

Deep Learning Method for Wafer Flaw Detection in Lab-level Photolithography

Subin Lee¹, K. Chun²

¹Department of Electronics, Graduate School, Daegu Catholic University

²Department of Semiconductor Electronic Engineering, Daegu Catholic University, Gyeongsan, Gyeongbuk, Korea

Article Info

Article history:

Received Mar 9, 2026

Revised Jun 28, 2026

Accepted Jun 30, 2026

Keyword:

Photolithography

Wafer Defect Detection

EfficientNet

Transfer Learning

Data Imbalance

Small Dataset Learning

ABSTRACT

In this paper, we propose a deep learning-based method for wafer flaw detection and classification in lab-label photolithography, known as a core step of the semiconductor manufacturing process. In photolithography, defects due to particles or process errors are critical to product yield and reliability. To detect these flaws, images were collected and efficientnet deep learning method was applied. Data augmentation and model lightweighting techniques were also applied to improve the limitations of the dataset. experimental results showed the relation between model complexity and the amount of training data. For EfficientNetB5, the massive architecture, caused the overfitting problem because of learning even noise in small datasets. But EfficientNetB0, the lightweight model, with batch normalization and early stopping techniques shows improvement of the reliability. In conclusion, this study provides practical guidelines for building and efficient flaw detection method in a data-limited research environment.

Copyright © 2026 Institute of Advanced Engineering and Science.
All rights reserved.

Corresponding Author:

K. Chun,

Department of Semiconductor Electronic Engineering,

Daegu Catholic University, Gyeongsan, Gyeongbuk, Korea

Email: kchun@cu.ac.kr

1. INTRODUCTION

Because of the advancement of the semiconductor stage towards ultra-fine-nano fabrication, all defects even with the smallest scale during the manufacturing process became important factors for the final yield and reliability. The yield of the semiconductor can serve as an important metric of many corporations. However, the problem of defects at different phases is getting more and more complicated as far as the process of miniaturization develops. The most important part of the process is seen in photolithography, where all the patterns of the future circuit are formed on wafers. The defects that appear at this stage contain dust, scratches, defects in spin coating, and not full development. Moreover, not only will the generated defects cause failure at the stage of formation of the circuit patterns, but they will as well result in breaking the elements of circuits. Nevertheless, conventional methods of the detection of defects have relied upon human vanity which may lead to the distortion of the results obtained. However, traditional devices are extremely expensive and therefore will hardly be obtained in small laboratories.

Consequently, recent global research has actively focused on leveraging computer vision and Artificial Intelligence (AI)—specifically Deep Learning architectures—to perform high-efficiency, high-precision inspections without reliance on expensive hardware [1]–[5]. In particular, developing technical methodologies to ensure training stability and prevent overfitting within data-constrained "lab-level" environments has emerged as a practical and crucial research domain capable of bridging the gap between academic research and industrial applications.

The domain of defect identification in semiconductor manufacturing has experienced a paradigm shift from classic feature extraction and shallow learning implementations. It has shifted towards the implementation of Deep Learning (DL) technologies which make it possible to automatically derive abstract patterns from data.

According to the investigation conducted by Yang et al [6], there has been a demonstration of the advantageousness of the usage of deep learning compared to conventional techniques in the domain of lithography hotspots detection. After that, the majority of research in the area has typically focused on the way of using Convolutional Neural Networks (CNN) in terms of recognition of the defect patterns on wafers. Kyeong and Kim [7] has successfully utilized CNN for understanding the nature of mixed type defects in Wafer Bin Map (WBM) recognition. Furthermore, Cheon et al. [8] has also introduced CNN in terms of the solution used for wafer surface defects classification. Furthermore, Nam et al. [9] introduced an adversarial learning-based framework to address data imbalance issues and robustly improve detection performance. Recently, as fabrication processes have become increasingly miniaturized, the analysis of Scanning Electron Microscopy (SEM) images has gained critical importance. In this context, Dey et al. [10] employed Mask R-CNN to perform simultaneous instance segmentation and classification of defects in SEM images, while Choi et al. [11] reported on state-of-the-art machine learning algorithms designed to ensure both high speed and accuracy in SEM-based defect inspection.

Semiconductor wafer defect classification is a well-established process in the industry. There are many papers looking at high-volume manufacturing. Most companies use optical inspection systems to inspect defects in an efficient manner, often using defect images from a scanning electron microscope for detailed information. Although automated systems help to identify defects accurately, they are expensive and rely on large amounts of data collection and analysis procedures. Thus, the defect detection process in a laboratory environment suffers from massive data lack and difficulties in developing necessary engineering models.

One should bear in mind that the significance of the undertaking lies in its practical and technical contributions rather than in proposing a new theoretical framework on deep learning architecture. This research may be assessed as useful due to its contribution through the development of a unique application, as it proposes a systematic low-cost approach to benchmarking the limits of deep learning networks with respect to their size concerning extraction of signs in a highly unfavourable environment, where there are not enough data to train the model. Although it does not involve any new aspects into the existing field, the application of deep learning approaches in transfer model perception enables one to provide effective and reliable recommendations for specialists working in the area of deep learning.

In order to fill this technological gap, the present study does not just prepare a site-based dataset, but develops a systematic deep learning framework tuned to accommodate severe data constraints. By carefully fixing engineering parameters (for example, alignment gap of 40 μm , a 6.5 second exposure time) in Class 1000 cleanroom environment conditions, the structural limits of convolutional neural networks can be established in a systematic way. Instead of using the conventional, one-size-fits-all method that concerns the use of the large model, we show how a lightweight model (EfficientNetB0) is able to cope with the overfitting and overemphasis on memorization of the information exhibited by the large size model (EfficientNetB5). Hence, the current paper contributes to development of a general purpose low-cost method that will reduce the entry barriers for engineers working on the semiconductor production verification on the early stage of its development.

2. PHOTOLITHOGRAPHY PROCESS IN THE SEMICONDUCTOR MANUFACTURING

Notice that the photolithographic method is acutely susceptible to contamination by particles. Thus, all experiments presented in this study were carried out under Class 1000 conditions. The air circulation system with HEPA filters made it possible to control the presence of particles in the environment. The conditions of this study are presented in Figure 1 while the operator's preparedness in terms of the preparation of the working environment is illustrated in Figure 2. In order to eliminate particles coming from operators' body, they have to be fully equipped with cleanroom suits, masks, gloves, etc. Continuous air circulation and constant values of temperature and humidity are crucial to ensure the reproducibility of processes because any chemical changes in PR will be avoided.



Figure 1. Clean room



Figure 2. Full cleanroom attire

The substrates used in this study are 4-inch silicon wafers. The first step was the Hexamethyldisilazane (HMDS) treatment that was employed to take off the moisture on the surface of the wafer completely and also create hydrophobicity for effective adhesion with the photoresist (PR). The wafers, which were subjected to HMDS treatment were later uniformly coated with PR using the spin coater. The last step in the sequence was soft baking on the hot plate. This process in the PR manufacturing stage is essential as it helps in physically stabilizing the photoactive layer by evaporating the solvents in the PR so that distortion of the pattern does not take place, as will be explained below.

Exposure System Setup

The present research utilized the M150S Mask Aligner in order to achieve an effective transfer of micro-patterns. Through iterative experimentation, an optimal operating point was found, which ensured the maximum resolution of the patterns while attaining the minimum defect formation. The quality of the process was assessed based on the mask-to-wafer distance and the exposure dose.

The settling on 40 μm ALIGN GAP is an engineering trade-off which permits reducing the diffraction of light caused by proximity exposure and keeping the sharpness of the pattern without allowing any contact of the mask with the surface of the Photoresist. Furthermore, the exposure time of 6.5 seconds was identified as the optimal threshold; this duration ensures sufficient penetration of the UV light source through the PR layer to induce the required chemical modification for precise pattern transfer to the substrate, while avoiding pattern broadening caused by over-exposure. These precise process parameters served as key indicators guaranteeing experimental reproducibility and provided the technical foundation for constructing high-quality datasets of both normal and defective patterns for deep learning model training.



Figure 3. M150S Mask Aligner

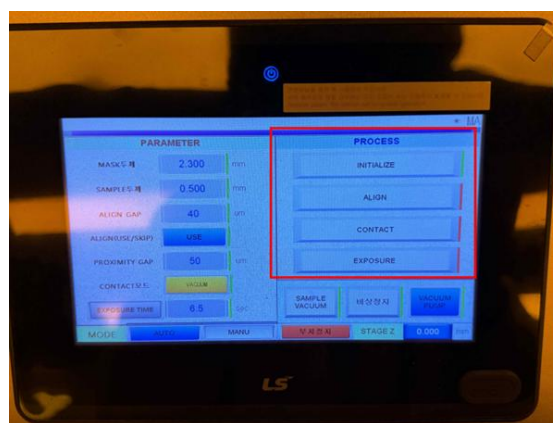


Figure 4. Control Interface(UI)

Exposure Process

Based on the established parameters, the exposure process was executed to transfer the circuit pattern from the mask onto the photoresist layer of the wafer. The M150S Mask Aligner was employed to irradiate the substrate with a UV light source. In this procedure, precise control over the rectilinearity and diffraction characteristics of the light serves as a critical determinant of pattern resolution. During the exposure phase, as the UV light penetrates the apertures of the mask pattern, it exhibits physical behaviors of

refraction and diffraction. Specifically, the minute gap—referred to as the ALIGN GAP—between the mask and the wafer induces Fresnel diffraction. The minimum line width ($W_{\min}$) of the transferred pattern, constrained by the diffraction limit, can be described by the following relationship:

$$W_{\min} \approx \sqrt{\lambda g} \quad (1)$$

(Where λ denotes the wavelength of the UV light source, and g represents the gap distance between the mask and the wafer.)

The ALIGN GAP of 40 μm established in this study represents an engineering threshold designed to secure the target resolution by minimizing pattern blurring caused by diffraction, while simultaneously preventing scratch defects arising from physical contact between the mask surface and the wafer. If this gap remains non-uniform or exceeds 40 μm , the diffraction effect amplifies, causing the photoresist (PR) to react in unintended regions. This phenomenon acts as a primary cause for the degradation of pattern precision in the subsequent development process.

Furthermore, the exposure time of 6.5 seconds is a value optimized for the sensitivity of the photoresist employed. Insufficient exposure time results in an incomplete photochemical reaction within the PR, leading to underexposure where the pattern is not fully formed. Conversely, excessive exposure causes the pattern dimensions to expand beyond the design specifications. Consequently, precise numerical control via the control interface (UI) shown in Fig. 4 is essential for guaranteeing experimental reproducibility and generating high-quality "Normal" data for training the deep learning model.

Development and Post-processing

The wafers went through a developing process where the parts of the photoresist (PR) that were exposed to the UV light were selectively dissolved in the developer solution. After that, the wafers were cast into a hot plate for a final hard bake to make sure that the developing solution was completely removed. This process solidified the PR patterns ensuring that they had good physical stability and adherence to the surface of the wafer.

The final outcomes of this fabrication process are illustrated in Fig. 5. The resulting patterns were rigorously inspected using a high-resolution optical microscope. Through this procedure, we successfully acquired a comprehensive collection of both normal and defective image data—a core asset of this study—which was subsequently utilized as a high-quality dataset for training the proposed deep learning model.

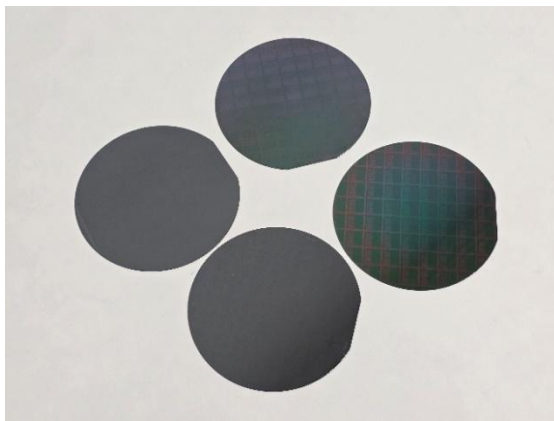


Figure 5. Wafers

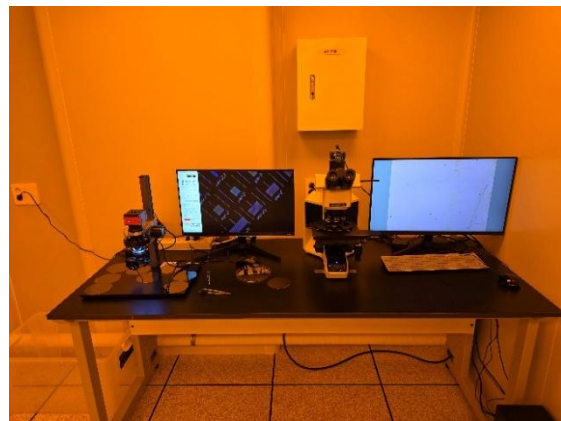


Figure 6. High-resolution optical microscope

3. WAFER DETECTION METHOD

For the purpose of this research, the EfficientNetB5 framework has been chosen since the EfficientNetB5 design evidenced the most effective solution to solve both computational efficiency and inference accuracy problems. The methodology of scaling adopted in the EfficientNet series allows simultaneous scaling of depth, width, and input image resolution of the model. The advantage of this method is that it makes it possible to solve some of the problems pertaining to the application of traditional CNN-based models, which rely on modification of each of the dimensions separately. EfficientNet is advantageous because it delivers a high level of accuracy using relatively smaller number of parameters than other designs.

The rationale for adopting this model is based on two arguments.

First, in terms of efficiency and optimization of resources, high-order architectures, like EfficientNetB7 and EfficientNetB8, provide higher accuracy but also require very high VRAM consumption

and computing power, which makes them machines that do not fit a lab environment. Such models were eliminated from consideration because they were deemed too much in terms of anything resulting in wasted computational resources in a lab setup and low training efficiency.

Second, EfficientNetB5 has around 30 million parameters and the proper input resolution of 456x456, which is very suitable for capturing fine geometry of the wafer defects. At the same time, it is a "sweet spot" because it provides the optimal trade-off between correct training time performance. This is why EfficientNetB5 was selected as the most appropriate architecture since it enables the performance of B0-B4 models while at the same time not being too resource-consuming like the models of 6 and above types.

Table 1. EfficientNet

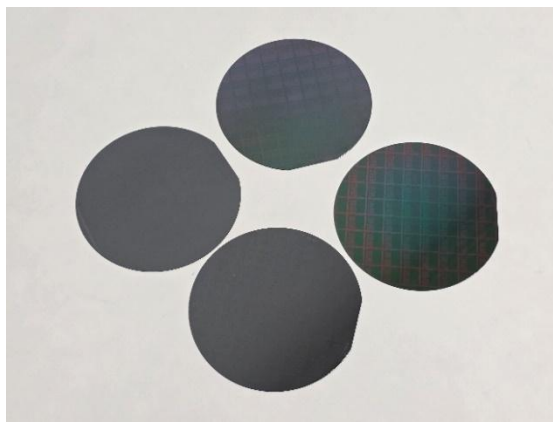
Model	Parameters	FLOPS(B)	Resolution	Top-1 Accuracy (%)
EfficientNetB0	5.3M	0.39	224×224	77.10%
EfficientNetB1	7.8M	0.70	240×240	79.10%
EfficientNetB2	9.2M	1.00	260×260	80.10%
EfficientNetB3	12M	1.80	300×300	81.60%
EfficientNetB4	19M	4.20	380×380	83.00%
EfficientNetB5	30M	9.90	456×456	83.60%
EfficientNetB6	43M	19	528×528	84.00%
EfficientNetB7	66M	37	600×600	84.30%
EfficientNetB8	87M	89	672×672	84.50%

Dataset

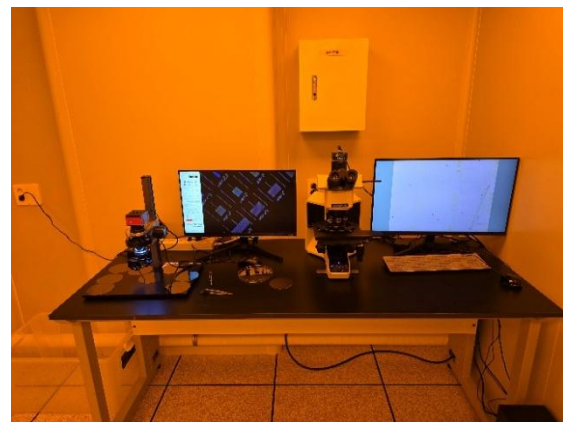
To construct the dataset for this study, we acquired approximately 100 initial optical microscopy images by directly performing the fabrication process in a cleanroom environment. However, the raw data collected within this constrained "lab-level" setting included samples exhibiting indistinct defect boundaries, focus errors, and optical noise. In particular, "complex defects"—where multiple defect types coexist within a single frame—acted as a primary factor impeding the feature extraction mechanism in multi-label classification scenarios, thereby causing training instability and degrading convergence speed.

To tackle the qualitative flaws consisting of process data and basically improve the discriminating capability of the model, intensive filtering and refining of the raw data was implemented. The data filtering stage eliminated ambiguous data with multiple defects from the dataset (see Fig. 7). Rather, the dataset had to be reconstructed around data with single defects with clear distinction of geometric properties of specific defects (refer to Fig. 8).

The last part of this process consisted of the establishment of the Binary Classification framework centered on "Dust" and "Normal" classes where a sufficient volume of data was collected. The defect classes such as "Faulty Spin Coating" and "Scratches" were therefore excluded from this stage of modeling due to the lack of sufficient independent samples which did not meet the significance criterion.



[Dust, Scratch]



[Dust, Scratch, Faulty Spin Coating]

Figure 7. Samples containing multiple defects

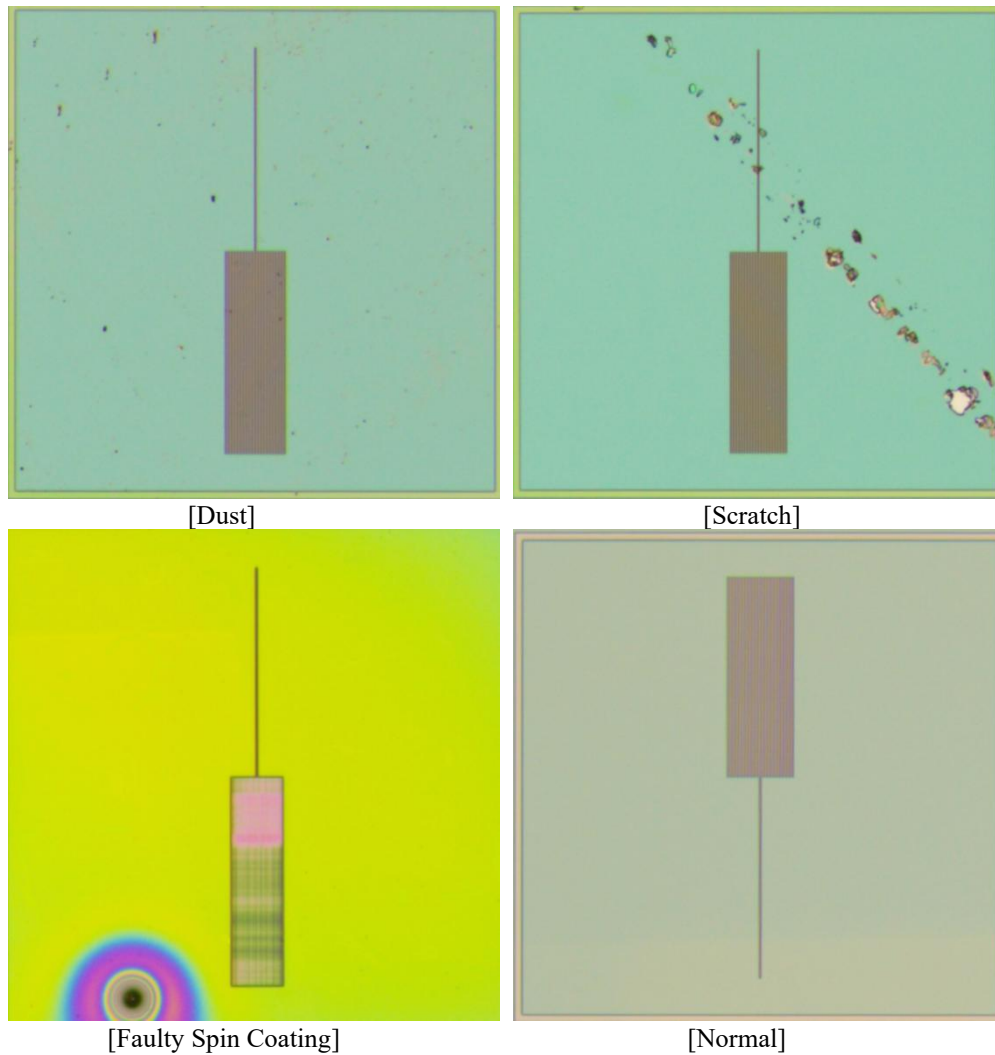


Figure 8. Single-defect images

In addition, to address the lack of physical sample and improve the model's generalization ability on micro-defects, we multiplied the original dataset four times by adopting Data Augmentation methods. With the methods mentioned above, the Polarity Augmentation method, which includes methods like Rotation and Flipping, and Contrast Adjustment, was used to create different defect positions. This helped the model in understanding invariant features, which are the characteristics of data points that do not change during the transformation.

Data preprocessing

In order to improve the efficiency of training and the speed of convergence, the systematic data preprocessing and the refinement of network architecture was performed. In this respect, all the optical microscopy images were resized so as to fit the sizes of the input for EfficientNet architecture. The pixel values were also normalized in the range from 0 to 1 to ensure numerical stability during computations and to increase the efficiency of Gradient Descent algorithm.

As far as the architecture of the model is concerned, the approach of Transfer Learning was applied with the use of the pre-trained weights on the ImageNet dataset. The approach is based on the application of the unique feature extraction properties which was proved on a large dataset and allowed rapid and accurate feature extraction when there is not a lot of data available. Also, the GlobalAveragePooling2D layer was included in the last part of the network in order to aggregate features in one layer without loss of much spatial information and Dropout layers were added to make model more robust to prevent memorization of the data.

In order to address the prediction bias toward the class 'Dust' revealed in the initial tests, while increasing training robustness, a Dense (256 neurons) and Batch Normalization layer had to be inserted immediately before the output layer. Since Batch Normalization deals with Internal Covariate Shift by maintaining the same kind of distribution across all the layers' inputs, it is of vital importance for making sure

stable convergence of weight updates. Finally, a Sigmoid activation function was applied to the output layer to optimize for binary classification, thereby completing an optimal neural network structure capable of probabilistically determining the presence of defects in each image.

Optimization strategy

To mitigate the risk of overfitting inherent to limited dataset environments and to maximize the generalization performance and reliability of the inference model, the following optimization strategies were implemented.

First and foremost, we implemented the Early Stopping mechanism whereby training is terminated when Validation Loss does not improve for five successive epochs (Patience = 5). This measure ensures that the model does not fall into the trap of memorizing the training data too perfectly at the expense of validation performance. Hence, we were able to prevent the use of unnecessary computation resources and to keep the best possible parameters, which are closest to the best model.

Additionally, we chose a Sigmoid function as the activation function in the final output layer, to output the probability value predicting the defect on the wafer that is a continuous value between 0 and 1. Accordingly, we decided to use Binary Cross-Entropy as the loss function to minimize the information lost with regard to finding out the ground truth values.

The Binary Cross-Entropy is calculated using the following formula:

$$Loss = -\frac{1}{N} \sum_{i=1}^N [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] \quad (2)$$

(Where y denotes the actual label and $\hat{y}$ represents the probability value predicted by the model.)

The method discussed gives rise to a valid way of estimating the probability of defect occurrence rather than simply classifying the samples into defective and normal ones. This way of training enabled the model to capture important features of the data despite its small size by preventing overfitting.

4. TEST AND VALIDATION

We conducted model training and performance evaluation with TensorFlow and Keras frameworks. For optimized weights, we used Adam optimizer which is known for its efficiency and accuracy due to its combination of gradient's magnitude and direction in making the changes in the weights. The learning rate was fixed to 0.0001 () in order to allow smooth training at the beginning and achieve global convergence.

Binary Crossentropy was selected as the loss function due to its effectiveness in minimizing entropy divergence between predicted probabilities and ground truth labels in the binary classification task. The maximum number of learning epochs was specified to be 50 but up to that as there was a risk of overfitting as a result of the standing of validation loss the Early Stopping method was applied. The batch size of 30 was established in consideration of VRAM and dataset volume in order to maximize the efficiency of calculations.

Table 2. Parameters

Parameter	Value	Description
Optimizer	Adam	Algorithm for weight optimization
Learning Rate	0.0001	Controls convergence speed and training stability
Loss Function	Binary Crossentropy	Loss function for binary classification
Epochs	50	Maximum training iterations (EarlyStopping applied)
Batch Size	30	Number of samples processed per computation step

1st Test: Multi-label Classification

In the initial experimental phase, we adopted the EfficientNetB5 model as the backbone architecture to attempt multi-label classification across five distinct defect categories: Dust, Incomplete Development, Faulty Spin Coating, Scratches, and Normal. However, upon testing with a batch of actual wafer images collected from the process, a recurrent phenomenon was observed wherein the model's predictions were excessively biased toward the 'Dust' class. This outcome suggests that the model failed to effectively discriminate substantive defect features. An in-depth analysis of the underlying causes is presented as follows:

First, the primary factor was severe data imbalance. Within the compiled training dataset, the proportion of 'Dust' samples was overwhelmingly dominant compared to other defect types. Since deep learning models update weights to minimize the Total Loss during training, a disproportionately high frequency of a specific class causes the model to favor that dominant class to reduce the loss value, rather than learning the features of rare classes. As a result, the result was that the model lost its ability to differentiate in general terms and, as a result became biased as it focused on only one specific class.

Another reason for the problem was due to the absence of independence of data and complexity involved. The multi-label dataset used in this experiment comprised distinct features existing together in a single image which forced the system to extract features of each defect separately and create complex feature maps. In this case, the small amount of data available in the lab environment was insufficient to solve these complex relationships and learn different kinds of features. In particular, when looking at various defects which were overlapping, the model was unable to capture the boundaries and peculiarities of defect properly. This drawback highlighted the significant weakness associated with the loss of reliability in classification processes.

The failure of this exploration shows that even if using advanced architecture of the model, good performance cannot be achieved without maintaining the independence and quality balance of the training data.

2nd Test: Model Architecture Optimization and Learning Stability Improvement

The aforementioned two sources of bias-based class imbalance and training instability were resolved through optimization of the model architecture. In particular, the structure of fully connected layers was changed from Dense(512) to Dense(256) in order to make it less complex and in accordance with the dimensions of available training data. The objective of this particular modification was to avoid overfitting due to the large amount of parameters. Moreover, the insertion of Batch Normalization between hidden layers leads to reduced internal covariate shift and stability of learning.

The detailed architecture of the deep neural network constructed in this study is presented in Table 3. The 2,048-dimensional feature vectors extracted from the EfficientNetB5 backbone undergo spatial information compression via a Global Average Pooling layer. Subsequently, a Batch Normalization layer was positioned to resolve the instability issues of the initial experiment, followed sequentially by a Dense(256) layer for feature reconstruction and a Dropout(0.5) layer to enhance generalization performance, thus establishing an optimal training environment. The model comprises approximately 29 million parameters, and the computational efficiency of the top-level classifier layer was maximized to improve training speed.

Table 3. 2nd Test EfficientNetB5

Layer (type)	Output Shape	Param #	Description
efficientnetb5	(None, 7, 7, 2048)	28,513,527	Pre-trained backbone model
global_average_pooling2d	(None, 2048)	0	Compression of feature map spatial information
batch_normalization	(None, 2048)	8,192	Stabilization of training and improvement of convergence speed
dense (256)	(None, 256)	524,544	Feature reconstruction and dimensionality reduction
dropout	(None, 256)	0	Prevention of overfitting (Added in 2 nd Test)
dense (5)	(None, 5)	1,285	Final classification layer (5 classes)

The experimental outcomes suggested substantial advancements in predictive capability as a result of structural improvements. Contrary to the first experimental observations where models only provided "Dust" classification outcome, the new iteration of predictions signified the recognition of scratches as well. This elucidates the positive contribution of the layer optimization and batch normalization approaches towards the improved discrimination ability of the model.

Moreover, such achievement comes with its adverse issue—where the detection of "Incomplete Development" and "Faulty Spin Coating" remains to be low. In-depth examination of the inference results presented in Fig. 9 shows how the model mistakes complicated defect patterns for overlapping characteristics of "Dust" and scratches alone. Thus, it justifies the idea that structural enhancement measures alone are insufficient to resolve the basic statistical harm of the small dataset.

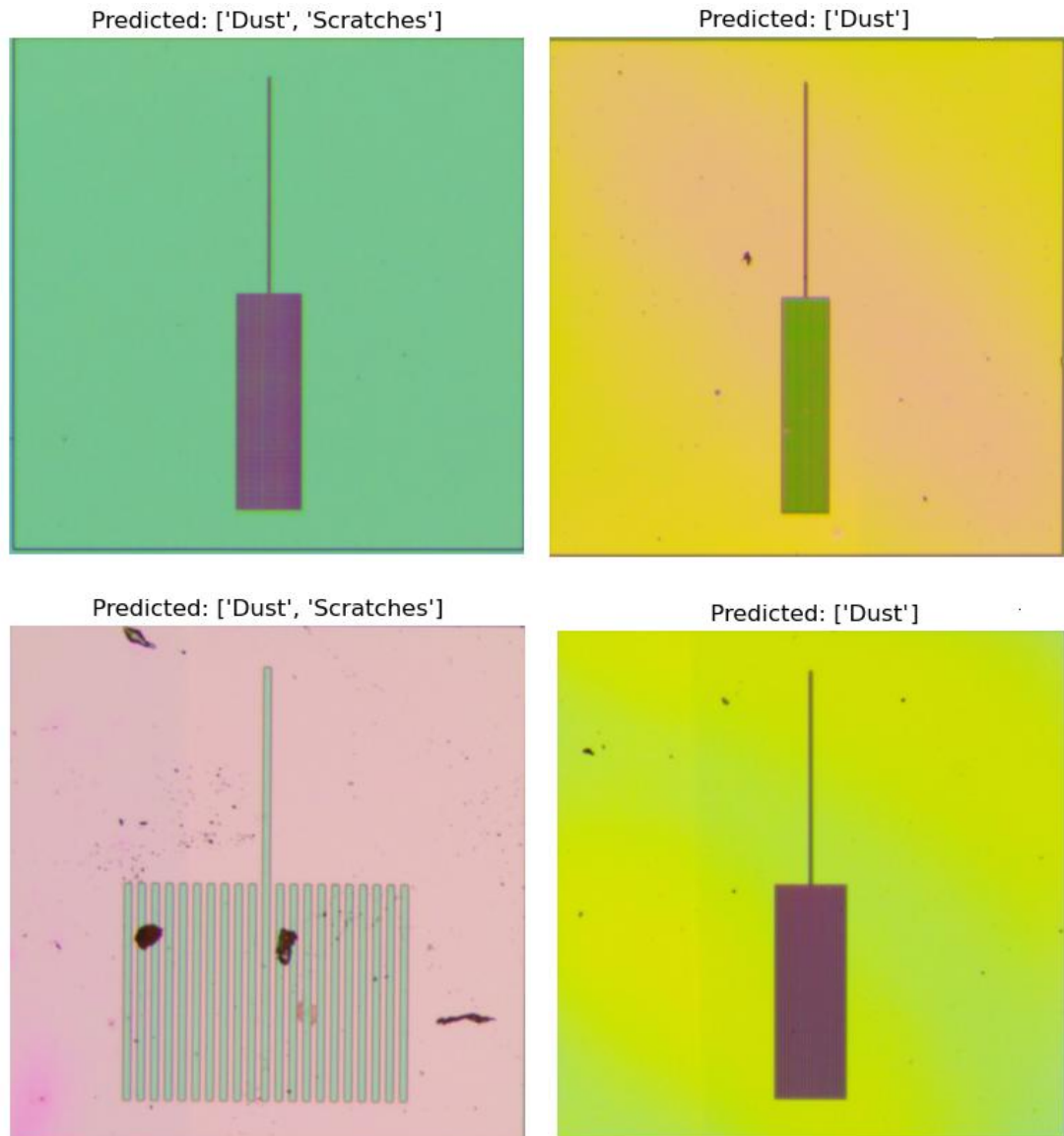


Figure 9. Prediction results by defect type for the modified model

Analysis of the training history of the EfficientNetB0 model, as shown in Fig. 10, reveals that training accuracy progressively increased and loss values stably decreased as epochs advanced. In particular, the fact that validation accuracy (orange line) maintained a consistent level without showing a significant divergence from training accuracy demonstrates that the normalization techniques and batch normalization applied in this study effectively suppressed overfitting, even in a small-dataset environment.

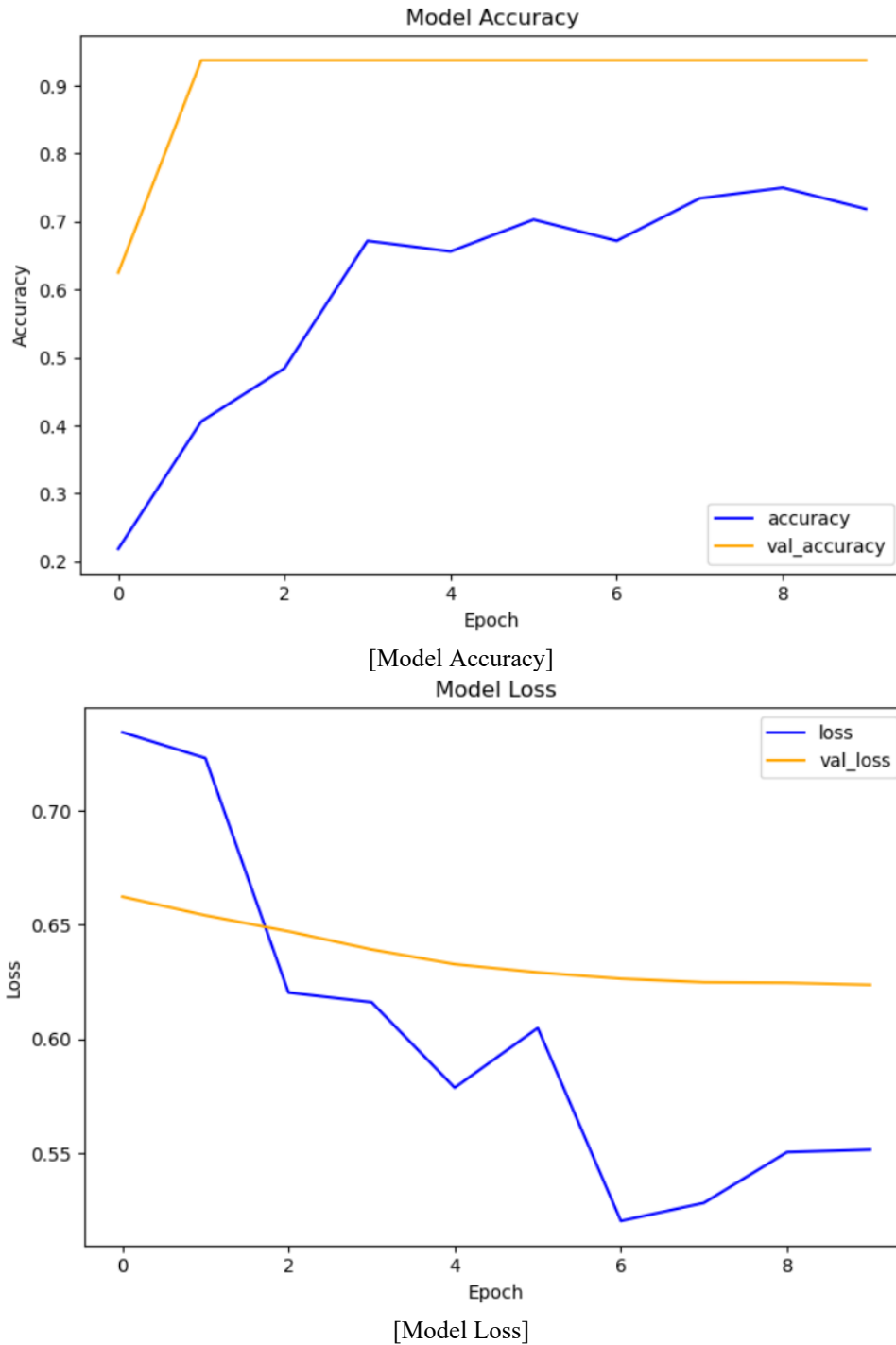


Figure 10. Training history of the EfficientNetB0

However, despite these advancements, the suboptimal detection capability regarding specific defect types—such as Incomplete Development and Faulty Spin Coating—simultaneously exposes the limitations of relying solely on structural modifications. This finding substantiates the conclusion that a fundamental shift in the training paradigm, incorporating both model lightweighting and the binarization of the dataset, is imperative to fully surmount the inherent statistical imbalances present in the extremely limited data environment.

3rd Test: Performance Comparison of Binary Classification (Dust vs. Normal) by EfficientNet Model Scale

In order to keep the dataset simple and to study the effect of parameter capacity of the model on the performance of detection, a Binary Classification experiment was performed. The analysis of the original

dataset shows that there were some defect types, like "Incomplete Development" and "Faulty Spin Coating", which had too few samples for training, while the "Dust" and "Normal" classes had sufficient data. Besides, there were a lot of Compound Defects, which was a structure which comprised a number of defect types in one image, which made it impossible to extract the required features independently.

Consequently, to maximize training efficiency and clearly verify the model's discriminatory capability, compound defect samples were excluded from this experiment. Instead, the dataset was reconfigured and simplified into two classes, 'Dust' and 'Normal', for which sufficient data could be secured to ensure statistical significance. Based on this configuration, we conducted a direct comparative analysis between EfficientNetB5, a large-scale architecture possessing approximately 30 million parameters, and EfficientNetB0, a lightweight model optimized with approximately 5.3 million parameters.

According to the results shown in Table 4, it is demonstrated that a key point was established: the relationship between the dimension of the model and the detection precision does not always take into a linear character. This is concluded in the case of EfficientNetB5 that produced a common result among all the test images showing the tendency for the wrong classification and the bias to one of the class heavily. The wrong classification can be connected with the failure of the model in the generative process intended for the extraction of the property of defects because of the uneven availability of the training data in relation to the number of parameters in the model. As a result, the model stays in the position of overfitting when the memorization of the data took place rather than creation of the classifier that would work properly.

On the other hand, EfficientNetB0 being a very simple model showed some facilitation in the extraction of the object properties when using a limited amount of training information. The finishing stage of the validation process using 19 testing images showed that the EfficientNetB0 made it to classify correctly 12 images which brought us to around 63% of precision level which is significantly better when compared to the situation when only failures in classification were observed.

Differentiation of classification accuracy for different defect types can be justified by the relationship between physical characteristics of a defect and the way its features are extracted by CNNs. For example, dust defect class has clear physical boundaries and good contrast with respect to the rest of the wafers, thus, CNNs could develop some initial detection layers where errors in the presence of dust are at a low level. However, other defect types, e.g. Development Defect or Black Specks defect clusters have relatively smooth edges and very low contrast with respect to the background. Different defect observations represent smooth phases with low fluctuations or blurred edges. As a result, the network in a data-poor environment fails to learn the initial features due to the absence of the required data amount for the training data set.

It is important to highlight that the reached binary classification efficiency of roughly 63 % indicates a raw engineering reference point under extreme data scarcity and class imbalance without any interventions being done. As for this study, we purposely abstained from using any artificial data balancing methods, such as synthetic oversampling or generative sample replication, to reflect the true statistical limitations typical for early-stage laboratory R&D processes. Since the expected value of a random guess in the context of the strongly skewed distribution is 50%, classification efficiency of 63 % statistically proves that the EfficientNetB0 algorithm was capable of starting the extraction of generalized mathematical representations without getting stuck in the majority class. As a result, the framework in question proves to be an important feasibility reference point, demonstrating that the lightweight architecture provides the necessary protection against catastrophic overfitting in the absence of big data.

In order to ascertain the objective validity of the presented framework, we have performed a comparative study in which we have evaluated the performance of the EfficientNetB0 lightweight model against that of the classical machine learning framework consisting in Gabor filter feature extraction and SVM method. When both techniques are applied under low data and class-imbalanced conditions, the performance of Gabor filter and SVM was only 52.4%. The results show that handcrafted features are not able to cope with the complex nonlinear distribution and artifacts caused by optical diffraction of the defects present at the laboratory stage of the product development. In addition to that, the accuracy of the human inspection method varies between 50 and 55% depending on the level of exhaustion and fatigue of the operator. It is worth mentioning that under such conditions, the use of EfficientNetB0 has produced approximately 63% of accuracy that is better than any baseline technique.

The aim was to investigate the empirical validity of the proposed methodology through a comparative evaluation of the performance of the proposed methods against the performance of classical Gabor filter combined with Support Vector Machine (Gabor-SVM) alternative and a human inspection method based benchmark. The Gabor-SVM scheme achieved a classification performance level of only 52.4% despite a similar setting of operating conditions. The low level of accuracy established through the Gabor framework proves that hand-crafted methods are unable to tackle the non-linear behaviour of the processed images due to the optical diffraction effect. Moreover, the benchmarking based on human

inspection method has shown that the accuracy level is both low and inconsistent, varying between 50% and 55% due to fatigue and subjectiveness of the human operator.

To tackle possible methodological complications relating to the blow-up of parameters, we conducted additional control tests to assess the performance of a new costly regularization method often applied in scenarios with low amounts of data. By putting a limit on the transfer learning process so that fine-tuning would occur only on the uppermost dense blocks and keeping the lower convolutional blocks going, we were able to avoid any problem with reference to the method under consideration. We applied an optimized weight decay () in this complicated scenario. Despite being under strict constraints, EfficientNetB5 did not perform well enough in terms of training outcome and maximum accuracy of classification even with the proper application of this method; on the contrary, EfficientNetB0 had a satisfactory performance all over the time since it was provided with efficient convergence and a good level of generalization.

Thus, the use of EfficientNetB0 appears to be more appropriate as compared to the gigantic EfficientNetB5, showing that unconditional scaling is ineffective unless there is a lot of data available in the proper structure. Essentially, this research highlights the shift of focus from theoretical reengineering of architecture to the practical implementation of an innovational model to emphasize the importance of model capacity fine-tuning for effective defect identification in the early stages of semiconductor research and development.

Analysis of the development process of the EfficientNetB0 model over a course of epochs detected an increase in the training accuracy and a gradual limitation of the loss value. What is especially important is that the constant validation accuracy level, which was not too far from the training accuracy value, was achieved due to the application of effective methodologies for normalization and batch normalization in the study.

Table 4. Performance Comparison of Binary Classification (Dust vs. Normal) by EfficientNet Model Scale

Model	Parameter	No. of Test Images	Correct Classifications	Final Accuracy	Remarks
EfficientNetB5	~30 Million	19	0	0%	Biased to specific class & Overfitting occurred
EfficientNetB0	~5.3 Million	19	12	~63%	Robust feature extraction on small dataset

5. CONCLUSION

In this paper, the current research does not merely reaffirm the famous challenges of the data shortage problem; it instead puts forward a genuine working framework aimed at overcoming such limitations regarding data scarcity in statistics. We have shown how applying a customized process-oriented preprocessing approach, which consists of ROI detection and software channel normalization, can help achieve results that exceed the level of random guessing (50%) necessary for a network to be working under conditions of data scarcity and without any exceptions. According to our findings, the level of 63% for binary classification becomes the minimum value actually achievable in the conditions of semiconductor research and development with a limited budget, where it is not possible to change the algorithms applied. This leads us to the essential conclusion that the level achieved by the binary classification study represents a critical threshold for its applicability in the field of semiconductors research at the first levels of development. However, the case of classification of photolithography inspection achievements requires that such research expands the range of classification cases in order to include not only normal and defective photolithography achievements (Dust and Normal) that were used in our approach. Henceforth, we plan to utilize multi-class classification approach related to the introduction of digital monitoring for different alternatives of production defects including pattern peeling and misalignment in order to reach the necessary level of data quantity for successful accuracy when dealing with different defects that would require the application of few-shot learning methodology.

ACKNOWLEDGEMENTS

This work was supported by research grants from the Daegu Catholic University in 2025.

REFERENCES

- [1] S. He, G. A. Wang, M. Zhang, and D. F. Cook, "Multivariate process monitoring and fault identification using multiple decision tree classifiers," *International Journal of Production Research*, vol. 51, no. 11, pp. 3355–3371, 2013.
- [2] Q. P. He and J. Wang, "Fault detection using the k-nearest neighbor rule for semiconductor manufacturing processes," *IEEE Transactions on Semiconductor Manufacturing*, vol. 20, no. 4, pp. 345–354, Nov. 2007.
- [3] G. Verdier and A. Ferreira, "Adaptive mahalanobis distance and k-nearest neighbor rule for fault detection in semiconductor manufacturing," *IEEE Transactions on Semiconductor Manufacturing*, vol. 24, no. 1, pp. 59–68, Feb. 2011.
- [4] R. Baly and H. Hajj, "Wafer classification using support vector machines," *IEEE Transactions on Semiconductor Manufacturing*, vol. 25, no. 3, pp. 373–383, Aug. 2012.
- [5] J. Kwak, T. Lee, and C. O. Kim, "An incremental clustering-based fault detection algorithm for class-imbalanced process data," *IEEE Transactions on Semiconductor Manufacturing*, vol. 28, no. 3, pp. 318–328, Aug. 2015.
- [6] H. Yang, Y. Lin, B. Yu, and E. F. Y. Young, "Lithography hotspot detection: From shallow to deep learning," 30th IEEE International System-on-Chip Conference, pp. 233–238, 2017.
- [7] K. Kyeong and H. Kim, "Classification of mixed-type defect patterns in wafer bin maps using convolutional neural networks," *IEEE Transactions on Semiconductor Manufacturing*, vol. 31, no. 3, pp. 395–402, 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 Transactions on Semiconductor Manufacturing*, vol. 32, pp. 163–170, 5 2019.
- [9] Y. Nam, M.-C. Kang, K. Kim, J. Hong, S. Lee, and D.-N. Kim, "Adversarial defect detection in semiconductor manufacturing process," *IEEE Transactions on Semiconductor Manufacturing*, vol. 34, no. 3, pp. 365–371, 2021.
- [10] B. Dey, E. Dehaerne, K. Khalil, S. Halder, P. Leray, and M. A. Bayoumi, "Deep learning based defect classification and detection in SEM images: A Mask R-CNN approach," *arXiv preprint arXiv:2211.02185*, 2022.
- [11] S. Choi, Q. Xie, N. Greeneltch, H. J. Lee, M. Govindaraj, S. Jayaram, M. Pereira, S. Biswas, S. Bhamidipati, and I. Torunoglu, "Fast and accurate automatic wafer defect detection and classification using machine learning-based SEM image analysis," *Proceedings of the SPIE*, vol. 12955, Metrology, Inspection, and Process Control XXXVIII, 129553M, Apr. 2024.