Solid-State Electronics 201 (2023) 108568

Contents lists available at ScienceDirect

Solid State Electronics

journal homepage: www.elsevier.com/locate/sse

A simulation physics-guided neural network for predicting semiconductor structure with few experimental data

QHwan Kim [a,*], Sunghee Lee [a], Ami Ma [a], Jaeyoon Kim [a], Hyeon-Kyun Noh [a], Kyu Baik Chang [a], Wooyoung Cheon [a], Shinwook Yi [a], Jaehoon Jeong [a], BongSeok Kim [b], Young-Seok Kim [b], Dae Sin Kim [a]

[a] Computational Science and Engineering Team, Samsung Electronics Co., Hwasung, Gyeonggi 18448, Republic of Korea
[b] Memory Metrology and Inspection Technology Team, Samsung Electronics Co., Hwasung, Gyeonggi 18448, Republic of Korea

ARTICLE INFO

The review of this paper was arranged by "Francisco Gamiz"

Keywords:
Optical spectrum
Critical dimensions
Ellipsometry
RCWA simulation
Physics-guided neural network

ABSTRACT

Prediction of semiconductor Critical Dimensions (CDs) from ellipsometry requires the machine learning model. However, proper training of a machine learning model is challenging because the measurement process of typical experimental CD data, which is mostly carried out with transmission electron microscopy (TEM), is a time- and cost-consuming process. To obtain a robust machine learning model with few experimental data, we propose a physics-guided neural network (PGNN) architecture. PGNN extracts spectrum-CD physics from simulation data and constructs physics-guided loss function for guiding the model optimization. The proposed algorithm has superior performance compared with other baseline algorithms and can be properly trained only with small experimental CD data, including label noise.

1. Introduction

As the semiconductor structure becomes complex and the length scale shrinks, the accurate estimation of ellipsometric measurement-semiconductor CD relationships with machine learning is crucial [1]. However, in an industrial field, the experimental data is minimal and include noise because measurement of the CD, which requires electron microscopy such as TEM [2], is a time- and cost-consuming process, while the spectrum can be relatively easily measured from optical metrology such as ellipsometry [3,4]. In this paper, we propose an end-to-end two-step PGNN algorithm [5], which uses a simulation-based model to train experimental data. In the first step, with the simulation data, we train the neural network, which can represent the physics function describing a general spectrum-CD relationship. In the second step, we train the experimental model whose loss function is guided by physics obtained from the first step. The proposed algorithms can avoid overfitting induced by small-sample and provide higher prediction accuracy than other benchmark algorithms.

2. Methods

We prepare six in-line semiconductor datasets of DRAM, which are composed of ellipsometry spectrum (x)-TEM CD (y) measurements (Table 1). Each dataset corresponds to a different CD from the sequential fabrication process. Train and test datasets are collected from different lots with different process conditions. We first use technical computer-aided design (TCAD) [6] with physical models and parameters such as etch rate to fit the simulation structure with TEM measurements. In the second step, the spectra are calibrated with ellipsometry measurement by controlling material properties. The spectra are calculated with rigorous coupled wave analysis (RCWA) [7]. After both structure and spectra are calibrated, we build the simulation data by changing physical parameters, which spends a day for the whole calculation. Most datasets include only 9–15 experimental points, which are not sufficient to train a trivial machine-learning model.

Fig. 1 shows the schematic of the proposed PGNN model. The model is composed of two sub-models, $f(\bullet)$ and $g(\bullet)$, which train the simulation and experimental data, respectively. The two sub-model structures are the same and composed of 1D convolutional layers, flattened layers, and fully-connected layers. The 1D convolutional layers are used

* Corresponding author.
E-mail address: qhwan.kim@samsung.com (Q. Kim).

https://doi.org/10.1016/j.sse.2022.108568

Available online 21 December 2022

Table 1
Details of prepared six benchmark datasets. The blue color denotes the number of data points. Train and test datasets are prepared using TEM. The simulation dataset is obtained via the RCWA of the TCAD structure.

	DATA1	DATA2	DATA3	DATA4	DATA5	DATA6
# of Train	12	12	293	15	9	9
Avg. (nm)	213.89	257.23	28.77	130.39	29.12	84.01
Std. (nm)	11.07	13.7	3.11	10.19	2.33	7.9
# of Test	18	18	123	6	10	14
Avg. (nm)	222.32	254.76	26.83	130.59	33.42	79.29
Std. (nm)	9.56	10.9	2.16	7.17	1.38	4.47
# of Simulation	1906	1906	1988	1988	1988	1988

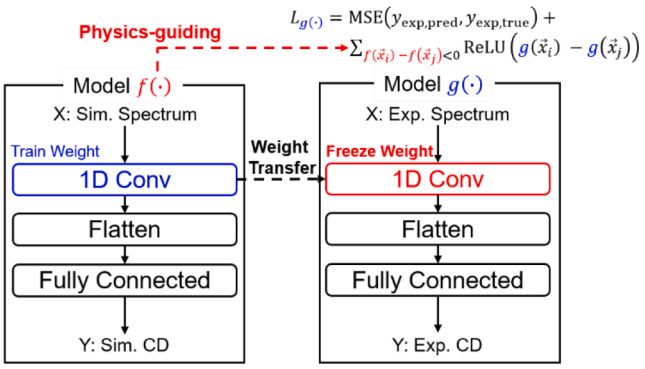

Fig. 1. A schematic of proposed PGNN architecture.

to obtain the spectrum representation, and the fully-connected layers act as a classifier. L2 regularization and dropout are used to regularize the model training.

$f(\bullet)$ is trained with the simulation data and represents the general spectrum-CD relationship. The trained model $f(\bullet)$ affects the training procedure of $g(\bullet)$ with experimental data in two ways. First, the weights of the $f(\bullet)$ 1D convolutional layers are transferred to the model $g(\bullet)$ and are frozen during further training. Second, $f(\bullet)$ provides a guideline to $g(\bullet)$ as a form of the loss function, which is defined [5] as follows

$$L_{g(\bullet)} = L_{\mathrm{MSE}} + L_{\mathrm{PHYS}} \quad = \mathrm{MSE}\left(y_{\mathrm{pred}}, y\right) + \sum_{f(x_i)<f\left(x_j\right)} \mathrm{ReLU}\left(g(x_i) - g\left(x_j\right)\right)$$

where $\mathrm{MSE}(\bullet)$ denotes mean squared error and $\mathrm{ReLU}(\bullet)$ denotes rectified linear unit function. The L_{MSE} represents typical empirical loss, while the L_{PHYS} represents regularization loss guided by physics from the $f(\bullet)$. L_{PHYS} directly guides to regularize $g(\bullet)$ to follow the spectrum-CD relationship defined from $f(\bullet)$. Note that we do not use any explicit physical equation in L_{PHYS}. Instead, we consider $f(\bullet)$ as the physical model of spectrum-CD relations because simulation data is already defined to follow Maxwell's equation. Also note that the additional training time from L_{PHYS} is about 20 % of training time only with, L_{MSE} which is acceptable for practical use.

For comparison, we compare our model against three baseline algorithms for regression: partial least squares (PLS), ridge (Ridge), and neural network (NN). PLS and Ridge are the linear regression models with regularizers and only train $g(\bullet)$ with experimental data. The NN model trains both $f(\bullet)$ and $g(\bullet)$, which use only empirical loss L_{MSE} and weight transfer of 1D convolution layers for training $g(\bullet)$.

3. Results

Table 2 lists the root mean square error (RMSE) of each algorithm on the train and the test data. Boldface represents the best-performed algorithm. The PGNN outperforms for predicting test data except for DATA2, which still shows a similar RMSE to Ridge. Note that PLS, Ridge, and NN perform better on train datasets, which indicates they are not generalized for unseen spectrum data and suffered from overfitting. Even though NN uses a similar architecture to PGNN, it does not consider L_{PHYS} and performance is comparable with that of PLS and Ridge. It denotes that the physics-guiding mechanism of PGNN is crucial for overcoming few-data problems and increasing the prediction accuracy of unseen spectrum data. The average test set RMSE of PGNN is 2.22 nm, which is improved by 51 % relative to the average RMSE of other baseline algorithms. Fig. 2 shows the scatter plots of PGNN and Ridge prediction results, where the X and Y axes indicate the predicted and true CDs, respectively. As shown in Fig. 2, PGNN shows significant improvements, especially in the test data prediction accuracies. Comparison with other baseline algorithms (PLS, NN) shows similar results. In the same wafer, spectrum-CD relationships can be approximated linear and the Ridge can well fit train data, but if the test spectrum of another lot, where the process condition is significantly changed, is entered, its locality easily breaks and the test prediction fails. It is well shown in the DATA6 result, where the test data prediction shows a constant offset from the train data prediction. Because the PGNN uses physics covering a broad range of parameters provided by the simulation data, it shows an accurate prediction of the test dataset and robust process condition variation.

Fig. 2 True and predicted spectrum – CD relationships of DATA6. Spectrums are decomposed by PCA algorithms and CDs are represented by colors. Note that increasing CD directions of Ridge and PGNN prediction are different. Only PGNN can correctly predict the true CD distribution component analysis (PCA) and CDs are represented by color maps. Fig. 3 shows that the Ridge correctly predicts only around the experimental data points. The physics of the spectrum-CD relationship can be simply denoted as the direction of increasing CD in the PCA map. The increasing CD direction from Ridge prediction is different to true value. It indicates that even though the Ridge can predict simulation data, this model violates the general physics function provided from simulation. However, PGNN algorithm can reproduce spectrum-CD relationships of simulation data correctly. It indicates that PGNN can make more physics-reliable model and is more robust to the unseen spectrum data if it is located in the broad range covered by simulation data.

To test the model robustness against various field conditions, we vary the data quality and estimate the change of RMSE in given algorithms.

Table 2
RMSE performance comparison of each algorithm on the train and test datasets. Boldface represents the best RMSE of the given dataset.

	Train Set RMSE (nm)				Test Set RMSE (nm)			
	PLS	Ridge	NN	PGNN	PLS	Ridge	NN	PGNN
DATA1	**0.583**	0.591	0.654	2.183	5.911	9.105	6.610	**3.073**
DATA2	1.13	**0.369**	1.085	2.585	3.442	**3.141**	3.162	3.155
DATA3	0.987	0.983	**0.982**	1.155	3.329	3.340	3.500	**1.847**
DATA4	1.03	1.031	**1.029**	1.39	3.099	3.095	2.954	**2.461**
DATA5	0.478	0.501	0.487	**0.269**	4.245	2.384	4.302	**1.647**
DATA6	0.407	0.407	0.408	**0.335**	6.241	6.292	6.325	**1.138**
DATA7	0.727	0.722	0.752	**0.437**	5.251	5.224	5.397	**3.206**

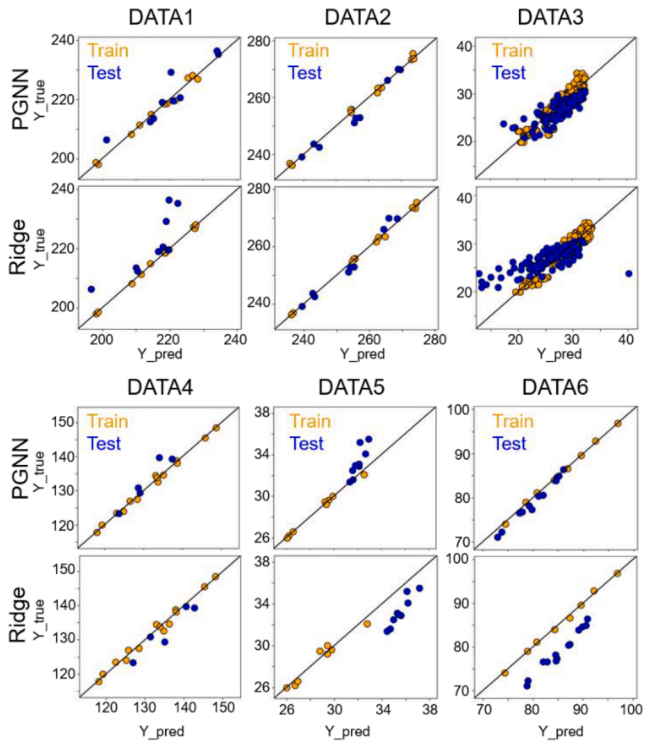

Fig. 2. Prediction results of Ridge and PGNN on the train (yellow dot) and test (blue) datasets. The X and Y axes represent the predicted and true CD values, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

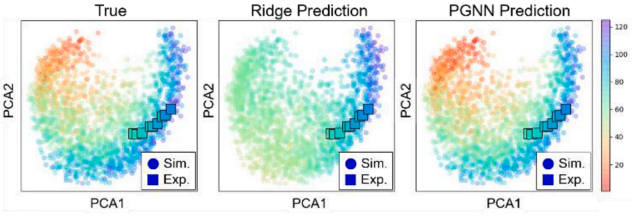

Fig. 3. Visualizes true and predicted spectrum-CD relationships of DATA6 with Ridge and PGNN. The dimension of spectrum features is reduced by principal.

First, we add the noise to the experimental CD values because the interpretation of TEM results is guided by human observation, which can exert bias on the estimation. We assume that the noise is random and controlled by the Label Noise $\times N(0, 1^2)$, where Label Noise is defined to be the percentage ratio of CD average of the given dataset and $N(0, 1^2)$ is the standard normal distribution. Fig. 4 shows the RMSEs of all algorithms as a function of the Label Noise level. Because the noise is chosen randomly, we carry out training 20 times for the given label noise and algorithms and average them to minimize variation in RMSE. Only proposed PGNN maintains relatively good performance regardless of varying data quality.

Second, we reduce the number of train experimental data to 70 % and estimate the change in performance. We average 50 training results for randomly chosen training data for a single result. Fig. 5 shows the PGNN show superior performance when the training data is reduced. Note that DATA1, DATA2, DATA5, and DATA6 have only 9–12 train data points. If we reduce 70 % of them, we use only 3 data points, which

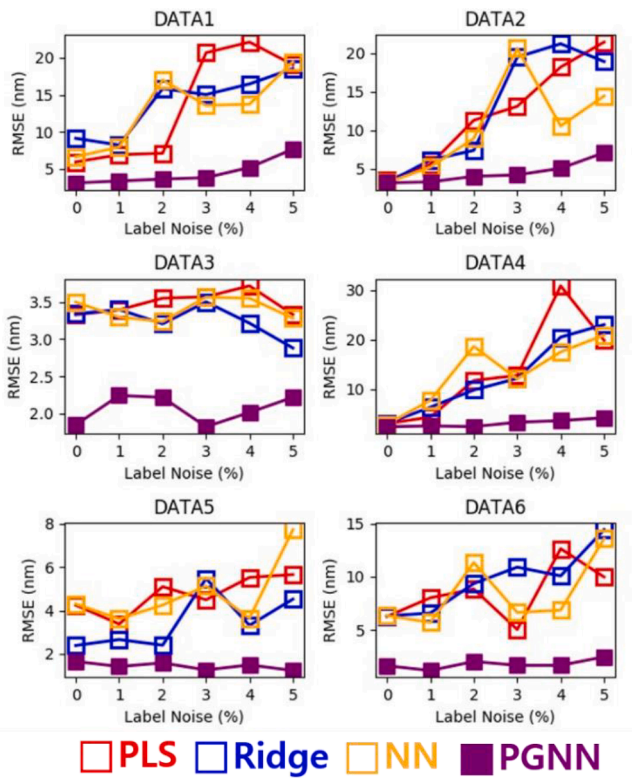

Fig. 4. The RMSE of algorithms as a function of the Gaussian label noise scale which is added to the train data.

requires only one wafer. The red-shaded regions of Fig. 5 show that the PGNN still can train a competitive prediction model with only one wafer, which can significantly reduce time- and cost-consumption during the metrology process in the semiconductor industry.

4. Conclusion

In this letter, we propose the PGNN algorithm, which uses the spectrum-CD relationship extracted from a simulation-based model to train a recipe with few experimental data. The proposed algorithms can avoid overfitting induced by small-sample and provide higher prediction accuracy than other benchmark algorithms. PGNN works properly even with three experimental data points from only one wafer, which can reduce the cost of data preparation drastically. The proposed algorithm is currently being used for the analysis of ellipsometry data of the DRAM, Logic, and Flash products. Furthermore, the algorithm architecture of PGNN can be applied to other metrology subjects where the simulation data is prepared.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

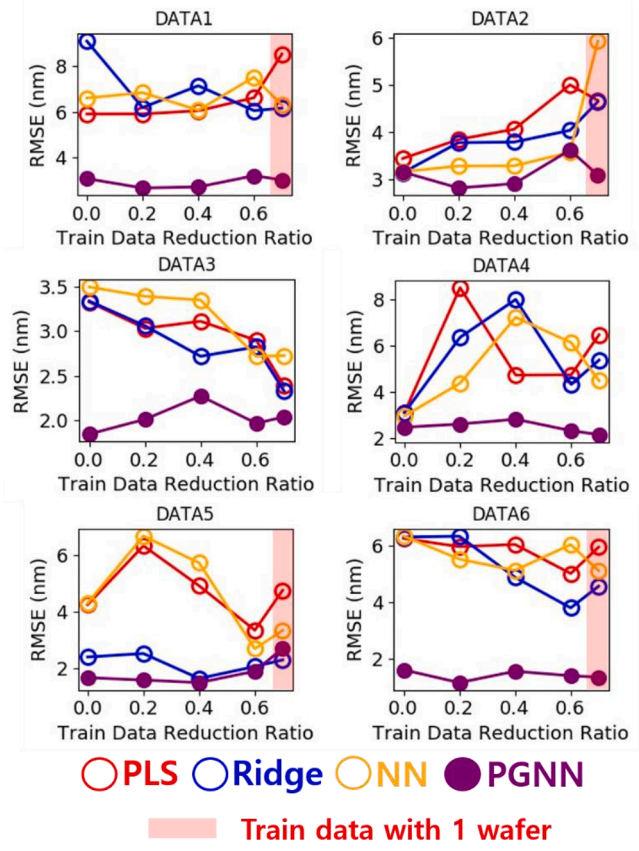

Fig. 5. The RMSE of the algorithms as a function of the train data reduction ratio. Red shaded region denotes the region where only one wafer is required to prepare the sample dataset. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

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] Liu J, Zhang D, Yu D, et al. Machine learning powered ellipsometry. Light Sci Appl 2021;10:55.
[2] Orji NG, Badaroglu M, Barnes BM, et al. Metrology for the next generation of semiconductor devices. Nat Electron 2018;1:532–47.
[3] Novikota T, De Martino A, Hatit SB, Drévillon B. Application of Mueller polarimetry in conical diffraction for critical dimension measurements in microelectronics. Appl Opt 2006;45:3688–97.
[4] Liu S, Chen X, Zhang C. Development of a broadband Mueller matrix ellipsometer as a powerful tool for nanostructure metrology. Thin Solid Films 2015;584:176–85.
[5] Daw A. Physics-guided Neural Networks (PGNN): An Application in Lake Temperature Modelling. 2017;*arXiv:1710.11431v3*.
[6] Schröter M, Rosenbaum T, Chevalier P, Heinemann B, Voinigescu SP, Preisler E, et al. SiGe HBT technology: future trends and TCAD-based roadmap. Proc IEEE 2017;105:1068–86.
[7] Huang HT, Terry FL. Spectroscopic ellipsometry and reflectometry from gratings (Scatterometry) for critical dimension measurement and in situ, real-time process monitoring. Thin Solid Films 2004;455:828–36.