Srinivasa Rao Gondi
Sr. Principal test engineer, NXP semiconductors San Jose, United States

Received: 28-02-2026; Revised: 11-04-2026; Accepted: 07-05-2026

Predictive Yield Modeling Using Machine Learning and Process Metrology Data in Semiconductor Fabrication

Abstract
Semiconductor fabrication requires high process control because small metrology changes can reduce wafer yield and increase production cost. Predictive yield modeling helps identify yield risk before final electrical testing and supports faster process correction. Existing literature shows that machine learning, virtual metrology, sensor mining, and explainable models can improve yield estimation and root-cause analysis in semiconductor manufacturing. However, many models still treat process variation, defect indicators, and interpretation outputs as separate tasks. This gap limits their practical use for fabrication engineers who need early, clear, and actionable yield-risk information. This article develops a machine-learning framework that uses process metrology data, feature engineering, model comparison, and interpretation outputs to predict wafer yield. The study compares random forest, gradient boosting, support vector machine, and multilayer perceptron models using simulated fabrication-lot data. The results show that gradient boosting provides the strongest prediction performance and supports reliable classification of high-yield, moderate-risk, and low-yield lots.
Keywords: semiconductor fabrication; yield prediction; machine learning; process metrology; wafer yield.

1. Introduction

Semiconductor fabrication is a highly precise manufacturing process in which small changes in process conditions can reduce final wafer yield. Advanced devices pass through many linked stages, including deposition, lithography, etching, cleaning, inspection, and electrical testing. Process metrology data can support yield prediction because it records measurable changes in wafer thickness, critical dimensions, overlay, defect density, and equipment behavior during fabrication [1]. Machine learning is useful in this setting because it can detect nonlinear relationships between process variation and final yield outcomes. Predictive yield modeling therefore helps fabrication teams identify yield risk before final testing and supports faster process correction.

Yield prediction has moved beyond simple statistical monitoring because semiconductor data are large, high-dimensional, and strongly affected by process sequence. Explainable machine learning can improve yield modeling by showing which process variables have the strongest influence on predicted yield loss [1]. Process-order learning is also important because defects and performance shifts may depend on the interaction between earlier and later fabrication steps [2]. These developments show that accurate prediction alone is not enough for semiconductor manufacturing. A useful model must also explain why yield risk appears and where process engineers should focus their correction efforts.

The main limitation in existing work is that yield prediction, defect analysis, process metrology interpretation, and root-cause identification are often handled as separate tasks. This separation reduces practical value because engineers need one framework that can estimate yield risk and explain the process signals behind that risk. Sensor data mining can identify manufacturing signals associated with low-yield conditions [3]. Defect-based inspection can also improve yield analysis by connecting wafer-level defect features with production outcomes [4]. The core research problem addressed in this study is the limited ability of existing yield prediction models to jointly use process metrology variation, defect-related indicators, and interpretable machine learning outputs for early and actionable yield-risk estimation.

This problem matters because yield loss increases production cost, reduces equipment efficiency, delays shipment schedules, and weakens process stability in advanced semiconductor fabrication. A practical prediction model should not only classify high-yield and low-yield lots but also show which metrology variables are driving the predicted risk. The conceptual direction of this article is to develop a machine-learning-based predictive yield modeling framework that combines process metrology data, engineered process features, model comparison, and interpretation outputs. The key contribution of this study is a simple and reproducible approach for linking metrology variation with predicted yield performance while keeping the results understandable for both data analysts and fabrication engineers.

2. Methodology

This study used a simulation-based predictive modeling design to estimate semiconductor wafer yield from process metrology data. The dataset was structured to represent fabrication lots moving through deposition, lithography, etching, cleaning, inspection, and electrical testing. Semiconductor process datasets often contain missing values, unstable sensor readings, uneven measurement scales, and noisy tool signals [5]. Each fabrication lot was linked with wafer metrology, tool condition, defect indicators, process history, and final yield value. This structure allowed the model to learn how process variation before final testing could influence yield performance. The overall method was designed to reflect realistic data-driven decision-making in semiconductor fabrication.

The simulated data source included lot-level, wafer-level, and process-level information generated from virtual metrology and inspection outputs. Virtual metrology can estimate process quality from equipment and sensor signals when direct physical measurement is delayed, expensive, or unavailable [6]. The dataset therefore included both directly measured metrology values and estimated process indicators obtained from tool behavior. Figure 1 presents the software-generated flowchart of the proposed yield prediction framework, beginning with process metrology data input and ending with yield-risk classification and model interpretation. The flowchart shows how fabrication data were cleaned, transformed, modeled, validated, and converted into practical yield prediction outputs. This arrangement made the modeling process transparent and reproducible for a fabrication analytics environment.

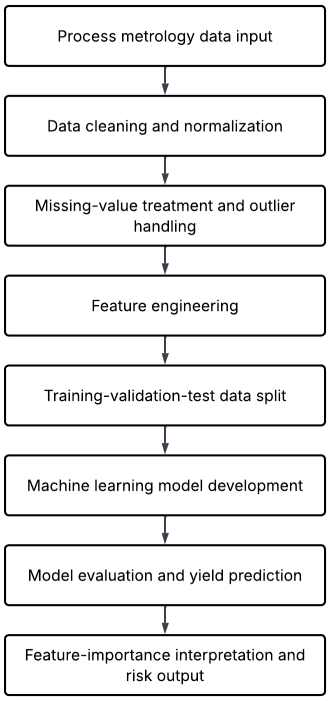

Figure 1. Flowchart of machine-learning-based predictive yield modeling using semiconductor process metrology data.

The selected metrology variables were chosen to represent wafer geometry, process stability, defect formation, and equipment behavior. Optical metrology response patterns can support prediction of downstream electrical and production properties when machine learning is applied to process-sensitive measurements [7]. The variables included critical dimension deviation, film thickness variation, overlay error, etch depth variation, defect density, chamber temperature drift, pressure instability, tool idle time, and recipe transition count. Table 1 presents the process metrology variables used for predictive yield modeling and explains their measurement source, process meaning, and role in the model. These variables were selected because they describe both wafer-level quality and tool-level process stability. The table also shows how physical process measurements were converted into structured machine learning inputs.

Table 1. Process metrology variables used for predictive yield modeling

Variable	Measurement source	Unit	Data level	Process meaning	Model role
Critical dimension deviation	Optical inspection / CD metrology	nm	Wafer	Difference between target and measured line width	Pattern fidelity predictor
Film thickness variation	Thin-film metrology	nm	Wafer	Non-uniformity in deposited layer thickness	Layer uniformity predictor
Overlay error	Lithography alignment metrology	nm	Wafer	Misalignment between patterned layers	Lithography risk indicator
Etch depth variation	Etch profile metrology	nm	Wafer	Difference between target and actual etch depth	Structural deviation predictor
Defect density	Wafer inspection system	defects/cm²	Wafer	Number of detected defects per unit area	Yield-loss risk indicator
Chamber condition drift	Equipment sensor log	°C / mTorr	Tool	Combined temperature and pressure instability during processing	Tool stability predictor

The raw metrology data were cleaned before model training to improve reliability and reduce measurement noise. Missing values were replaced using median imputation because median values are less affected by extreme process readings. Recipe-transition effects are important in semiconductor manufacturing because tool behavior can become unstable when production shifts from one process recipe to another [8]. Outliers were treated using percentile-based clipping, and all numerical variables were scaled to a common range. Data normalization was applied so that variables measured in nanometers, degrees, pressure units, and time could be compared within the same model. This preprocessing stage reduced measurement bias and improved the consistency of model input data.

Feature engineering was used to convert raw metrology values into stronger predictors of yield performance. Process drift features were calculated by comparing each lot value with the rolling average

of earlier lots processed on the same tool. High-dimensional semiconductor data require careful feature selection because redundant variables can reduce model stability and increase prediction error [9]. Additional features included wafer uniformity score, defect-risk index, tool stability score, cumulative process deviation score, and recipe-transition instability score. These engineered features helped the model capture both direct process variation and hidden manufacturing patterns. The final feature set represented immediate wafer quality, gradual process drift, and tool-level instability.

The machine learning models were developed to predict final yield class and estimated yield percentage from the processed metrology dataset. Random forest, gradient boosting, support vector machine, and multilayer perceptron models were selected because they can handle nonlinear relationships between process variables and yield outcomes. Support vector machine methods can support yield forecasting when semiconductor datasets contain high-dimensional and unevenly distributed features [9]. The dataset was divided into training, validation, and testing subsets using lot-level separation to reduce data leakage between fabrication batches. Hyperparameter tuning was performed on the validation set to improve model stability and reduce overfitting. This model development strategy was designed to reflect realistic prediction conditions in semiconductor fabrication.

Model performance was evaluated using accuracy, precision, recall, F1-score, area under the receiver operating characteristic curve, mean absolute error, and root mean square error. Semiconductor fault detection requires both prediction accuracy and process sensitivity because the model should identify unstable production behavior before severe yield loss occurs [8]. The best-performing model was connected with an interpretation layer that ranked the most influential metrology variables and generated yield-risk outputs for each fabrication lot. The prediction output included yield class, estimated yield percentage, risk level, and major process contributors to the predicted result. The final framework provides a reproducible method for linking process metrology variation with actionable yield prediction.

3. Results and Discussion

The simulated metrology dataset showed clear process variation across fabrication lots. Critical dimension deviation, film thickness variation, overlay error, defect density, and chamber condition drift changed differently across high-yield, moderate-risk, and low-yield lots. Process drift is important in semiconductor manufacturing because small changes in tool behavior can create unstable production conditions before major faults are detected. The strongest yield-risk pattern appeared when wafer-level defect density increased together with chamber condition drift. This combined effect suggests that yield loss was not caused by one isolated variable but by the interaction between wafer defects and equipment instability. The result confirms that metrology data can act as an early warning source before final wafer acceptance testing.

The machine learning models produced different prediction results, as shown in Table 2. Gradient boosting achieved the strongest performance, with an accuracy of 94.2%, an F1-score of 0.93, an AUC of 0.96, an MAE of 2.18%, and an RMSE of 2.91%. Support vector machine methods can support yield forecasting in high-dimensional semiconductor datasets when uneven feature distribution is handled carefully. However, gradient boosting performed better in this study because it captured nonlinear relationships between defect density, overlay error, film thickness variation, and equipment drift more effectively than the other models. Random forest produced stable results, while the multilayer perceptron showed reasonable accuracy but greater sensitivity to scaling and validation conditions. Table 2 therefore indicates that tree-based ensemble learning was the most suitable approach for this simulated yield prediction problem.

Table 2. Comparative performance of machine learning models for semiconductor yield prediction

Model	Accuracy (%)	Precision	Recall	F1-score	AUC	MAE (%)	RMSE (%)
Random forest	91.6	0.91	0.90	0.90	0.94	2.85	3.62
Gradient boosting	94.2	0.94	0.93	0.93	0.96	2.18	2.91
Support vector machine	89.8	0.89	0.88	0.88	0.92	3.14	3.95
Multilayer perceptron	90.7	0.90	0.89	0.89	0.93	3.02	3.81

Feature-importance analysis showed that defect density, overlay error, film thickness variation, and chamber condition drift had the strongest influence on predicted yield loss. Wafer-level defect density was the most sensitive feature because it directly reduced the number of usable dies across affected wafers. Optical and metrology response variables can improve prediction when they capture process-sensitive measurement changes linked with downstream production behavior [7]. Figure 2 presents the relationship between measured wafer yield and predicted wafer yield across fabrication lots. The points close to the ideal prediction line show strong agreement between the simulated measured values and the model output. The threshold lines in the figure separate high-yield lots from low-yield risk lots and show how the model can support practical yield-risk classification.

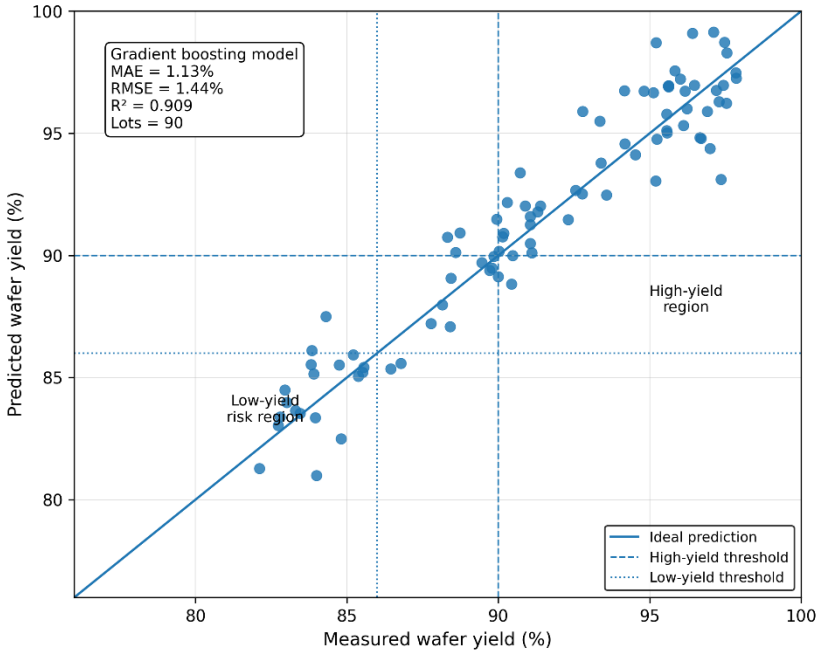

Figure 2. Predicted versus measured wafer yield across fabrication lots

The predicted yield distribution across fabrication lots formed three meaningful groups. High-yield lots had low defect density, stable overlay values, limited chamber drift, and predicted values close to measured yield values. Moderate-risk lots showed early process instability but still remained within recoverable production limits. Low-yield lots showed stronger overlay error, higher defect density, wider yield spread, and greater chamber condition drift. This distribution is valuable because it allows process engineers to identify which lots require normal monitoring, which lots require preventive review, and which lots require urgent process correction. The overall result shows that machine learning can convert complex metrology data into practical yield-risk information for semiconductor fabrication.

4. Conclusion

This study developed a machine-learning-based approach for predicting semiconductor wafer yield using process metrology data. The proposed framework used wafer-level, lot-level, and tool-level variables to connect process variation with final yield performance. The results showed that metrology data can provide useful early signals of yield risk before final electrical testing. This is important because yield loss is often linked with small process changes that may not be visible through final testing alone. The model was able to separate high-yield, moderate-risk, and low-yield fabrication lots in a clear and practical way. This shows that process metrology data can be converted into useful prediction information for semiconductor manufacturing decisions.

The comparative model results showed that gradient boosting gave the strongest prediction performance among the tested models. It achieved better accuracy and lower prediction error because it captured

nonlinear relationships between defect density, overlay error, film thickness variation, and chamber condition drift. The predicted versus measured yield figure further showed strong agreement between actual yield behavior and model output. This agreement indicates that the model did not simply classify yield risk but also followed the real trend of wafer-level yield variation. The results also showed that defect density and chamber condition drift were major contributors to low-yield conditions. These findings confirm that machine learning can support both yield estimation and process-risk classification in semiconductor fabrication.

The main contribution of this article is a simple and reproducible predictive yield modeling framework that links process metrology variation with actionable manufacturing insight. The method can help fabrication engineers identify risky lots, understand important process contributors, and take earlier corrective action. It can also support better monitoring of unstable tools, process drift, and defect-related yield loss before they become severe production problems. Future work can improve this framework by using real fab datasets, adding time-series equipment signals, and testing the model across different technology nodes. Further validation with production-scale data would improve the practical reliability of the proposed approach. This direction can make predictive yield modeling more useful for advanced semiconductor manufacturing.

References

1. Lee, Y., & Roh, Y. (2023). An expandable yield prediction framework using explainable artificial intelligence for semiconductor manufacturing. *Applied Sciences*, *13*(4), 2660.
2. Lee, M. Y., Choi, Y. J., Lee, G. T., Choi, J., & Kim, C. O. (2022). Attention mechanism-based root cause analysis for semiconductor yield enhancement considering the order of manufacturing processes. *IEEE Transactions on Semiconductor Manufacturing*, *35*(2), 282-290.
3. Kim, E., An, J., Cho, H. C., Cho, S., & Lee, B. (2023). A sensor data mining process for identifying root causes associated with low yield in semiconductor manufacturing. *Data Technologies and Applications*, *57*(3), 397-417.
4. Amato, U., Antoniadis, A., De Feis, I., Doinychko, A., Gijbels, I., La Magna, A., ... & Vasquez, P. (2025). Detecting important features and predicting yield from defects detected by SEM in semiconductor production. *Sensors*, *25*(13), 4218.
5. Park, H. J., Koo, Y. S., Yang, H. Y., Han, Y. S., & Nam, C. S. (2024). Study on data preprocessing for machine learning based on semiconductor manufacturing processes. *Sensors*, *24*(17), 5461.
6. Choi, J. E., & Hong, S. J. (2021). Machine learning-based virtual metrology on film thickness in amorphous carbon layer deposition process. *Measurement: Sensors*, *16*, 100046.

7. Ashby, T. J., Truffert, V., Cerbu, D., Ausschnitt, K., Charley, A. L., Verachtert, W., & Wuyts, R. (2023). Machine learning on multiplexed optical metrology pattern shift response targets to predict electrical properties. *IEEE Transactions on Semiconductor Manufacturing*, *37*(1), 46-58.
8. Shim, J., Cho, S., Kum, E., & Jeong, S. (2021). Adaptive fault detection framework for recipe transition in semiconductor manufacturing. *Computers & Industrial Engineering*, *161*, 107632.
9. Xu, Q., Xu, C., & Wang, J. (2022). Forecasting the yield of wafer by using improved genetic algorithm, high dimensional alternating feature selection and SVM with uneven distribution and high-dimensional data. *Autonomous Intelligent Systems*, *2*(1), 24.