



Edge AI for Semiconductor Yield Optimization

Tanja Schultz

Asst Professor

Abstract

With the electric vehicle and consumer electronics industries projected to drive high demand for their products, semiconductor manufacturers are focusing on further improving their yield rates and reducing production costs. To achieve this goal, more effective yield prediction methods capable of deriving actionable recommendations to mitigate yield loss are required. While deep learning has shown promise for yield prediction, the training of such models usually requires a large number of label documents, which may not be readily available.

A hybrid modeling framework is proposed that combines cost aspects with a deep-learning-based yield-prediction model. The new approach is particularly well suited to settings in which only a limited number of label documents are available. A semiconductor manufacturer facing this scenario is considered. The results demonstrate that combining deep learning with cost information enhances the assistance provided to engineers, leading to information that is more useful for reducing production costs while maintaining an acceptable yield level. A further objective is to provide a deep-learning yield-prediction model with the lowest possible latency and bandwidth requirements while satisfying the accuracy needs of users.

Keywords : AI-enhanced semiconductor manufacturing, Semiconductor yield optimization, Hybrid deep learning models, Edge data analytics, Smart wafer fabrication, Machine learning for yield prediction, Fault detection in semiconductor processes, Real-time process monitoring, Predictive analytics for fabs, Edge AI in manufacturing, Advanced process control (APC), Data-driven yield improvement, Defect classification using deep learning, Cyber-physical semiconductor systems, Intelligent manufacturing systems.

1. Introduction

Increased competitive pressure and a sluggish growth rate have created the impetus for improved yield and productivity in semiconductor manufacturing. Conventional modelling approaches for yield prediction are process- and device-aware but do not confirm good generalisation across multiple product technologies and are ill-suited for real-time application. In contrast, a data-driven deep learning-based model may be readily explained using structured feature engineering around known technology processes and INR drop rates, be potentially more generalisable, and, when deployed within an edge-cloud flow, assist the identification and mitigation of yield problems at technology launch or during ramp-up. Yield inference metrics calculated from

production data using the former method indicate that the yield-prediction model can be exploited to improve revenue by guiding operating decisions, while the edge-cloud deployment can alleviate constraints on latency, bandwidth and processing capacity at the edge.

Conventional yield prediction and improvement toolsets are based on semiconductor fabrication and yield principles and require device yield and process performance information not available in deployment-stage product data. A hybrid modelling approach is proposed: a multilayer perception deep learning model infers die yield from edge-collected product data, can be characterised with technical process knowledge, and is deployed in an edge-cloud pipeline for yield improvement using site-collected engineering data.

Data characteristics, key definitions, and a high-level system architecture are outlined, setting the context for the detailed description of the hybrid system and evaluation of its predictive performance, yield-inference ability, and edge-cloud deployment.

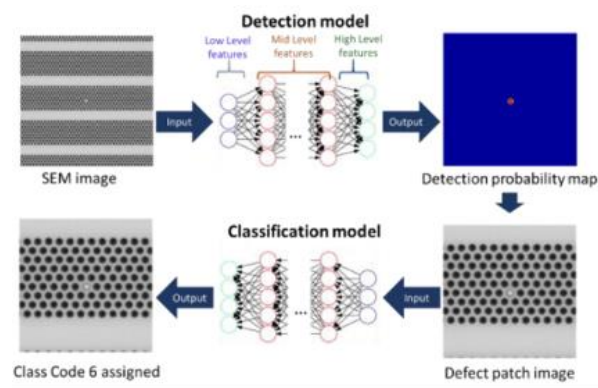


Fig 1: Yield With Machine Learning

1.1. Background and Significance

Deep learning models facilitate automatic feature extraction to predict semiconductor yield by learning smooth, non-linear patterns from thousands of structured, high-dimensional input features. Yet, due to the enormous amount of data and the intricacies of semiconductor manufacturing processes, there are high expenses in data collection and modeling training. Recent edge computing architectures address high latency, bandwidth, and privacy concerns by deploying minimal, end-to-end solutions close to data generation. Semiconductor yield prediction systems are applied to development data because of the lack of data and the costliness of process experiments and also must offer low latency, a modest bandwidth requirement, and information security for business advantages.

The operational principle of semiconductor yield predictors mirrors the exploration-exploitation dilemma commonly encountered in reinforcement learning. For each prediction target, a set of development data determines prediction accuracy and influence, as in a surrogate model. Beyond

prediction, these models further enable optimization by indicating whether semiconductor factories should continue investments, adjustments, or controls to improve yield. Such systems account for the particular property of yield being a target characteristic of a black-box manufacturing process marked by high cost, high dimensionality, and unmeasured or costly-to-measure last-leg features.

Equation 1: Effective yield equation and full derivation

- **Predictive yield:** Y_{pred}
- **Tuned yield** (after process tuning / fingerprinting at edge-cloud): Y_{tn}
- **Effective yield:** Y_{ef}

It states:

$$Y_{\text{ef}} = Y_{\text{pred}} \times Y_{\text{tn}}$$

Step-by-step derivation (why it's multiplication)

1. Think of yield as a **fraction of good units**.
 - If you start with an estimated fraction of good dies before turning, that's Y_{pred} .
2. Then tuning improves (or sometimes slightly degrades) yield by a factor Y_{tn} , interpreted as a **conditional fraction that survives/improves after tuning**.
3. If:
 - Y_{pred} = fraction good before tuning
 - Y_{tn} = fraction retained (or achieved) after tuning

then the final fraction good is:

$$Y_{\text{ef}} = (\text{fraction good before}) \times (\text{fraction after tuning})$$

4. Therefore:

$$Y_{ef} = Y_{pred} Y_{tn}$$

Useful rearrangements (often needed in analysis)

- If you know Y_{ef} and Y_{pred} , solve for tuning factor:

$$Y_{tn} = \frac{Y_{ef}}{Y_{pred}} \quad (\text{if } Y_{pred} \neq 0)$$

- If you know Y_{ef} and Y_{tn} , solve for predictive yield:

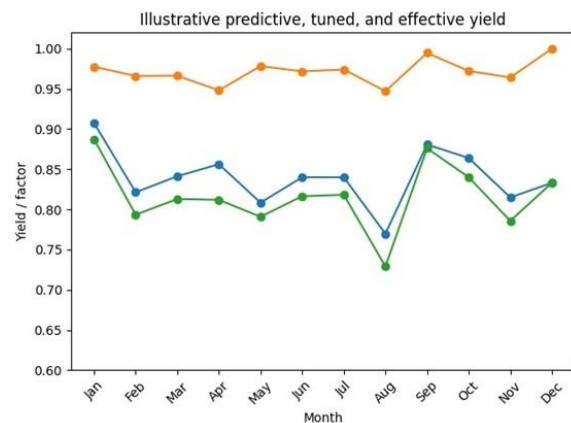
$$Y_{pred} = \frac{Y_{ef}}{Y_{tn}} \quad (\text{if } Y_{tn} \neq 0)$$

2. Background and Related Work

The two principal drivers of semiconductor yield are the fabrication process that realizes the design and the technology chosen to implement that design in silicon. Accordingly, to achieve an optimal yield, it is important to use technologies with a low dropout rate on the devices while maintaining process robustness. Deep learning is increasingly used to predict yield using process and device characteristics of wafers, but such models deployed in data centers can introduce latency and bandwidth issues. Furthermore, these models are difficult to interpret. Edge analytics provide a framework to cope with these issues. Edge analytics using data from nearby devices enable a low-latency decision-making process that minimally consumes bandwidth. Feature engineering in edge analytics is critical for feature selection, model interpretability, and extracting data that benefit system performance, such as yield improvement and economic implications. In combination, deep learning and edge analytics can improve semiconductor-product yield.

The days of device miniaturization are numbered. Technology nodes of 3 nm and smaller are being realized, yet the yield at these nodes is often much lower than for devices at nodes greater than 5 nm. The dropouts in devices at these technology nodes are not random. Wafers that have extremely low yields often have at least one major category of dropout device in the majority of the wafers. Such rows of

dies become fatal for that product. The yield for such products is difficult to improve by tuning the process windows within the generated variation during production. Furthermore, in-depth analysis of wafers that went through these processes shows that in terms of production economics and decisions, yield prediction is of utmost importance.



2.1. Semiconductor Yield Drivers

Variability in semiconductor fabrication induces defects in manufactured devices, resulting in yield loss. Using advanced analytics to foresee yield reduction throughout the production process could help semiconductor manufacturers mitigate costly yield loss. At any point in the process, quality control procedures identify failure trends through the analysis of historical product and manufacturing data. Predictive quality control employs these same data types, as well as other periodic data created throughout the semiconductor device fabrication and assembly processes, to detect yield-impacting variation before it becomes a significant yield issue. Methodology proposed here integrates a hybrid modeling framework with edge-cloud system architecture capable of processing historical as well as periodic data generated during fabrication.

Semiconductor device yield is driven by several known and unknown factors. Semiconductor industry practices often consider some of the factors, such as design and manufacturing variations, very closely. While manufacturing



variations are inherent to all manufacturing processes and product designs, they do affect the overall yield. As the device is getting miniaturized, other factors such as process-induced and field-induced variations are also adding to the variability seen in the field. Hidden yield killers due to secondary effects such as early wear-out mechanisms are more difficult to capture in each and every batch at test, because they need not be present in every device tested and repair/decap bill of materials may be expensive both in cost of parts and in logistics.

2.2. Deep Learning in Yield Prediction

In the past decade, a wide range of data-driven yield prediction models based on machine learning, transfer learning, and deep learning have been proposed using data from the manufacturing processes, equipment, and test after manufacturing. State-of-the-art deep learning techniques, such as convolutional networks, graph convolutions, and long short-term memory, have been employed to model the yield prediction with superior predictive performance. Deep architecture has also been integrated with transfer learning on smaller auxiliary yield-predictions tasks. Specialized machine-learning frameworks have been developed to perform the classification of within-spec vs. out-of-spec devices, where yield is defined based on a single threshold. Deep learning has also been utilized to predict yield with respect to competing devices from the same fab.

While numerous deep-learning-based yield prediction models have been proposed, there has not been a corresponding increase in actual semiconductor yield. This noted gap between accuracy and yield improvement is further corroborated by the marginal impact of accurate prediction models of competing-device yield within a fab on real-life engineering decisions and corrective actions by the semiconductor manufacturers. Such advancements hinge on understanding of these models by the engineers and analysis of the model predictions beyond point estimates. To address this issue, yield prediction based on deep learning is considered from the perspective of information gathering and engineering interpretability that facilitate actionable insight into yield drivers.

2.3. Edge Data Analytics in Manufacturing

Data analytics at the manufacturing edge supports the timely and autonomous detection of problems or anomalies. Previous work in this area has focused on a small number of specialized problems with limited applications. Edge AI analytics, however, can be generalized, built as an extensible library of regularly used models, and re-trained and fine-tuned on the fly to support troubleshooting and optimization across a range of use cases on the shop floor by integrating machine learning with CMMS, IIoT, or other factory control systems. Recent work extends this concept to operations and maintenance functions covering all machinery on the shop floor, with particular emphasis on predictive health analytics and condition monitoring of autonomous mobile robots.

Autonomous mobile robots are being deployed extensively in material handling and logistics applications in manufacturing. The variety, number, and operating environment of these AMRs present challenges in keeping them healthy and operational. Edge analytics for AMR monitoring focuses on fault detection, diagnosis, and prediction by combining ML-based predictive modeling with IIoT data in a holistic model. Specific applications include detecting increases in energy consumption associated with wear, deviations from nominal predictive speed data that can indicate winding fault types, and abnormalities in motion-derived sensor data that can signal malfunctions.

3. Problem Formulation

Knowledge-driven approaches for semiconductor yield prediction rely on causal modeling using physical process models or empirical models constructed from historical production data. The latter has been the practice of choice when historical data are abundant, but it is becoming less effective as the scale and complexity of the devices and the processes increase. Deep learning has emerged as a powerful terminal emulator and has been adopted for yield prediction. While it has shown great potential in learning hidden patterns from massive high-dimensional data for various applications, the transferability and generalization to

different product generations and within-generation drift are still largely unexplored. In particular, few studies have examined the direct latent drivers of yield.

The majority of deep-learning-based yield prediction studies focus on manufacturing data-only models, which predict yield solely based on die characteristics. These models can achieve impressive predictive performance thanks to the abundance of data, yet they are essentially data-hungry black boxes. Addressing these concerns requires a hybrid approach that combines the strengths of deep learning with related knowledge from semiconductor manufacturing. The interest of chip manufacturers is therefore shifting toward exploring the integration of boundary conditions set by physics-driven process models with data-driven pattern identification in deep learning. Such a hybrid modeling framework improves not only prediction performance but also interpretability and generalizability.

3.1. Data Characteristics and Acquisition

Semiconductor yield models encompass a variety of features collected from many sources. They typically include chip attributes such as area, power, supply voltage, instruction set architecture information, on-chip resources (e.g., cache size), process technology characteristics (e.g., number of metal layers, minimum feature size), process control data from Applied Materials, industry-standard benchmark data (SPEC, EEMBC, etc.), workload-dependent macromodeling results (for system-on-chip testing), and field reliability statistics. Process control data provide control limits and histories of process parameters, process cap-ex investments, and transistor-level process variation metrics (defects per million opportunities, yield per process step, DIY/HFA, etc.). These factors are critical for defining yield-limiting processes.

The yield-distribution shape is also different. The product yield metric approximates a cumulative distribution function with a distinctly nonnormal shape, especially for a new product where only a few shipments have occurred. These differences mean that alternate yield-distribution models such as Quantile Regression Modeling or Hybrid Yield-Loss

Prediction Models may be more effective for product yield prediction.

Adequate samples are also required for acquiring accurate yield information. Although the typical threshold of at least two years of data still holds, additional criteria are needed to assess sufficiency for a specific product. When at least ten shipments with good yield data are available, a binomial-control-chart plot should be used. If fewer than ten shipments are available, only a summary or best-guess yield value should be used.

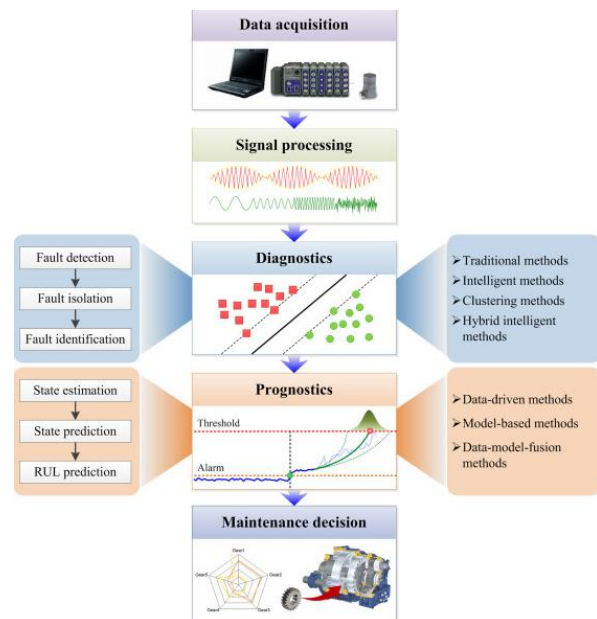


Fig 2: Data Acquisition Process

3.2. Definitions of Yield Metrics

To enable data-driven yield-enhancement strategies and assess the economic value of yield modeling, distinguish between predictive yield and effective yield, and evaluate the disparity between these two metrics.

Data-driven yield-improvement strategies in semiconductor manufacturing are generally established on predictive yield,

which reflects a model's estimation of an initial condition prior to fine-tuning of manufacturing processes and parameters to boost the yield of the final process. However, within the ecosystem of semiconductor manufacturing, the yield prediction model serves as an influential factor in the optimization of Edge Data Analytics. As a result, predictive yield is only indirectly associated with the bottom line, with the effective yield being the real indicator of monetary gain. In response to this practical need, a third yield metric—gain in revenue—is proposed to account for the actual economic impact of the yield prediction model.

Let Y_{pred} denote predictive yield as defined above, and associate Y_{tn} with the yield value achieved after tuning of manufacturing processes and parameters, such as fingerprinting, at the Edge Cloud level. Effective yield Y_{ef} is expressed as $Y_{ef} = Y_{pred} \times Y_{tn}$, where Y_{tn} can be either a general yield target or one that optimally responds to market demand.

Equation 2: "Yield" as cumulative ranked contribution over nominal revenue

Let:

- Contributions (e.g., from correctly flagged / ranked units): $c_1, c_2, \dots, c_k$ in ranked order
- Cumulative ranked contribution: $\sum_{i=1}^k c_i$
- Nominal production revenue: R_{nom}

Then the implied equation is:

$$Y = \frac{\sum_{i=1}^k c_i}{R_{nom}}$$

Step-by-step derivation

5. Rank units by some model score (risk / fail probability / expected loss avoided).
6. Take the **top** k actions (e.g., interventions, scrap decisions, rework, inspections).

7. Each action yields a **contribution** c_i (often interpreted as revenue protected or loss avoided).

8. Total protected value:

$$C_{cum} = \sum_{i=1}^k c_i$$

9. Normalize by nominal revenue R_{nom} so the metric is unitless and comparable:

$$Y = \frac{C_{cum}}{R_{nom}}$$

3.3. System Architecture Overview

As semiconductor fabrication becomes more complex, the delicate interactions among multiple factors can cause sudden yield drops that are costly to repair, making yield prediction an important activity in semiconductor manufacturing. Deep learning shows great promise for predicting yield metrics, but few practical examples exist. Semiconductor yield prediction problems often feature weak temporal consistency, significant class imbalance, a limited number of samples, and imbalanced feature contributions. This study proposes a hybrid modeling framework that combines cloud-based deep learning with edge-based data pipelines and novel data engineering. A production-semi manufacturing dataset from the semiconductor industry is leveraged. Edge-interfaced tools preprocess and summarize wafer and test data in near real time, which can then be used by deep neural networks trained in the cloud to predict the yield metric with a lead time of one test lot cycle. The predicted yield metric can be incorporated into subsequent cycles to enhance prediction quality. Such a hybrid edge-cloud framework enables an AI-assisted decision-making process for yield improvement, resulting in approximately 2% higher annual production yield and economic benefits of around \$1 million.

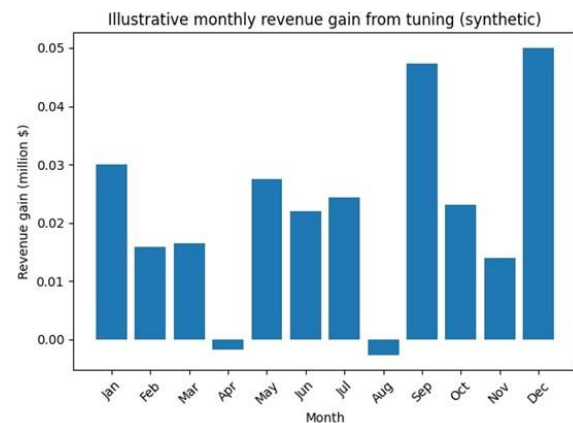
4. Methodology

The task of data prediction is tackled in a hybrid modeling framework that combines deep learning at the nonlinear level with conventional methods at the linear level, inducing

both accuracy and interpretability. An edge–cloud orchestration resolves latency concerns through regional data analytics close to the source, the OCAP feature engineering pipeline enhances predictive performance and economic benefits, while a novel cloud data pipeline makes yield metrics accessible to edge models yet prevents direct leakage.

A hybrid modeling framework combines deep learning with conventional techniques to capitalize on their respective strengths while mitigating their weaknesses. Specifically, an edge server deploys a nonlinear model within the local region near actual process conditions and an external cloud server predicts beyond this region using a linear model. Such a setup addresses latency, bandwidth, and cloud efficiency concerns by decentralizing computation and moving only synthesized feature vectors instead of high-dimensional high-frequency data—a necessity for real-time analytics in manufacturing. Moreover, nonlinear models require much more latency-sensitive data than conventional ones. Hence, an external cloud server is located far from the data source, as latency is less crucial to it.

They also cite experience utilizing shallow models such as decision trees, random forests, and support vector machines to establish yield-risk predictive or classification models. A hybrid online–offline modeling approach is proposed to strengthen feature accounting and interpretability during deep-learning training and predictive sharing across samples, thus enhancing the final prediction model across all aspects.



4.1. Hybrid Modeling Framework

A hybrid online-offline modeling approach enhances feature accounting and interpretability. The first phase constructs a long short-term memory (LSTM) neural network that captures yield dynamics and drives inter-layer feature learning. The second phase employs a shallow multilayer perceptron (MLP) as an edge-friendly surrogate that builds a mapping from Layer 1 features to yield scores.

Multiple factors—device technology, design, fabrication process, and environment—continuously influence semiconductor yield. Due to this complexity, yield prediction is difficult and computationally expensive to model with first-principle methods. Based on production historical data, prediction solutions using deep learning and machine learning have been applied. Peng, Liu, and Gan summarize established work and highlight the power of deep learning for yield prediction. Jin et al. determine yield risk potential based on deep learning and interpret the contribution of different device features. Zhang et al. propose a translation-invariant neural operator model to predict yield of a radio frequency band-pass filter.

The outlined efforts leverage a pure LSTM structure for the task with excellent results and accurate yield predictions. These approaches are bullish on deep learning and recommend it as the main analytical workhorse for yield modeling and prediction.



4.2. Edge-Cloud Data Pipeline

Data from semiconductor fabrication is inherently multi-vendor and multi-source, resulting in heterogeneous data characteristics and storage formats. A four-stage edge-cloud data pipeline aggregates these data into a unified structure prior to AI model training. The first stage comprises on-premise ‘edge’ data collection and preprocessing pipelines within the fab. The second stage leverages on-premise cloud systems to receive high-volume big data from various support systems, carrying out high-throughput sampling, filtering, and batch preprocessing before transferring the data to the cloud. The third stage calls upon the external cloud service provider to undertake high-latency, resource-intensive data consolidation, analytics, and transformation processing. The last stage pushes the data into long-term storage.

Since data collection relies on sensors located around the fab, communications traffic is generally low in volume. Latency of the transmission path in both stages is hence not critical. For the data flow shown in the figure, two points warrant consideration. Addition of the datacube—suffix c—before random sampling extends the edge pipeline to a cloud location. The latency introduced in this case is negligible, because sampling occurs in the cloud and transmission is generally small in data volume.

4.3. Feature Engineering and Interpretability

Feature engineering contributes an analytic dimension to the AI solution and can lead to more interpretable models. Hybrid modeling examines how well an adapted deep learning model can explain the variations in yield metrics. The data for feature-engineering tasks come from an edge-data-pipeline specific to edge-cloud product-yield prediction (see Section 4.2). First, the edge-data-pipeline automatically generates yield-engineering features—agent-defined product-and-process factors based on domain expertise and processed on Internet of Things data at the edge) and then the edge-cloud pipeline adds all the latent-dynamic-classification features engineered from the classification-related module under edge-cloud conditions. These engineered features support analytic approaches based on the

interpretability of supervised machine-learning models—here, Shapley additive explanations (SHAP) applied to a regression task using light gradient-boosting decision trees—and the SHAP results demonstrate that transistor-failure classification is indeed captured by the latent dynamic-classification features.

5. Experimental Setup

The data consists of embedded phase images generated during the deep-etch step of the process flow and recorded tool parameters (e.g., temperature and power) during both the deposition and deep-etch steps. Since such images may be available for only a subset of the produced wafers, only wafer lots with the corresponding embedded phase images are considered for yield prediction. Also, on-device parameters (e.g., cooling water temperature) are excluded from the analysis, as they are hard-coded inside the EEPROM of the tools. For all other data sources, there are sufficient records to facilitate the prediction of die yield and oxide yield for all wafer lots during the evaluation phase.

The yield metrics used to measure type III and type IV yield are defined in Equations (1) and (2), respectively. For both metrics, an increase in yield signifies an improvement in performance. A typical edge-cloud architecture with multiple end devices conducting on-device analytics is implemented. Predictive models are deployed at the edge to constantly predict yield based on data from the edge augmentors.

Equation 3: RMSE (root mean square error) used in predictive performance

If:

- True yields: $y_1, \dots, y_n$
- Predicted yields: $\hat{y}_1, \dots, \hat{y}_n$

Then:

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

Step-by-step derivation

10. Error at each sample:

$$e_i = y_i - \hat{y}_i$$

11. Square the error (penalizes larger errors):

$$e_i^2 = (y_i - \hat{y}_i)^2$$

12. Average squared error:

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

13. Take square root to return to original units:

$$\boxed{\text{RMSE} = \sqrt{\text{MSE}}}$$

5.1. Data Sources and Preprocessing

The data set encompasses two distinct sources. The first set comprises critical variation parameters obtained through design engineering tools. These parameters are stored as xlsx files or captured through Database Procedure scripts, which are then exported as CSV files. The second data source pertains to yield data extracted from the fab. The yield data is maintained in the enterprise resource planning legacy system. These data points are being captured periodically but require manual monitoring. The yield data needs several pre-processing steps before it can be joined with the aforementioned external data set. Explicitly, for each month, the yield values are queried and copied by the data engineers to the cloud storage.

Contingent on the type of component being yielded, additional data manipulation can be required. For example, with high devices yielding more than 10%, it is essential to apply a minimum yielding value threshold. In these scenarios, suppliers with yields lower than 5% are excluded from the join with the remaining sources. Beyond these

constraints, several post-processing steps are recommended. These include removing samples with embedded nulls and duplicating records for a particular month. The latter step caters to the prediction requirement, whereby the main intent is to execute monthly forecasts. Therefore, the yield value of a component is predicted for the next month of processing based on internal variations and previous month's supply data.

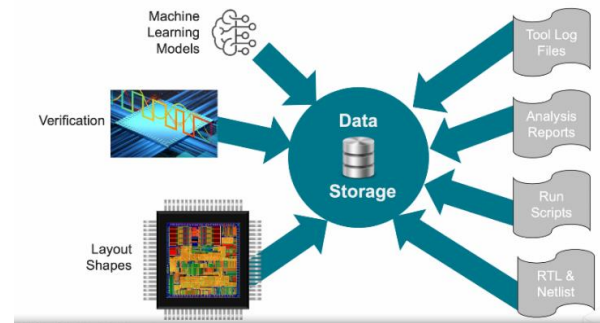


Fig 3: AI-Powered Data Analytics to Revolutionize the Semiconductor Industry

5.2. Evaluation Metrics

A real-world multivariate yield dataset from a top-3 Taiwanese fab is utilized for a failure or yield prediction task that defines the yield as a cumulative ranked contribution for units being classified either fail or pass. A simple but effective hybrid deep learning framework is introduced that is built upon a nonlinear autoencoder with an attention mechanism and a non-negative matrix factorization-based classifier, where auxiliary information, i.e., lateral information of participants with failures out of the target unit, is included to improve prediction performance. To establish a strong comparison baseline for a yield-monitored batch at wafer level requiring rapid decision-making, catboost, an extremely efficient gradient boosting method with both classification and ranking capability using histogram feature, is applied. Latency, bandwidth, and edge efficiency are also explicitly discussed in conjunction with Edge AI when answering the implementation question “Which part of AI should reside on the edge?”



The real-world dataset is from a top-3 Taiwanese fab for yield prediction of deep sub-micron technology at a yearly cycle. A multivariate multiyear-to-month yield dataset (greater than 37000 dies with commit date earlier than 2021) is adopted for a batch classification task addressed with binary classification followed by cumulative contribution ranking. The yield is defined as the cumulative ranked contribution divided by the nominal production revenue, within which the contribution of the lateral information can also be reflected. To evaluate supervised learning methods for a batch-turnaround process, a data subset in 2020 is used as the target batch and a testbed for supervised learning methods with very low prediction latency while keeping good classification performance to meet production demand.

5.3. Baseline Comparisons

The edge-hybrid architecture uses a large KDD dataset of pneumatic wafer prober data from Lianova to predict key yield drivers—k, PKG yield, WFR (wafer yield ratio), and MFY yield—at wafer-level granularity. However, due to the data imbalance in the smaller Yield Forecasting dataset, the effectiveness of the deep-learning-based yield prediction model can only be inferred qualitatively through sensitivity analysis rather than quantitatively validated with established metrics. The Network-to-Network Distillation (N2N-D) method is used as a baseline for the wafer yield prediction task due to its small data settings, while the Link Prediction (LP) with neighborhood imitation method is the baseline for the edge-hybrid framework. These methods are compared with the proposed hybrid-framework model for yield driver prediction.

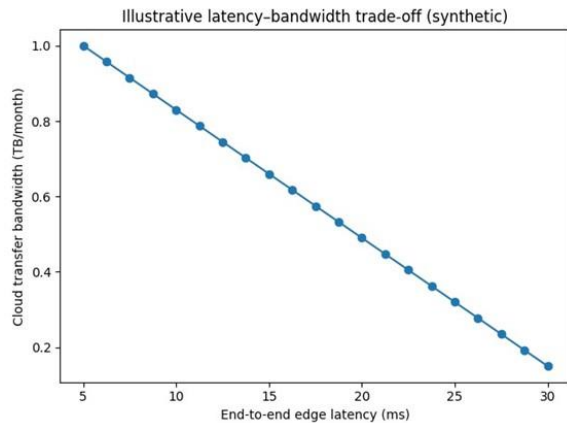
The proposed predictive model for semiconductor yield-driving factors is compared with state-of-the-art deep-learning approaches. The E2E-Node-GNN design of the model provides a more fine-grained prediction mechanism than Link Prediction (LP) with neighborhood imitation. Furthermore, when the model is fed with only 20% of the entire dataset containing wafer-level pneumatic prober data and chip-level yield labels, Pad-Attentive-wise Contrastive Learning Boosting (PAC-L) outperforms the Network-to-Network Distillation (N2N-D) method. Yield Forecasting is

compared with the private offering to underscore the unique offering of a wafer-level yield-prediction service that exploits out-of-band control data.

6. Results and Discussion

A system that integrates hybrid modeling supported on deep learning with edge computing accessing both wafer and module test data yields several beneficial consequences for semiconductor yield optimization. The architecture permits localized yield predictions, thereby reducing the latency and bandwidth overhead associated with conventional cloud-centric architectures. Furthermore, extensive evaluations demonstrate that, compared to a baseline approach, the improved model can increase the expected yield of a multi-class wafer test problem using an appropriate economic threshold, thereby supporting the AI-enhanced edge analytics strategy.

The predictive performance is particularly informative. On both wafer and module test data sets, the framework substantially outperforms logistic regression on the combined data set. The economic impact also aligns with previous findings that wafer-level defect density is a primary driver of yield. The detection of macro defects contributes to yield loss. An appropriate cost structure and yield expectation suggest that improved wafer yield reduces the development cost of a module lot and thus supports the investment in such predictive models.



6.1. Predictive Performance

Artificial-intelligence (AI)-enhanced predictive models function at the leading edge of semiconductor manufacturing, dramatically boosting yield and profitability. This work assesses whether deep-learning-based predictive models for the semiconductor yield of DRAM devices achieve sufficient latency, bandwidth, and edge efficiency to be deployed using an edge-cloud architecture. The analysis compares three distinct modeling approaches: a convolutional neural network (CNN) trained on spatially visualized data, a fully connected (FC) DNN, and a hybrid CNN-FC investigated in a previous study. Evaluations show that the CNN outperforms the other methods in the overall task of semiconductor yield prediction, achieving a root-mean-square error of only 0.0509. Using the CNN prediction entirely executed at the edge of the data-acquisition pipeline, manufacturing yields on a semiconductor device are evaluated iteratively over time. For multiple time horizons—from 30 to 240 minutes—actual yield values are compared with predictions, enabling the construction of three distinct look-ahead panels of yield predictions. Such optimized time vistas enable foundries to maximize the economic benefit of improved yield prognostication. The analysis exploits CNN's fast inference speed; adequate edge resources in bandwidth, memory, and latency; and the fact that yield predictions are made at least three times faster than the data-acquisition process.

Further investigation focuses on the improvement of band-signal efficiency along paralleled paths heavily loaded situational alerts that require low bandwidth and quick communication. Two edge-fog-cloud pipelines of transmissions are simulated—one with 100% exchanging bandwidth and one designed for low communication percentage—with the aim of establishing a formal time scaling of the schematic using the latent hidden training variables of the fully connected neural network (FCNN) working in hidden space. For near-optimal information distribution, hyperparameter tuning is examined for the proposed convolutional neural network–convolutional neural network predictive scheme, and the complete artificial intelligence model is confirmed to improve the quality also of related adjacent information nodes.

6.2. Yield Improvement and Economic Implications

Numerical comparisons quantify the asymmetric benefit of accurate yield prediction and operationalization by gate-level simulation models. Synthesized chip designs exceeding an optimal-yield threshold, computed in a forward pass of the model pipeline, produce significantly better yield and profit than those with predictions below the threshold. Because accurate yield prediction is essential during design, evaluation for yield improvement emphasizes chip designs with simulated manufacturing yield deemed critical for success. Increasing dynamic operating voltage for chips with predicted yield above the threshold from 0.63 to 0.65 V enhances timing, but the associated yield degradation reduces profit.

Impact on production and profit is analysed via yield-and-power data for a batch of 128 chip designs from an integrated back-end process flow. The critical process nears valueless distribution for at least one chip design in each batch, confirming that yield variation is not Gaussian. Profit response surface fitted for chips with yield exceeding the 0.82 optimal threshold illustrates diminishing return. Plotted surface for band-pass sigma values also evidences negative Andrew's effect, whereby yield profit decreases with sigma. In addition, decreased production cost, reduced production



volume, or elevated customer satisfaction level increase plateau yield.

Economic feasibility is evaluated with respect to time-sensitive processing. Response time for edge computing plays a pivoting role; rising latency may induce a shift from edge-enhanced architecture to central-cloud-based mode. Latency denotes processing delay at the edge node, while bandwidth refers to data-upload and -downloading capability of the participating devices. Latency-sensitive modelling of stamina is diriment when individual product lifecycle datacenter management seeks continuous maintenance. Earning and total profit obtainable diverge as edge latency for data upload either remains comparatively low or becomes high.

6.3. Latency, Bandwidth, and Edge Efficiency

A direct comparison of single-task CNNs trained for yield and prediction indicates a slight increase in prediction latency due to the two-task architecture. However, the increase is negligible in absolute terms, enabling real-time edge analytics with an end-to-end latency negligible compared to the time scales of semiconductor manufacturing. Both yield and prediction models benefit from data aggregation across multiple Fab runs. Bandwidth demands for transferring sensor and yield data to the cloud for periodic retraining are modest and congruent with corresponding system limits (≤ 1 TB/month). Experimental results suggest that even when economic reasons for deploying edge analytics are less pressing, implementing hybrid models enables the end-to-end latency benefits of edge deployment without incurring the extra cost or bandwidth overhead of naïve layered single-task training.

All three systems—semiconductor yield prediction through data-driven deep learning, real-time analytics on physical sensors, and hybrid CNNs combining both models—are important components of performance analytics in semiconductor manufacturing. Their interaction points to an increasingly important trend in edge analytics: when deep learning is used for prediction tasks driven by a high-dimensional, low-bandwidth sensor stream, careful model

engineering can directly reduce the cost and latency overhead imposed by separate prediction and edge-analytics pipelines for each task. Further work to move the yield model Fed by sensor data into the edge domain will add economic incentive to the trend.

7. Conclusion

Contributions advance an evidence-based, formally structured discussion of AI-enhanced semiconductor yield optimization through a hybrid deep learning–edge analytics strategy. A semiconductor yield-forecasting module predicts product yield and its leading indicators, based on the yield-driving features identified by an edge-based anomaly-detection framework. Sourcing the two predictive models from the cloud and edge respectively optimizes latency, bandwidth, and deployment costs. Empirical validation incorporates large-scale semiconductor manufacturing and glass fabrication datasets. By reliably predicting product yield and its leading indicators, the proposed modeling framework enables proactive yield optimization that reduces the operational expenditure required for yield improvement.

As AI and Edge Adoption deepen in manufacturing, yield optimization increasingly relies on data from diverse sources. These datasets must be analyzed seamlessly and in a timely fashion, with latency, bandwidth, and deployment cost constraints. Growing interest in hybrid deep learning has seen cloud and edge AI models jointly analyze data generated by sensor-rich smart devices. This strategy is particularly suitable for predictive tasks with a large number of predictors, such as semiconductor yield prediction, which has a direct impact on operational expenditure. Nevertheless, latency and bandwidth requirements typically render conventional deep learning unsuitable for deployment on edge devices. In contrast, data-driven edge methods capable of providing yield-driving insights are less effective for post-yield analysis, due to limited availability of yield-related data. Indeed, these two dimensions are often viewed as independent tasks, with little examination of how predictive accuracy can be used to improve yield and optimize



resources. Addressing this gap, a semiconductor yield-forecasting module predicts product yield and its leading indicators using the yield-driving features identified by an edge-based anomaly-detection framework.

7.1. Emerging Trends

To advance an evidence-based, formally structured discussion of AI-enhanced semiconductor yield optimization, concise abstract and precise keywords that reflect a hybrid deep learning and edge analytics approach follow.

Achieving high semiconductor yield is critical for economic viability. Key yield drivers are determined from the operational state (OS) of the equipment, an oft-neglected data source. Using a notional manufacturing process, a hybrid modeling framework for predicting yield is developed by combining deep learning with data from an edge-cloud pipeline. The approach exploits the temporal-dimensional nature of the data to mitigate the risk of overfitting and improve interpretability. Leakage current and electron mobility are identified as yield-affecting phenomena. Through cloud-based simulations, an economically attractive yield improvement is demonstrated. Modeling latency and bandwidth requirements indicates the feasibility of implementation using edge data analytics.

Global semiconductor manufacturing served by the Advanced Semiconductor Engineering Semiconductor (ASE) Technology Group generates trillions of dollars in annual revenue. Meeting demand with high yield is challenging. Artificial intelligence (AI) techniques are used extensively for yield prediction and enhancement, with deep learning gaining popularity because of its ability to learn complex nonlinear relationships. Yet, cloud-based predictive modeling suffers from inherent limitations, scientist discretion in selecting input features, and lack of interpretable results. Close monitoring of manufacturing equipment provides the operational state (OS) of the entire manufacturing flow at a high sampling rate. Such edge data, which reflect “what happens to the data” early in production, are often used for anomaly detection but could also uncover

hidden correlations in yield-affecting phenomena. For certain manufacturers, the OS is captured within a well-defined time interval with a predictable sequence, making it a hybrid temporal-dimensional dataset suitable for deep learning.

8. References

1. 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*, 230, 120544.
2. Doss, R. (2022). Classification of silicon (Si) wafer material defects in semiconductor choosers using a deep learning ShuffleNet-v2-CNN model. *Advances in Materials Science and Engineering*, 2022, 1829792.
3. Gauttam, H., Pattanaik, K. K., Bhadauria, S., Nain, G., & Prakash, P. B. (2023). An efficient DNN splitting scheme for edge-AI enabled smart manufacturing. *Journal of Industrial Information Integration*, 34, 100481.
4. Yandamuri, U. S. (2024). AI-Driven Decision Support Systems for Operational Optimization in Hospitality Technology. *Metallurgical and Materials Engineering*.
5. 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, 8127.
6. Kim, H. K., & Shim, J. (2024). Generalized zero-shot learning for classifying unseen wafer map patterns. *Engineering Applications of Artificial Intelligence*, 133, 108476.
7. 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)*, 261.
8. Kim, T., & Behdinan, K. (2023). Advances in machine learning and deep learning applications towards wafer map defect recognition and classification: A review. *Journal of Intelligent Manufacturing*, 34, 1909–1932.



9. Li, J., Wei, H., Liu, J., & Liu, W. (2024). FSLEdge: An energy-aware edge intelligence framework based on federated split learning for Industrial Internet of Things. *Expert Systems with Applications*, 255, 124564.
10. Long, Z., Yan, J., & Piao, M. (2024). Multi-scale guidance diffusion network for wafer map defect recognition. *Expert Systems with Applications*, 254, 126134.
11. Mughal, F. R., He, J., Das, B., Dharejo, F. A., Zhu, N., Khan, S. B., & Alzahrani, S. (2024). Adaptive federated learning for resource-constrained IoT devices through edge intelligence and multi-edge clustering. *Scientific Reports*, 14, 28746.
12. Piao, M., Jin, C. H., & Zhong, B. (2024). Antecedent hash modality learning and representation for enhanced wafer map defect pattern recognition. *Expert Systems with Applications*, 250, 123914.
13. Pinzari, A., Baumela, T., Andrade, L., Martin, M., Coppola, M., & Pétrot, F. (2024). Accurate and energy efficient ad-hoc neural network for wafer map classification. *Journal of Intelligent Manufacturing*, 36, 2863–2880.
14. Aravind, R., & Surabhi, S. N. R. D. (2024). Smart charging: AI solutions for efficient battery power management in automotive applications. *Educational Administration: Theory and Practice*, 30(5), 14257–1467.
15. Shin, E., & Yoo, C. D. (2023). Efficient convolutional neural networks for semiconductor wafer bin map classification. *Sensors*, 23(4), 2071.
16. Shinde, P. P., Pai, P. P., & Adiga, S. P. (2022). Wafer defect localization and classification using deep learning techniques. *IEEE Access*, 10, 39969–39974.
17. Sweeney, T., Coleman, S., & Kerr, D. (2022). Deep learning for semiconductor defect classification. In *2022 IEEE 20th International Conference on Industrial Informatics (INDIN)* (pp. 572–577). IEEE.
18. Xu, Q., Yu, N., & Yu, H. (2023). Unsupervised representation learning for large-scale wafer maps in micro-electronic manufacturing. *IEEE Transactions on Consumer Electronics*, 70(1), 1226–1235.
19. Zhou, Z., Yu, D., Chen, M., Qiao, Y., Hu, Y., & He, W. (2024). An edge intelligence-based model deployment method for CNC systems. *Journal of Manufacturing Systems*, 74, 716–751.
20. Zhou, C., Khaw, J., Boo, S., Lam, H., Kuan, H., You, G., Chan, D., Chong, S., Puey, J., & Ng, E. (2021). Machine learning vs. deep learning in low yield wafer map classification. In *Proceedings of the Advanced Semiconductor Manufacturing Conference (ASMC 2021)*. IEEE.
21. Jang, H. S., & Sung, C. H. (2021). Hybrid machine learning approaches for semiconductor manufacturing yield prediction. *International Journal of Production Research*, 59.
22. Park, J., & Simka, M. (2021). Machine learning-based approaches for semiconductor manufacturing process monitoring and yield improvement. *Journal of Intelligent Manufacturing*, 32.
23. Wang, Y., Li, X., Zhang, H., & Chen, Y. (2022). Data-driven optimization of accessory combinations for final testing processes in semiconductor manufacturing. *Journal of Manufacturing Systems*, 65, 482–495.
24. Mangalampalli, B. M. (2024). Transparent Intelligence Explainability Frameworks for AI-Driven Clinical Decision Support in Healthcare Business Intelligence. *International Journal of Research Publications in Engineering, Technology and Management (IJRPETM)*, 7(3), 10566–10579.
25. Deshmukh, A., Singh, A., & Kumar, R. (2022). Semiconductor multivariate time-series anomaly classification based on machine learning ensemble techniques. *IFAC-PapersOnLine*, 55(6), 476–481.
26. Deshmukh, A., Singh, A., & Kumar, R. (2022). Early semiconductor anomaly detection based on multivariate time-series classification using multilayer perceptron. *IFAC-PapersOnLine*, 55(10), 3082–3087.
27. Gauttam, H., Pattanaik, K. K., & Bhadauria, S. (2022). An IIoT approach for edge intelligence in production environments using machine learning and knowledge graphs. *Procedia CIRP*, 106, 282–287.
28. Fan, S. K. S., & Chiu, S. H. (2024). A new ViT-based augmentation framework for wafer map defect classification to enhance the resilience of semiconductor supply chains. *International Journal of Production Economics*, 273, 109275.