International Journal of Precision Engineering and Manufacturing
https://doi.org/10.1007/s12541-026-01493-w

Online ISSN 2005-4602
Print ISSN 2234-7593

REGULAR PAPER

Advanced Wafer Hotspot Detection through Image Segmentation and Stacked Model

Ali Raza[1,2] · Julio Cesar Martinez Espinosa[3,4,5,6] · Debora Libertad Ramirez Vargas[3,4,7,8] · Muhammad Shafiq[9] · Imran Ashraf[9]

Received: 8 October 2025 / Revised: 19 February 2026 / Accepted: 27 February 2026

Abstract

The wafer map is a data visualization of a thin semiconductor fabric made of crystalline silicon, such as defects or test results. The wafer map is a base for creating electronic coordinate circuits and photovoltaic cells. During the wafer map production, any fault results in a product failure. The wafer map faults are undetectable to the naked eye, which is a big challenge. Hotspot detection in wafer maps is significantly important to evaluate the manufacturing process and. improve product yield. The hotspot detection in the wafer maps is the primary aim of this research. A novel wafer map hotspot detector (WHD) is proposed based on three stack fully connected conventional neural network layers and a dense layer. Data augmentation uses the segmented images of the wafers to build the proposed model. The proposed model is evaluated through several evalua-tion parameters and state-of-the-art studies comparative analysis. The proposed model achieved a 94% training and 90% testing performance accuracy for hotspot detection and shows better results than existing approaches. This study helps semiconductor engineers improve wafer manufacturing designs and efficiency in the semiconductor industry.

Keywords Precision manufacturing · Hotspot detection · Machine learning · Stacked model · Image processing

1 Introduction

A semiconductor is a material whose electrical conductivity lies between that of con-ductors and insulators, making it suitable for controlled conduction of electric current. Silicon, an elemental semiconductor, is one of the most widely used materials in semiconductor manufacturing. Semiconductors enable precise control over electrical conductivity, which is essential in the design and function of modern electronic devices [1]. A wafer map is created using

✉ Muhammad Shafiq
shafiq@ynu.ac.kr

✉ Imran Ashraf
imranashraf@yu.ac.kr

Ali Raza
ali.raza.scholarly@gmail.com

Julio Cesar Martinez Espinosa
julio.martinez@unini.edu.mx

Debora Libertad Ramirez Vargas
debora.ramirez@unini.edu.mx

1 Department of Software Engineering, University of Lahore, Lahore 54600, Pakistan

2 Department of Precision Medicine, Sungkyunkwan University School of Medicine, Suwon 16419, Republic of South Korea

3 Universidad Europea del Atlantico. Isabel Torres 21, Santander 39011, Spain

4 Universidad Internacional Iberoamericana, Campeche 24560, Mexico

5 Universidad Internacional Iberoamericana, Arecibo, Puerto Rico 00613, USA

6 Universidade Internacional do Cuanza, Cuito, Bie, Angola

7 Fundacion Universitaria Internacional de Colombia, Bogota, Colombia

8 Universidad de La Romana, La Romana, Dominican Republic

9 School of Computer Science and Engineering, Yeungnam University, Gyeongsan-si 38541, Republic of Korea

silicon semiconductors and serves as a data visual-ization tool representing the layout and characteristics of thin semiconductor wafers made from crystalline silicon. The wafer map plays a crucial role in identifying defec-tive integrated circuits and failures in chips during the manufacturing process [2, 3]. A single wafer can contain hundreds of individual chips. With advances in semiconduc-tor manufacturing, these chips are becoming increasingly dense and complex, which raises the likelihood of surface defects appearing on the wafer map.

Wafer hotspot detection is the process of identifying abnormal temperature regions on the surface of a wafer, which can cause problems during the manufacturing of microelectronic devices [4]. Using segmented images based on a stacked model is one approach to detecting hotspots. This approach involves segmenting the images into different regions and using a machine learning model, such as a stacked model, to identify regions with elevated temperatures [5]. The stacked model could be composed of multiple individual models, such as Convolutional Neural Networks and Random Forests, that are combined to make a prediction, allowing for the consideration of multiple features and patterns in the data.

The traditional wafer defect detection techniques during manufacturing are time-consuming, outdated, and require experienced industry engineers to analyze [6]. The semiconductor industry has automated methods for analyzing and observing wafer map defects during manufacturing using electron microscopes. However, these methods exhibit low detection performance [7]. The traditional manual ways can lead to more errors during the inspection of defects in wafers. An advanced deep learning(DL)-based model is needed for efficient and early detection of hotspots during wafer map manufacturing.

Wafer hotspot detection using segmented images involves using computer vision techniques to identify and locate areas on a wafer that are unusually hot, which can indicate a malfunction or defect [3, 8, 9]. One way to do this is to first segment the wafer images into different regions and then apply algorithms to analyze the temperature data in each region. This can be done using machine learning techniques, such as training a convolutional neural network to classify regions as either normal or hotspot [10, 11]. The analysis results can be used to identify and address potential issues with the wafer before they cause problems in the manufacturing process. Segmented images divide an image into multiple regions or segments [12, 13]. This division can be based on various criteria, such as color, texture, or object boundaries. Segmenting an image simplifies the analysis process and allows for the consideration of individual segments, which can have distinct features or patterns [14, 15].

In the context of wafer hotspot detection, using segmented images based on a stacked model refers to the division of the wafer images into segments and the use of a stacked model to identify the segments that contain hotspots. This approach leverages the strengths of image segmentation and stacked models to improve the accuracy and efficiency of the hotspot detection process. A stacked model is a machine learning (ML) technique that combines the predictions of multiple individual models to make a final prediction [16]. In this study, each model is trained on a different subset of the data, and the predictions of these models are combined to make the final prediction. This allows for integrating multiple features and patterns in the data and can improve accuracy compared to using a single model [17].

Artificial neural networks (ANNs) can be applied to detect the wafer map hotspot defects efficiently [18]. The wafer map defect patterns can be learned and recognized accurately using artificial neural network methods. The convolutional neural network (CNN)-based image processing technique is applied in this study for wafer map hotspot detection. The CNN is specifically utilized for the problem of processing pixel data and image recognition [19]. The digital wafer map images are the binary representation of visual data. The CNN is the best choice for learning patterns from visual data and achieving high performance in image recognition.

Hotspots are a major problem in the semiconductor industry, leading to yield loss and reduced device performance. Accurately detecting hotspots is critical for improving yield and reducing costs. Existing methods have limitations, including low accuracy and difficulty detecting hotspots in localized regions. This study proposes a novel approach to detect hotspots on semiconductor wafers using image segmentation and a stacked model. The proposed approach involves segmenting the wafer image into different regions using image segmentation techniques. Later, a stacked model is used to identify hotspots in segmented regions.

1.1 Research Innovation

The proposed approach addresses the limitations of low accuracy by using image segmentation to identify hotspots in localized regions, a three-stack fully connected conventional neural network layers, and a family of dense layers to improve detection accuracy. Results demonstrate the effectiveness of the approach on a real-world dataset of wafer images, achieving a detection accuracy of 94.3% and a recall rate of 94.8%.

1.2 Contributions

The primary contributions of this research study are as follows.

- A novel wafer hotsopt detection approach is designed in this study based on the stack of three fully-connected conventional neural network layers and a family of dense layers. The proposed approach is utilized to detect the hotspot in the wafer map based on segmented images.
- The data augmentation technique is applied to enhance the image dataset for better model training. The data augmentation technique reduces the proposed model's overfitting problem and enhances the model's performance scores.
- The proposed model hyperparameter tuning and optimization are applied to achieve high-performance scores in detecting wafer map hotspots.

The remaining research study is assembled as follows: Section 2 contains the wafer map-related work analysis. Section 3 is based on the study methods for wafer map hotspot detection. Section 4 provides the results, validations, and detailed discussion. The study is summarized in Section 5.

2 Related Work

The related literature on wafer map hotspot detection is analyzed in this section. The applied techniques and performance evaluation metrics in the existing studies are described.

The wafer map failure classification using DL was proposed in [20]. The one-channel-based wafer map images were used to conduct experiments. Eight types of wafer map failure were used for training and classification. The VGG16 and AlexNet models were applied to detect faults. The applied model was hyperparameter-tuned. The data augmentation techniques on wafer map images were applied to achieve good performance scores. A ML-based approach was proposed in [21] for wafer map defect classification. The CNN, support vector machine (SVM), AdaBoost, and XGBoost were the applied models for classification. The hyperparameter tuning techniques were applied to determine the best-fit parameters of each model. The study experiments show that the CNN achieved good performance compared to other models.

The variational autoencoder-based enhanced DL approach was proposed for wafer defect classification in [22]. The imbalanced wafer maps images dataset was used for conducting the experiments. Experiments show that the autoencoder is a lightweight model with better results. The data augmentation techniques were applied to enhance the proposed model. The authors [23] proposed a CNN for wafer map fault detection. The mixed-type and single-type wafer patterns were used for training the proposed model. The results show that the proposed approach can detect mixed and single-type patterns accurately.

For lithographic hotspot mining, an ML-based technique was presented in [24]. Innovatively, contour information initially assists hotspot identification, making it easier to detect topological aberration brought on by complex process effects. More significantly, based on novel techniques, the suggested approach offers great efficiency and is suitable for large-scale production. CNN achieved more accurate detection in [25]. From self collected dataset, a wafer image is divided into hundreds of images of silicon grains using digital image preprocessing. The comparison tests are carried out on a dataset with more than ten thousand grain pictures using three models. The find-ings demonstrate that the WDD-detection Net's speed is 105.6FPS, five times quicker than VGG-16 and MobileNet-v2, and it shows better performance.

A fast post-processing method for the overlapping issue of wafer map detection was proposed in [26]. The fast-soft non-maximum suppression model is proposed for this purpose. The experiments were conducted using single-stage and two-stage detectors. The logistic function was used during the detection. The penalty distribution function optimization updated the detection box score. Superior results are reported in this study.

Traditional image segmentation methods, such as thresholding and edge detection, may not be effective for detecting hotspots due to the complexity of the wafer images. On the other hand, DL methods have shown promise in improving hotspot detection accuracy. The main problem in wafer hotspot detection is accurately and efficiently identifying the hotspot regions on the wafer surface. There are several challenges in this process.

- **Complex surface patterns**: Wafer surfaces can have complex patterns and textures, making distinguishing between normal and abnormal regions difficult.
- **Data variability**: The temperature data collected from wafers can vary greatly due to factors such as the type of wafer, the manufacturing process, and the measurement equipment used for collection.
- **Falsepositives**: It is important to minimize the number of false positives (i.e., regions identified as hotspots but are actually normal) in order to reduce the cost and time of re-inspecting and re-testing the wafers.

Such challenges necessitate a more accurate and efficient approach. The gap in the current state-of-the-art in wafer hotspot detection using segmented images is to develop a more accurate and efficient method for detecting hotspots that can effectively handle the challenges mentioned above. A stacked model-based approach may address these challenges by leveraging the strengths of multiple models and allowing for the integration of multiple features and patterns in the data.

3 Methods

This study proposes an approach for detecting hotspots on the surface of wafers. This method involves dividing the wafer images into multiple segments and using a stacked model to identify the hotspot segments. The use of image segmentation allows for the consideration of individual segments, which can have distinct features or patterns, and simplifies the analysis process. On the other hand, the stacked model combines the predictions of multiple individual models to make a final prediction, allowing for the integration of multiple features and patterns in the data.

The wafer map hotspot detection approach is illustrated in Fig. 1. The data augmentation technique is applied to enhance the wafer map data. The data augmentation results in high performance and helps to overcome overfitting problems. A novel WHD is proposed based on three stack fully connected CNN layers. 90% of the dataset is utilized for training, while 10% of the data is utilized to validate the performance of the proposed model.

The proposed approach comprises the following steps.

- **Step1:**The data exploration is carried out first. The wafer map segmented images are analyzed and imported. The sample plotting of the wafer maps by each defect is analyzed.
- **Step2:**Data augmentation is carried out. During the data augmentation, the newly generated wafer map images are created by rotation, shift by width, shift by height, rescaling, shear, zooming, and horizontal flipping.
- **Step3:**The proposed model is fine-tuned concerning various parameter config- urations to obtain the best performance. The model-building stack contains the convolutional layers, flatten layers, max-pooling layers, dropout layers, and dense layers.
- **Step4:**The trained model is tested with the test data. The results are compared with existing approaches for performance validation.

3.1 Wafer Map Segmented Images Data

For experiments, the wafer image dataset is used and is publicly available on Kaggle [27]. The dataset is based on the balanced wafer map segmented images. The nine categories are used to classify the wafer map images. The dataset categories include near-full, edge-ring, center, edge-loc, none, donut, scratch, random, and loc. The 160 MB is the total size of the dataset. The dataset is carefully analyzed for wafer map hotspot detection. The sample wafer maps by each category are visualized in Fig. 2. The experimental dataset used in this study contains a total of 3,681 wafer map samples. The data are distributed across nine defect categories plus one "normal" class, with each class containing approximately 409 samples, ensuring a balanced dis-tribution among categories. The wafer map images have sufficient resolution for defect recognition, and the dataset does not contain noise or artifacts.

3.2 Wafer Map Image Augmentation

DL models required a significant amount of data for training. The smaller datasets lead the DL models towards overfitting problems and low performance. Data augmentation is applied to increase the number of samples in the dataset. Data augmentation results in high-performance accuracy scores and helps to overcome overfitting problems. The parameter used for data augmentation is analyzed in Table 1.

3.3 Proposed Approach

A novel WHD is proposed based on a stack of fully connected CNN layers. The archi-tectural analysis of the proposed model is presented in Figure 3. The wafer map segmented images are input to the proposed model stack with shape formate (64,65,4). The three-stack conventional

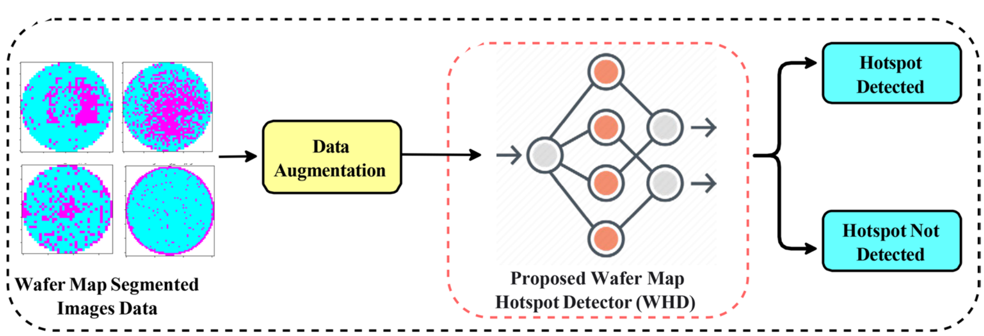


Fig. 1 The study methods and working flow analysis for wafer map hotspot detection

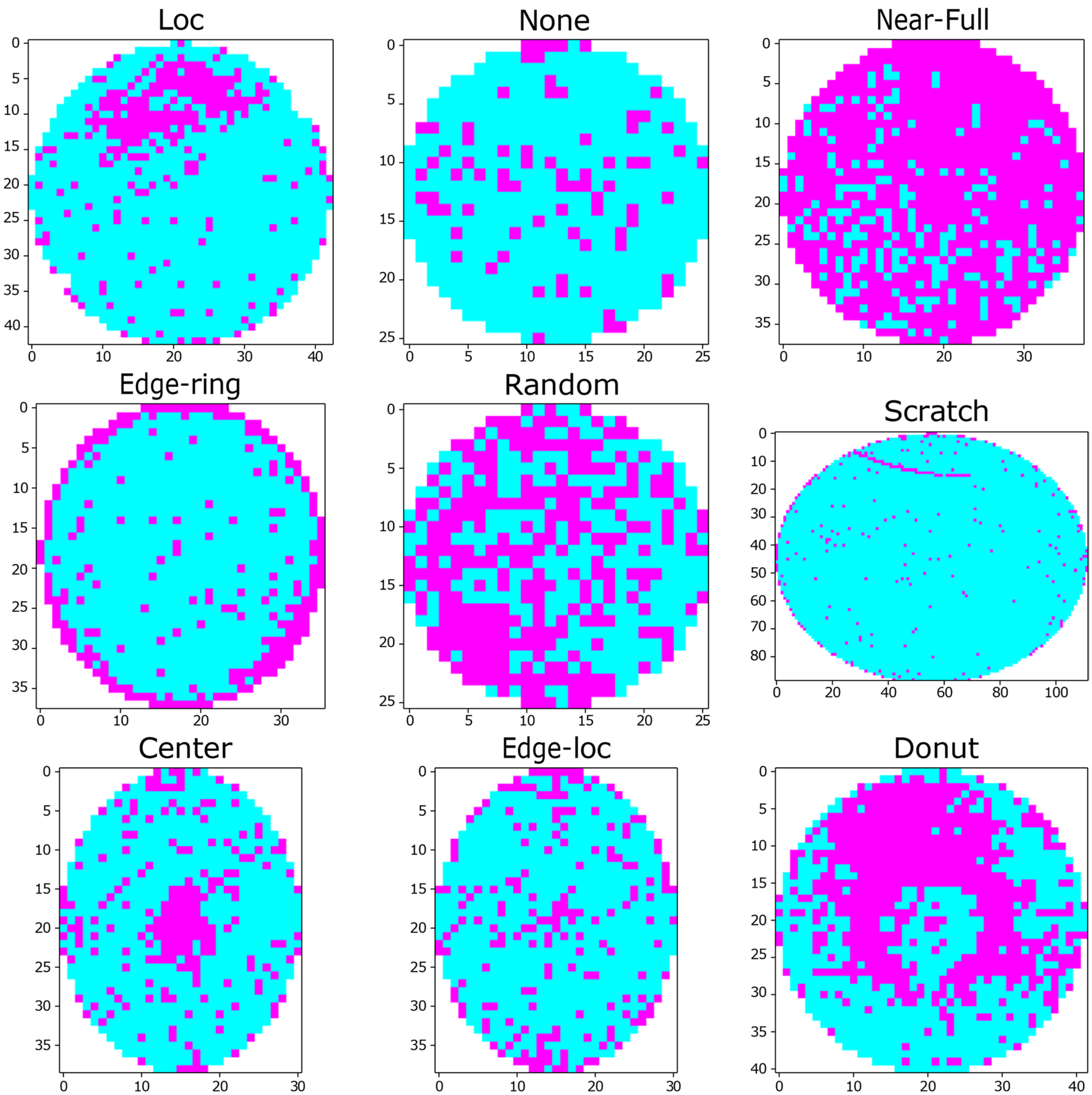


Fig. 2 The sample wafer map segmented images analysis by each class

Table 1 The parameters used for data augmentation

Parameter	Value
Rotation range	20
Width shift range	0.10
Height shift range	0.10
Rescale	1/255
Shear range	0.1
Zoom range	0.1
Horizontal flip	True
Fill mode	Nearest

neural network layers are involved as automatic feature extractors from the wafer map images. The stack includes 2D conventional neural network layers and 2D pooling layers. The extracted high-involvement features are flattened and input into the dense layers family. The fully connected dense layers are used as a classifier. The SoftMax activation function is used to classify wafer images in a multi-label output.

The choice of three convolutional layers was guided by both empirical evaluation and architectural considerations. During

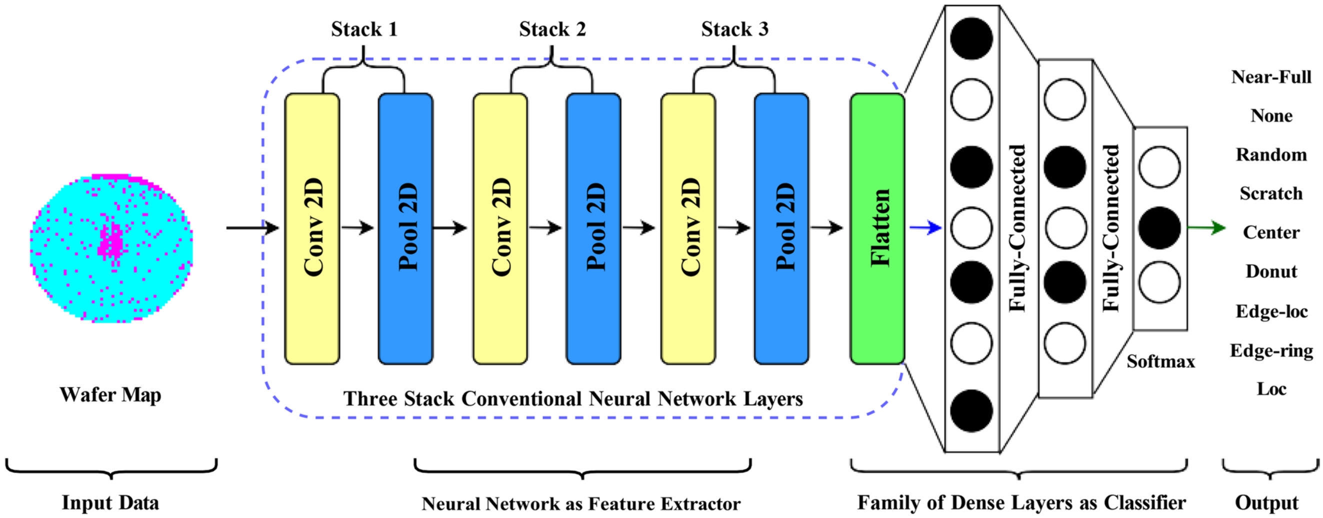


Fig. 3 The novel proposed approach's layered architecture analysis is examined

preliminary experiments, we tested different depths (two, three, and four convolutional layers). A three-layer stack provided the best balance between model complexity and performance.

The proposed model takes segmented wafer map images as input. Each wafer map is represented as a multi-channel tensor with a shape of (64, 65, 4), where 64×65 corresponds to the spatial resolution of the segmented wafer map and 4 denotes the number of input channels generated during preprocessing.

- **Featurerepresentation**: Three layers were sufficient to progressively capture low-, mid-, and high-level spatial features from wafer map images without overfitting.
- **Performanceenhancement**: Deeper stacks beyond three layers did not yield a noticeable improvement in classification accuracy.
- **Overfitting prevention**: Shallower networks (one or two layers) lack the represen-tational power to extract complex defect patterns, whereas deeper networks tend to overfit given the dataset characteristics.

Therefore, the three-layer stack was selected as it consis-tently enhanced per-formance, improved feature extrac-tion, and maintained generalization ability with manageable computational requirements.

3.3.1 Proposed Model's Hyperparameters and Configuration Analysis

The hyperparameter tuning of the proposed model is given in Table 2. The best-fit hyperparameters are determined by the recursive training and testing of the model. The analysis is based on the deep learning model parameters on which the proposed model achieved a high-performance accuracy score.

Table 2 The proposed model hyper-parameters analysis

Name	Value
Loss function	Categorical cross entropy
optimizer	ADAM
metrics	Accuracy
activation	Softmax

Table 3 analyzes the proposed model architectural details. The configuration is based on the layers used to build the model and the relevant parameters. The architec-ture base layers involve the 2D convolutional neural network and max pooling layers. Then, dense layers are involved in per-forming multi-label classification.

In the proposed approach, hyperparameter tuning and optimization are carried out using a systematic search-based optimization strategy (grid search) is applied during the model training phase. The optimization process is car-ried out using the Adam optimizer during network train-ing. Adam is employed to update the model parameters by minimizing the training loss through adaptive learning rate estimation based on first- and second-order moments of the gradients.

During optimization, different combinations of hyperpa-rameters, for example, learning rate, number of layers/neu-rons, batch size, etc., are investigated, and the.

configuration yielding the best validation performance is selected as optimal. Hyperpa-rameter tuning focuses on the network architecture parameters, including the number of layers and the incorporation of max pooling layers, which are adjusted experimen-tally to improve feature extraction and reduce overfitting. The objective function of the optimi-zation process is the loss function (e.g., cross-entropy loss), which is minimized during training to enhance the model's predictive performance.

Table 3 Parameters configuration of the proposed model for wafer hotspot detection

Layers	Unit	Kernel	Activation	Output shape	Parameters
The 2D Convolutional layer	32	4 × 4	RELU	(None, 61, 62, 32)	2080
The 2D Max Pooling layer	0	2 × 2	N/A	(None, 30, 31, 32)	0
The 2D Convolutional layer	64	4 × 4	RELU	(None, 27, 28, 64)	32,832
The 2D Max Pooling layer	0	2 × 2	N/A	(None, 13, 14, 64)	0
The 2D Convolutional layer	128	4 × 4	RELU	(None, 10, 11, 128)	131,200
The 2D Max Pooling layer	0	2 × 2	N/A	(None, 5, 5, 128)	0
The Flatten layer	0	0	N/A	(None, 3200)	0
The Dense layer	64	0	RELU	(None, 64)	204,864
The Dense layer	32	0	N/A	(None, 32)	2080
The Output Dense layer	9	0	SOFTMAX	(None, 9)	297

The proposed approach involves image segmentation and stacked model analysis to detect hotspots on semiconductor wafers. The process begins with segmenting the wafer image into different regions using image segmentation techniques. This allows for the identification of hotspots in specific regions of the wafer. The segmented regions are then analyzed using a stacked model analysis, combining multiple models to improve hotspot detection accuracy.

The proposed approach is evaluated on a real-world dataset of wafer images, and the results showed a significant improvement in hotspot detection accuracy compared to existing methods. Overall, the proposed method provides a promising approach to hotspot detection on semiconductor wafers and has the potential to significantly improve yield and reduce costs in the semiconductor industry.

3.4 Train, Test, and Validation Composition

For model building we have used, 80% of the data were used for training, of which 10% were reserved for validation during the training process, and the remaining 10% of the data were used for testing.

4 Results

Experimental results and performance validations are analyzed in this section. The results and discussions of the proposed model are precisely described. The results val-idation includes performance metrics, confusion matrix analysis, time series analysis, and state-of-the-art model comparison.

4.1 Experimental Setup

The Python programming language is utilized for conducting all research experiments. The modules Tensorflow with version 2.8.2 and Keras with version 2.8.0 are utilized for building the proposed model on Google Colab based Jupyter Notebook environments hosted on Google's cloud servers. The module Sklearn with version 1.0.2 is utilized for performance metric results evaluations. The recall, precision, and F1 scores are the performance metrics.

Recall and precision metrics are used to assess the effectiveness of the proposed model in detecting hotspots. The precision is the proportion of wafer map hotspots the model found adequately as having a defect out of all wafers. The recall metric is the percentage of total relevant results correctly identified by the proposed model. The F1 score is another model accuracy performance measure. The F1 score value is the harmonic mean of the proposed model's recall and precision scores.

4.2 Proposed Model Performance Evaluation

The epoch-wise comparative analysis of the proposed stacked model during the train-ing is analyzed in Table 4. The comparative analysis contains the results of time computations (seconds), loss score, and performance accuracy score. The performance analysis demonstrates that the loss score is very high at the first epoch, and time computations are also high. The accuracy score at the first epoch is also low. With the increase in the epoch, the loss is decreased, and the accuracy performance score is increased. During the complete neural network model building, the minimum accu-racy score achieved is 0.44. The highest accuracy performance score achieved is 0.94 for detecting wafer map hotspots.

The class-wise performance analysis of the proposed model is presented in Table 5; Fig. 4. The analysis demonstrates that the proposed model achieved a high precision score of 100% for the hotspot class 'edge-ring'. The 'edge-loc' and 'none' classes achieved a low precision score in comparison. The hotpot 'donut' and 'near-full' achieved a 100% recall score compared to other classes. The 'edge-loc' class achieved a lower recall score. The 'donut' and 'near-full' class achieve a 99% F1 score, which is superb. The

Table 4 The epoch-wise performance metrics analysis of the proposed stacked model during the training

Epoch	Time (Sec.)	Loss Score	Accuracy	Epoch	Time (Sec.)	Loss Score	Accuracy
1	137s	1.4958	0.4414	26	46s	0.2701	0.9043
2	48s	0.9239	0.6688	27	48s	0.2719	0.9100
3	47s	0.6616	0.7672	28	46s	0.2549	0.9139
4	48s	0.6052	0.7802	29	48s	0.2577	0.9085
5	46s	0.5596	0.7998	30	46s	0.2632	0.9121
6	48s	0.5259	0.8068	31	45s	0.2534	0.9146
7	47s	0.5162	0.8149	32	47s	0.2529	0.9133
8	49s	0.4531	0.8309	33	48s	0.2258	0.9251
9	46s	0.4437	0.8481	34	45s	0.2388	0.9136
10	48s	0.4190	0.8524	35	48s	0.2381	0.9185
11	46s	0.4204	0.8527	36	45s	0.2196	0.9278
12	48s	0.3716	0.8696	37	52s	0.2165	0.9212
13	48s	0.3830	0.8626	38	48s	0.2126	0.9233
14	46s	0.3844	0.8678	39	47s	0.2239	0.9209
15	48s	0.3499	0.8795	40	46s	0.2220	0.9251
16	47s	0.3444	0.8807	41	47s	0.2129	0.9221
17	46s	0.3183	0.8880	42	48s	0.2239	0.9233
18	48s	0.3397	0.8825	43	46s	0.2137	0.9248
19	46s	0.3044	0.8934	44	48s	0.2116	0.9272
20	48s	0.3188	0.8856	45	48s	0.2116	0.9257
21	48s	0.3171	0.8916	46	46s	0.2270	0.9245
22	48s	0.2902	0.9007	47	48s	0.1847	0.9378
23	48s	0.2811	0.9004	48	46s	0.1899	0.9336
24	46s	0.2791	0.9067	49	48s	0.2049	0.9239
25	46s	0.2896	0.8995	50	46s	0.1777	0.9429

Table 5 The performance metric evaluation results of the proposed model in detecting wafer map hotspots

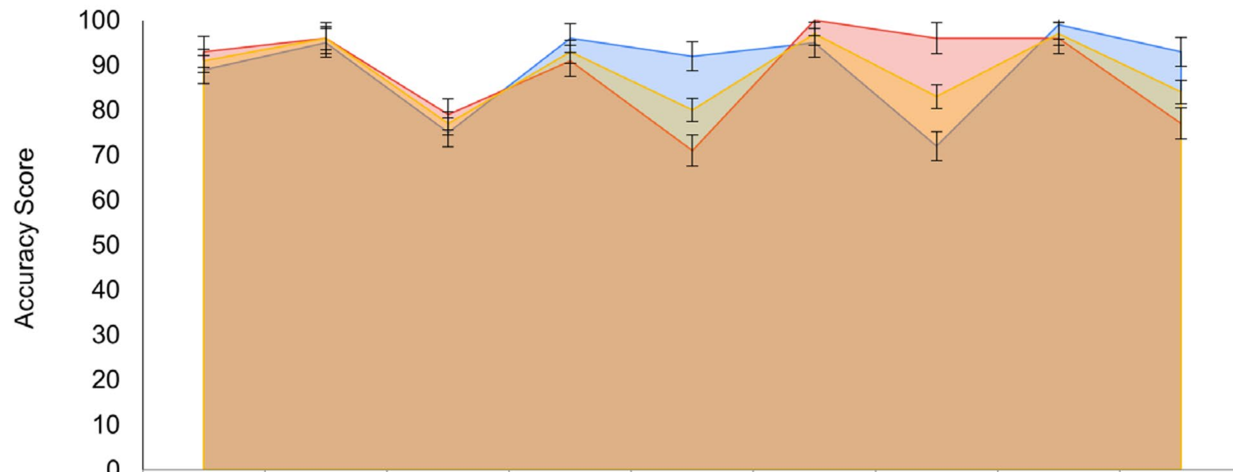
Category	Precision Score (%)	Recall Score (%)	F1 Score (%)	Support Score
Center	91	98	94	41
Donut	98	100	99	41
Edge-loc	76	61	68	41
Edge-ring	100	88	94	41
Loc	89	78	83	41
Near-Full	95	100	98	41
None	77	98	86	41
Random	95	95	95	41
Scratch	86	88	87	41
Average	**90**	**89**	**89**	**369**
Accuracy	**90**			

analysis shows that the average accuracy for all metrics is 90%. This comparative analysis indicates that the proposed approach achieves high performance for wafer map hotspot detection.

The relatively lower precision values observed for the Edge-Loc (76%) and None (77%) categories can be attributed to several dataset- and model-related factors.

- First, both Edge-Loc and None exhibit high visual similarity to other classes, which increases the likelihood of false positives. In particular, Edge-Loc shares spatial characteristics with Edge-Ring and partially with Center defects, as hotspots may appear sparsely near the wafer boundary without forming a complete ring. This overlap in spatial distribution makes it challenging for the segmentation model to draw a strict boundary between these classes, leading to occasional misclassification and reduced precision.
- Second, the None class represents wafers without distinct defect patterns, yet it often contains random noise or minor pixel-level variations introduced during manu-facturing or preprocessing. These subtle variations can resemble weak or early-stage defect patterns (e.g., Random or Loc), causing the model to incorrectly label some defective wafers as None or vice versa. As a result, the number of false positives increases, directly impacting precision.
- Finally, the segmentation-based approach focuses on pixel-level learning. For classes where defects are sparse, weakly localized, or irregular, minor segmentation errors can propagate into the classification stage of the stacked model, further reducing precision.

Figure 4 shows the classification performance of the proposed approach across mul-tiple conditions using accuracy, F1 score, etc. Overall, 'near-full' and'center' classes



	Center	Donut	Edge-loc	Edge-ring	Loc	Near-Full	None	Random	Scratch
Precision	89	95	75	96	92	95	72	99	93
Recall	93	96	79	91	71	100	96	96	77
F1	91	96	77	93	80	97	83	97	84

Fig. 4 The area chart-based performance metric evaluation results of the proposed model

achieve the highest and most consistent results across all metrics, indicating robust model performance. The‘random’ and‘scratch’ classes show moderate results, while‘edge-left’ and ‘edge-ring’ perform relatively weaker, suggesting positional sensitivity. The close alignment of accuracy and F1 score indicates a balanced precision-recall trade-off, demonstrating stable and well-generalized model behavior across most configurations.

Figure 5 represents the model’s training and validation patterns. The analysis demonstrates that during the model’s first five epochs, the loss score is high, and the accuracy score is low. After epoch five, the loss is decreased, and the accuracy is increased. The analysis shows that the proposed model performed excellently in the wafer map hotspot detection.

The plot in Fig. 6 shows the accuracy (blue line) and loss (red line) of the proposed model over 50 epochs. Overall validation accuracy increases while the loss score decreases, indicating improved model performance and reduced overfitting.

4.3 Confusion Matrix Analysis

The confusion matrix analysis is based on predicted class and actual class and is shown in Fig. 7. The confusion matrix is utilized to determine the effectiveness of the proposed model. The confusion matrix analysis compares the proposed model’s prediction with actual values. The analysis demonstrates that the proposed model achieved high-performance scores for each class. The proposed model achieved a lower prediction error rate in the confusion matrix analysis.

4.4 In-depth Error Analysis

Table 6 reveals that most misclassifications occur between ‘edge-loc’, ‘loc’, and ‘none’ classeswhich suggests overlapping patterns and subtle distinctions in their features. For instance, ‘edge-loc’ often gets confused with ‘loc’ (4 cases) and ‘none’ (8 cases), indi-cating the model struggles with differentiating edge-adjacent hotspots and blank areas. Similarly, ‘loc’ and ‘edge-ring’ exhibit cross-misclassifications, likely due to similarities in hotspot distribution.

Furthermore, ‘scratch’ was misclassified in multiple categories, possibly because scratches may share geometric characteristics with other patterns like ‘center’ or ‘donut’. Addressing these issues may involve augmenting the dataset with more diverse examples or refining feature extraction techniques to capture distinct patterns better. Figure 8 visualizes the learned features from the model. The figure illustrates three convolutional layers (Layer 1–3) visualizing feature activations in the proposed model. Layer 1 captures broad, low-level patterns, highlighting general intensity contrasts and edges. Layer 2 refines these into

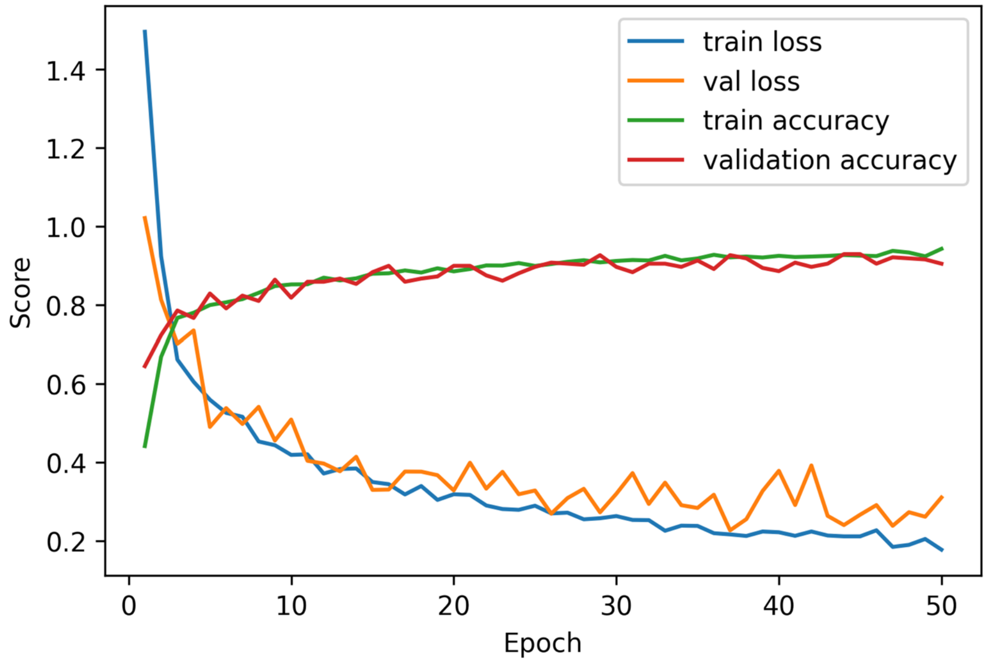


Fig. 5 Training and validation curves for detecting wafer map hotspots

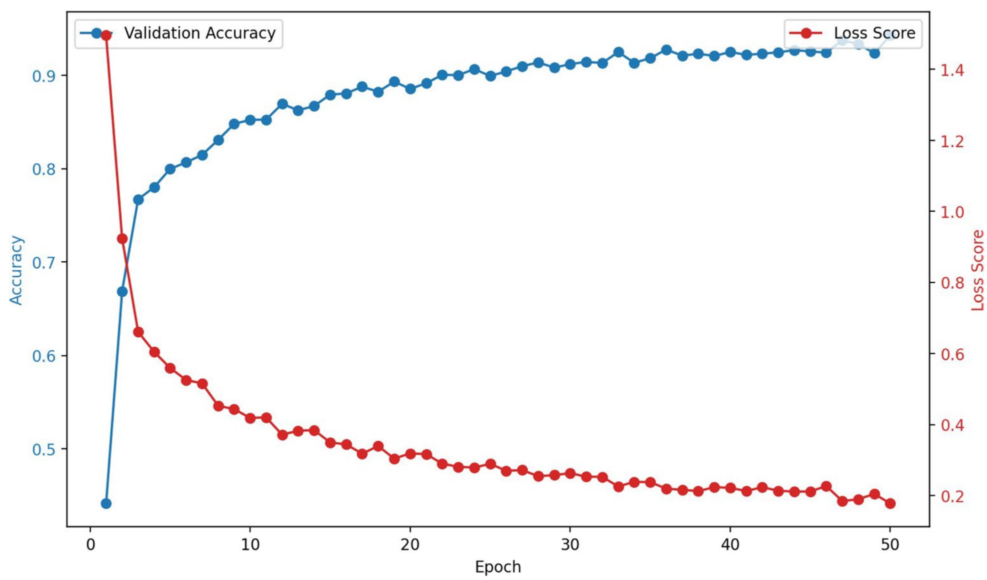


Fig. 6 The model accuracy and loss score validations across 50 epochs

more structured, localized features, emphasiz-ing regions of higher activity. Lastly, Layer 3 presents more abstract representations, showing enhanced separation of key spatial zones while suppressing background noise. Collectively, these layers demonstrate the progressive extraction of hierarchical spa-tial features as the model deepens, leading to improved pattern recognition and discrimination.

4.5 State-of-the-Art Studies Comparison

The comparative performance analysis of the proposed approach with other state-of-the-art studies is shown in Table 7. The recent studies based on image classification are considered for this comparison. The state-of-the-art studies analysis demonstrates that the proposed approach achieved high performance compared to existing studies. The proposed approach shows better accuracy in comparison to these studies.

The proposed approach has the potential to improve the accuracy and efficiency of wafer hotspot detection by addressing the challenges of complex surface patterns, data variability, and false positives. This approach can also be integrated into the wafer manufacturing process to allow for real-time hotspot detection and to reduce the cost and time associated with re-inspecting and re-testing wafers.

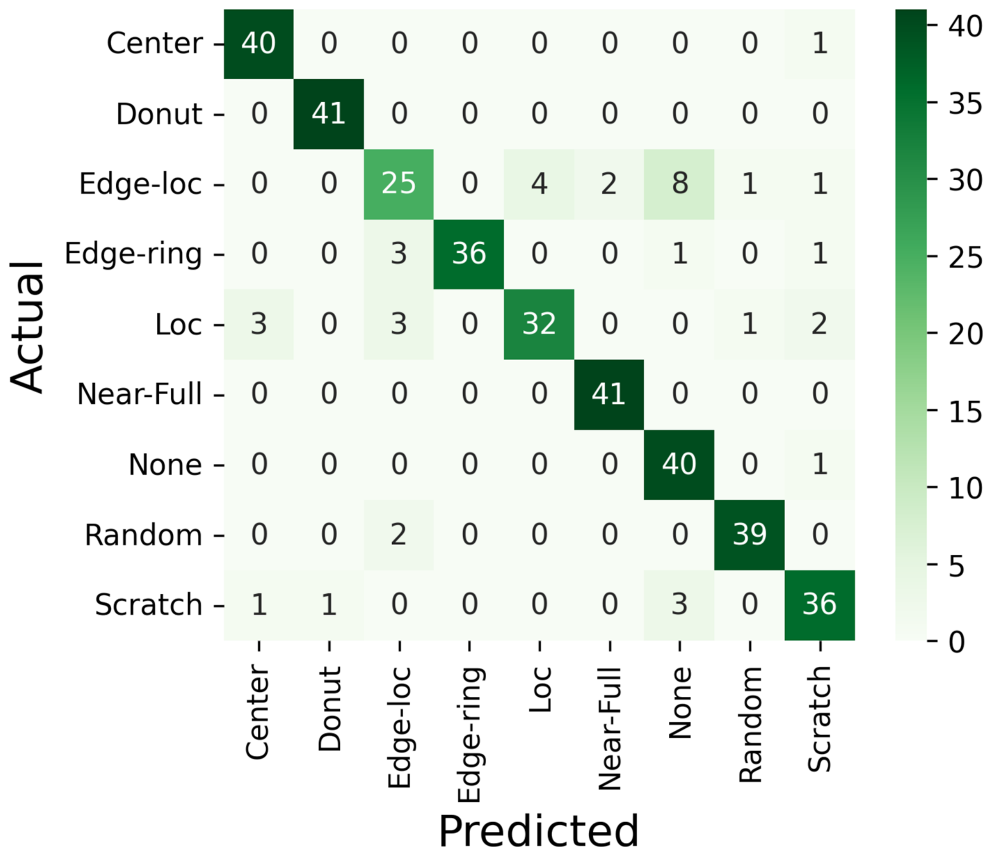


Fig. 7 The confusion matrix analysis of the proposed model in detecting wafer map hotspots

4.6 Alternative Method Validation

To assess model generalization and validation performance, we employed an alter-native method, MobileNetV2 [33]. The model is trained and evaluated on the study dataset under the same experimental protocol to ensure a fair and consistent compar-ison. The validation results are summarized in Table 8. Overall, the model achieved an accuracy of 64%. Performance varied across defect categories. Notably, the Near-Full and Donut classes demonstrated strong predictive performance, achieving high precision and recall values, whereas classes such as Loc and None exhibited relatively low recall, indicating challenges in correctly identifying those categories. These results highlight both the strengths and limitations of the applied MobileNetV2 model in capturing class-specific characteristics within the dataset.

4.7 Ablation Experiments

To further analyze the impact of data augmentation on model performance, a set of ablation experiments was conducted by comparing accuracy across different epoch ranges with and without data augmentation. As shown in Table 9, the model achieved an average accuracy of 0.68 in the early training phase (epochs 1–5) without aug-mentation. When augmentation was introduced (epochs 6–10), the accuracy improved significantly to 0.82, indicating that augmentation enhanced the model's generalization ability. Subsequent ranges (epochs 11–20) also demonstrated consistent improve-ments, with augmentation leading to slightly higher accuracy than training without it. By epoch 21–25, the model stabilized at an average accuracy of 0.89, further con-firming that augmentation contributed positively to learning efficiency and overall performance.

4.8 Study Limitations

The dataset has certain limitations with respect to representing real industrial sce-narios. In actual semiconductor manufacturing, defect occurrences are typically highly imbalanced, with certain defect types being rare and others much more frequent. Addi-tionally, real wafer map data may contain measurement noise, incomplete patterns, or irregular wafer shapes. Therefore, while the WaferMap dataset is suitable for bench-marking and model comparison, its controlled and balanced nature may limit direct generalization to raw industrial datasets.

4.9 Discussion

This study presents a DL-based approach to detecting hotspots on semiconductor wafers using image segmentation and a stacked model. The methodology for detecting

Table 6 False predictions, errors, and possible causes

Actual	Predicted	Count	Possible Causes
Center	Scratch	1	Similar patterns or features with scratches.
Edge-loc	Loc	4	Confusion due to overlapping hotspot areas.
Edge-loc	Near-Full	2	Marginal differences in density.
Edge-loc	None	8	Subtle patterns difficult to classify.
Edge-loc	Random	1	Variability in random patterns.
Edge-loc	Scratch	1	Shared texture or geometric features.
Edge-ring	Loc	3	Proximity of hotspot locations.
Edge-ring	None	1	Patterns appearing blank-like.
Edge-ring	Scratch	1	Similar edge artifacts with scratches.
Loc	Center	3	Incorrect positional identification.
Loc	Edge-loc	3	Features confused with edge hotspots.
Loc	Near-Full	1	Marginal pattern differences.
Loc	Scratch	2	Irregular hotspot patterns.
None	Scratch	1	Subtle artifacts interpreted as scratches.
Random	Edge-loc	2	Random artifacts overlapping with edges.
Scratch	Center	1	Overlap in detected shapes or textures.
Scratch	Donut	1	Misinterpreted symmetrical features.
Scratch	None	3	Subtle scratch patterns misclassified.

hotspots involves segmenting the wafer image into different regions and then using a stacked model to identify hotspots in those regions. Experimental results show that the proposed approach outperforms existing methods for hotspot detection in terms of accuracy, precision, and recall. The effectiveness of the approach is shown using a real-world dataset of wafer images, achieving a detection accuracy of 94.3% and a recall rate of 94.8%.

Table 7 The comparative performance analysis of the novel proposed approach with other state-of-the-art studies

Ref	Year	Learning Type	Technique	Accuracy (%)
[28]	2022	Deep Learning	U-Net model	80
[29]	2021	Deep Learning	Convolutional and Recurrent Neural Network	85
[30]	2021	Deep Learning	MobileNet and LSTM	85
[31]	2024	Deep Learning	CNN	88
[32]	2025	Deep Learning	TransCNN	89
Proposped	**2025**	**Deep Learning**	**Novel WHD approach**	**90**

Table 8 The performance validation results of the applied MobileNet model on this study dataset

Class	Precision	Recall	F1-score
Center	48%	53%	50%
Donut	78%	100%	88%
Edge-loc	39%	40%	40%
Edge-ring	78%	95%	85%
Loc	57%	10%	17%
Near-Full	100%	100%	100%
None	50%	07%	13%
Random	85%	97%	91%
Scratch	41%	78%	53%
Accuracy	64%		

Table 9 Ablation experiment results across various epochs

1–5	Without augmentation	With augmentation
	0.68	0.74
6–10	0.72	0.79
11–15	0.78	0.83
16–20	0.82	0.86
21–25	0.86	0.89

The results are significantly important because hotspot detection is a major problem in the semiconductor industry, leading to yield loss and reduced device per-formance. Accurately detecting hotspots is critical for improving yield

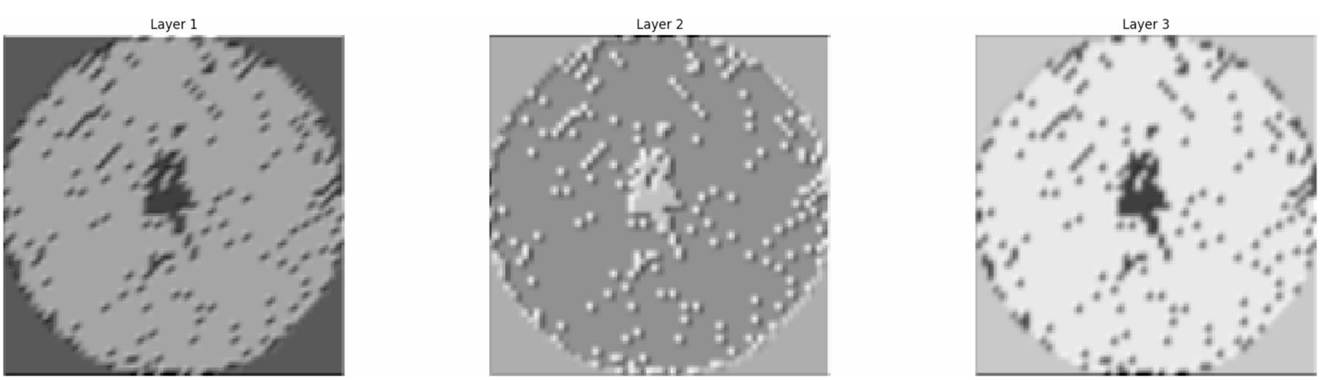


Fig. 8 The CNN layers learned features

and reducing costs, and the proposed approach has the potential to significantly improve hotspot detection accuracy and efficiency.

One of the strengths of the proposed approach is the use of image segmentation, which allows for the identification of hotspots in specific regions of the wafer. This is important because hotspots can occur in localized regions due to factors such as process variations, and identifying these regions can help to pinpoint the root cause of the hotspot and improve process control. Another strength of the approach is the use of a stacked model, which combines multiple models to improve hotspot detection accuracy. The proposed approach outperforms existing methods for hotspot detection, demonstrating the effectiveness of the approach.

One potential limitation of the approach is that it requires a large amount of data to train the models, which may not be readily available for all semiconductor manufacturers. However, the approach can be effective even with a limited amount of training data, and the use of transfer learning techniques could further improve performance.

4.10 Future Work

Keeping in view the challenges of the proposed approach, the following directions can be followed for future research.

- Further improving the accuracy of the stacked model by incorporating additional features such as texture and color information.
- Evaluating the performance of the stacked model on large datasets to demonstrate its scalability.
- Developing a real-time implementation of the ho tspot detection system to allow for the integration of the method into the wafer manufacturing process.
- Exploring the use of transfer learning approaches, which are pre-trained on larger datasets and often show better results even when a smaller number of training samples are available.

5 Conclusions

The wafer map hotspot detection based on a stacked model is proposed in this study. The traditional hotspot detection methods during semiconductor production in the industry are outdated and have high time computation costs. The presented image processing-based model achieved high performance in hotspot detection compared to traditional methods. The wafer map segmented images are used for evaluating the performance of the proposed model. Nine wafer map defect classes are used in the training and evaluation of experiments. The proposed model outperformed other state-of-the-art studies with better accuracy and sensitivity, and specificity. The use of image segmentation and stacked model analysis for wafer hotspot detection has the potential to significantly improve the accuracy and efficiency of hotspot detection. By dividing the images into segments and using a combination of multiple models, this approach can effectively handle the challenges of complex surface patterns, data variability, and false positives. For the future, we intend to enhance the dataset and perform experiments involving the use of transfer learning-based techniques.

Author Contributions AR conceptualization, data curation, writing manuscript. JCME formal analysis, methodology, writing manuscript. ELRV investigation, funding acquisition, visualization.
MS visualization, investigation, software.
AR supervision, validation, writing and editing the manuscript. All authors reviewed the manuscript.

Funding This research was funded by the European University of Atlantic.

Data Availability The dataset used in this study is available from corresponding authors on request.

Declarations

Ethics Approval Not applicable.

Consent to Participate Not applicable.

Consent for Publication Not applicable.

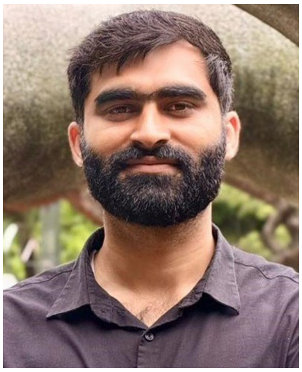
Clinical Trial Number Not applicable.

Conflict of interest The authors declare no conflict of interest.

References

1. Chatterjee, A., Stevenson, P., De Franceschi, S., Morello, A., Leon, N. P., & Kuem-meth, F. (2021). Semiconductor qubits in practice. *Nature Reviews Physics, 3*(3), 157–177.
2. Shim, J., Kang, S., & Cho, S. (2021). Active cluster annotation for wafer map pattern classification in semiconductor manufacturing. *Expert Systems with Applications, 183*, Article 115429.

3. Song, Y., Song, J., & Qu, J. (2022). Optimal image based information hiding with one-dimensional chaotic systems and dynamic programming. *International Arab Journal of Information Technology, 19*(1), 1–10.
4. Lee, T., Jung, S., Kwon, S., Kim, W., Park, J., Lim, H., & Lee, J. (2019). Formation of inter-stitial hot-spots using the reduced gap-size between plasmonic microbeads pattern for surface-enhanced raman scattering analysis. *Sensors (Basel), 19*(5), 1046.
5. Goudelis, G., Lazaridis, P. I., & Dhimish, M. (2022). A review of models for photovoltaic crack and hotspot prediction. *Energies, 15*(12), Article 4303.
6. Systems, I. S. waferMAP. https://www.semiconductoronline.com/doc/ wafermap-0001 Accessed 2022-10-29.
7. Borisov, V., & Scheible, J. (2018). Lithography hotspots detection using deep learning. In: 2018 15th International Conference on Synthesis, Modeling, Analysis and Simula-tion Methods and Applications to Circuit Design (SMACD), pp. 145–148 IEEE.
8. Zhou, K., Zhang, K., Liu, J., Liu, Y., Liu, S., Cao, G., & Zhu, J. (2021). An imbalance aware lithography hotspot detection method based on hdam and pre-trained googlenet. *Measurement Science and Technology*, *32*(12), 125008.
9. Liao, L., Li, S., Che, Y., Shi, W., & Wang, X. (2022). Lithography hotspot detection method based on transfer learning using pre-trained deep convolutional neural network. *Applied Sciences, 12*(4), Article 2192.
10. Xie, Z., Huang, Y. H., Fang, G. Q., Ren, H., Fang, S. Y., Chen, Y., & Hu, J. (2018). Routenet: Routability prediction for mixed-size designs using convolutional neu-ral network. In: 2018 IEEE/ACM International Conference on Computer-Aided Design (ICCAD), pp. 1–8 IEEE.
11. Donegan, S. P., Kumar, N., & Groeber, M. A. (2019). Associating local microstructure with predicted thermally-induced stress hotspots using convolutional neural networks. *Materials Characterization, 158*, Article 109960.
12. Abualigah, L., Diabat, A., Sumari, P., & Gandomi, A. H. (2021). A novel evolutionary arith-metic optimization algorithm for multilevel thresholding segmentation of COVID-19 ct images. *Processes, 9*(7), Article 1155.
13. Abualigah, L., Almotairi, K. H., & Elaziz, M. A. (2022). Multilevel thresholding image seg-mentation using meta-heuristic optimization algorithms: Comparative analysis, open challenges and new trends. *Applied Intelligence*. https://doi.org/10.1007/s10489-022-04064-4
14. Liu, Q., Li, N., Jia, H., Qi, Q., & Abualigah, L. (2022). Modified remora optimization algo-rithm for global optimization and multilevel thresholding image segmentation. *Mathematics, 10*(7), Article 1014.
15. Lin, S., Jia, H., Abualigah, L., & Altalhi, M. (2021). Enhanced slime mould algorithm for multilevel thresholding image segmentation using entropy measures. *Entropy (Basel, Switzerland), 23*(12), Article 1700.
16. Khan, I. U., Javeid, N., Taylor, C. J., Gamage, K. A., & Ma, X. (2021). A stacked machine and deep learning-based approach for analysing electricity theft in smart grids. *IEEE Transactions on Smart Grid, 13*(2), 1633–1644.
17. Rajadurai, H., & Gandhi, U. D. (2020). A stacked ensemble learning model for intrusion detection in wireless network. *Neural Computing and Applications*. https://doi.org/10.1007/s00521-020-04986-5
18. Yoon, S., & Kang, S. (2022). Semi-automatic wafer map pattern classification with con-volutional neural networks. *Computers & Industrial Engineering, 166*, Article 107977.
19. Li, Z., Liu, F., Yang, W., Peng, S., & Zhou, J. (2021). A survey of convolutional neu-ral networks: analysis, applications, and prospects. IEEE transactions on neural networks and learning systems.
20. Gong, J., & Lin, C. (2019). Wafer map failure pattern classification using deep learning. Computer Science.
21. Yuan-Fu, Y. (2019). A deep learning model for identification of defect patterns in semi-conductor wafer map. In: 2019 30th Annual SEMI Advanced Semiconductor Manufacturing Conference (ASMC), pp. 1–6 IEEE.
22. Wang, S., Zhong, Z., Zhao, Y., & Zuo, L. (2021). A variational autoencoder enhanced deep learning model for wafer defect imbalanced classification. *IEEE Transactions on Components, Packaging and Manufacturing Technology, 11*(12), 2055–2060. https://doi.org/10.1109/TCPMT.2021.3126083
23. Byun, Y., & Baek, J. G. (2020). Mixed pattern recognition methodology on wafer maps with pre-trained convolutional neural networks. In: ICAART (2), pp. 974–979.
24. Zhou, T., Xu, B., Li, C., Diao, X., Yan, Y., Chen, S., Zhao, Y., Zhou, K., Zhou, W., Zeng, X. (2021). : Mining lithography hotspots from massive sem images using machine learning model. In: 2021 China Semiconductor Technology International Conference (CSTIC), pp. 1–3 IEEE.
25. Chen, X., Chen, J., Han, X., Zhao, C., Zhang, D., Zhu, K., & Su, Y. (2020). A light-weighted cnn model for wafer structural defect detection. *IEEE Access, 8*, 24006–24018.
26. Li, Y., & Wang, J. (2021). A fast postprocessing algorithm for the overlapping problem in wafer map detection. *Journal of Sensors*. https://doi.org/10.1155/2021/2682286
27. Shawon, A. WaferMap — Kaggle. https://www.kaggle.com/datasets/shawon10/wafermap Accessed 2022-10-29.
28. Jimenez Gaona, Y., Castillo Malla, D., Vega Crespo, B., Vicuna, M. J., Neira, V. A., Davila, S., & Verhoeven, V. (2022). Radiomics diagnostic tool based on deep learning for colposcopy image classification. *Diagnostics (Basel, Switzerland), 12*(7), Article 1694.
29. Vorontsova, D., Menshikov, I., Zubov, A., Orlov, K., Rikunov, P., Zvereva, E., Flitman, L., Lanikin, A., Sokolova, A., Markov, S., et al. (2021). Silent EEG-speech recog-nition using convolutional and recurrent neural network with 85% accuracy of 9 words classification. *Sensors, 21*(20), Article 6744.
30. Srinivasu, P. N., SivaSai, J. G., Ijaz, M. F., Bhoi, A. K., Kim, W., & Kang, J. J. (2021). Clas-sification of skin disease using deep learning neural networks with mobilenet v2 and lstm. *Sensors, 21*(8), Article 2852.
31. Mushtaq, S., & Singh, O. (2024). A deep learning based architecture for multi-class skin cancer classification. *Multimedia Tools and Applications*, *83*(39), 87105–87127.
32. Sun, S., Wang, Z., Yu, M., Zhao, Y., He, Y., & Zhao, L. (2025). Automated classification of rheumatoid arthritis and knee synovitis from hyperspectral reflectance data. *Journal of Biophotonics*. https://doi.org/10.1002/jbio.202500197
33. Tuteja, Manchanda, G., Badhani, R., Jaitly, P., Gupta, G., & Gupta, S. (2025). MobileNetV2-Based Framework for Scalable and Accurate Remote Sensing Image Classification. ICE2CPT.



Ali Raza Ali Raza received the Bachelor of Science and M.S. degrees in computer science from the Department of Computer Science, Khwaja Fareed University of Engineering and Information Technology (KFUEIT), Rahim Yar Khan, Pakistan, in 2021 and 2023, respectively. He is currently pursuing a Ph.D. degree in the Department of Precision Medicine at Sungkyunkwan University, Suwon, South Korea. He was a Lecturer with the Faculty of Information Technology, Department of Software Engineering, The University of Lahore, Pakistan. He has published several articles in reputed journals. His current research interests include data science, artificial intelligence, data mining, natural language processing, machine learning, deep learning, and image processing.

Julio Cesar Martinez Espinosa Julio Cesar Martinez Espinosa is working as a Professor at the Universidad Europea del Atlantico, Spain. He is also affiliated with Universidad Internacional Iberoamericana, US and University Autonoma De Campeche, Mexico. His research interests include use of machine learning for automated and decision support systems.

Debora Libertad Ramirez Vargas Debora Libertad Ramirez Vargas is currently a Professor at the Universidad Europea del Atlantico, Spain. She is also affiliated with Universidad Internacional Iberoamericana, US. Her works include automated systems, image processing and machine learning.

Muhammad Shafiq Muhammad Shafiq received the master's degree in information technology (IT) from the University of the Punjab, Gujranwala, Pakistan, in 2006, the M.S. degree in computer science from the University Institute of Information Technology, Arid Agriculture University, Rawalpindi, Pakistan, in 2010, and the Ph.D. degree in information and communication engineering from Yeungnam University, South Korea, in February 2018. He was with the Faculty of Computing and IT, University of Gujrat, Gujrat, Pakistan, as a Faculty Member, from 2010 to 2014, and formerly held the same position with the Department of Computer Science and IT, Federal Urdu University, Islamabad, Pakistan. His research interests include the Internet of Things (IoT), cognitive radio-based IoT networks-architecture and design, mobile ad-hoc networks, wireless sensor networks, performance, management, and security, 5G cellular networks, admission control, mobility management, device-to-device communications, medium access control protocols, internet routing protocols, spectrum trading and auctions, information systems, design, and access control, and human–computer interaction.

Imran Ashraf Imran Ashraf received his Ph.D. in Information and Communication Engineering from Yeungnam University, South Korea in 2019, and the M.S. degree in computer science from the Blekinge Institute of Technology, Karlskrona, Sweden, in 2010 with distinction. He has worked as a postdoctoral fellow at Yeungnam University, as well. He is currently working as an Assistant Professor at the Information and Communication Engineering Department, Yeungnam University, Gyeongsan, South Korea. His research areas include machine learning, image processing, industrial automation and smart sensors.