2022 IEEE 18th International Conference on Automation
Science and Engineering (CASE)
August 20-24, 2022. Mexico City, Mexico

2022 IEEE 18th International Conference on Automation Science and Engineering (CASE) | 978-1-6654-9042-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/CASE49997.2022.9926636

Imbalanced Wafer Map Dataset Classification with Semi-Supervised Learning Method and Optimized Loss Function

Jianchuan Huang, Kuo-Yi Lin, Jia Xu, Lili

***Abstract*—Wafer is the crucial raw material of semiconductor devices. In wafer production, impurities cannot be removed entirely, which will cause various wafer map defects. Quickly and precisely classifying wafer map defects can help engineers track failures in the semiconductor manufacturing process. However, different wafer map defect patterns occur randomly and irregularly, and labeling work is labor-intensive and time-consuming. Therefore, the wafer map dataset is usually imbalanced and consists of many unlabeled data. In this paper, we utilize unlabeled data by using semi-supervised learning methods and alleviate the imbalance problem by optimizing the loss function to increase the accuracy of wafer map classification. The performance of the proposed method is illustrated with the WM-811K dataset which consists of real-world wafer maps.**

I. Introduction

Wafer, a thin slice of semiconductor, is used as the substrate for microelectronic devices. With the increasing demands of the semiconductor market as well as higher quality and performance requirements of the semiconductor manufacturing process, wafer production capacity becomes insufficient. The wafer production is complicated, expensive, and time-consuming. It consists of many microfabrication processes, such as doping, ion implantation, etching, thin-film deposition of various materials, and photolithographic patterning. Therefore, even using modern, highly automated, precisely positioned equipment in a nearly particle-free environment, the factory cannot produce an utterly faultless wafer map [1]. An example of a wafer bin map is shown in Fig. 1.

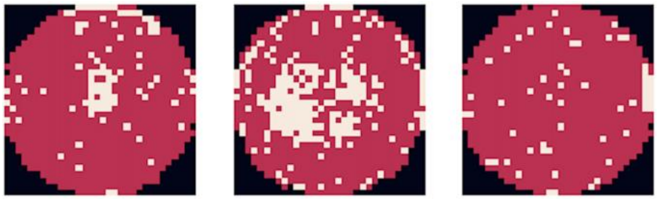

Figure 1. The examples of defective wafer maps

Wafer map comes from the wafer probe, presenting different electrical functions on dies. The wafer map defects are usually clustered into several patterns, and different patterns contain essential information that can identify different defects in the production line, which can help engineers repair the production line in time and increase production efficiency. Hence, how to put on classifying different wafer patterns quickly and precisely is the emphasis. After the production, the inspectors randomly select wafers to examine the die's locations, sizes, and electrical properties using the probe station [1]. However, the inspectors cannot handle the increasing number of wafer data. Therefore, it is critical to developing novel methods to analyze enormous wafer data to improve the efficiency of the wafer manufacturing process [2].

Many approaches have been proposed to classify the types of patterns. Most of the approaches are based on supervised learning. Chien et al. [3] constructed a system combing spatial statistics and neural networks to classify defect patterns. Fan et al. [4] used the ordering point to identify the cluster structure (OPTICS) clustering method to remove anomaly defects and extract features based on density and geometry. Wu et al. [5] integrated Radon- and geometry-based features to form a new representation of each wafer map and used a support vector machine (SVM) to identify wafer failure patterns. Deep learning methods have also been used to determine pattern types and achieved great success. Nakazawa et al. [6] use convolutional neural networks (CNNs) for defect pattern classification and image retrieval. Kyeong et al. [7] utilized CNN to classify mixed-type defect patterns in WBMs in the framework of an individual classification model for each defect pattern. Cheon et al. [8] combined CNN with k-NN to classify defect patterns and detect unknown classes. Shim et al. [9] propose a cost-effective classification system of wafer map patterns based on the active learning of a CNN.

However, labeled wafer map defect samples are usually few and imbalanced since the wafer map data are hard to collect and label. The performance of CNNs is highly affected by the limited and long-tailed wafer data. Unsupervised learning is a promising method to process a large amount of unlabeled data because it is able to learns patterns from unlabeled data and reconstruct the data distribution. Yu et al. [10] developed an unsupervised version of a joint local and nonlocal linear discriminant analysis (JLNDA) for defect detection of wafer maps. Nakazawa et al. [11] used a deep convolutional encoder-decoder neural network to detect unseen patterns.

Semi-supervised learning, utilizing both labeled and unlabeled data, can alleviate the problem caused by limited data used to train the classifier. The first achievements made by combing deep learning methods and semi-supervised learning are based on generative models such as denoising autoencoders [12], variational inference [13], and generative adversarial nets (GAN) [14], [15]. Kong et al. [16] utilize a semi-supervised ladder network to minimize the sum of supervised and unsupervised costs simultaneously. The

*This work is supported in part by National Natural Science Foundation (NNSF) of China under Grant 62003245 and Grant 72171172, in part by Shanghai Municipal Science and Technology Major Project under Grant 2021SHZDZX0100, and in part by Shanghai Municipal Commission of Science and Technology under Grant 19511132100.

J. Huang, K. Lin, J. Xu, L. Li are with the Department of Control Science and Engineering, Tongji University, Shanghai 201804, China (e-mail: 2033800@tongji.edu.cn, 19603@tongji.edu.cn, 615xujia@tongji.edu.cn, lili@tongji.edu.cn)

978-1-6654-9042-9/22/$31.00 ©2022 IEEE

imbalance problem can also be eased up by optimizing loss function. Focal loss proposed by Lin et al. [17] lowered the weights of negatives that is easy to classify to make the classifier focus on the complex examples and prevent over-fitting. Cao et al. [18] proposed a theoretically-principled label-distribution-aware margin (LDAM) loss to encourage larger margins for minority classes as regularization.

This study proposes a framework based on semi-supervised learning and various optimized losses. This framework uses the wafer dataset with labeled and unlabeled data to increase the classification accuracy. We hypothesize that adding pseudo-labeled data to train the dataset can reconstruct the distribution of the dataset and enhance the performance of the model. We also make the model focus on hard-classified samples by altering the loss function.

Our contribution is as follows:

(1) We propose a deep learning-based framework to classify the wafer map dataset. We use Pseudo-labeling to extract the additional information of unlabeled data and LDAM loss to alleviate the imbalance problem.

(2) We compare the framework with other methods in the original dataset without data enhancement methods to better illustrate the improved performance of our framework. The effectiveness is illustrated with the experiment results.

The remainder of the paper is organized as follows. In section II we elaborate on the proposed methods. In section III we present the results tested on WM-811K datasets and analyze them. In section IV we summarize the article.

II. Methodology

This section presents our approach to preprocess the wafer data and the classification methods in detail. We attempt to solve the problem caused by the overwhelming number of none-pattern samples, which decreases the accuracy of defect pattern classification significantly. The semi-learning methods and different loss functions are also discussed.

A. Data Preprocessing

The height and width of the wafer map in the WM-811K dataset [5] are mostly around 26, and we resize all wafer maps into (26, 26) to make the wafer map easy to classify. Undersampling is a simple way to reduce the number of none-pattern samples. Undersampling is a method that keeps all the data in the minority class and decreases the majority class's size to balance uneven datasets. Inspired by the elbow method used in K-means [19] to determine the k value, we gradually reduce none-pattern samples to observe how defect pattern accuracy changes. The result is shown in Fig.2. Based on the result, the accuracy grows slowly when we use around 20000 none-type wafer maps in training set and we determine the value of none-pattern samples as 14000, nearly 1/10 of its original number.

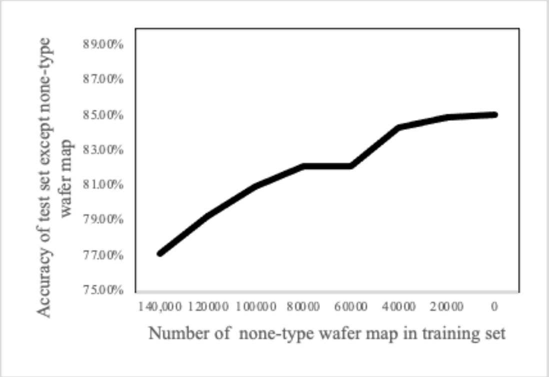

Figure 2. The accuracy of defective wafer maps

B. Residual neural network (ResNet)

As the neural network becomes deeper and deeper, the accuracy would be decreased and the gradients would be vanishingly small, which may stop the network from further training. Residual neural network (ResNet) [20] utilized shortcuts to skip some layers to resolve the vanishing gradients problem and relieve the degradation problem. The identity shortcut connections add no extra parameters and computational complexity and allow data flow across some layers. The adjustment of weights is more sensitive to the variation of outputs from the residual neural network, which makes the backpropagation gradient larger and the training easier.

C. Semi-supervised Learning

WM-811K dataset contains an enormous quantity of unlabeled data. Yang et al. [21] suggested that the label bias can be reduced in a semi-supervised manner with more unlabeled data, which significantly improves the performance of the final classifier. Pseudo-labeling is a simple semi-supervised learning technique to utilize unlabeled data. The procedure of pseudo labeling is shown in Fig.3.

D. Focal Loss

To solve the problem of class imbalance, Lin et al. [17] proposed a new loss function: focal loss, which is modified based on standard cross entropy loss. Focal loss can make the model focus more on hard-to-classify samples during training by reducing the weight of easy-to-classify samples. Focal loss is defined as:

$$FL(p_t) = -(1-p_t)^{\gamma}\log(p_t) \tag{1}$$

where γ is the focusing parameter which smoothly adjusts the rate at easy examples. Variable p_t is defined as:

$$p_t = \begin{cases} -\log(p) & if\ y=1 \\ -\log(1-p) & if\ y=-1 \end{cases} \tag{2}$$

where $y \in \{\pm 1\}$ specifies the ground-truth class and $p \in [0,1]$ is model's estimated probability for the class with label $y = 1$.

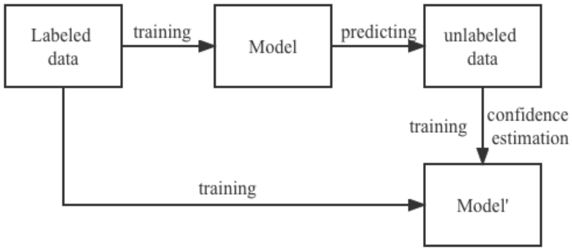

Figure 3. The procedure of pseudo-labelling

E. LDAM Loss

To further gain more information about the minority classes, Cao et al. [18] extended the existing soft margin loss [22] by encouraging the minority classes to have larger margins. They also developed a deferred re-balancing training procedure to apply a re-weighted LDAM loss with a lower learning rate after certain training epochs. LDAM loss is defined as:

$$\begin{aligned} L_{LDAM} &= -\log \frac{e^{z_y-\Delta_y}}{e^{z_y-\Delta_y}+\sum_{j\neq y} e^{z_j}} \\ where\ \Delta_j &= \frac{C}{n_j^{1/4}}\ for\ j \in \{1,\ldots,k\} \end{aligned} \tag{3}$$

III. Experience Results

In this section, the whole framework for wafer map classification proposed is presented, and we test and compare the above methods with other methods. We used a platform with an Xeon E5-2678 v3 @ 2.50 GHz (12 cores), 62 GB memory, and a NVIDIA GeForce RTX 2080 Ti GPU.

A. Dataset Description

WM-811K is a real word wafer map that contained 3.1% wafers (25,519 wafers) with real failure patterns and 18.2% none-pattern wafers (147,431 wafers), and unlabeled 78.7% wafers (638,570 wafers). Among labeled wafer maps, the failure type classes include None, Center, Donut, Edge-Loc, Edge-Ring, Loc, Random, Scratch, Near-full. After the data preprocessing, training set, validation set, and test set are shown in Table I.

TABLE I. Training set, validation set and test set distribution

Type	training set	validation set	test set
Center	2546	855	893
Donut	323	111	121
Edge-Loc	3124	1043	1022
Edge-Ring	5822	1945	1913
Loc	2151	725	717
Random	522	180	164
Scratch	722	249	222
Near-full	92	31	26
none	8410	2765	2825

B. Model architecture

We have tested plain and residual CNN networks to classify wafer map defect patterns. The plain CNN network consists of three convolutional layers using a 3x3 kernel with a 2x2 max-pooling layer and a Rectified Linear Unit (Relu). A fully connected (FC) layer with 256 nodes follows the convolutional layers, and the class probability is calculated through the last softmax layer. We also test residual CNN networks to classify wafer map defect patterns. We use ResNet18 as our residual CNN network shown in Table II.

To better illustrate the performance of the framework, we calculate all the accuracy in the experiments using only defect pattern samples since the accuracy of none-pattern samples is always high enough, and we care more about the defect pattern samples. Table III shows the comparison between plain CNN networks and residual CNN networks. We validated the effectiveness of the residual CNN network that shows better performance in classifying most defect patterns.

Fig.4 shows the pseudo-labelling and optimized loss function classification flowchart used in this article. After the data preprocessing, the labeled data of WM-811K dataset was split into training set and test set. The unlabeled data was utilized with pseudo-labelling to generate more labeled data. New training set was formed by the generated labeled data and the original labeled data. The classifier based on ResNet18 was trained combined with optimized loss function using new training set and test set.

C. The performance of LDAM Loss

Compared to cross entropy loss and Focal loss, utilizing LDAM loss could provide better classification performance to manage the imbalanced problem. The classifier using LDAM loss improves the classification accuracy in most wafer defect patterns, especially the scratch type. The results are shown in Table IV.

TABLE II. ResNet18 Structure

Layer Name	Output Size	ResNet-18
conv1	112×112	7×7, 64, stride 2
conv2_x	56×56	3×3 max pool, stride 2
		$\begin{bmatrix} 3\times3 & 64 \\ 3\times3 & 64 \end{bmatrix}\times 2$
conv3_x	28×28	$\begin{bmatrix} 3\times3 & 128 \\ 3\times3 & 128 \end{bmatrix}\times 2$
conv4_x	14×14	$\begin{bmatrix} 3\times3 & 256 \\ 3\times3 & 256 \end{bmatrix}\times 2$
conv5_x	7×7	$\begin{bmatrix} 3\times3 & 512 \\ 3\times3 & 512 \end{bmatrix}\times 2$
average pool, 1000-d fc, softmax		

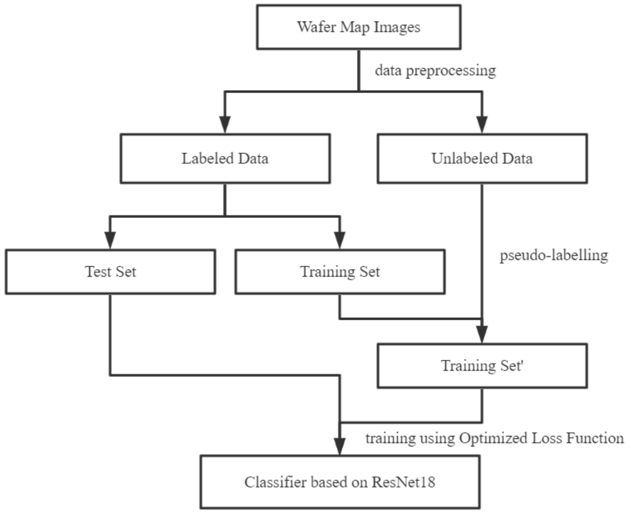

Figure 4. The flowchart of wafer map classification

TABLE III. Accuracy comparison between plain CNN networks and residual CNN networks (The best results are highlighted in red)

Type	**Center**	**Donut**	**Edge-Loc**	**Edge-Ring**	**Total**
plain CNN network	92.50%	79.34%	84.15%	97.91%	84.84%
	Loc	**Random**	**Scratch**	**Near-full**	
	60.67%	90.24%	20.27%	96.15%	
Type	**Center**	**Donut**	**Edge-Loc**	**Edge-Ring**	**Total**
residual CNN network	93.84%	81.82%	86.11%	98.17%	86.73%
	Loc	**Random**	**Scratch**	**Near-full**	
	67.36%	88.41%	25.68%	92.31%	

TABLE IV. Accuracy comparison between CE Loss, Focal Loss and LDAM loss (The best results are highlighted in red)

Type	**CE**	**Focal**	**LDAM**
Center	93.84%	93.84%	95.63%
Donut	81.82%	77.69%	77.69%
Edge-Loc	86.11%	86.01%	86.99%
Edge-Ring	98.17%	97.07%	98.69%
Loc	67.36%	69.32%	69.46%
Random	88.41%	89.02%	82.32%
Scratch	25.68%	25.23%	46.85%
Near-full	92.31%	88.46%	92.31%
Total	86.73%	86.45%	88.34%

D. *The performance of Pseudo-labelling*

Pseudo-labeling was added to alter the distribution of the dataset and to provide more information for the classifier. The new training set was formed by the generated and original samples. The Classifier trained by the new training set significantly improved the accuracy in some defect patterns, e.g., Random and Near-full. The results are illustrated in Table V.

TABLE V. Accuracy comparison between CE Loss, Focal Loss and LDAM loss with Pseudo-labeling (The best results are highlighted in red)

Type	**CE Loss**	**Focal loss**	**LDAM loss**
Center	95.07%	94.40%	95.63%
Donut	76.86%	77.69%	78.51%
Edge-Loc	85.03%	83.27%	87.28%
Edge-Ring	98.27%	97.75%	98.38%
Loc	67.64%	68.76%	70.57%
Random	90.85%	90.24%	86.59%
Scratch	36.94%	34.23%	45.95%
Near-full	92.31%	96.15%	96.15%
Total	87.26%	86.65%	88.58%

IV. CONCLUSION

In this work, we propose a semi-supervised learning framework combined with LDAM loss and investigate WM-811K dataset classification to evaluate the performance of the proposed approach. The label bias problem arises when negative samples vastly outnumber positive samples in the training data. A ResNet classifier determines different defect patterns after undersampling of the none-pattern samples. The results show that the proposed semi-supervised framework with LDAM loss can alleviate the label bias problem and show better classification performance than other methods.

REFERENCES

[1] T. Yuan, W. Kuo, and S. J. Bae, "Detection of spatial defect patterns generated in semiconductor fabrication processes," *IEEE Trans. Semicond. Manuf.,* vol. 24, no. 3, pp. 392-40, Aug. 2011.

[2] F. L. Chen, and S. F. Liu, "A neural-network approach to recognize defect spatial pattern in semiconductor fabrication," *IEEE Trans. Semicond. Manuf.,* vol. 13, no. 3, pp. 366-373, Aug. 2000.

[3] C. F. Chien, S. C. Hsu, and Y. J. Chen, "A system for online detection and classification of wafer bin map defect patterns for manufacturing intelligence," Int. J. Prod. Res., vol. 51, no. 8, pp. 2324-2338, Feb. 2013.

[4] M. Fan, Q. Wang, and B. van der Waal, "Wafer defect patterns recognition based on OPTICS and multi-label classification," in *2016 IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC)*, 2016, pp. 912-915.

[5] M. J. Wu, J. S. R. Jang, and J. L. Chen, "Wafer map failure pattern recognition and similarity ranking for large-scale data sets," *IEEE Trans. Semicond. Manuf.,* vol. 28, no. 1, pp. 1-12, Feb. 2015.

[6] T. Nakazawa, and D. V. Kulkarni, "Wafer map defect pattern classification and image retrieval using convolutional neural network," *IEEE Trans. Semicond. Manuf.,* vol. 31, no. 2, pp. 309-314, May 2018.

[7] K. Kyeong, and H. Kim, "Classification of mixed-type defect patterns in wafer bin maps using convolutional neural networks," *IEEE Trans. Semicond. Manuf.,* vol. 31, no. 3, pp. 395-402, Aug. 2018.

[8] S. Cheon, H. Lee, C. O. Kim, and S. H. Lee, "Convolutional neural network for wafer surface defect classification and the detection of unknown defect class," *IEEE Trans. Semicond. Manuf.,* vol. 32, no. 2,

pp. 163-170, May 2019.

[9] J. Shim, S. Kang, and S. Cho, “Active learning of convolutional neural network for cost-effective wafer map pattern classification,” *IEEE Trans. Semicond. Manuf.,* vol. 33, no. 2, pp. 258-266, May 2020.

[10] J. Yu, and X. Lu, “Wafer map defect detection and recognition using joint local and nonlocal linear discriminant analysis,” *IEEE Trans. Semicond. Manuf.,* vol. 29, no. 1, pp. 33-43, Feb. 2016.

[11] T. Nakazawa, D. V. Kulkarni, “Anomaly detection and segmentation for wafer defect patterns using deep convolutional encoder–decoder neural network architectures in semiconductor manufacturing,” *IEEE Trans. Semicond. Manuf.,* vol. 32, no. 2, pp. 250-256, May 2019.

[12] A. Rasmus, H. Valpola, M. Honkala, H. Valpola, and T. Raiko, “Semi-supervised learning with ladder networks,” arXiv preprint arXiv:1507.02672, 2015.

[13] D. P. Kingma, S. Mohamed, D. J. Rezende, and M. Welling, “Semi-supervised learning with deep generative models,” in *Advances in neural information processing systems,* 2014, pp. 3581-3589.

[14] A. Odena, “Semi-supervised learning with generative adversarial networks,” arXiv preprint arXiv:1606.01583, 2016.

[15] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and F. F. Li, “Imagenet large scale visual recognition challenge,” *Int. J. Comp. Vis.*, vol. 115, no. 3, pp. 211-252, April 2015.

[16] Y. Kong, and D. Ni, “Semi-supervised classification of wafer map based on ladder network,” in *2018 14th IEEE International Conference on Solid-State and Integrated Circuit Technology (ICSICT),* 2018, pp. 1-4.

[17] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, “Focal loss for dense object detection,” in *Proc. the IEEE international conference on computer vision,* 2017, pp. 2980-2988.

[18] K. Cao, C. Wei, A. Gaidon, N. Arechiga, and T. Ma, “Learning imbalanced datasets with label-distribution-aware margin loss,” arXiv preprint arXiv:1906.07413, 2019.

[19] C. F. Chien, W. C. Wang, and J. C. Cheng, “Data mining for yield enhancement in semiconductor manufacturing and an empirical study, “*Exp. Syst. Appl.*, vol. 33, no. 1, pp. 192-198, July 2007.

[20] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. the IEEE conference on computer vision and pattern recognition,* 2016, pp. 770-778.

[21] Y. Yang, and Z. Xu, “Rethinking the Value of Labels for Improving Class-Imbalanced Learning,” in *Advances in Neural Information Processing Systems,* 2020, pp. 19290-19301.

[22] F. Wang, J. Cheng, W. Liu, and H. Liu, “Additive margin softmax for face verification,” *IEEE Signal Process. Lett.*, vol. 25, no. 7, pp. 926-930, Jan. 2018.