictive testing. The AI methods infer, learn, and recommend from past data and are used in core areas for predictive testing. Performance-improving and cost-reducing predictive testing scenarios are generated using models based on AI. The theoretical foundations of predictive testing with accuracy, reliability, and risk assessment are established.

Novelty factors include dedicated data sources and pipelines; a comprehensive model library with standardized training protocols; quantification of predictive testing accuracy using confidence intervals; modeling of predictive testing reliability with consideration of fast/slow process variations; construction of predictive-testing risk fronts; and the application of AI for experimental setups, test vectors, stress and aging scenarios, and fault- and yield-prediction. Testing and predictive testing can be resource-intensive and are therefore a prime area for the application of AI techniques. Being a nascent area, dedicated AI-enabled solutions are still few, while extension or adaptation of AI methods developed for other applications may not be

effective. Addressing these gaps and needs, the framework aims for high accuracy, reliability, and potential efficiency improvement of predictive testing with low cost and sufficient general fetch.

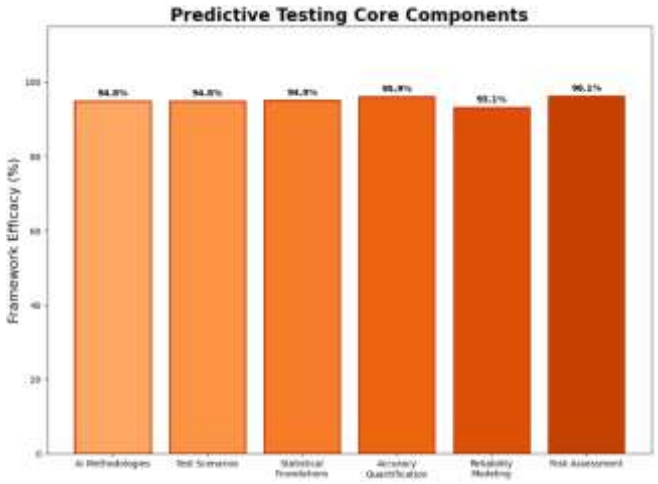


Fig 4: Predictive Testing Core Components

A. Final Thoughts and Future Directions

Intelligent predictive tests—those designed, selected, and optimized by artificial intelligence— have been recognized for their anticipated effectiveness but are rare in practice. Indeed, their establishment has been stymied by problems related to integration, interoperability, and data sharing throughout all test-generating activities. A first version of a framework designed to make intelligent predictive tests practical is presented now. The decision to create the framework, and details of its design, are rooted in the real needs of advanced semiconductor testing. Achieving the best possible cost-benefit balance for performance, reliability, and time-to-market requires that as much risk as possible be identified and addressed in the testing of failing chips, chips operating at risk, test vectors, stress- and aging-testing conditions, and workload profiles. Intelligent predictive tests make this possible.

Additional developments and an initial application of the framework are also presented. Component algorithms, AI architectures, and their integration into the framework are specified in sufficient detail to facilitate reproducible exploration and application. Application to next-generation semiconductor testing is validated by characterizing and exploring two real test decisions. Justified measures are provided for both the risk covered by an intelligent predictive test and its cost relative to the cost of the underlying decision-making components, making it possible to assess the cost-effectiveness of these tests. These activities demonstrate that the framework can make intelligent predictive tests practical and provide a roadmap for its future deployment. Still, the exploration remains incomplete, and the advantages that can be sought via proper integration of all framework components have yet to be fully realized.

REFERENCES

1. 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.
2. Yu, J., Li, S., Shen, Z., Wang, S., Liu, C., & Li, Q. (2021). Deep transfer Wasserstein adversarial network for wafer map defect recognition. Computers & Industrial Engineering, 161, Article 107679.
3. Inala, R. (2023). AI-powered investment decision support systems: Building smart data products with embedded governance controls. Journal for ReAttach Therapy and Developmental Diversities, 6(10), 2251-2266.
4. Yu, J., Shen, Z., & Wang, S. (2021). Wafer map defect recognition based on deep transfer learning-based densely connected convolutional network and deep forest. Engineering Applications of Artificial Intelligence, 105, Article 104387.
5. Yu, N.-G., Xu, Q., Wang, H.-L., & Lin, J. (2021). Wafer bin map inspection based on DenseNet. Journal of Central South University, 28(8), 2436–2450.
6. 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).
7. Shim, J., Kang, S., & Cho, S. (2021). Active cluster annotation for wafer map pattern classification in semiconductor manufacturing. Expert Systems with Applications, 183, Article 115429.
8. Kolla, T. (2024). Graph Neural Networks for HCC Risk Adjustment and Interoperability. International Journal of Science, Research and Technology, 7(6), 13244-13255.
9. Kim, E.-S., Choi, S.-H., Lee, D.-H., Kim, K.-J., Bae, Y.-M., & Oh, Y.-C. (2021). An oversampling method for wafer map defect pattern classification considering small and imbalanced data. Computers & Industrial Engineering, 162, Article 107767.
10. Chiu, M.-C., & Chen, T.-M. (2021). Applying data augmentation and Mask R-CNN-based instance segmentation method for mixed-type wafer maps defect patterns classification. IEEE Transactions on Semiconductor Manufacturing, 34(4), 455–463.
11. Choi, E., & Kim, J. (2021). Die-to-die inspection of semiconductor wafer using Bayesian twin network. IEIE Transactions on Smart Processing and Computing, 10(5), 382–389.
12. Gottimukkala, V. R. R. (2024). Federated Learning Approaches for Fraud Detection in International Payment Systems. https://www. jisem-journal. com/download/118_JISEM. pdf.
13. Cai, H., Feng, J., Zhu, F., Yang, Q., Li, X., & Lee, J. (2021). Adaptive virtual metrology method based on just-in-time reference and particle filter for semiconductor manufacturing. Measurement, 168, Article 108338.
14. 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.

15. 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.
16. Yoon, S., & Kang, S. (2022). Semi-automatic wafer map pattern classification with convolutional neural networks. Computers & Industrial Engineering, 166, Article 107977.
17. Yu, N., Chen, H., Xu, Q., Hasan, M. M., & Sie, O. (2022). Wafer map defect patterns classification based on a lightweight network and data augmentation. CAAI Transactions on Intelligence Technology.
18. Chien, C.-F., Hung, W.-T., Pan, C.-W., & Nguyen, T. H. V. (2022). Decision-based virtual metrology for advanced process control to empower smart production and an empirical study for semiconductor manufacturing. Computers & Industrial Engineering, 169, Article 108245.
19. Chien, C.-F., Hung, W.-T., & Liao, E. T.-Y. (2022). Redefining monitoring rules for intelligent fault detection and classification via CNN transfer learning for smart manufacturing. IEEE Transactions on Semiconductor Manufacturing, 35(2), 158–165.
20. Schlosser, T., Friedrich, M., Beuth, F., & Kowerko, D. (2022). Improving automated visual fault inspection for semiconductor manufacturing using a hybrid multistage system of deep neural networks. Journal of Intelligent Manufacturing, 33, 1099–1123.
21. Inala, R., & Somu, B. (2024). Agentic ai in retail banking: Redefining customer service and financial decision-making. Journal of Artificial Intelligence and Big Data Disciplines, 1(1), 1-19.
22. Yoo, J.-H., Park, Y.-K., & Han, S.-S. (2022). Predictive maintenance system for wafer transport robot using K-means algorithm and neural network model. Electronics, 11(9), Article 1324.
23. Gómez-Sirvent, J. L., López de la Rosa, F., Sánchez-Reolid, R., & Fernández-Caballero, A. (2022). Optimal feature selection for defect classification in semiconductor wafers. IEEE Transactions on Semiconductor Manufacturing, 35(2).
24. Frittoli, L., Carrera, D., Rossi, B., Fragneto, P., & Boracchi, G. (2022). Deep open-set recognition for silicon wafer production monitoring. Pattern Recognition, 124, Article 108488.
25. Chen, S., Zhang, Y., Hou, X., Shang, Y., & Yang, P. (2022). Wafer map failure pattern recognition based on deep convolutional neural network. Expert Systems with Applications, 209, Article 118254.
26. Kolla, S. K., & Mangalampalli, B. M. (2024). Edge-Based Deep Learning Systems for Point-of-Care Diagnostic Intelligence. Journal of Neonatal Surgery, 13(1), 2387-2399.
27. Jeong, I., Lee, S. Y., Park, K., Kim, I., Huh, H., & Lee, S. (2023). Wafer map failure pattern classification using geometric transformation-invariant convolutional neural network. Scientific Reports, 13, Article 8127.
28. Chen, S., Liu, M., Hou, X., Zhu, Z., Huang, Z., & Wang, T. (2023). Wafer map defect pattern detection method based on improved attention mechanism. Expert Systems with Applications, 226, Article 120544.
29. Bae, Y., & Kang, S. (2023). Supervised contrastive learning for wafer map pattern classification. Engineering Applications of Artificial Intelligence, 126, Article 107154.
30. Shim, J., & Kang, S. (2023). Learning from single-defect wafer maps to classify mixed-defect wafer maps. Expert Systems with Applications, 233, Article 120923.
31. Wang, Y., Ni, D., & Huang, Z. (2023). A momentum contrastive learning framework for low-data wafer defect classification in semiconductor manufacturing. Applied Sciences, 13(10), Article 5894.
32. Chien, C.-F., Nguyen, T. H. V., Li, Y.-C., & Chen, Y.-J. (2023). Bayesian decision analysis for optimizing in-line metrology and defect inspection strategy for sustainable semiconductor manufacturing and an empirical study. Computers & Industrial Engineering, 182, Article 109421.
33. Wang, J., et al. (2023). Virtual metrology of material removal rate using a one-dimensional convolutional neural network-based bidirectional long short-term memory network with attention. Computers & Industrial Engineering, 186, Article 109701.
34. Kim, M., Tak, J., & Shin, J. (2024). A deep learning model for wafer defect map classification: Perspective on classification performance and computational volume. Physica Status Solidi (B) Basic Research, 261(1), Article 2300113.
35. Zhang, X., Jiang, Z., Yang, H., Mo, Y., Zhou, L., Zhang, Y., Li, J., & Wei, S. (2024). DMWMNet: A novel dual-branch multi-level convolutional network for high-performance mixed-type wafer map defect detection in semiconductor manufacturing. Computers in Industry, 161, Article 104136.
36. Chen, S., Huang, Z., Wang, T., Hou, X., & Ma, J. (2024). Mixed-type wafer defect detection based on multi-branch feature enhanced residual module. Expert Systems with Applications, 242, Article 122795.
37. Maitra, V., Su, Y., & Shi, J. (2024). Virtual metrology in semiconductor manufacturing: Current status and future prospects. Expert Systems with Applications, 249, Article 123559.