Spatial Modeling with Automated Machine Learning and Gaussian Process Regression Techniques for Imputing Wafer Acceptance Test Data

Ming-Chun Wei[1], Chun-Wei Shen[1], Hsun-Ping Hsieh*[1,2]
Department of Electrical Engineering[1]
Academy of Innovative Semiconductor and Sustainable Manufacturing[2]
National Cheng Kung University, Tainan, Taiwan
N26100692@gs.ncku.edu.tw, N26120595@gs.ncku.edu.tw, hphsieh@mail.ncku.edu.tw(corresponding author)

***Abstract*—The Wafer Acceptance Test (WAT) is a significant quality control measurement in the semiconductor industry. However, because the WAT process can be time-consuming and expensive, sampling test is commonly employed during production. This makes root cause tracing impossible when abnormal products have not been tested. Therefore, in our study, we focus on establishing a reliable method to estimate WAT results for non-tested shots, including both intra and inter-wafer prediction. Notably, we are the first to combine the use of Chip Probing data with WAT to improve the predictions. Our proposed method first extracts valuable features from Chip Probing test results by using the Automated Machine Learning technique. We then employ Gaussian Process Regression to capture the spatio-temporal correlation. Finally, we adopted the linear regression model to ensemble two components and proposed a SMART-WAT model to effectively estimate the wafer acceptance test data. Our method has been tested on a real-world dataset from the semiconductor manufacturing industry. The prediction results of four key WAT parameters indicate that our proposed model outperforms the state-of-the-art methods in both intra and inter-wafer prediction.**

***Index Terms*—Semiconductor process modeling, Wafer acceptance test, Spatial correlation, Fined-grained prediction**

I. Introduction

Controllable process variations are increasingly challenging in modern semiconductor manufacturing, where smaller component sizes make even slight variations critical, leading to economic losses. Therefore, effectively monitoring process variations is crucial for integrated circuit manufacturing.

Wafer Acceptance Test plays a crucial role in yield monitoring by verifying manufacturing parameters and guiding process adjustments. However, due to cost, WAT typically conducts sampling in actual testing procedures, as shown in Figure 1. In each lot of products, only a few wafers will have sparse shots subjected to WAT measurements. While this approach identifies variations across product lots, it overlooks within-wafer and wafer-to-wafer variations, making it difficult to trace e root cause of the process impact outside sampling points. Addressing this limitation requires a reliable method to estimate the untested WAT data from the sparse available data.

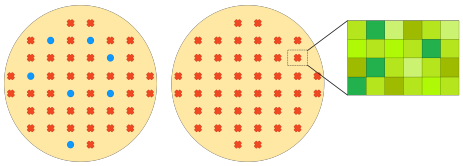

Fig. 1. An example of Wafer Acceptance Test sampling. The left wafer undergoes WAT, with blue circles representing the fixed measurement shots, each containing several dies as depicted in the grid on the far right. Red X on both wafers indicate the untested area, signifying the absence of WAT data.

Various methods for WAT data estimation have been proposed. Virtual Probe (VP) [1] model wafer-level variation using Discrete Cosine Transform, while Gaussian Process (GP) [2] improves intra-wafer interpolation accuracy compared to VP. Further studies [3]–[5] improved the basic GP method by adding radius features, optimizing kernel functions, and considering inter-wafer relationships. The comprehensive GP model has achieved significant improvements in WAT estimation accuracy, laying the foundation for our research.

In addressing the WAT data estimation problem, we encountered two challenges. First, WAT sampling points are sparsely distributed. We address this by including Chip Probing(CP) results, which correlate with WAT. However, this brings up a second challenge: how to align the scales and measurement parameters between die-level CP and shot-level WAT.

To overcome these challenges, we propose a novel framework that leverages wafer-level spatial correlations and integrates CP data to enhance WAT estimation for non-tested shots. CP is a die-level test performed on each wafer and individual die before dicing. Integrating CP data significantly improves our understanding of the wafer's properties and performance. We address the alignment issue by extracting CP results from each shot using percentage sampling and applying Automated Machine Learning (AutoML) to predict WAT results based on the CP values. AutoML model provide highly accurate estimations of the WAT results for the non-tested shots. Our hybrid approach combines AutoML predictions with the GP

978-3-9826741-0-0/DATE25/© 2025 EDAA

model, capturing complex spatio-temporal relationships between different shots on the wafer and accounts for underlying variations in the manufacturing process.

Our framework, tested on the four key indicators identified through discussions with domain experts from the industry's high-volume manufacturing (HVM) data, demonstrated a 30% average error reduction compared to existing solutions. This improvement showcases our approach's potential to enhance process monitoring, improve yields, and reduce economic losses in semiconductor manufacturing.

II. Related Works

A. Statistical Methods

Common methods for estimating untested WAT results include Nearest Neighbors, averaging, and historical data usage, often lack accuracy and fail to provide actionable insights for process adjustments. Inverse Distance Weighting [6] assumes that the relevance of a data point is inversely proportional to its distance from the point of interest. However, by focusing on linear relationships, it overlooks complex data patterns. Statistical methods overlook complex wafer-level spatial variations, require advanced techniques to improve estimation accuracy.

B. Virtual Probe

The Virtual Probe approach proposed by Li et al. [1], has emerged as a promising method in recent literature. This technique models spatial variations using a Discrete Cosine Transform (DCT) that projects spatial statistics into the frequency domain. Specifically, it collects measurements from a sparse subset of data points on each wafer and trains statistical spatial models to predict performance outcomes at unobserved locations. While the VP approach has shown potential, it does have certain limitations.VP approach only considers x and y coordinates, which restricts it to reasoning from the domain of compressed sensing and overlooks other factors that might influence the wafer's behavior. Furthermore, the VP approach is focused on single-wafer estimation, neglecting the beneficial insights that could be captured from a cross-wafer evaluation.

C. Gaussian Process Regression

Gaussian Process Regression model offers a robust solution for estimating untested WAT. The Gaussian process is a collection of random variables, where any finite subset has a joint Gaussian distribution [7]. The entire process is defined by its mean function and a kernel-based covariance function, offering a robust framework for modeling complex, non-linear relationships. Initially applied to interpolation of semiconductor data by Liu [2]. This model has been further refined with several advanced features such as radial feature inclusion, multiple kernel evaluation, and the addition of a regularization parameter [3]–[5]. A significant feature of the GP model is its capability to expand the correlation model across wafers within the same lot by taking the wafer's production order as a temporal feature. The resulting GP model has demonstrated significant improvements in prediction accuracy and computational efficiency compared to the VP method, making it our chosen starting point for further research.

D. Machine Learning in WAT

Machine learning has proven effective in improving WAT processes. Fan et al. [11] focuses on diagnosing faults for the WAT and CP using logistic regression, SMOTE, and PCA, improving fault detection by identifying critical operations and features. Wang and Yang [12] integrates control charts, clustering methods, and association rule mining to pinpoint critical stations and equipment impacting yield during WAT. Xu et al. [13] proposes a hybrid feature selection method, combining mRMR and GA-DBN, to manage data dimensionality and improve yield prediction. Additionally, Cheng et al. [14] employed Support Vector Machines, Logistic Regression, and neural network models to identify test-induced defects in semiconductor wafer testing, enhancing manufacturing yield by distinguishing these from fabrication-induced faults.

Our work, however, addresses a different challenge, estimating WAT results for non-tested shots and lots by integrating CP data. This innovative approach is especially advantageous for predicting parameters in both intra and inter-wafer scenarios, enabling manufacturers to identify potential issues even when WAT tests are not performed on those samples.

III. Preliminaries

A. Chip Probing

Chip Probing also known as Circuit Probing, is a crucial process that identifies defective dies in a wafer based on their electrical specifications. CP is performed on each individual wafer and die to ensure that they meet the required specifications for electrical and functional performance. This process helps optimize yield, reduce production costs, and ensure high-quality semiconductor devices. During the CP process, probe needles contact specific points on individual dies to measure electrical parameters. By analyzing the electrical responses, manufacturers can detect abnormalities or defects that may affect the final product's functionality or reliability.

B. Wafer Acceptance Test

WAT is a key quality control measure in the semiconductor industry. By testing on specific samples of wafers to ensure that they meet the required electrical and functional performance. WAT identifies defects or variations that could impact product quality. Due to the high cost and time involved, WAT typically uses sampling. Instead of examining each point on every wafer, a select few wafers that represent the entire lot are chosen for testing. These chosen wafers are subjected to various measurements and evaluations to assess their compliance with specific standards. The testing electronic devices have been strategically placed in the wafer's scribe lines to maximize space utilization. These structures enable the measurement of intra-wafer variations at the shot level, providing insights into the uniformity and consistency of the semiconductor devices across the wafer.

C. Problem Definition

Our research focuses on estimating missing WAT results. In our study, each lot contains j wafers, and each wafer includes h shots. However, the WAT has only been performed on k shots from m wafers, where $km << jh$.

We define this problem as a prediction task, where Y_t represents the WAT results for the shots that have undergone the test, and we seek to predict the WAT results, Y_u, for the untested shots. To achieve this, our model utilizes a combination of different input features. Firstly, we use the die-level CP test results for each shot, which includes n dies, symbolized as X^c. Secondly, the order of the wafer within each lot, denoted as X^t, is considered as a temporal feature. Finally, we include spatial features represented as X^s. Formulating the optimization objective for our proposed model, we aim to minimize the difference between the predicted $\hat{Y}_u$ and actual WAT results Y_u for the shots. This can be formally stated as follows:

$$f(X^c; X^t; X^s; Y_t) \rightarrow \hat{Y}_u. \tag{1}$$

The optimization objective of the model is to minimize the error between Y_u and $\hat{Y}_u$, i.e., $\|Y_u - \hat{Y}_u\|$. In doing so, our proposed model strives to improve the accuracy of predicting the WAT results for the untested shots, effectively addressing the problem of missing WAT values.

IV. The Proposed Method

In this section, we provide the details of our proposed framework, SMART-WAT. The architecture is illustrated in Fig. 2. Our method comprises two key components: Gaussian Process Regression (GPR) for capturing spatio-temporal correlation and Automated Machine Learning for learning from CP Data. The final ensemble model is trained using a linear regression model to combine these two approaches. The mathematical representation of SMART-WAT are shown as Eq. 2.

$$\beta_1 * Y_{gp} + \beta_2 * Y_{ml} \rightarrow \hat{Y}, \tag{2}$$

Where Y_{gp} and Y_{ml} are prediction of GPR and AutoML from Eq. 3 and Eq. 4. The weights β_1 and β_2 are determined by linear regression, combining both components into an ensemble.

The GPR component is represented as follows:

$$\mathcal{GP}(X^t; X^s; Y_t) \rightarrow Y_{gp}, \tag{3}$$

The AutoML component is represented as follows:

$$\mathcal{ML}(X^t; X^c; X^s; Y_t) \rightarrow Y_{ml}, \tag{4}$$

Further details of the GPR and AutoML components are discussed in IV-A and IV-B, respectively.

A. Gaussian Process Regression

Gaussian process regression [7] is a supervised learning method grounded in Bayesian statistics. It models complex, nonlinear relationships using a prior distribution over functions. GPR learns a distribution over functions by determining the mean (Eq. 5) and covariance (Eq. 6) functions.

$$m(x) = \mathbb{E}[f(x)], \tag{5}$$

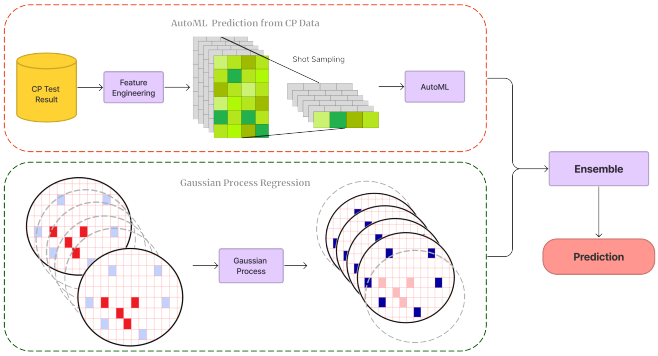

Fig. 2. The architecture of our SMART-WAT framework.

$$k(x, x') = \mathbb{E}[(f(x) - m(x))(f(x') - m(x'))], \tag{6}$$

Mean function $m(x)$ gives the expected value of the output variable $f(x)$ for a given input x. The covariance function $k(x, x')$, measures how outputs at two different inputs, x and x' vary together. Gaussian process can be written as

$$f(x) \sim \mathcal{GP}(m(x), k(x, x')), \tag{7}$$

By representing the covariance function as an inner product, we can utilize the kernel trick and express it as a kernel function $k(x, x')$. In previous studies, the Radial Basis Function (RBF) kernel, also known as the squared exponential kernel, has been the most commonly used option. The RBF kernel is given by

$$k(x, x') = exp(-\frac{d(x, x')^2}{2l^2}), \tag{8}$$

Where d is the Euclidean distance. Length-scale l is a hyper-parameter determining the function's smoothness.

In our work, we also experiment with Rational Quadratic kernel (Eq. 9) which can be seen as a mixture of RBF kernels with different characteristic Length-scales. This flexibility can be particularly beneficial in semiconductor test estimates, where the data may contain patterns at different length scales.

$$k(x, x') = (1 + \frac{d(x, x')^2}{2al^2})^{-a}, \tag{9}$$

Additionally, our methodology enhances the modeling process by employing data from the inner four shots of the first and last wafer in each lot. We utilize the x and y coordinates, the radius of each shot, and the wafer sequence in the lot to capture more detailed patterns in the dataset.This approach is particularly aimed at predicting the characteristics of the remaining six shots in the middle wafer, allowing for a comprehensive analysis of wafer quality across the entire lot.

B. Automated Machine Learning Prediction from CP Data

In this section, we explore the application of Machine Learning technology to extract valuable insights from CP test results and predict WAT parameters. Our objective is to establish an automated method that effectively utilizes Machine Learning techniques to improve the accuracy and reliability of WAT parameter predictions based on CP test data.

C. CP data preprocessing

CP data preprocessing involves key steps: removing missing values and low-variance columns to improve data quality, creating new features by calculating differences between stations under varying testing temperatures, and aggregating features by computing slopes of the same parameters measured at different scales. These preprocessing techniques improve the usefulness and relevance of CP data for subsequent prediction tasks.

D. Shot Sampling

To align the die-level CP data with the shot-based WAT target, we employ a percentage sampling method. In collaboration with industry experts, we select suitable values, such as 95%, 90%, 10%, and the average, which represent the CP information for each shot. This method helps bridge the gap between the CP measurements and the desired WAT predictions.

E. Automated Machine Learning

With the increasing automation in semiconductor factories, it is essential to develop methods that can quickly predict various WAT parameters. To address this challenge, we leverage the power of Automated Machine Learning technology.

AutoML offers several advantages, including automated model selection, which identifies the best model for each WAT parameter, saving time and improving accuracy. Additionally, AutoML uses ensemble methods to combine multiple models, enhancing prediction robustness and performance.

Another key benefit of AutoML is feature pruning, which removes irrelevant or redundant CP data features by ranking their importance and applying recursive elimination. It results in a more streamlined and focused set of features, enhancing prediction accuracy and improving the robustness of the model.

In our work, we implement AutoML by AutoGluon [8]. Overall, AutoGluon streamlines the entire process of utilizing CP data for WAT prediction in the semiconductor manufacturing industry. It saves time, improves accuracy, and ensures robust and reliable results, contributing to increased productivity and quality in the manufacturing processes.

F. Final Ensemble Model

The final ensemble model we propose combines GPR and AutoML prediction, inspired by Huang et al. [9]. In their work, Huang et al. utilized regression techniques to enhance semiconductor testing by combining alternate, less comprehensive test parameters with spatial correlation data across wafers.

We extends this concept by incorporating GP regression to capture not only spatial correlations within each wafer but also temporal correlations across different wafers. Additionally, AutoML extract die-level information from CP data, thereby enhancing prediction accuracy. To integrate these models, we train a linear regression model [10] that optimally weighs the GPR and AutoML predictions. Linear regression is simple yet effective for handling two inputs, providing clear contributions from each model while ensuring computational efficiency.

The final ensemble model combines the strengths of both approaches. The GPR component captures the intricate patterns, while the AutoML component leverages automated model selection, feature pruning and ensemble methods for improved accuracy and robustness. The ensemble model provides more accurate predictions, streamlines the process, and enhances efficiency in the semiconductor industry.

V. Experiments

Our method was tested using real-world manufacturing data. The comparison focused on four key WAT parameters: V1, voltage value that is linked to the performance of the end product; LV1 and LV2, two critical threshold voltage parameters that are often subjected to testing in the standard procedures of production; and R1, representing the resistance value.

These key indicators were identified as primary targets through consultations with semiconductor manufacturing experts. These indicators were extracted from the HVM WAT data. The dataset include 155 lots, with each lot containing 6 tested wafers. Each wafer has 10 fixed shots that underwent WAT measurements, resulting in a total of 9,300 data points of WAT data. For CP data, each shot contained 24 dies, and 1,270 parameters were collected under various testing conditions.

Specifically, we aim to answer three research questions:

RQ 1. In the task of predicting WAT parameters by using data not originating from the same wafer and shot as the target. Can our framework successfully minimize prediction errors?

RQ 2. Can our proposed shot sampling method be effective in capturing the necessary information to enhance the model's predictive capabilities?

RQ 3. How does each component contribute to the overall effectiveness and performance of the final ensemble model within the SMART-WAT framework?

A. Experimental Setup

In each lot, we first select inner 4 shots from 10 extracted from the first and last wafers out of 6. This subset is used to train the AutoML and GPR models. The goal is to evaluate the framework's effectiveness in minimizing prediction errors for unseen wafers and shots. An ensemble model is then trained using the validation data. To evaluate the performance of this ensemble model, we construct a test set from the last 16 lots of the dataset. The remaining 139 lots are further divided into two subsets: 124 lots for training and 15 lots for validation.

For our implementation details, we use Gaussian Process with different kernels and length scales for specific targets: a Rational Quadratic kernel with a length scale of 1 is used for both V1 and LV1, a length scale of 3 for LV2, and a RBF kernel with a length scale of 3 for R1. For AutoML, we employ AutoGluon's default settings, including neural networks, LightGBM, CatBoost, XGBoost, Random Forest, Extra Trees, and KNeighbors, with a feature pruning ratio of 10%.

B. Baselines

In order to verify the effectiveness of our method. We select the following 5 methods as competitors:

- **Average**: Semiconductor manufacturing processes usually yield consistent wafer properties. To establish a baseline

for comparison, we average the inner four shots from the first and last wafer of each lot.

- **Inverse Distance Weighting(IDW)** [6]: This method calculates the predictive values for the outer six shots of the first and last wafers in each lot. The average value derived from this computation serves as the predicted result for the other four wafers in the lot. This computation is based on the inverse square of the distance from the four inner shots of the same wafers.
- **Virtual Probe(VP)** [1]: Virtual Probe algorithm is used to calculate the predictive values. It computes the values of the outer six shots from the first and last wafer of each lot and uses the average of two wafers for prediction.
- **Gaussian Process Regression(GPR)** [3]–[5], [7]: As a component of our proposed method. This approach applies Gaussian Process Regression to predict the values of the outer six shots. The method utilizes data from the inner four shots from the first and last wafer of each lot, incorporating the x and y coordinates, radius, and the order of the wafer within the lot as a temporal feature.
- **AutoML**: Another key component, AutoML is employed to extract valuable insights from the CP test results. In this approach, the CP data from the inner four shots of the first and last wafer of each lot is used to predict the WAT parameters. The AutoML process is implemented using the AutoGluon framework [8].

C. Evaluation metrics

To assess the performance of our proposed model against the baseline methods, we adopt two commonly employed regression metrics: Root Mean Square Error (RMSE) and Mean Absolute Error (MAE), as defined in Eq. 10 and Eq. 11.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}, \tag{10}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|Y_i - \hat{Y}_i|, \tag{11}$$

D. Experimental Results

Comparison with Baseline Models: From the results in Table I, we can summarize the following two points:

1) Our proposed method, SMART-WAT, outperforms all baseline models, providing a more accurate prediction across the four WAT parameters as shown in the lowest RMSE and MAE values. Compared to the most effective previous solution, SMART-WAT achieved improvements of 29% for V1, 11% for LV1, 35% for LV2, and an impressive 43% for R1. These significant gains demonstrate the effectiveness of our proposed method in minimizing prediction errors in WAT parameter estimation.
2) When compared individually with GPR and AutoML, the results from our ensemble method further validate its effectiveness. GPR is good at modeling nonlinear relationships and capturing spatiotemporal correlations, while AutoML extracts valuable die-level information from CP data. However, each approach has its limitations. For example, AutoML generally performs well but exhibited a drop in accuracy for LV1. In contrast, our ensemble method still maintained high performance, demonstrating its robustness even when individual components may under-performed. By mitigating the individual limitations of GPR and AutoML, we effectively leverage their combined strengths, resulting in better prediction accuracy for untested WAT parameters.

The results highlight the effectiveness of our framework in minimizing prediction errors and incorporating diverse features. Our method is more accurate and reliable in estimating untested WAT parameters in semiconductor manufacturing.

Evaluation of Shot Sampling Methods: To address research question RQ 2, we conducted an experiment with various shot sampling percentages. The performance evaluated on the validation set is presented in Table II.

We can derive the following insights:

1) The method used in our work mostly performs the lowest error among the various shot sampling percentages. This combination of shot sampling percentages enables our model to capture a greater amount of information, which contributes to more accurate predictions.
2) Method (b), using average values, consistently underperforms. This shows that shot sampling percentages provides more valuable information than using mean values alone, leading to more accurate predictions.

Overall, these findings support the effectiveness of our shot sampling method in capturing more information and improving prediction accuracy, answering RQ 2 positively.

Weights within the ensemble model of SMART-WAT: Table I demonstrates the robustness of our ensemble method. To further explore research question RQ 3, we analyzed the Linear Regression weights in the ensemble segment of SMART-WAT, as presented in Table III. The weights reveal that the AutoML component plays a more significant role in the final ensemble, particularly in predicting V1, which relies heavily on AutoML. However, GPR holds a higher weight for LV1. This disparity can be attributed to the absence of directly correlated features in CP data, which increases GPR's influence on this target. The weights provide insights into the individual contribution of each component to the final ensemble model, offering a clearer understanding of the prediction process and valuable data-driven insights for Production Engineers.

E. Real-World Case Study

We used 7 completely tested wafers from 3 different lots for the experiment. Excluding the originally 10 sampled shots, each wafer contained 60 WAT test results that would not be tested during the standard procedures. Providing a total of 420 shots of WAT data for our experiment.

Experiment Configuration: For the IDW and VP approaches, calculations were first done on the six wafers with ten shots data. Then using the average value of six wafers as the predicted

TABLE I
PREDICTION ERROR OF SMART-WAT AND BASELINES

	Model	V1		LV1		LV2		R1	
		RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
Previous Solutions	Average	0.4276	0.3529	0.0118	0.00941	0.00780	0.00630	6.7401	5.6514
	IDW	0.4304	0.3562	0.0116	0.00932	0.00763	0.00617	6.7221	5.6595
	VP	0.3973	0.3278	0.0126	0.01006	0.00765	0.00619	7.3294	6.1978
	GPR	0.4083	0.3377	0.0115	0.00922	0.00761	0.00616	6.3393	5.3383
	AutoML	0.2844	0.2284	0.0127	0.00934	0.00492	0.00399	3.7034	2.9616
	SMART-WAT	**0.2822**	**0.2268**	**0.0103**	**0.00821**	**0.00482**	**0.00391**	**3.6234**	**2.8980**
	Improvement	28.97%	30.82%	10.96%	10.98%	35.24%	35.39%	42.84%	45.71%

TABLE II
PREDICTION ERROR OF SMART-WAT USING DIFFERENT SHOT SAMPLING METHODS

Sampling Description	V1		LV1		LV2		R1	
	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
(a) 95th, 90th, 10th percentiles, and overall average	**0.2928**	0.2385	**0.0091**	**0.00696**	**0.00449**	**0.00348**	3.2530	2.6215
(b) Overall average only	0.3005	0.2453	0.0111	0.00713	0.00461	0.00358	3.7852	2.6238
(c) 90th, 10th percentiles, and overall average	0.2933	**0.2373**	0.0093	0.00726	0.00458	0.00359	**3.1457**	2.5395
(d) 95th, 10th percentiles, and overall average	0.3099	0.2508	0.0094	0.00706	0.00460	0.00358	3.2974	**2.5147**
(e) 95th, 90th, 10th percentiles	0.3394	0.2731	0.0094	0.00729	0.00462	0.00359	3.3327	2.6090
(f) 95th, 90th percentiles, and overall average	0.3264	0.2637	0.0093	0.00727	0.00456	0.00354	3.5051	2.8205
(g) 95th, 90th, 50th, 10th, 5th percentiles, and overall average	0.3372	0.2717	0.0091	0.00723	0.00459	0.00360	3.3900	2.7168

TABLE III
WEIGHT OF EACH COMPONENT IN SMART-WAT FOR VARIOUS TARGETS

	V1	LV1	LV2	R1
AutoML	0.956	0.324	0.874	0.838
GPR	0.044	0.676	0.126	0.162

TABLE IV
RMSE IN REAL-WORLD CASE

Model	**V1**	**LV1**	**LV2**	**R1**
Average	0.3867	0.0133	0.00604	7.8202
IDW	0.3760	0.0130	0.00592	7.6969
VP	0.4011	0.0140	0.00648	7.7488
GPR	0.3434	0.0121	0.00553	5.8154
AutoML	0.3364	0.0123	0.00512	3.4428
SMART-WAT	**0.3335**	**0.0115**	**0.00503**	**3.3139**

values for each shot. In the GPR model, we trained using the ten sampled WAT shots from the same lots. Concurrently, we used the ten shots of sampled WAT data from the prior lots to construct the AutoML model. The implementation details for these two models are identical to the experiment described in section V-A. For the ensemble model, we utilized linear regression with the parameters trained on prior lots.

Based on the results shown in Table IV and V, we can make the following observations:

TABLE V
PREDICTION MAE IN REAL-WORLD CASE

Model	**V1**	**LV1**	**LV2**	**R1**
Average	0.2804	0.01061	0.00408	6.5507
IDW	0.2741	0.01027	0.00396	6.4752
VP	0.3054	0.01093	0.00467	6.4568
GPR	0.2462	0.00923	0.00355	4.5829
AutoML	0.2380	0.00897	0.00337	2.7130
SMART-WAT	**0.2352**	**0.00846**	**0.00326**	**2.5956**

Experimental Results:

1) SMART-WAT consistently outperforms all prediction targets. Even when AutoML shows competitive results, particularly with R1, SMART-WAT achieves lower error rates, demonstrating its effectiveness.
2) The outcomes align with section V-D. While AutoML's performance dropped for LV1 and GPR for R1, SMART-WAT maintained strong performance. This underscores its robustness and reliability, even when individual components may not perform as well.
3) Traditional IDW and VP methods have higher error rates. This underlines the advantages of machine learning, particularly SMART-WAT, in handling cross-wafer prediction in semiconductor manufacturing.

In conclusion, these findings show the robustness and superior performance of SMART-WAT in real-world scenarios, establishing its potential as a effective tool for estimating untested WAT results in the semiconductor manufacturing industry.

VI. CONCLUSION

In this paper, we propose SMART-WAT, a novel predictive framework designed to enhance Wafer Acceptance Test estimation accuracy in semiconductor manufacturing. As the semiconductor industry continues to evolve, the need for precise process monitoring grows ever more critical. Therefore, our work addresses a significant gap in the current manufacturing process by providing a comprehensive method for accurately estimating WAT results for untested shots based on existing sparse information. In summary, SMART-WAT provides an automatic and reliable tool for improving the overall WAT estimation quality in semiconductor manufacturing.

ACKNOWLEDGMENT

This work was supported by National Science and Technology Council under Grants 112-2221-E-006 -150 -MY3.

REFERENCES

[1] X. Li, R. R. Rutenbar, and R. D. Blanton, "Virtual Probe: A Statistically Optimal Framework for Minimum-Cost Silicon Characterization of Nanoscale Integrated Circuits," *Proceedings of the International Conference on Computer-Aided Design (ICCAD '09)*, San Jose, California, 2009, pp. 433–440, DOI: 10.1145/1687399.1687481.

[2] F. Liu, "A General Framework for Spatial Correlation Modeling in VLSI Design," *2007 44th ACM/IEEE Design Automation Conference*, 2007, pp. 817–822.

[3] N. Kupp, K. Huang, J. Carulli, and Y. Makris, "Spatial estimation of wafer measurement parameters using Gaussian process models," *2012 IEEE International Test Conference*, 2012, pp. 1–8, DOI: 10.1109/TEST.2012.6401545.

[4] N. Kupp, K. Huang, J. M. Carulli, and Y. Makris, "Spatial correlation modeling for probe test cost reduction in RF devices," *2012 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, 2012, pp. 23–29.

[5] A. Ahmadi, K. Huang, S. Natarajan, J. M. Carulli, and Y. Makris, "Spatio-temporal wafer-level correlation modeling with progressive sampling: A pathway to HVM yield estimation," *2014 International Test Conference*, 2014, pp. 1–10, DOI: 10.1109/TEST.2014.7035325.

[6] D. Shepard, "A Two-Dimensional Interpolation Function for Irregularly-Spaced Data," *Proceedings of the 1968 23rd ACM National Conference (ACM '68)*, 1968, pp. 517–524, DOI: 10.1145/800186.810616.

[7] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*, Adaptive Computation and Machine Learning series, MIT Press, 2006, ISBN: 026218253X.

[8] N. Erickson et al., "AutoGluon-Tabular: Robust and Accurate AutoML for Structured Data," *arXiv preprint arXiv:2003.06505*, 2020.

[9] K. Huang, N. Kupp, C. Xanthopoulos, J. M. Carulli, and Y. Makris, "Low-Cost Analog/RF IC Testing Through Combined Intra- and Inter-Die Correlation Models," *IEEE Design & Test*, vol. 32, no. 1, pp. 53–60, 2015, DOI: 10.1109/MDAT.2014.2361721.

[10] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to Linear Regression Analysis*, John Wiley & Sons, 2021.

[11] S.-K. S. Fan, C.-W. Cheng, and D.-M. Tsai, "Fault Diagnosis of Wafer Acceptance Test and Chip Probing Between Front-End-of-Line and Back-End-of-Line Processes," *IEEE Transactions on Automation Science and Engineering*, vol. 19, no. 4, pp. 3068–3082, Oct. 2022, DOI: 10.1109/TASE.2021.3106011.

[12] C.-C. Wang and Y.-Y. Yang, "A Machine Learning Approach for Improving Wafer Acceptance Testing Based on an Analysis of Station and Equipment Combinations," *Mathematics*, vol. 11, no. 7, Article no. 1569, 2023, DOI: 10.3390/math11071569.

[13] H. Xu, J. Zhang, Y. Lv, and P. Zheng, "Hybrid Feature Selection for Wafer Acceptance Test Parameters in Semiconductor Manufacturing," *IEEE Access*, vol. 8, pp. 17320–17330, 2020, DOI: 10.1109/ACCESS.2020.2966520.

[14] K. C. -C. Cheng et al., "Machine Learning-Based Detection Method for Wafer Test Induced Defects," *IEEE Transactions on Semiconductor Manufacturing*, vol. 34, no. 2, pp. 161-167, May 2021, doi: 10.1109/TSM.2021.3065405.