Classification of Wafer Defects with Optimized Deep Learning Model

Koon Hian Ang[1], Koon Meng Ang[1], Mohd Rizon Bin Mohamed Juhari[1], Chin Hong Wong[2,3], Abhishek Sharma[4], Chun Kit Ang[1], Sew Sun Tiang[1,*], Wei Hong Lim[1,*]

[1]*Faculty of Engineering, Technology and Built Environment, UCSI University, Kuala Lumpur 56000, Malaysia*
[2]*Maynooth International Engineering College, Maynooth University, Maynooth, Co Kildare, Ireland*
[3]*Maynooth International Engineering College, Fuzhou University, Fujian, 350116, China*
[4]*Department of Computer Science and Engineering, Graphic Era Deemed to be University, Dehradun 248002, India*
E-mail: 1001850063@ucsiuniversity.edu.my, 1001436889@ucsiuniversity.edu.my, mohdrizon@ucsiuniversity.edu.my, chinhong.wong@mu.ie, abhishek15491@gmail.com, angck@ucsiuniversity.edu.my, tiangss@ucsiuniversity.edu.my, limwh@ucsiuniversity.edu.my

Abstract

Wafer defect inspection is one of the crucial semiconductor processing technologies because it can help to identify the surface defects in the process and eventually improve the yield. Manual inspection using human eye is subjective and long-term fatigue can lead to erroneous classification. Deep learning technology such as convolutional neural network (CNN) is a promising way to achieve automated wafer defect classification. The training of CNN is time consuming and it is nontrivial to fine tune its hyperparameters to achieve good classification performance. In this study, Arithmetic Optimization Algorithm (AOA) is proposed to optimize the CNN hyperparameters, such as momentum, initial learn rate, maximum epochs, L2 regularization, to reduce the burden brought by trial-and-error methods. The hyperparameters of a well-known pretrained model, i.e., GoogleNet, are optimized using AOA to perform wafer defects classification task. Simulation studies report that the AOA-optimized GoogleNet achieves promising accuracy of 91.32% in classifying wafer defects.

Keywords: Arithmetic optimization algorithms, wafer defects classification, convolutional neural networks, hyperparameters optimization.

1. Introduction

Emergence of Industrial Revolution 4.0 (IR4.0) era has led to the growing demands in semiconductor industries to produce sophisticated integrated circuit chips. More integrated circuit components are patterned and etched onto semiconductor wafers to meet various requirements of chips in terms of its accessing speed, lifespan, memory storage, size and etc. Likelihood of having manufacturing process-based defects on the wafer surface tends increase with the pressures of satisfying the growing demands of customer sides and this undesirable scenario can reduce the yields. One of the important steps used to address the manufacturing yield issue is to identify and classify the wafer defect patterns that can be associated with different steps of manufacturing process. Some typical issues of chip fabrication process include flow leakages, robot handoffs, contamination and etc. Engineers can improve the yield by locating the manufacturing problems of chips based on wafer defect patterns observed[1]. For most semiconductor industries, the wafer defect identification is still performed through human visual inspection. This manual process has undesirable drawbacks such as lack of objectivity and the high tendency of making false classifications due to long term fatigue issue.

To address the drawbacks of human inspection, an automated machine vision system incorporated with optimized deep learning model is designed for wafer defects classification in reliable manners. Convolutional neural network (CNN) is a popular deep learning method used to solve real-world problems[2],[3],[4],[5],[6],[7], including wafer defect classification[8],[9], motivated by its promising ability to learn the nonlinear relationships of input and output based on useful information extracted from raw data. Tremendous success of deep learning has driven the designs of various popular CNN architectures (e.g., AlexNet[10], ResNet-50[11], GoogleNet[12], VGG-16[13], etc.). Transfer learning is used to train these CNN networks to solve new tasks with lesser datasets and time. A crucial factor that governs the performance of pretrained network to solve new tasks is hyperparameter settings used during the training process.

Conventionally, CNN hyperparameters are manually tuned in trial-and-error basis but it is time-consuming. Given their strong global search ability, metaheuristic search algorithms (MSAs) inspired by various natural phenomena[15] (e.g., evolution theory, animal behaviors, physics principles and human activities) are used to solve many complex optimization problems[16],[17],[18],[19],[20],[21], including the hyperparameter tuning of CNN. Arithmetic optimization algorithm (AOA)[14] is an emerging MSA inspired by the distribution behaviors of four major arithmetic operators (i.e., addition, subtraction, multiplication and division). In this study, GoogleNet is selected as a pretrained network and trained with new datasets via transfer learning for solving wafer defect classification task. AOA is used to optimize four hyperparameters if CNN, i.e., momentum, initial learn rate, maximum epochs and L2 regularization. Performance of optimized CNN model in classifying wafer defects is investigated.

2. Related Works

2.1. *Conventional CNN and GoogleNet*

A typical CNN model consists of three different layers, i.e., convolutional layer, pooling layer and fully connected layer[22]. Convolution layer is mainly used to extract desired features from input images. Pooling layer is applied to reduce the size of the feature maps produced by the convolution layer. Subsequently, a fully connected layer acts as the classification module of a CNN model. Table 1 presents an example of a typical CNN model.

GoogleNet was proposed by Szegedy et al.[12] and it has 7 million parameters. The network architecture of GoogleNet consists of nine inception modules, four convolutional layers, four max-pooling layers, three average pooling layers, five full-connected layers, and three SoftMax layers for the main auxiliary classifiers in network. GoogleNet also has the dropout regularization and ReLU activation functions in its fully connected layers and convolutional layers.

Table 1. Architecture of a typical CNN.

Layer	Type of Layer	#Feature maps	Feature map size	Filter size
1	Input	1	14 × 14	-
2	Convolutional 1	6	7 × 7	5 × 5
3	Pooling 1	6	4 × 4	2 × 2
4	Convolutional 2	16	4 × 4	5 × 5
5	Pooling 2	16	2 × 2	2 × 2
6	Fully connected 1	1	120	-
7	Fully connected 2	1	84	-

2.2. *Basic AOA*

AOA was proposed by Abualigah, *et al.*[14] in year 2020 and its search mechanism was inspired by the distribution behaviors of four popular arithmetic operators, known as Addition (*A*), Subtraction (*S*), Multiplication (*M*) and Division (*D*). The search process AOA is implemented with three main phases, i.e., initialization, exploration and exploitation phases to solve an optimization problem.

During initialization phase, a set of possible solutions are generated randomly within the predefined boundaries of each dimensional component. A Math Optimizer Accelerated (*MOA*) that controls the search behavior of algorithm in either exploration phase or exploitation phase, is then calculated as follow:

$$MOA(C_{Iter}) = Min + C_{Iter} \times (\frac{Max - Min}{M_{Iter}}) \qquad (1)$$

where $MOA(C_{Iter})$ is the function value at *t*-th iteration; C_{Iter} is current iteration number; M_{Iter} is the predefined maximum number of iterations; *Min* and *Max* indicate the minimum and maximum values of accelerated function, respectively. If *MOA* is smaller than a random number with value between 0 to 1 that are produced by a uniform, the AOA population is assigned with exploration phase. Otherwise, exploitation phase is performed.

During exploration phase, both *M* and *D* operators are applied to increase the coverage of AOA population in search space with good solution diversity given their high-distributed values with high dispersion. The *d*-th dimension for the new position of each *i*-th AOA solution is updated in exploration phase as:

$$x_{i,j}(C_{Iter}+1) = \begin{cases} best_j \div (MOP + \epsilon) \times \left((UB_j - LB_j) \times \mu + LB_j\right), & r1 < 0.5 \\ best_j \times MOP \times \left((UB_j - LB_j) \times \mu + LB_j\right), & otherwise \end{cases} \qquad (2)$$

where $x_{i,j}(C_{Iter}+1)$ refers to the j-th dimension of i-th solution in next iteration of $C_{Iter}+1$, where $j=1,\dots,D$, $i=1,\dots,I$ with I refers to population size, $C_{Iter}=1,\dots,M_{Iter}$ with M_{Iter} refers to maximum iteration number of algorithm; $best_j$ is the j-th dimension of the best solution; UB_j and LB_j are the upper and lower boundary limits in the j-th dimension, respectively; $r1$ is a random number within 0 to 1 generated by uniform distribution; μ is a parameter used to adjust the range of search process. Given the values of C_{Iter} and M_{Iter}, a Math Optimizer Probability (*MOP*) can be formulated as:

$$MOP(C_{Iter}) = 1 - \frac{C_{Iter}^{\frac{1}{\alpha}}}{M_{Iter}^{\frac{1}{\alpha}}} \tag{3}$$

where $MOP(C_{Iter})$ is a function value at t-th iteration; C_{Iter} is the current number iteration; α is a sensitive parameter that used to define exploitation accuracy.

For exploitation phase, *S* and *A* operators are applied to fine tune the promising solution regions around global optimum by leveraging its high-dense results with low dispersion. The *d*-th dimension for the new position of each *i*-th solution is updated in exploitation phase as:

$$x_{i,j}(C_{Iter}+1) = \begin{cases} best_j - MOP \times \left((UB_j - LB_j)\times\mu + LB_j\right), & r2 < 0.5 \\ best_j + MOP \times \left((UB_j - LB_j)\times\mu + LB_j\right), & otherwise \end{cases} \tag{4}$$

where *r*2 is a random number within 0 to 1 generated by uniform distribution. The exploration and exploitation processes of AOA are repeated iteratively until the stopping criteria are satisfied. The best solution produced by AOA at the end of the optimization process is returned to solve the given problem.

3. Optimization of CNN Models Using AOA

3.1. *Dataset Preprocessing*

In this study, the WM-811K dataset is used to perform the training and evaluation processes of the optimized deep learning model. It is notable that the data distributed across all classes are heavily unbalanced. A total of 30,000 None defect images is extracted from the WM-811K dataset and compiled with all other defect images[7]. Table 2 summarizes the number of all defect and non-defect images observed in the newly formed dataset.

Table 2. Number of defect and non-defect images

Types of Defects	# Labeled Images
Center	4,294
Donut	555
Edge-Loc	5,189
Edge-Ring	9,680
Loc	3,593
Near-full	149
Random	866
Scratch	1,193
None	30,000

3.2. *AOA-Optimized GoogleNet*

Transfer learning used to train GoogleNet with updated WM-811K datasets presented in Table 2, enabling this pretrained network to solve wafer defects classification problems. In order to perform the new classification task effectively, several modifications are made on the hyperparamters of GoogleNet. Particularly, the size of its input layer is modified from the original value of 224 × 224 × 3 to 64 × 64 × 1. Besides that, the padding size of first convolution layer is changed from its original value of 3 × 3 × 3 × 3 to be same as the input size. Finally, the output size of fully connected layer is changed from its original value of 1000 to be 9 (i.e., defect types).

To further enhance the performance of GoogleNet in wafer defect classification task, AOA is implemented to optimize the hyperparameters of GoogleNet during the transfer learning process. Each *i*-th AOA solution vector of AOA is encoded with four decision variables known as momentum, initial learn rate, maximum epochs and L2 regularization, where their boundary limits are presented in Table 3. Fitness function used to evaluate the quality of each AOA solution can be defined based on the classification accuracy obtained. The overall framework of the proposed AOA-optimized GoogleNet used for wafer defect classification is described in Fig 1.

Table 3. Search range of four hyperparameters.

Hyperparameters	Lower Boundary	Upper Boundary
Momentum	0.5	0.9
Initial Learn Rate	0.01	0.1
Maximum Epochs	5	10
L2 Regularization	1×10^{-4}	5×10^{-4}

4. Performance Evaluations

4.1. *Simulation settings*

The pre-processed WM-811K dataset is randomly split for training, validating and testing purposes with the ratio of 70%, 10% and 20%, respectively. All the input images

Algorithm 1: AOA-Optimized GoogleNet

Input: N, D, UB_j, LB_j

```
Initialize C_Iter = 0;
for i =1 to I do
    Randomly generate solution x_i;
    C_Iter = C_Iter + 1;
end for
while C_Iter ≤ M_Iter do
    Decode the hyperparameters from x_i;
    Evaluate the accuracy f(x_i) of GoogleNet;
    Calculate MOP using Eq. (3);
    Calculate MOA using Eq. (1);
    for i =1 to I do
        for j =1 to D do
            if rand > MOA then
                Update x_i,j(C_Iter + 1) with
                Eq. (2);
            else
                Update x_i,j(C_Iter + 1) with
                Eq. (4)
            end if
        end for
        Decode the hyperparameters from x_i;
        Evaluate the accuracy f(x_i) of
        GoogleNet;
        Update x_i, f(x_i), best, f(best)
        C_Iter ← C_Iter + 1;
    end for
end while
```

Output: *best*

Fig.1 Pseudocode of optimizing GoogleNet using AOA.

are resized into 64 × 64 × 1. Besides that, the minimum batch size is set as 32.

4.2. *Performance Comparisons*

The performance of unoptimized GoogleNet and AOA-Optimized GoogleNet used to classify the wafer defects are evaluated and compared based on the values of recall, accuracy, precision, F1 score and area under curve (AUC). Recall measures the model's ability to detect positive samples. Accuracy reports the numbers of data are predicted correctly to the class it supposed to be. Precision refers to the accuracy of CNN in classifying a sample as positive. F1 score is the harmonic mean of CNN's precision and recall. AUC represents the area under receiver operating characteristic curve. Table 4 compares the classification performance of GoogleNet and AOA-Optimized GoogleNet in quantitative manner based on five performance metrics described. Qualitative performance analyses are also conducted based on the confusion matrices produced by unoptimized GoogleNet and AOA-Optimized GoogleNet as shown in Fig 2 and Fig 3, respectively.

Table 4 reports that the proposed AOA-optimized GoogleNet shows significantly better performance over the unoptimized GoogleNet, in terms of recall, accuracy, precision, F1 score and AUC, when solving the wafer defects classification problems. According to Fig 2 and Fig 3, GoogleNet without hyperparameters optimization is reported to suffer severe issue by misclassifying the scratch defect to be Loc defects (34.3%) and non-defected (50.2%). This misclassification behavior has been rectified by optimizing the hyperparameters of GoogleNet using AOA as illustrated in Fig 3.

Table 4. Quantitative performance comparison.

Performance	Unoptimized-GoogleNet	AOA-Optimized GoogleNet
Recall	0.7300	**0.8676**
Accuracy	0.8936	**0.9132**
Precision	0.8185	**0.8523**
F1 Score	0.7000	**0.8566**
AUC	0.7735	**0.8151**

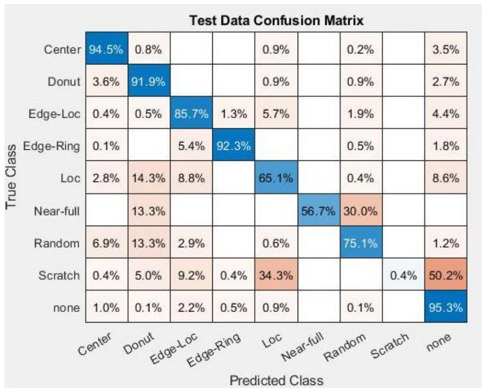

Fig.2 Confusion matrix of unoptimized GoogleNet.

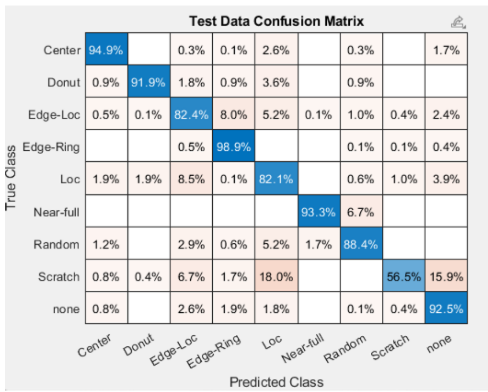

Fig.3 Confusion matrix of AOA-Optimized GoogleNet.

5. Conclusions

An optimized deep learning model is proposed to solve wafer defect classification problems more effectively. In particularly, transfer learning process is performed on the GoogleNet based on the wafer defects datasets and its hypermeter settings are optimized by AOA. Simulation results show that AOA-optimized GoogleNet has better performance than unoptimized GoogleNet to solve wafer defects classification by producing more competitive values of recall, accuracy, precision, F1 score and AUC.

Acknowledgements

This research is supported by Ministry of Higher Education Malaysia (MOHE) under the Fundamental Research Grant Scheme with the project codes of FRGS/1/2019/TK04/UCSI/02/1 and FRGS/1/2020/TK0/UCSI/02/4. This research is also supported by UCSI Research Excellence & Innovation Grant (REIG) with project code of REIG-FETBE-2022/038.

References

1. Chien, J.C., Wu, M. T., Lee, J. D. Inspection and Classification of Semiconductor Wafer Surface Defects Using CNN Deep Learning Networks. *Applied Sciences*, **10**, 5340 (2020).
2. Berghout, T., Benbouzid, M., Muyeen, S. M., Bentrcia, T., Mouss L. Auto-NAHL: A Neural Network Approach for Condition-Based Maintenance of Complex Industrial Systems. *IEEE Access* **9**, 152829 – 152840 (2021).
3. Berghout, T., Benbouzid, M., Muyeen, S. M. Machine learning for cybersecurity in smart grids: A comprehensive review-based study on methods, solutions, and prospects. International Journal of Critical Infrastructure Protection, **38**, 100547 (2022).
4. Alrifaey, M., Lim, W. H., Ang, C. K. A Novel Deep Learning Framework Based RNN-SAE for Fault Detection of Electrical Gas Generator. *IEEE Access* **9**, 21433-21442 (2021).
5. Alrifaey, M. *et al.* Hybrid Deep Learning Model for Fault Detection and Classification of Grid-Connected Photovoltaic System. *IEEE Access* **10**, 13852-13869 (2022).
6. Jdid, B., Hassan, K., Dayoub, I., Lim, W. H. & Mokayef, M. Machine Learning Based Automatic Modulation Recognition for Wireless Communications: A Comprehensive Survey. *IEEE Access* **9**, 57851-57873 (2021).
7. Jdid, B., Lim, W. H., Dayoub, I., Hassan, K. & Juhari, M. R. B. M. Robust Automatic Modulation Recognition Through Joint Contribution of Hand-Crafted and Contextual Features. *IEEE Access* **9**, 104530-104546 (2021).
8. Batool, U., Shapiai, M. I., Tahir, M., Ismail, Z. H., Zakaria, N. J., Elfakharany, A. A Systematic Review of Deep Learning for Silicon Wafer Defect Recognition, *IEEE Access*, **9**, 116572-116593 (2021).
9. Bhatt, D., et al. CNN Variants for Computer Vision: History, Architecture, Application, Challenges and Future Scope. *Electronics*, **10**, 2470, (2021) .
10. Krizhevsky, A., Sutskever, I., Hinton., G. E. ImageNet Classification with Deep Convolutional Neural Networks. *Advances in neural information processing systems* (2012).
11. He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770-778 (2016).
12. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1-9 (2015).
13. Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).
14. Abualigah, L., Diabat, A., Mirjalili, S., Abd Elaziz, M., & Gandomi, A. H., "The Arithmetic Optimization Algorithm", *Computer methods in applied mechanics and engineering* **376**, 113609, (2021).
15. Ahmad, M. F., Isa, N. A. M., Lim, W. H. & Ang, K. M. Differential evolution: A recent review based on state-of-the-art works. *Alexandria Engineering Journal* **61**, 3831-3872 (2022).
16. A. Sharma, A. Sharma, V. Jately, M. Averbukh, S. Rajput, B. Azzopardi. A Novel TSA-PSO Based Hybrid Algorithm for GMPP Tracking under Partial Shading Conditions," *Energies* **15**, 3164 (2022).
17. A. Singh, A. Sharma, S. Rajput, A. K. Mondal, A. Bose, and M. Ram. Parameter Extraction of Solar Module Using the Sooty Tern Optimization Algorithm. *Electronics* **11**, 564 (2022) .
18. Shaari, M. et al. Supervised evolutionary programming-based technique for multi-DG installation in distribution system. *IAES International Journal of Artificial Intelligence (IJ-AI)* **9**, 11 (2020).
19. Tee, Z. Y., Yeap, S. P., Hassan, C. S. & Kiew, P. L. Nano and non-nano fillers in enhancing mechanical properties of epoxy resins: a brief review. *Polymer-*

Plastics Technology and Materials **61**, 709-725 (2022).
20. Hassan, C., Durai, V., Sapuan, S., A.A, N. & Mohamed Yusoff, M. Z. Mechanical and Crash Performance of Unidirectional Oil Palm Empty Fruit Bunch Fibre-reinforced Polypropylene Composite. *Bioresources* **13**, 8310-8328 (2018).
21. Jamaludin, F. A. et al. Considering the effects of a RTV coating to improve electrical insulation against lightning. *2016 33rd International Conference on Lightning Protection (ICLP)*, 1-5 (2016).
22. Ang K. M. et al. Optimal Design of Convolutional Neural Network Architectures Using Teaching–Learning-Based Optimization for Image Classification. *Symmetry* **14**, 2323 (2022).

Authors Introduction

Mr. Koon Hian Ang

He is currently pursuing Bachelor of Mechatronics Engineering with Honours as final year student in Faculty of Engineering, Technology and Built Environment, UCSI University, Malaysia. His research interests are machine learning, deep learning, and optimization algorithm.

Mr. Koon Meng Ang

He received the B.Eng. degree in Mechatronic Engineering with Honours from UCSI University, Malaysia, in 2019. He is currently pursuing Ph.D. degree in UCSI University, Malaysia. His research interests are swarm intelligence, machine learning and deep learning.

Prof. Dr. Mohd Rizon Mohamad Juhari

He is a Professor in Faculty of Engineering at UCSI University in Malaysia. He received his PhD in Engineering from Oita University, Japen in 2002. His research interests are face analysis, pattern recognition and vision for mobile robot.

Dr. Chin Hong Wong

He is a Lecturer in Maynooth International Engineering College at Fuzhou University in China. He received his PhD in Electrical and Electronic Engineering from Universiti Sains Malaysia in 2017. His research interests are Energy harvesting and control system.

Dr. Abhishek Sharma

He is a Research Assistant Professor at Graphic Era Deemed to be University in India. He received his PhD from University of Petroluem & Energy Studies in 2022. His research interests are artificial intellgence and power electronics.

Dr. Chun Kit Ang

He is the Dean and Associate Professor in Faculty of Engineering at UCSI University in Malaysia. He received his PhD in Mechanical and Manufacturing Engineering from Universiti Putra Malaysia in 2014. His research interests are artificial intelligence, soft computing, robotics and mechatronics.

Dr. Sew Sun Tiang

She is an Assistant Professor in Faculty of Engineering at UCSI University in Malaysia. She received her PhD in Electrical and Electronic Engineering from Universiti Sains Malaysia in 2014. Her research interests are optimization and antenna design.

Dr. Wei Hong Lim

He is an Associate Professor in Faculty of Engineering at UCSI University in Malaysia. He received his PhD in Computational Intelligence from Universiti Sains Malaysia in 2014. His research interests are optimization and artificial intelligence.