DLADC: Deep Learning Based Semiconductor Wafer Surface Defects Recognition

Charissa Phua, and Lau Bee Theng, *Member, IAENG*

***Abstract*— In semiconductor manufacturing, surface defects on wafers must be classified accurately for better yield management. To manage the increasing chip demand in speed and scale, automatic defect classification (ADC) system has been introduced. Most existing ADC systems utilize machine learning-based algorithms that require manual feature extractions and manual intervention such as human-based classification for accuracy and consistency. These methods are labour-intensive, unreliable, and highly prone to human error. Therefore, by leveraging on deep learning technologies, this paper proposes *DLADC* - an ADC system using a deep convolutional neural network (CNN) architecture for detecting and classifying semiconductor wafer surface defects. The proposed system takes Scanning Electron Microscope (SEM) images as input and outputs the defect's class and location. The proposed system also sub-classifies particle-type defects into various sizing groups. Identification of defect types that occurred on wafer surfaces allows for better defect root cause analysis, and the additional information of defect size further serves as an essential indication of the origin of machine failure. The proposed *DLADC* promotes 2x time saving while achieving an improved accuracy of 93.69% based on experimental results with a real semiconductor defect dataset. Not only does *DLADC* outperforms the 70% classification performance of trained operators, but it also surpasses the 90% classification performance of industrially pragmatic defect classification.**

***Index Terms*— automatic defect classification (ADC), convolutional neural network (CNN), deep learning (DL), Review-SEM defects, semiconductor manufacturing**

I. Introduction

THE growing demand for semiconductors motivates manufacturers to develop wafers with better capabilities, are defect-free, and have high reliability to avoid field failures [1]. To meet these requirements, the wafers must be flat and particle-free so that the final chip products do not incorporate any defects. However, defects are sometimes generated during the lengthy and complex Metal-Oxide-Semiconductor (MOS) technology fabrication process [2]. Therefore, wafer defect management is an important process in semiconductor manufacturing to ensure that sources of defects are captured by associating defects to specific steps in the fabrication process and, after that, quickly eliminated before reaching the final production stage. The classified defects are also used for excursions containment to reduce the loss in wafer fabrication time and cost [3].

One of the common practices in defect management is establishing in-line inspection stations along with the most important processing steps and/or machines with a high probability of defect deposition to observe abnormalities on wafer dies. These inspection tools produce wafer maps, which is a visualization created based on the detected abnormal locations on the wafers. Wafer maps are widely used to understand the process abnormalities, facilitating the diagnosis of low-yield problems [4]. A step further to pinpoint defects is to take Scanning Electron Microscope (SEM) images of the wafer surface after identifying the defect locations in wafer maps. The SEM images capture any particles, flaws or irregular connections caused by misaligned electronic circuits stacked on the many layers of a wafer [18]. The traditional method applied in semiconductors is visual inspection. However, this manual approach has proven to be time-consuming and expensive. Hence, automatic defect classification (ADC) systems have been introduced as an alternative.

Ideally, utilizing high-resolution microscopes, an ADC system involves scanning the wafer surface and collecting information (e.g., location, shape, and size) on detected defects. The inspection system uses automatic defect detection technology and the classification of the defects using visualizations of wafer maps, optical images or SEM images [6][7]. Despite the success of applying the ADC approach in the semiconductor industry where existing ADC systems display good defect detection, they have low accuracy (intelligence) in defect classification for both present and newly discovered defects in varied wafer designs. Consequently, human intervention in defect classification is unavoidable and must be performed, producing unreliable and inconsistent results due to variation in expertise level. For the reasons stated above, it is clear that research is required for an advanced and powerful ADC system to improve defect classification accuracy and efficiency.

Various approaches to addressing this problem statement of a high-performance ADC system have been proposed in the literature. In particular, machine learning techniques, such as clustering, support vector machines, and artificial neural networks, have been proposed to deal with the problem using both unsupervised and supervised learning paradigms [1,8]. These approaches rely on a series of features to capture the spatial properties of wafers. Conversely, these wafers can be intuitively interpreted as images, with defect patterns serving as spatial features. Therefore, using images as the realistic representation of wafers will best preserve the native spatial characteristics of the defect patterns [9]. Defect

Manuscript received June 05, 2021; revised December 27, 2021.
The work of C. Phua was supported by the the domain experts from the Defect Management team and AIDA team in X-FAB Sarawak Sdn. Bhd..
Charissa Phua is a data scientist in Global Artificial Intelligence & Data Analytics Department, X-FAB Sarawak Sdn. Bhd., Malaysia. (e-mail: Charissa.Phua@xfab.com).
Lau Bee Theng is an associate professor in ICT and the Director School of Research, Swinburne University of Technology Sarawak Campus, Malaysia. (e-mail: blau@swinburne.edu.my).

detection and recognition in semiconductor manufacturing can then be tackled as an automatic image classification problem by utilizing various machine learning models with the application of image processing techniques. In correspondence to this approach, deep learning models have attracted much interest in recent years because of the demonstrated excellent successes in machine vision-related tasks. Motivated by this success, we propose using deep learning techniques to study the detection and classification of wafer surface defects.

In contrast to conventional ADC which consists of two separate steps in the pipeline: manual feature extraction and classification, a convolutional neural network (CNN) has the abilities to effectively extract significant defect features directly from the image inputs using convolutional features and classify them into respective defect classes in a single architecture. Furthermore, each semiconductor often undergoes hundreds to thousands of SEM review scans every day, resulting in a large number of pictures to be evaluated at various magnification levels and defects of varied sizes to be classified. We incorporate defect size into our proposed system since it is a vital element for root cause analysis to discover the specific problematic machinery tool from the lengthy manufacturing process. Additionally, to the best of found knowledge, there has been no effort specifically on the smart automation of the detection and recognition for defects and defect sizes at metal layers of the back-end-of-line (BEOL) process, wherein these steps the metallization is applied [10]. Therefore, this paper proposes a novel autonomous deep-learning-based defect classification (*DLADC*) system for the metal layers as early as the first metal layer coating stage with a methodology that uses a CNN-based object detection network architecture to locate defects and scale bars in the images, allowing for autonomous measurement of exact defect sizes relative to their real-life dimensions. This novel approach has three contributions:

i. *DLADC* is specifically designed with the goal of achieving high classification performance with the CNN architecture.
ii. *DLADC* includes a feature that calculates defect size.
iii. *DLADC* increases operation efficiency, which saves 2x time over manual review and reclassification.

The remainder of this paper is organized as follows: Section II describes the related works. Section III presents the proposed *DLADC* system, and Section IV evaluates its performance with real semiconductor defect dataset. And finally, Section V concludes the paper and discusses future works.

II. Related Works

In the current semiconductor manufacturing industrial scene, defect inspection is still heavily reliant on manual labour. This prevailing method of defect inspection has proven to be unreliable with high inconsistency resulting in many false identifications of defects, mainly due to the lack of objectivity of engineers and fatigue from excessive hours of the mundane task of looking at defect images. The following section covers the need for autonomous defect classification for semiconductor wafer surface defects, followed by the introduction of convolutional neural network architecture in the interest of its widespread achievements in tackling vision-based tasks, especially on the topic of defect classification. This section also discusses multiple variants of the CNNs and several established CNN-based object detection architectures.

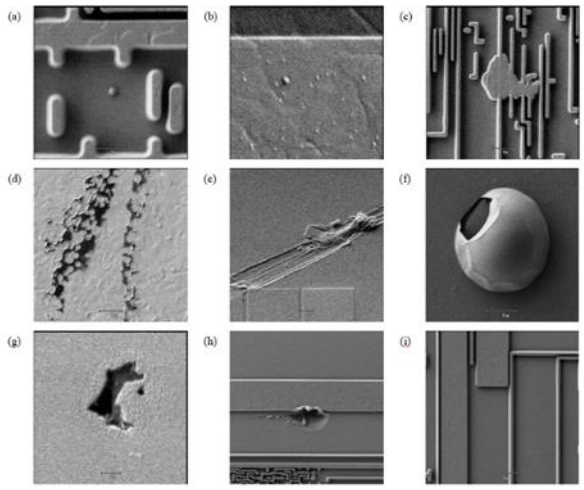


Fig. 1. Examples of SEM images of wafer surface defects.

A. Automatic Defect Classification

Automatic defect classification (ADC) is one of the mechanisms helping defect engineers in their quest for early defect detection and quick analysis. Before ADC was introduced, semiconductors mainly depended on manual labour for defect inspection, review, and classification. Most process engineers rely on their domain knowledge and expertise to determine the unique features of wafer defects and relate them to fabrication processing steps. As a result, the accuracy of identifying the root causes of defects is highly reliant on the process engineers' experience [2].

According to a whitepaper by Intel in 2018 [11], the authors reported that it could take up to 9 months to train operators or engineers to classify defects with 90% accuracy manually. Even so, the trained operator will over time has a drop in performance to only 70% to 85% accuracy due to reasons such as the repetitive nature of the task, lack of insight, advancements in fabrication processes, and plain difficulty in classification purely based on human eye and brain.

As a result, semiconductors have introduced the use of ADC to improve defect inspection throughput as well as data integrity. The ADC system uses a die-to-die comparison, otherwise known as golden template method to detect defects, and measures multiple spatial and textural features of the defects to determine the defect classes. Either a rule-based system or machine learning algorithm is applied to classify the defects into pre-defined defect classes based on training samples. Despite the use of ADC systems, existing ADC systems claim a high defect detection rate but low accuracy (high false alarms, high erroneous classifications), necessitating further manual intervention. Based on significant study, previous works mostly focused on defect pattern detection and classification in semiconductor wafer maps. Nevertheless, there has been emerging research over the recent few years on defect classification using SEM images which would be the methodology used in this study.

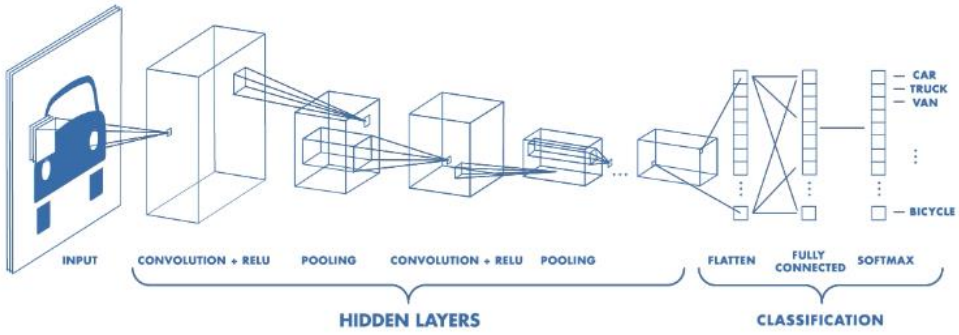


Fig. 2. Example of a Convolutional Neural Network.

B. Review-SEM Defect Classification

One of the methodologies of defect classification is done using scanning electron microscope (SEM) images of the wafer surface, which are taken after the completion of each circuit layer (particularly between the etching and deposition steps). It is almost implausible for any SEM image to contain more than one defect due to the SEM's microscopic sensing field. Yet, the defect classification accuracy by conventional ADC systems is poor, and a manual inspection process is often required to ensure accurate classification. Fig. 1 visualizes the sample SEM images of wafer surface defects.

Many studies concentrate on defect control on the wafer map level, but limited literature focuses on defect detection and classification of microscopic defects, particularly those that use deep learning techniques. Deep learning has recently acquired popularity because to its ability to extract compact features from complicated and high-dimensional data (e.g., pictures and sensor signals) automatically. The convolutional neural network (CNN) model, in particular, was widely used due of its superior performance in identifying picture input. In [12], Cheon et al. proposed a CNN-based ADC method to address the lack of application of CNN-based models for wafer surface defect classification. Recognizing the limitation of the CNN model being a supervised model that requires a sufficient amount of data for each defect class, the authors designed their CNN architecture to achieve high classification performance for all known defect classes. Besides, by utilizing the feature space created by the CNN training result, the authors included a k-nearest neighbours (k-NN) algorithm to classify unknown defects as an 'Unknown' class without retraining the CNN and hence avoiding the time-consuming process of collecting new data for the unseen defects. In the study, Cheon et al. explored the proposed CNN architecture on a dataset with five defect classes exclusive of the 'Unknown' class. The authors reported an average of 96.2% classification accuracy of the trained CNN for the testing subset of 319 images and it showcased the CNN's superiority in wafer surface defect classification performance in the comparison experiments against other classifiers such as MLP, SVM, and stacked autoencoder (SAE).

In a separate study, Leary et al. [13] explored an ADC approach to classify the chemical composition of particle defects on semiconductor wafers based on combined information of the SEM image and energy-dispersive x-ray (EDX) spectroscopy data. The proposed method consisted of a CNN which was trained using SEM images and then merged EDX spectra data using fully connected layer. The main intention of this approach is to enable engineers to identify the defect source(s) of the classified defects within the manufacturing process by connecting the source to the semiconductor equipment chamber. With eight defect classes used in the study, the authors reported that the combined SEM image and EDX spectra CNN yielded a greater than 99% Top-3 accuracy and 82.1% Top-1 accuracy on the testing subset of 1172 images. CNNs trained with either SEM image only or EDX spectra data have also been explored and yielded Top-1 accuracy of 51.0% and 70.6%, respectively. Therefore, the authors concluded that CNN trained with both SEM image and EDX spectra data is able to classify semiconductor defects with higher accuracy.

In another study done by Yang and Sun (2020) [14], the authors explored a double feature extraction method based on CNN for defect recognition. The proposed model combined Radon transform with CNN's feature and input these features into the last dense layer to obtain the final classification output. The proposed model, which was named as RadonNet, achieved a recognition rate of 98.5%, averaging across a total of 11 defect types.

Similar to the studies above, leveraging the emergence of new technologies such as computing resources, we propose using deep learning technology for this vision-based task, particularly a CNN-based object detection architecture for the detection and classification of semiconductor wafer surface defects. In addition, we integrate a defect-sizing computation feature to enhance the CNN's classification results and provide insights into mechanisms that cause various defect modes. The methodology of the proposed *DLADC* system will be explored in length in Section III. In the next sub-sections, we will first learn about the CNN operating principle, CNN-based classifier and CNN-based object detector.

C. Convolutional Neural Network

While many types of artificial neural networks (ANN) have been developed and adapted to research and industrial fields, CNNs have been exclusively emphasized to handle image recognition problems with outstanding performance on the MNIST, CIFAR-10 and ImageNet datasets. CNNs are end-to-end auto-learning models with minimal requirements for any human intervention as their architecture enables the functionalities combinations of a feature extractor functions and a classifier [15]. For this reason, CNNs have been the most sought after as an efficient image recognition method.

Most CNNs are made up of four-layered concepts, namely convolution, ReLU (activation), pooling, and fully connectedness (see Fig. 2). After several convolutional and

pooling operations, the high-level reasoning in the neural network is done via fully connected (FC) layers as this FC layer converts feature maps of the input data into a feature vector. For multiclass problem as in our case here, a SoftMax function (i.e., a generalized form of logistic regression) is commonly used to transform the feature vector from FC layer into classification probabilities.

D. CNN-based Classification Architecture

The advance of deep neural networks in computer vision applications is primarily contributed by the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) [16], which tested a few generations of large-scale image classification systems. We will discuss two significant image classification systems that have hugely impacted the ImageNet Challenge, specifically, the VGG Network (VGG-Net) and the Residual Net (ResNet).

VGG-Net was created in 2014, and its main contribution being the comprehensive assessment of networks of increasing depth. Simonyan & Zisserman [17] introduced an architecture with tiny convolutional filters (3 by 3), which showed that depth could be extended to 16 to 19 weight layers compared to prior-art configurations. The width of convolutional layers, i.e., the number of channels, is relatively small, increasing by a factor of 2 after each max-pooling layer; it starts from 64 in the first layer to 512 in the last layer. As a result, during the time it was proposed , VGG-Net reported significantly more accurate performance amongst other CNN architectures. It reported high accuracy on ILSVRC 2014 classification task with a 7.3% error rate. The researchers demonstrated that the representation depth is beneficial for classification accuracy.

ResNet, invented in late 2015 by Microsoft Research Asia, is a 152-layer network architecture that has won ILSVRC 2015 with an error rate of 3.6%. Instead of learning unreferenced functions, the creators had explicitly reformulated the layers as learning residual functions concerning the layer inputs [18]. The authors demonstrated that these residual networks are easier to optimize, and accuracy can be gained from the significantly increased depth.

E. CNN-based Object Detection Architecture

In this subsection, we will discuss some region-based CNNs for object detection on images. Girshick et al. proposed a simple and scalable detection algorithm by combining region proposals with CNNs, later named R-CNN [19]. The proposed approach employed a function called Selective Search. Selective Search performs the function of generating 2000 bottom-up region proposals that have the highest probability of containing an object. These proposals are then warped into an image size that can be fed into a trained CNN that extracts a feature vector for each region. This vector is then used to input class-specific linear SVMs and outputs a classification result for each region. The same vector also gets fed into a bounding box regressor to obtain the most accurate coordinates. Non-maxima suppression is then used to suppress bounding boxes that have a significant overlap with each other. The introduction of the R-CNN model architecture provided performance breakthroughs in object detection tasks. It reported a mean average precision (mAP) of 53.3% on PASCAL VOC 2012 dataset and 31.4% on ILSVRC 2013 detection dataset.

Despite the notable achievements in object detection tasks, R-CNN suffers from few drawbacks due to its multi-stage training process, which is computationally expensive and extremely slow. Hence, in [20], Girshick proposed Spatial Pyramid Pooling networks (SPPnets) to speed up R-CNN by sharing computation of the convolutional layers between different proposals. With this algorithm that utilizes R-CNN and SPPnet, the authors called this method Fast-RCNN, mainly derived from its comparatively fast speed to train and test. The authors showed the higher detection quality of Fast R-CNN by reporting 65.7% mAP in VOC 2012.

Later, the authors in [21] introduced a Region Proposal Network (RPN) after the last convolutional layer. This network takes the full-image convolutional features from the detection network, thus producing region proposals as an outcome. This architecture that combats the complex training pipeline of both R-CNN and Fast R-CNN is dubbed as Faster R-CNN. The authors reported high object detection accuracies on various datasets, including VOC 2012 and MS COCO. Proven by its excellent object detection performance, Faster R-CNN has become the standard and foundation for many object detections today.

In 2016, an architecture named SSD was introduced to reduce the computational power required at model inference time. In the paper SSD: Single Shot MultiBox Detector, authored by W. Liu et al., [22], the authors presented a method for detecting objects in images using a single deep neural network. As the name of this architecture highlights, object localization and classification tasks are done in a single forward pass of the network where the MultiBox technique is used for bounding box regression. In the SSD approach, a base network from high-quality image classification architecture such as VGG-16 built the early network layers. Then, instead of the original fully connected layers used for classification, a set of auxiliary convolutional layers were added to the network to enable feature extractions at multiple scales and the progressive decrease in input size to subsequent layers. The SSD approach has reported a significant improvement in speed and accuracy for detection problems where it achieved 74.3% mAP on VOC 2007 at 59 frames per second (FPS) while Faster R-CNN reached 73.2% with 7 FPS.

As discussed above, CNN has numerous network architectures that can be chosen dependent on the fields in which it is applied. For ADC, we propose using a CNN from the object detection architecture to take advantage of its ability to generate multiple regions of interest (i.e., bounding boxes) and class labels for detected objects in each image input, which is essential for defect size measurement on top of defect recognition for our research.

III. PROPOSED METHOD

In this section, we discuss on the methodology of the proposed *DLADC*, starting with the dataset collection process, and thereafter continuing on to the pre-processing step. Then, we discuss about the model architecture design and transfer learning technique.

TABLE I
NUMBER OF DEFECTS PER CLASS

Defect Class	Count
sp	1792
ed	1191
pd	3513
res	277
scr	868
bub	1056
po	1908
rare	476
False Count	1735
Total Defects	***11081***
Total Data	***12816***

TABLE II
NUMBER OF DEFECTS PER CLASS FOR "SP" AND "ED" CLASSES
small: ≤1μm; large: >1μm

Defect Class	Defect Class with Sizing	Count	Total Count
sp	sp_small	1023	1792
	sp_large	769	
ed	ed_small	524	1191
	ed_large	667	

A. Dataset Collection and Details

The dataset used in this paper contains real defect images obtained from our industrial partner. The dataset consists of grayscale 480 × 480 pixels Review-SEM defect images at metal layers of the BEOL process on 8-inch wafers. The number of defect classes differs across each semiconductor since the wafer defect class set is not closed, i.e., categories of defect class are determined based on the widely available process recipes for each semiconductor's production environment. In our study, the metal layers defects are grouped into nine main classes. Among the nine classes, eight of them are actual defect classes. The remaining one class represents the "False Count" class because images can be wrongly captured as having surface defects in a real production setting.

For this study, our industrial partner, X-FAB, provided a total of 11,081 metal layers defect images. Inclusive of the "False Count" class, 12,816 images were used to train, validate, and test the performance of the proposed approach for the ADC system.

Table I shows the distribution of the number of defects per class. For example, defect class "sp" and "ed" are the particle-type defect classes, respectively also known as Surface Particles and Embedded Defects. The defects from these classes require accurate size measurements. Table II shows the distribution of the number of defects for classes "sp" and "ed".

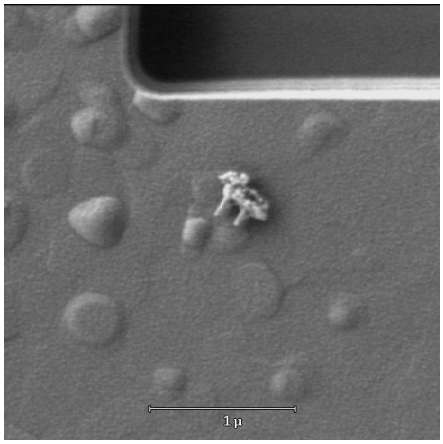


Fig. 3. An example of SEM image of defect class "sp" and size 0.38μm.

These SEM defect images are often captured in a micrometre scale, and the actual defect size can be measured by referencing the scale bar at the bottom of each defect image (see Fig. 3). The defects in this dataset have sizes ranging from 0.18μm to 200μm. In the defect management process, precise and accurate measurement of defects is critical for determining the best relevant root cause. Defect sizes can be an important indicator of which process step or tool out of the many complex BEOL processes resulted in defects on the wafer. Besides, there are recipe issue, equipment issue, and delay in queue time, all of which cause defects. Therefore, it is important to accurately locate the defects and scale bars in the images; the images are labelled by having bounding boxes drawn on the respective region of interest. Note that both data collection and defect labelling are mainly manual processes. In other words, some mislabeled defects may have been used in model training. This mislabeling may degrade the accuracy performance of the proposed model. Nevertheless, the expectation is that only less than 2% of total data are possibly mislabeled.

B. Dataset Pre-processing and Partitioning

The provided SEM defect images are initially at the size of 480 × 480 pixels, and the images are then resized to 256 × 256 pixels during the model training process. In an effort to help with model generalization which enables robust learning, image data augmentation techniques such as adding noise or randomly jittering the brightness, contrast, blurring, saturation, and hue of the image, were applied.

The values of input images are normalized from a scale of [0, 255] to [0, 1] to minimize noise and speed computation during model training. Moreover, the values of input images are also normalized with the mean and standard deviation pixel values from the ImageNet dataset to take full advantage of transfer learning.

With the collected total of 12,816 images, the dataset is partitioned for the purpose of training, validation, and testing. The dataset partitioning ratio used in this study is 80% for training, 10% each for validation and testing. The defect class distribution is kept consistent during dataset partitioning.

C. Model Architecture

The proposed *DLADC* system comprises three main phases. The first phase detects and classifies the Review-SEM defect image samples accordingly into their

defect classes. In this phase, *DLADC* applies a Faster R-CNN detector with a ResNet-50 classifier backbone trained on the collective labels from the nine defect classes and the scale bars. The reason for selecting such architecture is that it has a proven top-1 error rate of 20.74% on the ImageNet validation set [18] and achieved great performance in other major object detection datasets [21,23]. Instead of a bigger ResNet of more layers, the selection of ResNet with 50 layers is to avoid an overfitting scenario whilst acknowledging the comparatively smaller dataset size compared to ImageNet.

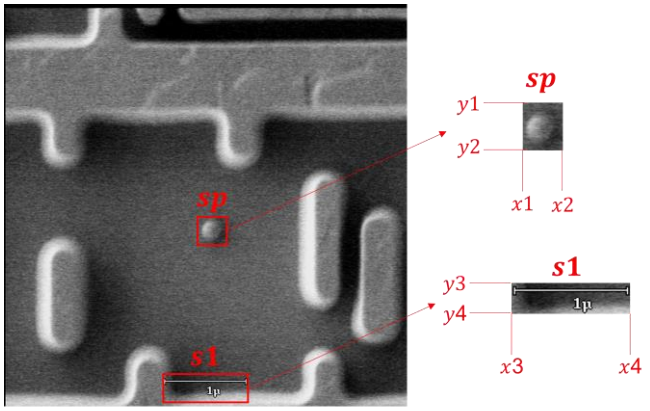
In the second phase, if the classification by the CNN model in the first phase is found to be belonging to either "sp" or "ed" class, then it is sent for defect size calculation whereby the (x, y)-coordinates from the bounding boxes of the localized defect and the localized scale bar are used to calculate the actual defect size in micrometre. Equation (1) defines the defect size calculation operation:

$$defect\ size = \frac{max\big((width_d), (height_d)\big)}{(width_s)} \times m\ , where$$

$$width_{d/s} = x_2 - x_1$$

$$height_d = y_2 - y_1 \quad (1)$$

d represents defect and s represents scale bar. m represents scale measurement in µm unit. The bounding box is rectangular, which is determined by the x and y coordinates of the upper-left corner of the rectangle and the such coordinates of the lower-right corner (i.e., x_1, x_2, y_1 and y_1). Subsequently, with the calculated defect size, *DLADC* sub-classifies the defect into their respective sizing categories. Fig. 4 visualizes the defect calculation for a "sp" defect.



$$size\ of\ \text{"sp"} = \frac{max\big((x2 - x1), (y2 - y1)\big)}{(x4 - x3)} \times 1$$

Fig. 4. "sp" defect size calculation.

The third (last) phase either finalizes the defect class from the first phase or finalizes once it sub-classify the image into individual defect classes with sizing from the second phase. The layout of the proposed *DLADC* system is depicted in Fig. 5.

To summarise, *DLADC* is designed to autonomously and continuously process each image input. The object detection model, Faster R-CNN with ResNet50 backbone in phase 1 is chosen after comparison study against other CNN

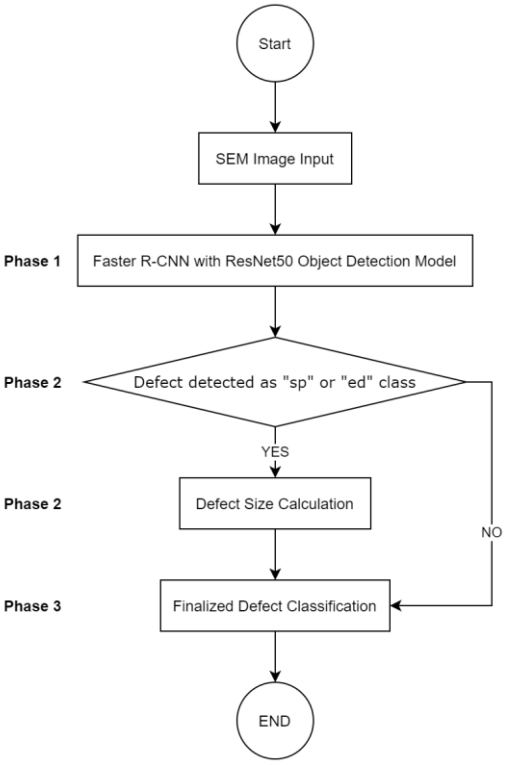


Fig. 5. DLADC System Overview.

architectures and contributes to the high defect classification performance. In addition to defect classification, phase 2 of *DLADC* contributes the novel feature of an ADC system that calculates accurate defect size. Overall, the *DLADC* design improves SEM review efficiency in semiconductor operations.

D. Transfer Learning

Transfer learning is a machine learning approach where a model developed for a task is redeployed as the starting point for a model on a second related task [24], particularly popular in deep learning for computer vision and natural language processing tasks. The proposed method in this study exploited the benefits of transfer learning by loading

pre-trained ResNet-50 on COCO object detection dataset, which yields a 40.2 box AP result. Since COCO dataset contains images of common objects from complex everyday scenes [25], we re-define the output layer to fit the computer vision task of defect classification, i.e., nine defect classes, and lastly, the models are tuned to achieve greater defect recognition performance.

IV. Findings And Discussions

Firstly, we evaluate the performance of the proposed *DLADC* on the validation and testing datasets to better understand the CNN's detection and classification capabilities, particularly for defect classes of varied sizes. In this paper, validation dataset is used to assess the classification accuracy of the trained Faster R-CNN in *DLADC* system after each update of the Faster R-CNN parameters using the augmented training dataset. In general, the Faster R-CNN should not demonstrate the best performance for the training dataset, but for the validation

TABLE III
VALIDATION DATASET OVERALL ACCURACY RESULT.

Fold	Dataset	Accuracy: Total Data	Accuracy: Total Defects
1	Validation	0.9354	0.9259
2	Validation	0.9308	0.9173
3	Validation	0.9352	0.9249
average	Validation	0.9339	0.9227

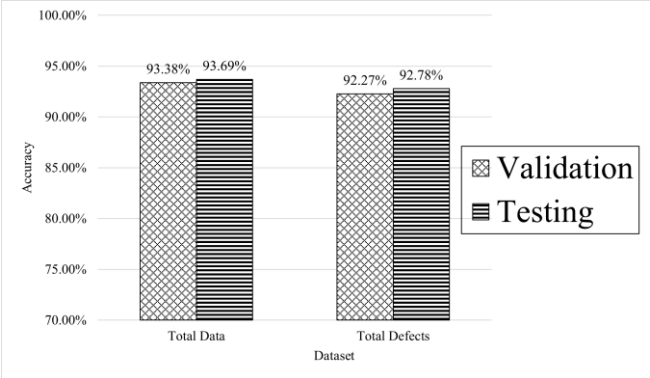


Fig. 6. Accuracy results for total data and total defects for validation and testing datasets.

TABLE IV
TESTING DATASET AVERAGE ACCURACY RESULT FOR DEFECT CLASS WITH SIZING.

Defect Class	Defect Class with Sizing	Accuracy	Overall Accuracy
sp	sp_small	0.9699	0.9327
	sp_large	0.9051	
ed	ed_small	0.8911	0.8918
	ed_large	0.8936	

TABLE V
RESULTS OF THE COMPARISON EXPERIMENT.

CNN-based Model	Testing Accuracy for Total Data	Training Hours	Inference Speed
DLADC	0.9369	< 1 hour	0.0881 second/image
SSD with VGG16	0.9417	8 hours	0.1297 second/image
SSD with ResNet50	0.9152	8 hours	0.1297 second/image

dataset, in order to construct a object detector with robust performance. After the training process is completed, the test dataset is used to conduct the final performance evaluation. The final solution performance is calculated using a procedure called stratified k-fold cross-validation where, in our case, k=3 and each fold has the same class distribution.

The accuracy of each of the three folds for the validation dataset are reported in Table III. The validation dataset records an overall accuracy for total data of 93.38%, while overall accuracy for all the defects is at 92.27%. We report the final performance measure on the testing dataset using the mean of the values computed in the loop of the three cross-validation folds because stratified k-fold cross-validation is commonly used to demonstrate the robust model's performance and because the evaluation done on validation datasets has proven the model's generalization performance.

With reference to the above, the testing dataset records an overall accuracy for total data of 93.69%, while overall accuracy for all the defects is at 92.78%. Fig. 6 visualizes the accuracy results for total data and total defects between the validation and testing datasets.

Table IV shows the accuracy for defect classes with sizing for the testing dataset. The result indicates that accuracy for each sizing class varies, ranging from 80% to 100%. However, when evaluating the defect class as a whole, Class "sp" records 93.27% while Class "ed" records 89.18% as an average across the three cross-validation folds.

A comparison experiment has also been conducted with two different CNN-based object detection models: SSD with VGG16 backbone and SSD with ResNet50 backbone. The two CNNs similar accept full-size 480 × 480 pixels images as inputs and output nine classes in the final fully connected layer. The comparison results show that SSD-VGG16 model has an accuracy of 94.17% and SSD-ResNet50 model has an accuracy of 91.52%. Besides comparing testing accuracy results, we compare the average training hours and inference time between the 2 SSD-based models against the proposed *DLADC*. Both SSD networks require up to 8 hours for each training round with 0.1297 second per image during the inferencing stage. However, the *DLADC* only requires less than 1 hour for each training round with 0.0881 second per image during inferencing. The comparison results are provided in Table V, in which we conclude that *DLADC* showcases consistent and great recognition performance for wafer surface defects and trains much faster and has more efficient inference speed than other CNNs.

Moving forward, we analyze the testing dataset result based on the confusion matrix. The confusion matrix reveals that several classes are more accurately classified as compared to other classes. Table VI shows the confusion matrix for the *DLADC* performance on the testing dataset. The confusion matrix shows that 9 out of 11 classes are classified with greater than 90% accuracies, including those with sizing calculation.

For further understanding, we study the effects of the count of training samples on the model's performance. We also investigate the relationship between average testing accuracy and the percentage count of training samples per defect class. The study suggests that even though Class "res" and Class "rare" have the lowest percentage count of training samples, the model achieved 96.30% and 93.62% testing accuracy for these classes, respectively. Comparatively, Class "ed" performed the worst amongst the other classes while having a fair number of training samples (9.29%). Hence, the study concludes that the count of training samples does not have an apparent effect on the model performance. However, since the distribution has been adjusted when performing

TABLE VI
CONFUSION MATRIX FOR TESTING DATASET.

		Predicted										
		sp_small	sp_large	ed_small	ed_large	pd	scr	res	bub	po	rare	False Count
Actual	sp small	**98.2%**	0.0%	0.0%	1.8%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	sp large	0.0%	**95.1%**	0.0%	0.0%	3.7%	0.0%	0.0%	0.0%	0.0%	0.0%	1.2%
	ed small	12.1%	0.0%	**84.8%**	0.0%	3.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	ed large	0.0%	8.9%	0.0%	**80.4%**	10.7%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	pd	0.3%	1.4%	0.3%	2.3%	**91.6%**	0.3%	0.9%	0.6%	0.9%	0.3%	1.2%
	scr	0.0%	0.0%	0.0%	0.0%	1.2%	**94.1%**	0.0%	0.0%	3.5%	0.0%	1.2%
	res	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	**96.3%**	0.0%	3.7%	0.0%	0.0%
	bub	0.0%	0.0%	0.0%	0.0%	1.0%	0.0%	0.0%	**96.0%**	1.0%	0.0%	2.0%
	po	0.5%	0.0%	0.0%	0.0%	0.0%	0.5%	0.0%	0.5%	**96.8%**	0.5%	1.1%
	rare	0.0%	2.1%	0.0%	0.0%	2.1%	0.0%	0.0%	0.0%	2.1%	**93.6%**	0.0%
	False Count	0.0%	0.0%	0.0%	0.0%	0.6%	0.6%	0.0%	0.0%	0.0%	0.0%	**98.9%**

augmentation to increase training samples, this study proposed that dataset balancing is essential in ensuring model performance robustness.

In summary, the experiments conducted using actual semiconductor's Review-SEM defect images have demonstrated that the proposed ADC system, *DLADC*, can detect and classify wafer surface defects without manual feature extractions and with an improved accuracy performance at 93.69%. In addition, previous related works approach the same problem statement using CNNs as image classifiers. In contrast, this paper presents its novelty by approaching CNN as an object detector which is far more challenging since it combines two tasks: object localization and image classification. Nevertheless, the proposed *DLADC* achieves an average accuracy of 93.69% using only SEM images as input, outperforming both the existing manual classification benchmark of 70% and the industry-ready product benchmark of 90%. *DLADC*'s autonomous design and quick inference speed allow it to cut the time required for manual review and reclassification in half, resulting in a reduction in miss or delay tool inhibitions due to shortened review cycletime in a real-world semiconductor manufacturing environment.

V. CONCLUSION

This research proposed a deep-learning-based defect classification model (*DLADC*) using CNN-based object detection architecture to detect and classify wafer surface defects from metal layers. The proposed *DLADC* system firstly takes Review-SEM images as input to the Faster R-CNN model for defect detection and scale bar detection. With the detected objects (defect and scale bar), sizing calculation is further applied for Classes "sp" and "ed".

Based on our investigation, the proposed *DLADC* system represents the one of the first examples of deep learning based ADC, which considers defect sizing and focuses on metal layers out of the other many semiconductor layers. This proposed ADC system can classify semiconductor surface defects with high and production-ready accuracy when tested on actual semiconductor data. *DLADC*'s autonomous architecture also allows for a 2x time savings over manual SEM review and reclassification. The primary limitation is the small and non-uniform dataset.

Moreover, the high accuracy of the proposed *DLADC* suggests that the solution might be extended to all other wafer layers as well, with only Review-SEM images as the training dataset. On the other hand, the overall classification accuracy may be improved by experimenting with better-advanced CNNs with larger training datasets and more uniform distribution across defect classes. Concerning the applicability of transfer learning, the study reveals that it can be adopted to increase the model's training efficiency and shorten training time for the defects classification problems in semiconductor wafers.

ACKNOWLEDGMENT

The authors would like to thank both X-FAB and Swinburne University of Technology in supporting this research.

REFERENCES

[1] G. Tello, O. Y. Al-Jarrah, P. D. Yoo, Y. Al-Hammadi, S. Muhaidat and U. Lee, "Deep-Structured Machine Learning Model for the Recognition of Mixed-Defect Patterns in Semiconductor Fabrication Processes," in *IEEE Transactions on Semiconductor Manufacturing*, vol. 31, no. 2, pp. 315-322, May 2018, doi:

10.1109/TSM.2018.2825482.
[2] K. Taha, "An Effective Approach for Associating the Sources of Defect Signatures to Process Zones," *in IEEE Transactions on Semiconductor Manufacturing*, vol. 30, no. 2, pp. 176-184, May 2017, doi: 10.1109/TSM.2017.2679714.
[3] T. Nakazawa and D. V. Kulkarni, "Wafer Map Defect Pattern Classification and Image Retrieval Using Convolutional Neural Network," in *IEEE Transactions on Semiconductor Manufacturing*, vol. 31, no. 2, pp. 309-314, May 2018, doi: 10.1109/TSM.2018.2795466.
[4] Chia-Yu Hsu, "Clustering Ensemble for Identifying Defective Wafer Bin Map in Semiconductor Manufacturing", *Mathematical Problems in Engineering*, vol. 2015, Article ID 707358, 11 pages, 2015. https://doi.org/10.1155/2015/707358
[5] M. Wu, J. R. Jang and J. Chen, "Wafer Map Failure Pattern Recognition and Similarity Ranking for Large-Scale Data Sets," in *IEEE Transactions on Semiconductor Manufacturing*, vol. 28, no. 1, pp. 1-12, Feb. 2015, doi: 10.1109/TSM.2014.2364237.
[6] M. Saqlain, B. Jargalsaikhan and J. Y. Lee, "A Voting Ensemble Classifier for Wafer Map Defect Patterns Identification in Semiconductor Manufacturing," in *IEEE Transactions on Semiconductor Manufacturing*, vol. 32, no. 2, pp. 171-182, May 2019, doi: 10.1109/TSM.2019.2904306.
[7] H. Corporation., 2019. 6. Review SEM - What is a Review SEM? Hitachi High-Technologies GLOBAL. [online] *Hitachi-hightech.com.* Available at: <https://www.hitachi-hightech.com/global/products/device/semiconductor/review-sem.html> [Accessed: 12 December 2019].
[8] F. Adly et al., "Simplified Subspaced Regression Network for Identification of Defect Patterns in Semiconductor Wafer Maps," in *IEEE Transactions on Industrial Informatics*, vol. 11, no. 6, pp. 1267-1276, Dec. 2015, doi: 10.1109/TII.2015.2481719.
[9] M. B. Alawieh, D. Boning and D. Z. Pan, "Wafer Map Defect Patterns Classification using Deep Selective Learning," *2020 57th ACM/IEEE Design Automation Conference (DAC)*, 2020, pp. 1-6, doi: 10.1109/DAC18072.2020.9218580.
[10] P. Stich, M. Wahl, P. Czerner, C. Weber and M. Fathi, "Yield prediction in semiconductor manufacturing using an AI-based cascading classification system," *2020 IEEE International Conference on Electro Information Technology (EIT)*, 2020, pp. 609-614, doi: 10.1109/EIT48999.2020.9208250.
[11] "Faster, More Accurate Defect Classification using Machine Vision", *Intel.com*, 2020. [Online]. Available: https://www.intel.com/content/dam/www/public/us/en/documents/best-practices/faster-more-accurate-defect-classification-using-machine-vision-paper.pdf.
[12] 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," in *IEEE Transactions on Semiconductor Manufacturing*, vol. 32, no. 2, pp. 163-170, May 2019, doi: 10.1109/TSM.2019.2902657.
[13] J. O'Leary, K. Sawlani and A. Mesbah, "Deep Learning for Classification of the Chemical Composition of Particle Defects on Semiconductor Wafers," in *IEEE Transactions on Semiconductor Manufacturing*, vol. 33, no. 1, pp. 72-85, Feb. 2020, doi: 10.1109/TSM.2019.2963656.
[14] Y. Yuan-Fu and S. Min, "Double Feature Extraction Method for Wafer Map Classification Based on Convolution Neural Network," *2020 31st Annual SEMI Advanced Semiconductor Manufacturing Conference (ASMC)*, 2020, pp. 1-6, doi: 10.1109/ASMC49169.2020.9185393.
[15] S. Zhou, Y. Chen, D. Zhang, J. Xie, and Y. Zhou, "Classification of surface defects on steel sheet using convolutional neural networks, " in *Materials and Technology*, vol. 51, no. 1, pp. 123–131, 2017. Available: 10.17222/mit.2015.335.
[16] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M. and Berg, A.C., 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3), pp.211-252.
[17] Simonyan, K. and Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556.*
[18] He, K., Zhang, X., Ren, S. and Sun, J., 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).
[19] Girshick, R., Donahue, J., Darrell, T. and Malik, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 580-587).
[20] Girshick, R., 2015. Fast R-CNN. *In Proceedings of the IEEE international conference on computer vision* (pp. 1440-1448).
[21] Ren, S., He, K., Girshick, R. and Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497.*
[22] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y. and Berg, A.C., 2016, October. SSD: Single shot multibox detector. In *European conference on computer vision* (pp. 21-37). Springer, Cham.
[23] Y. Wu, A. Kirillov, F. Massa, W. Lo and R. Girshick, "facebookresearch/detectron2", *GitHub*, 2019. [Online]. Available: https://github.com/facebookresearch/detectron2.
[24] I. Goodfellow, Y. Bengio, and A. Courville. Deep Learning. Cambridge, MA, USA: *MIT Press*, 2016.
[25] Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P. and Zitnick, C.L., 2014, September. Microsoft COCO: Common objects in context. In *European conference on computer vision* (pp. 740-755). Springer, Cham.