A Few-Shot Machine Learning-Based OCD Metrology Algorithm with Anomaly Detection and Wafer-Level Data Augmentation

Minkyu Kim*[a],QHwan Kim[a], Kyu-Baik Chang[a], Jaehoon Jeong[a], Sunghee Lee[b], Seonghui Mo[b], Dahan Kang[b], Jinkook Park[b], Young-Seok Kim[b], Yongdeok Jeong[b], Dae Sin Kim[a]
[a]Computational Science and Engineering Team; [b]Metrology and Inspection Technology Team, Samsung Electronics Co., Hwasung, Gyeonggi, 18448, Korea

ABSTRACT

With the increasing complexity of 3-D semiconductor structures, the use of optical critical dimension (OCD) metrology has become a popular solution due to its accuracy and fast inference time. Machine learning has been widely adopted in this field to further improve the efficiency and precision of OCD metrology. Especially for high aspect ratio structures such as DRAM and VNAND, where the required computing power for physical modeling increases exponentially, the importance of machine learning with reference data is crucial. However, one significant challenge of the machine learning-based metrology under rapidly changing process condition is the limitation of available labeled data, which causes overfitting and decreases recipe reliability in the manufacturing process as the cost of wafer consumption increases. To utilize machine learning algorithms in mass production, the development of robust algorithms that can be optimized with few-shot data is required. In this paper, we propose a few-shot machine learning algorithm that includes i) wafer-level statistical information-based data augmentation and ii) anomaly detection to automatically remove data with measurement errors. The proposed algorithm shows superior accuracy, repeatability, and in-wafer uniformity compared to the benchmark algorithm in tests with manufacturing phase data. Additionally, this robustness can be sustained with the minimum amount of data in metrology, as only 9 reference training data are used on three design of experiment (DoE) wafers. The proposed optimized solution is expected to contribute to the reduction of measurement costs and production yields of highly complicated 3D semiconductor structures.

Keywords: OCD metrology, Few-Shot Machine Learning, Anomaly Detection, Data Augmentation

1. INTRODUCTION

The demand of machine learning (ML) in the semiconductor metrology is growing as the complexity of semiconductor structures increases [1-3]. As semiconductors become more integrated and perform better with technological advancements, the increasingly fine-tuned nano-structures pose a challenge in metrology [4]. It is important to ensure that the structures produced through processes have the appropriate dimensions. In each process step, an accurate measurement process must be performed to determine whether the Module Target Spec (MTS) has been manufactured within the standard value. Ellipsometry-based OCD measurement is widely used as a method to perform inline measurement without interfering with the chips being produced. OCD measurement predicts MTS by shooting polarized light onto a wafer chip and analyzing the intensity of each phase and wavelength of the light reflected from the repeating structure (Fig. 1a). It is necessary to match the OCD spectra generated for each process step with the actual semiconductor structure. Several methods are used in the process of matching spectra and structures. In some cases, various structural libraries are created through simulation or emulation, and the structure most similar to the measured spectrum is selected [5]. In other cases, a regression model between MTS and the spectra is trained using ML or DL methods. The preparation of structure data uses wafer-destructive method such as electron microscopy, which limits the amount of available training data. Taking Transmission Electron Microscope (TEM) as an example, usually only three images can be obtained from one wafer of 12-inch diameter. TEM requires thin samples to transmit electrons and uses a destructive technique. Therefore, only limited references can be obtained from TEM due to the increasing cost and typically available number of reference data are smaller than 100, which is not sufficient to train standard ML models. A larger amount of data is needed to train an ML structure prediction model, and as more data is available for learning, the accuracy of the structure prediction model improves. However, this causes more time and cost to obtain data for training the model. In real-world industries, due to time and cost constraints, it is prefered to use small amounts of structure data as learning references to train structure prediction ML models.

Metrology, Inspection, and Process Control XXXVIII, edited by Matthew J. Sendelbach, Nivea G. Schuch, Proc. of SPIE Vol. 12955, 129550U · © 2024 SPIE
0277-786X · doi: 10.1117/12.3009965

It is often believed that the problem of using small amounts of data can be easily solved by simply investing a lot of resources. On the other hand, the statistical problems that arise between OCD spectra and TEM are unavoidable issues that cannot be resolved even with the investment of resources, and they become an obstacle to measurement through the training of an accurate structural prediction model. Figure 1b shows a schematic diagram of the difference in size of the area where the OCD beam is irradiated to the sample and of the sample produced as a specimen for TEM imaging [6]. The size of the spot where the OCD beam is irradiated is an area with a diameter of approximately 50 um, and within this area, there are tens to hundreds of thousands of nanostructures. OCD measures MTS, a representative value of these repeating nanostructures. TEM is imaged by taking a sample at the same coordinates as the OCD spot. Since it is the same coordinate in the same wafer, it is assumed that the value represented by the OCD spectrum and the MTS obtained from TEM are the same. This is based on the assumption that the distribution of nanostructures within the OCD beam spot converges to 0 to form an almost perfect repeating structure. However, in reality, structural dispersion inevitably occurs due to various factors in the process, which causes statistical problems.

Figure 1c shows basic inferential statistics, which allow us to make inferences about a population based on a sample taken from it. Since it is mostly impossible to survey the entire population for which we want to obtain statistical information, we extract some samples from the group and calculate statistics on that sample to infer properties of the entire population. The confidence interval of the population mean changes depending on the size of the sample, and as the sample size increases, the confidence interval decreases, allowing for more accurate inference [7]. TEM images corresponding to data samples are collected to infer MTS information of a set of repetitive structures within the area where the OCD beam is irradiated. However, because the sample size is significantly smaller than the number of data in the population, the confidence interval for the inference cannot be sufficiently narrowed, so the sample mean may deviate significantly from the population mean that we are trying to infer. To reduce these errors, we need larger sample sizes, but this is often impractical due to the challenges of obtaining more TEM specimens. Errors in these training data cause a decrease in the prediction performance of the structural prediction model. As the number of data used for training becomes smaller, the impact of individual data errors becomes more significant. Therefore, it is important to properly identify and exclude outliers from the training data, especially those in the high dispersion sections.

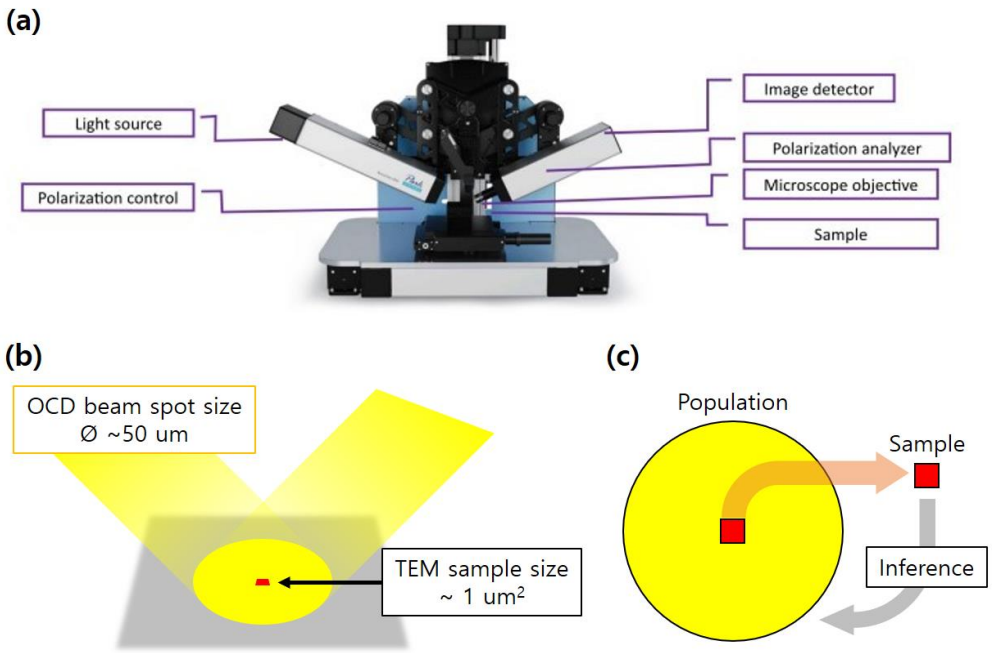

Figure 1. a) In an ellipsometer device, the polarized light emitted from the light source is reflected by the sample in the center, and then reaches the detector by passing through the polarizer. b) The schematic of the area where the OCD beam is reflected from the sample and the TEM sample area. Compared to the size of the spot where the beam is irradiated, the sample area that can be obtained from TEM is thousands of times smaller. c) A schematic diagram of general inferential statistics that infers a population based on samples extracted from the population. Population is the area where the OCD beam is irradiated, and the sample matches the TEM specimen.

In this paper, we propose an anomaly detection algorithm based on the relationship between OCD spectrum and MTS as a way to mitigate the decline in consistency of structure prediction ML models caused by statistical errors. Linearity occurs between the OCD spectra and MTS, and an algorithm was constructed based on the engineer's empirical process of classifying data that does not follow this trend as anomaly. By performing Principal Component Analysis (PCA) on the OCD spectra, we identified a vector that represents the trend of MTS on the two-dimensional principal component plane. We then detected data points that violated the linear trend between the vector and MTS, and extracted refined data for use in learning. To overcome the resulting decrease in the number of learning data, we introduced a wafer-level data augmentation method. By generating and adding virtual data using the median value of the spectra for each DoE wafer, we were able to double the amount of learning data and prevent overfitting. The proposed algorithm can train accurate and robust ML recipe with 9 points obtained from 3 spectra per 3 DoE wafers. An average of about 42% improvement in consistency and wafer uniformity was confirmed in four data sets targeting DRAM and FLASH memory.

2. WORKFLOW

The figure 2 shows the overall workflow of proposed ML using an anomaly detection and data augmentation algorithm. In the Data Preparation Step, training spectra are collected from ellipsometry measurement tool and corresponding training references are measured from the TEM. The proposed workflow includes the anomaly detection and data augmentation algorithms in the Training ML Recipe step. These two algorithms use the domain information about spectra measurements and semiconductor and can train robust and accurate recipe with few-shot data. All learning models presented here are Ridge, multinomial linear regression model which prevent overfitting through normalization of coefficients, was used. We used only minimum of 9 to a maximum of 55 data points for the learning, the use of DL is not possible.

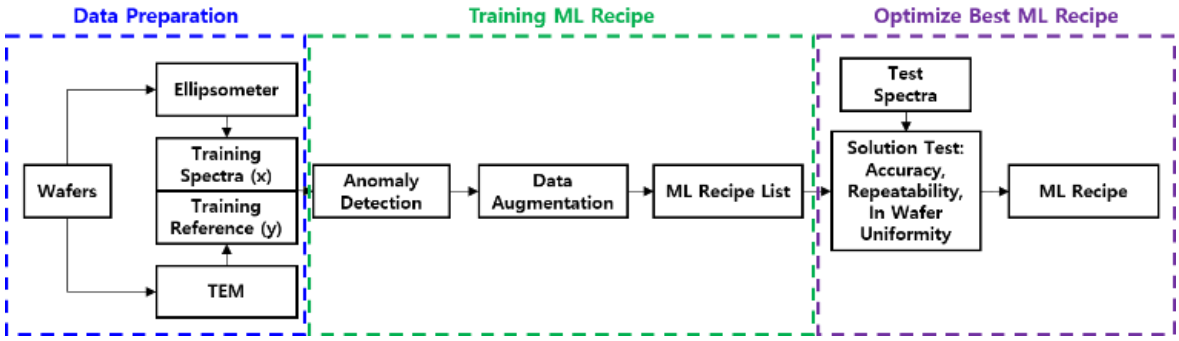

Figure 2. Overall workflow of proposed ML. Anomaly detection and data augmentation are performed before ML learning.

The information about the statistical errors that inevitably occur during the TEM measurement process was theoretically clearly defined. It was acknowledged that incorrect references would inevitably exist due to errors in population mean estimation that occur in the process of inferring the population from a sample that is significantly smaller than the population. Therefore, we first perform a data refinement process to detect samples with large errors and remove them from the learning data group. This was carried out by working engineers acquiring data empirically and excluding data with inconsistent spectral-TEM trends from learning. We also performed wafer-level data augmentation to alleviate the negative effects of data reduction due to anomaly refinement and increase the number of data used for learning. These two methods were performed before ML learning and were improved to learn theoretically more robust recipes. We will provide a detailed explanation of each method in the next two chapters.

3. SPECTRA-MTS TREND BASED ANOMALY DETECTION

3.1 Spectra-MTS Trend in PCA map

In the process of acquiring data for learning a structural prediction model, engineers have recognized the existence of unsuitable data and excluded it. Although there was no clear awareness of the statistical errors that occur between OCD spectra and TEM, efforts have been made to reduce these errors based on empirical knowledge. Among the two representative methods, the first is to simply exclude unusual data from the spectrum group, and the second is to utilize the characteristics of the structure generated by the process. Figure 3a is a representative example. It was widely known that the structure of the three sites measured within one wafer in the process had a chevron-shaped distribution. Based on this empirical knowledge, the engineer considered data that did not follow the distribution to be anomaly and excluded them from learning. So the two data that went against the expected trend (red dotted line) from the two wafers #14 and #16 were excluded and used (Fig. 3a). Although it was possible to improve the consistency of some structural prediction models through this empirical method, the following disadvantages exist for general use. First, it can only be used in structures in processes where certain trends occur widely. It cannot be used in cases where this trend changes due to process factors or does not appear clearly. Additionally, it is difficult to apply it easily in processes that require new measurement model setup because there is no domain knowledge. Second, there are no clear standards. Because there is no standard that can be confident that the measurement value is correct, it is impossible to objectively determine which data has a large error. Lastly, because the engineer's subjective judgment is required, a decision-making process involving human resources is necessary, and there is a risk that different results may appear depending on the engineer performing the task.

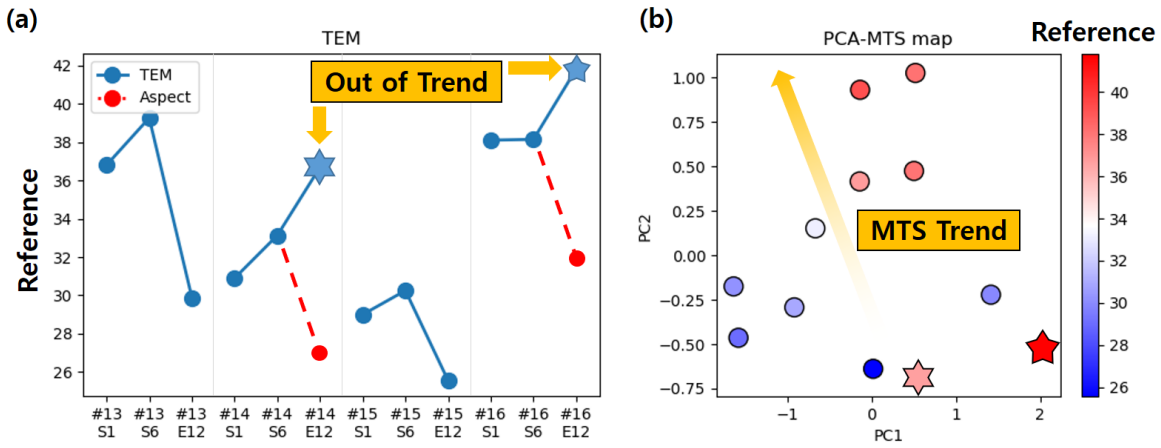

Figure 3. (a) Distribution of reference TEM values measured within each wafer (blue line). According to the engineer's domain knowledge, two data that went against the expected trend (red dotted line) from the two wafers #14 and #16 were excluded and used. (b) Reference data displayed in color on a two-dimensional PCA map of the spectra. While the reference value shows a tendency to increase in the direction of the indicated arrow, the two data points marked with star markers do not follow the trend.

Figure 3b is a graph reflecting the reference MTS value in color on the plane where the OCD spectra was analyzed through 2D PCA map. In general, PCA of spectra has been widely used to check whether data is suitable for learning a structure prediction model. When linearity between the spectrum and MTS is clearly evident, the trend of MTS on the PCA plane is also clearly evident, which means that the data is suitable for learning a structure prediction model. While the data also shows a clear trend between the spectrum and MTS, the two data marked with star markers do not follow this trend and go against the major color gradient. The data perfectly matched the anomalies that engineers had previously determined empirically, and through this, an existing series of refinement processes performed by engineers were established as an objective standard and laid the foundation for automation based on an algorithm.

3.2 Anomaly Detection

Anomaly detection analyzes the optimal trend of the reference that appears on Spectra's PCA map and goes through the process of selecting data that does not follow this trend as an anomaly. Figure 4a shows a simple flowchart of this process. One dataset produces a PCA-MTS map, and verification is performed using searching vectors in all 360-degree directions to find the optimal trend of MTS on this plane. Figure 4b is a PCA-MTS map graph with a searching vector applied in the central angle of θ direction. One index is calculated to analyze the linearity formed by the searching vector and MTS in the corresponding PCA map. First, the inner product of the two-dimensional principal component and the unit vector with the angle of θ is calculated to make a new coordinate system based on the distance when the position of each point is projected onto the corresponding searching vector. Therefore, the scaled projection distance, which is the most important index obtained from the PCA-MTS map, is calculated as

$$Projection\ Distance(PD) = PC \cdot serching\ vector \tag{1}$$

And

$$Scaled\ Projection\ Distance(SPD) = min(MTS) + \frac{max(MTS) - min(MTS)}{max(PD) - min(PD)} PD. \tag{2}$$

SPD is the min-max scaling value of PD to the MTS range for analysis of linearity between the position information of the main component and MTS. Figure 4c is the result of plotting a scattering graph with SPD as the x-axis and reference MTS as the y-axis, and the linearity between the two items can be visualized. In the graph, data that matches the main trend is distributed close to the y=x graph, and anomaly data that runs counter to the main trend is distributed far away. To index this, a value based on the remaining data and cosine distance is calculated for each data. Data within the cosine distance standard that can be determined by the user are classified as normal data used for learning, and data outside the standard are considered anomaly and removed.

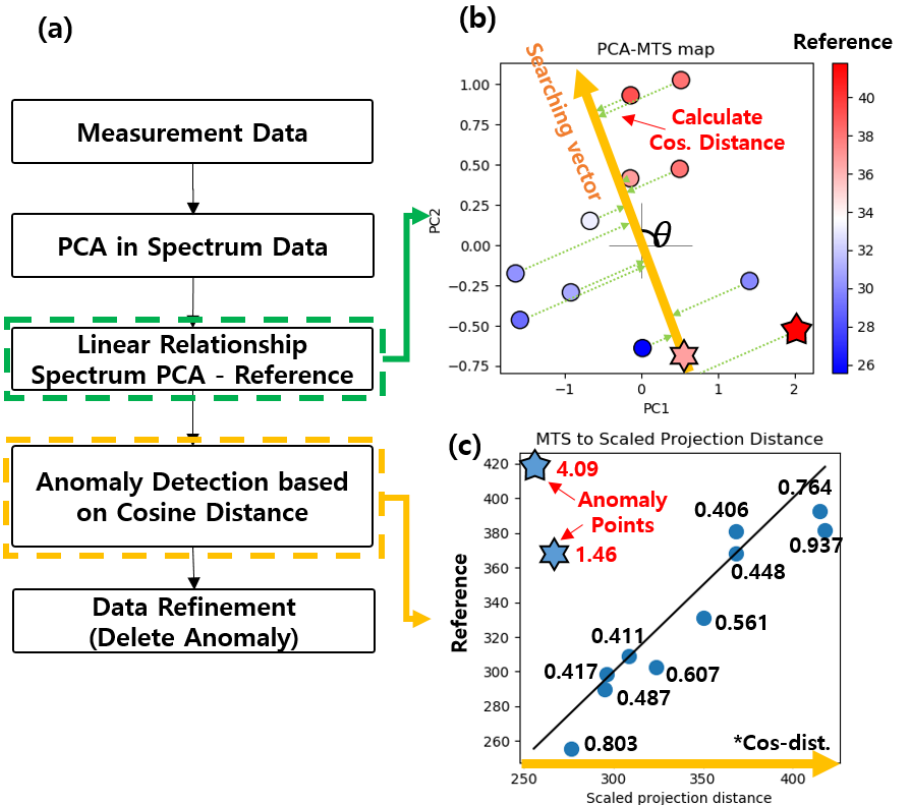

Figure 4. (a) Anomaly detection algorithm flow. (b) Spectra-reference relationship in principal component plane. Color denotes reference value. (c) Scaled cosine distance–reference plot. Two points lowest linearity are selected as the anomaly points.

The above series of processes is repeated in all 360-degree directions, and all cases of outliers that violate the trend in each direction are verified. The most optimal trend is selected as the trend when the average value of the cosine data of the remaining data sets after anomaly are removed is the minimum, and at this time, the anomaly are determined as the final anomaly. The data set from which anomaly have been removed is a data set that minimizes the error in population mean estimation that occurs between the OCD spectrum and TEM, and improves consistency by removing overfitting factors in the structure prediction model.

4. WAFER-LEVEL DATA AUGMENTATION

Training an ML model with few DoE data can easily lead to overfitting. Augmented data generated statistically can prevent overfitting. Statistical data is defined as the median value of the spectrum within the same wafer, and the reference is the same as the original. By training with statistical and original data together, the number of data is doubled, reducing the proportion of anomalies, preventing overfitting, and improving homeostasis. DoE wafers intentionally adjust process factors to expand the numerical range of structures included in the reference data, thereby allowing the learning model to learn based on information from a wider range of structures. The difference in MTS average values between wafers is generally larger than the distribution within a wafer. Therefore, augmenting the data on a per-wafer basis ensures that the data is distributed uniformly over the entire area without being biased to one side. Additionally, as the median value of each spectrum matches the original MTS value on a wafer basis, it acts as a kind of noise generation effect. This prevents overfitting to the small number of data given for learning and has the effect of responding to input data outside the learning area.

5. RESULTS

The four datasets of spectra-reference pair are prepared to test the accuracy and robustness of proposed ML algorithm. Each datasets correspond to the key structure parameters in DRAM and VNAND devices which is hard to make a physical model with a simulation respectively. The number of reference data obtained from DoE wafers varies between 9 ~ 55 pts. The amount of reference is small and specialized machine learning algorithm developed for few-shot data is required. As a benchmark ML, to capture the effect of anomaly detection and data augmentation, we exclude them in proposed ML and only use PCA and Ridge algorithms.

Table 1. R2 accuracy of propose ML for four datasets.

	Data Points	**Data Points After Anomaly Detection**	**R2 (Proposed ML)**	**R2 (Benchmark ML)**
Dataset1	**55 pts**	**43 pts**	**0.851**	0.589
Dataset2	**35 pts**	**33 pts**	**0.878**	0.829
Dataset3	**12 pts**	**10 pts**	**0.989**	0.881
Dataset4	**9 pts**	**8 pts**	**0.741**	0.359

Table 1 shows R2 accuracy of proposed ML and benchmark ML. Proposed ML approach improves the accuracy of all four datasets. Because Dataset1 and Dataset2 secured a relatively sufficient number of learning data points, it was possible to separate the test set for verification of the structural learning model, and the indicated R2 indicates the consistency of the test set. In the case of Dataset3 and Dataset4, since they only have about 10 learning data points, it is impossible to separate

the test set from them and proceed with verification, so the entire data was used for learning, and Train R2 was confirmed. In particular, Dataset4 is the evaluation result of a blind data set measured several months after the recipe was applied to the line and the measurement was performed. Despite changes in various processes that continuously affect the semiconductor structure and the possibility of recipe deterioration over time, the recipe actually maintains an R2 of over 0.7 for the blind test set. This means that a more consistent structural prediction model was able to be learned through training data refinement and augmentation. In the recipe learning stage of Dataset4, the train R2 of the proposed ML was 0.954, but the train R2 of the benchmark ML was 0.97, indicating that it was overfitted for one data point with a large error in the learning stage.

Detailed accuracy and in wafer uniformity profiles of Dataset3 are shown figure 5. The distribution of in-wafer uniformity as well as accuracy of proposed ML improves, which indicates that the overfitting is reduced and model robustness is confirmed.

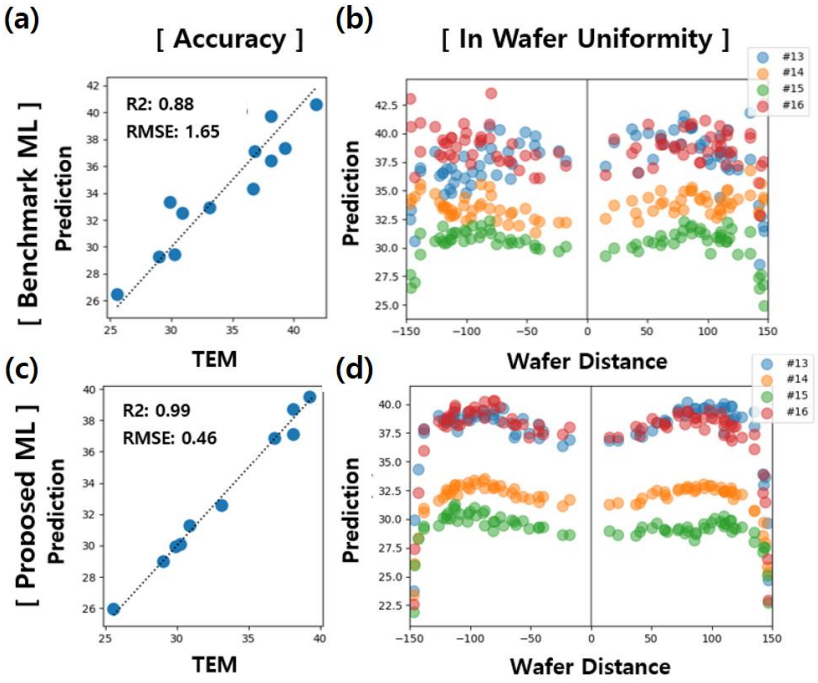

Figure 5. (a), (b) TEM-prediction graph of benchmarked ML and inference result graph for the x-direction wafer distance of a wafer full shot. (c), (d) TEM-prediction graph of proposed ML and inference result graph for the x-direction wafer distance of a wafer full shot. Compared to the benchmarked ML, the proposed ML shows more stable in wafer uniformity results even though the number of data used for learning through the refinement process is 2 less.

Figures 5a and 5c show the train data consistency of benchmark ML and proposed ML, respectively. Since two of the training data of the proposed ML were determined to be outliers, they were trained with the removed data, and since there is no separate test data set, verification can be questioned. However, the question can be resolved by looking at figures 4b and 4d, which are graphs plotting the results inferred by each ML for the OCD spectra measuring the entire area of four DoE wafers against the x-direction wafer distance. In the results inferred using proposed ML compared to the results inferred using benchmark ML, the gap between DoE wafers occurs more clearly, and the tendency for wafer distance also occurs clearly. This is a distribution of a structure that is believed to occur more appropriately in the structure formed through the process, meaning that the proposed ML was not overfitted for the 10 learning data points.

6. CONCLUSION

In this study, we demonstrate the effectiveness of our proposed ML approach with few-shot data on the OCD metrology. Our approach is applied to four different use cases, and achieves R2 performances from 0.85 to 0.99, which are satisfactory for the criterion of the semiconductor manufacturing. In addition, we show that the proposed method not only improves the consistency of learning, train, and blind data, but also enables the learning of a more consistent model by obtaining a more uniform distribution for the inference results of the wafer full shot. To improve the performance of the structure prediction model, it is essential to understand the statistical errors that may arise during the acquisition of training data and to apply appropriate preprocessing techniques to overcome them. We expect that this study could contribute to the improvement of inferential model recipes and reduction of total production costs based on ML in mass production environments.

7. ACKNOWLEDGEMENTS

We greatly appreciate the fruitful discussion and data support from Computational Science and Engineering Team and Memory Metrology and Inspection Technology Team, Samsung Electronics Co

REFERENCES

[1] F. J. Wong, Y. Hao, W. Ming, P. Žuvela, P. Teh, J. Shi, and J. Li, “Methods to overcome limited labeled data sets in machine learning-based optical critical dimension metrology,” SPIE Advanced Microlithography, Proc. SPIE, vol. 11611, 2021.Booth, N. and Smith, A. S., [Infrared Detectors], Goodwin House Publishers, New York & Boston, 241-248 (1997).
[2] I. Kim, S. Gwak, Y. Bae, T. Jo, “Optical spectrum augmentation for machine learning powered spectroscopic ellipsometry” Optics Express, 30, 16909 (2022)
[3] Q. Kim et al., “A simulation physics-guided neural network for predicting semiconductor structure with few experimental data” Solid-State Electronics, 201, 108568 (2023)
[4] H. Choi, et al., “Sensitivity enhancement in OCD metrology by optimizing azimuth angle based on the RCWA simulation, Solid-State Electronics”, Volume 200, 108574 (2023)
[5] B. Ahn et al., "Improvement of on-cell metrology using spectral imaging with TCAD modeling", Solid-State Electronics, Volume 201, 108578 (2023)
[6] Wurstbauer, U. et al., “Imaging ellipsometry of graphene. Applied Physics Letters”, 231901 (2010)
[7] Kumar, S. “Improved estimation of population mean in presence of nonresponse and measurement error”. J Stat Theory Pract 10, 707–720 (2016).