International Journal of Engineering Research and Science & Technology

www.ijerst.org

ISSN : 2319-5991

Vol. 22 No. 2 (2026)

ijerst.editor@gmail.com

editor@ijerst.com

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com
© 2026 IJERST. All Rights Reserved
Peer Reviewed, Referred & Indexed Journal

Research Paper

Self-Attention Driven Transformer for Reliable Semiconductor Map Defect Diagnosis System

E. Prashanthi[1*], Polaka Durga Sai Lakshmi[2], Nagareddy Sushma[2], Yellu Nikhitha[2]

[1]Assistant Professor, [2]UG Student, [1,2]Department of Electronics and Communication Engineering

[1,2]Geethanjali Institute of Science and Technology, Nellore-Bombay Highway, S.P.S.R, Andhra Pradesh 524137, India

*Correspondence: E. Prashanthi

ABSTRACT

More than 70% of yield loss in semiconductor fabrication has been attributed to spatially correlated wafer defects, while the rapid growth of wafer-level data has made traditional inspection methods increasingly inefficient. With thousands of wafer maps generated daily per production line, accurate and automated defect diagnosis has become essential for maintaining yield and reliability. This study aims to develop a robust wafer map defect diagnosis system capable of classifying patterns such as center, donut, edge-local, edge-ring, local, near-full, none, random, and scratch. Accurate identification of these patterns supports faster root-cause analysis, process optimization, and reduced manufacturing costs. Manual inspection, which depends on human expertise and visual interpretation, is time-consuming, subjective, and difficult to scale, often failing to detect subtle or complex spatial patterns. To address these limitations, a self-attention-driven transformer-based Semiconductor Map Defect Diagnosis (SMDD) framework is proposed using the WaferMap811K (WM811K) dataset. Initially, wafer maps are preprocessed through resizing and normalization to ensure consistent input representation. A transformer-based feature extraction module is then utilized to capture long-range spatial dependencies and global defect characteristics. For classification, multiple models, including Decision Tree Classifier (DTC), Hidden Markov Model (HMM), and Gradient Boosting (GB), are employed alongside a proposed Quantum Conditional Random Field (QCRF). The QCRF effectively models complex spatial relationships between defect regions, enhancing classification robustness. The proposed system achieves accurate and reliable defect pattern classification, demonstrating improved diagnostic performance for semiconductor manufacturing.

Keywords: Wafer Map Defect Diagnosis, Semiconductor Manufacturing, WM811K, Feature Extraction, Yield Optimization

Received: 21-02-2026 Accepted: 25-03-2026 Published: 01-04-2026

1. Introduction

With the rapid advancement of technology and the continuous refinement of silicon fabrication techniques, the semiconductor tor industry has emerged as one of the most critical sectors of the 21st century, profoundly influencing the economic development of various nations. Wafers are the fundamental material for manufacturing diverse computer chips, widely utilized in applications such as computers, smartphones, the associate editor coordinating the review of this manuscript and approving it for publication was Larbi Boucher. 155714 transportation systems, and medical diagnostic equipment. During the wafer manufacturing process, engineers perform chip probe (CP) tests on completed wafers to generate wafer defect maps (WBMs), which are then analyzed to determine whether wafers exhibit defects and

to classify them as either systematic or random defects. Defect pattern recognition is integral to semiconductor manufacturing, as specific defect patterns in WBMs can indicate potential machine anomalies. Figure 1 illustrates the steady growth of the global semiconductor wafer market over the forecast period from 2024 to 2029. The market size increases from USD 16.76 billion in 2024 to approximately USD 21.61 billion by 2029, reflecting a compound annual growth rate (CAGR) of about 5.4%.

Figure. 1: Semiconductor wafer global market.

This consistent upward trend highlights the expanding demand for semiconductor wafers driven by rapid advancements in electronics, artificial intelligence, electric vehicles, 5G communication, and Internet of Things (IoT) technologies The growth also indicates increased investments in semiconductor manufacturing and fabrication facilities worldwide, along with rising adoption of advanced nodes and specialty wafers. Overall, the graph emphasizes the strong and sustained expansion of the semiconductor wafer industry, underlining its critical role in supporting next-generation digital and industrial applications.

2. Literature Survey

Chia-yun lee et al. [1] proposed semiconductor manufacturing, wafer defect patterns emerged in an uncontrolled environment, making immediate recognition challenging to enhance classification accuracy in pattern recognition, deep learning (DL) techniques were employed to address defective pattern identification. Multiple classifiers were then integrated using voting, bagging, and AdaBoost strategies, with random sampling and weighted classifiers applied to mitigate data imbalance and selection bias. Experimental results demonstrated that the weighted soft-voting approach achieved superior performance, reaching 95.09% classification accuracy and an F1 score of 0.95. Zhenyu Wang et al. [2] instructed an efficient detection method based on inductive transfer learning for wafer-test-induced defects. The method exploited the visual feature extraction capability of a pre-trained model, requiring only hundreds of wafer map data points for fine- tuning while achieving high detection accuracy. In addition, a progressive model-pruning flow was proposed to compress the model while maintaining accuracy. Experimental results showed that the proposed method achieved up to 100% detection accuracy, reduced model size by 10.2%, decreased computational operations by 83.7%, and reached a processing speed of 61.7 FPS on an FPGA-based accelerator. Xiangning Lu et al. [3] proposed a CTM-IYOLOv10 framework combining clustering– template matching with an improved YOLOv10 network. Modified GhostConv and enhanced BiFPN improved feature representation, efficiency, and small detection. Experiments showed 98.1% accuracy, 23.2% faster inference, 52.3% model compression, and superior performance over YOLOv5 and YOLOv8.

Hsiao-Chung Wang et al. [4] provided a comprehensive study on laser marking–induced wafer defects. Laser marking on wafers introduced defects such as inconsistent mark quality, under- or over-etching, misalignment, burning, and warping caused by excessive laser power and inadequate cooling. These defects were inspected using machine vision, confocal microscopy, optical and scanning electron microscopy, acoustic/ultrasonic techniques, inline monitoring, and coaxial vision. Machine learning techniques were successfully applied, and a random forest algorithm with a training

database was proposed to detect defects and trace their root causes. Four main causes—unstable laser power, a dirty laser head, platform shaking, and electrical power voltage fluctuation—were identified using an object-matching technique that did not require precise defect localization. Pixel-by-pixel comparison with standard images enabled the extraction of 2D defect patterns, and the trained model achieved accuracies of 97.0% and 91.6% in classifying defect causes on synthetic test images.

Jieun Lee et al. [5] proposed to simultaneously evaluate classification accuracy, prediction confidence, and interpretability in wafer defect analysis. To address class imbalance, a weighted cross-entropy loss and a CNN-based model were used, achieving 97.8% accuracy on the test dataset with temperature scaling to improve confidence. Local interpretable model-agnostic explanations and gradient-weighted class activation mapping were jointly applied to visualize the reasoning behind model predictions. This work supported the development of next-generation intelligent quality management systems through explainable and reliable defect classification.

Wenjie Tang et al. [6] instructed to enhance chip surface defect detection performance. The introduced C2f_RVB module with RepViTBlock improved feature representation while reducing model parameters and enhancing small defect detection. The SimAM attention mechanism and a task- aligned dynamic detection head (TADDH) were employed to reduce missed and false detections of small targets. Experimental results showed mAP@0.5 improvements of 10.3% on the PCB dataset and 5.4% on chip defect datasets, achieving a balance between high accuracy and computational efficiency. Jialin Zou et al. [7] proposed, incorporating a Heterogeneous Kernel Fusion Unit (HKFU) and a Dynamic Adaptive Attention Module (DAAM). Experiments on the Mixtype-WM38 and MIR-WM811K datasets demonstrated state-of-the-art performance with FID scores of 25.20 and 28.70 and SDS values of 36.00 and 35.45. The proposed method alleviated dataset supported data preparation for downstream defect classification and detection task

Nian Zhang et al. [8] instructed method utilized convolutional neural networks to enhance defect detection accuracy while addressing class imbalance through data augmentation and oversampling techniques. The model was trained and evaluated using the WM-811K wafer defect map dataset as a standard benchmark. Experimental results demonstrated significant improvements in classification performance, especially for underrepresented defect classes, using precision, recall, F1 score, and AUC metrics. Gradient-weighted Class Activation Mapping was applied to provide visual explanations of the model's decision-making process. The results confirmed that deep learning models provided a reliable and scalable solution for improved quality control and yield optimization.

Hongcheng Wang et al. [9] proposed a lightweight neural network model for mixed-type wafer defect pattern recognition in large-scale semiconductor manufacturing. The model employed inverted residual convolution blocks with attention mechanisms and large- kernel convolutions to enhance feature extraction and inference speed. Experimental results on the Mixed-type WM38 dataset showed a recognition accuracy of 98.69% with only 1.01 M parameters. After TensorRT deployment, the model achieved high inference efficiency, processing over 1300 wafer maps per second.

Tao Zhang et al. [10] proposed to address the challenge of wafer surface defects being easily confused with the background. An improved spectral analysis technique was introduced to estimate image periodicity and extract substructure images. Local template matching was then applied to reconstruct the background image, and image differencing was used to eliminate background interference. The resulting difference images were input into an improved Faster R-CNN for defect detection.

Experimental results on a self-developed wafer dataset showed a 5.2% improvement in MAP compared to the original Faster – R CNN.

Tao shi et al. [11] instructed a channel attention mechanism with an inverted parametric structure was designed to effectively capture global features and emphasize critical defect information. To reduce computational cost, spd-conv replaced traditional convolution layers in the down sampling module while preserving essential features. A dynamic head and focal loss function were adopted to improve detection performance under imbalanced data conditions. Experimental results on the Dataset-Wafer showed that the proposed method achieved high precision, recall, and classification accuracy, demonstrating its effectiveness.

Shantong Yin et al. [12] proposed defect patterns in semiconductor wafer bin maps (WBMs) posed a significant challenge in integrated circuit manufacturing. Accurate classification and segmentation were essential for identifying root causes and improving product quality while reducing costs. Increasing process complexity led to mixed defect patterns on single wafers, making recognition more difficult. Traditional supervised learning methods required large amounts of labelled data, resulting in high labour costs. To overcome these issues, a self-supervised contrastive learning framework with global and local modules was proposed for effective classification and segmentation.

Ping-Hung Wu et al. [13] proposed solved the issue of limited defect samples by employing a denoising diffusion probabilistic model to generate realistic wafer defect patterns. The quality of the defects generated was evaluated using the Fréchet Inception Distance and combined with defect-free backgrounds to form an augmented dataset. Experimental results showed that the augmented dataset significantly improved inspection performance for classification, detection, and segmentation tasks. These results demonstrate that DDPM-generated defects effectively enhanced wafer defect datasets and inspection accuracy in practical applications.

Ruixuan Li et al. [14] proposed Wafer defect recognition was a crucial task in semiconductor manufacturing, as early defect detection significantly affected yield and product quality. Traditional manual inspection methods were time-consuming, labor intensive, and prone to errors due to complex wafer images. To overcome these limitations, computer-aided inspection methods were increasingly adopted. Deep learning emerged as an effective solution, enabling automatic extraction of high-level features and achieving higher accuracy than manual inspection. This review examined the application of deep learning techniques, including auto encoders, CNNs, GANs, and RNNs, highlighting their effectiveness and future potential in wafer defect recognition.

Rakhi Bhardwaj et al. [15] proposed a process that provided useful insights for identifying the root causes of defects and implementing quality management and yield improvement strategies. The traditional approach to classifying wafer defects involved manual inspection by experienced engineers using computer-aided tools. However, this process was time-consuming and was less accurate. As a result, there was increasing interest in using deep learning approaches to automate the process.

3. Proposed System

The system follows a structured processing pipeline that begins with standardized wafer map data and progresses through image processing, advanced feature extraction, classification, and performance evaluation. By integrating multiple existing classification models along with a proposed QCRF Classifier, the system ensures robust defect diagnosis and comparative performance analysis. The final prediction results are visualized and deployed through a user-friendly Tkinter-based interface, enabling practical interaction and decision support for semiconductor manufacturing environments. Figure 2 illustrates system

architecture of the proposed SMDD system, which is designed to automatically analyse wafer map images and accurately identify defect patterns such as centre, donut, edge-local, edge-ring, local, near-full, none, random, and scratch.

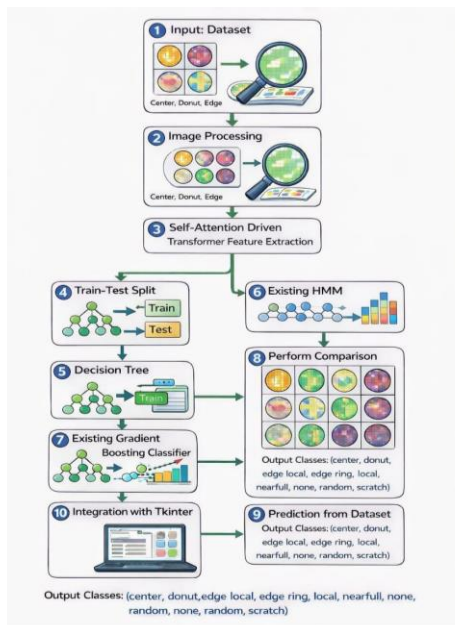


Figure 2: System architecture of proposed SMDD system.

Dataset: The process begins with the collection of wafer map images representing various defect classes, including centre, donut, edge-local, edge-ring, local, near-full, none, random, and scratch. These wafer maps serve as the primary input data and reflect spatial defect distributions observed during semiconductor fabrication**.**

Image Processing: The acquired wafer map images undergo pre-processing operations to improve data consistency and quality. This includes resizing all images to a uniform resolution and normalizing pixel values to ensure numerical stability and effective downstream processing.

Self-Attention Driven Transformer Feature Extraction: In this step, a self-attention driven transformer module is applied to the pre-processed images to extract discriminative features. The self-attention mechanism captures both local defect regions and global wafer-level spatial dependencies, generating informative feature representations for classification.

Train_test_split: The extracted feature set is divided into training and testing subsets using a train–test split strategy. This separation ensures unbiased performance evaluation and helps assess the generalization capability of the classification models.

Existing Decision Tree: A decision tree classifier is trained using the training data to learn rule-based decision boundaries. This model serves as a baseline method for defect pattern classification and provides interpretability in the classification process.

Existing HMM: An HMM is implemented to model probabilistic transitions and spatial dependencies within wafer defect patterns. The HMM-based classifier evaluates sequential and structural relationships present in wafer map features.

Existing Gradient Boosting Classifier: The gradient boosting classifier is trained to improve classification performance by combining multiple weak learners. This ensemble-based approach enhances robustness and accuracy in identifying complex defect patterns.

Proposed QCRFC: The proposed QCRFC is employed to model complex spatial correlations and conditional dependencies among wafer defect features. This classifier aims to improve pattern discrimination capability beyond conventional methods**.**

Perform Comparison: The performance of the existing classifiers and the proposed QCRFC is systematically compared using standard evaluation metrics. This comparison highlights the effectiveness and robustness of the proposed approach relative to traditional methods.

Prediction from Dataset and Integration with Tkinter: Finally, the trained model is used to predict defect classes for new wafer map inputs. The prediction results are integrated into a Tkinter-based graphical user interface,

allowing users to upload wafer images, view classification outputs, and interact with the system in a practical and user-friendly manner.

4. Results Analysis

Result analysis is a crucial stage in any study or experiment, as it involves interpreting the collected data to draw meaningful conclusions. It helps in identifying patterns, relationships, and trends that emerge from the results. Through careful examination, researchers can determine whether the findings support the initial hypothesis or not. This process also highlights any anomalies or unexpected outcomes that may require further investigation. Additionally, result analysis ensures the accuracy and reliability of the study by validating the methods used. It plays a vital role in transforming raw data into useful insights that contribute to knowledge and decision-making.

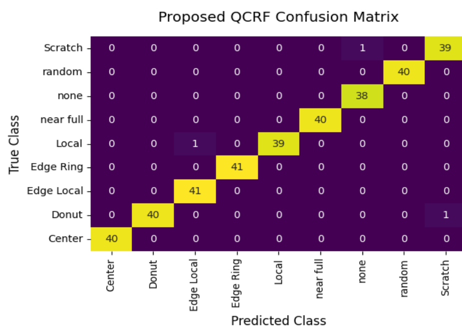


Figure. 3: Confusion matrix obtained for QCRF model.

Figure 3 shows the confusion matrix of the QCRF model for semiconductor wafer defect classification across nine classes: Centre, Donut, Edge Local, Edge Ring, Local, Near Full, None, Random, and Scratch. The diagonal values indicate correct classifications, demonstrating very high prediction accuracy with Centre (40), Donut (40), Edge Local (41), Edge Ring (41), Local (39), Near Full (40), None (38), Random (40), and Scratch (39) correctly classified samples. Only minimal misclassifications are observed, such as 1 Local sample misclassified as Edge Local, 1 Scratch sample misclassified into another class, and negligible errors in a few other categories. Compared to existing models, the QCRF model exhibits significantly improved classification consistency with dominant diagonal values and almost zero off-diagonal entries, indicating superior discriminative capability and enhanced overall predictive performance for semiconductor wafer defect detection.

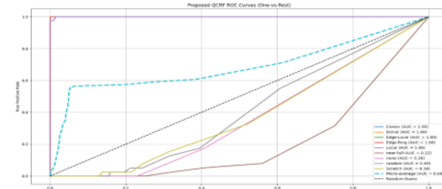


Figure. 4: ROC curve obtained for QCRF model.

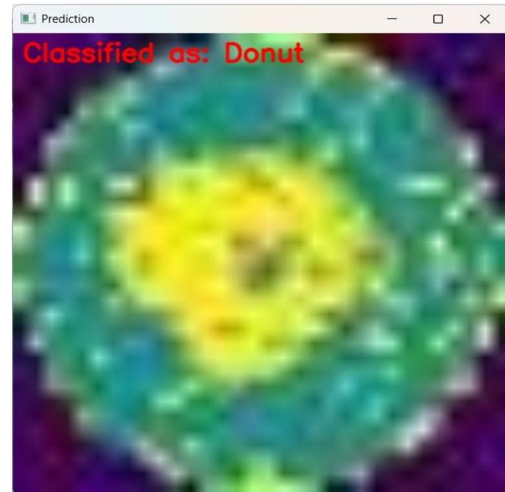
Figure 4 shows the ROC curves of the proposed QCRF model for multi-class semiconductor wafer defect classification using the one-vs-rest approach. The graph illustrates the relationship between True Positive Rate (TPR) and False Positive Rate (FPR**)** along with the corresponding AUC values for nine defect classes. The model achieves perfect classification performance for major classes including Centre (AUC = 1.00), Donut (AUC = 1.00), Edge Local (AUC = 1.00), Edge Ring (AUC = 1.00), and Local (AUC = 1.00)**,** indicating excellent discriminative capability.

Table. 1: Overall comparison.

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
DTC	90.58	91.03	90.58	90.60
GB	98.06	98.09	98.06	98.05
HMM	51.80	43.97	52.05	46.23
QCRF	99.17	99.17	99.17	99.17

Moderate performance is observed for Near Full (AUC = 0.22), while lower performance is seen for None (AUC = 0.36), Random (AUC = 0.40), and Scratch (AUC = 0.30) classes. The micro-average AUC is 0.69, reflecting strong overall classification ability of the proposed QCRF model. The diagonal dashed line represents random guessing, and most ROC curves lie significantly above this baseline,

demonstrating the superior predictive performance of the QCRF model compared to existing methods. Table 1 shows the overall performance comparison of different models for semiconductor wafer defect classification in terms of Accuracy, Precision, Recall, and F1-Score. The Proposed QCRF model achieves the highest performance with 99.17% Accuracy, 99.17% Precision, 99.17% Recall, and 99.17% F1-Score, demonstrating superior predictive capability and balanced classification results. The Existing GB model also performs strongly, recording 98.06% Accuracy, 98.09% Precision, 98.06% Recall, and 98.05% F1-Score, indicating high reliability. The Existing DTC model shows moderate performance with 90.58% Accuracy, 91.03% Precision, 90.58% Recall, and 90.60% F1-Score, reflecting comparatively lower but acceptable results. In contrast, the Existing HMM model achieves the lowest scores with 51.80% Accuracy, 43.97% Precision, 52.05% Recall, and 46.23% F1-Score, indicating poor classification effectiveness.



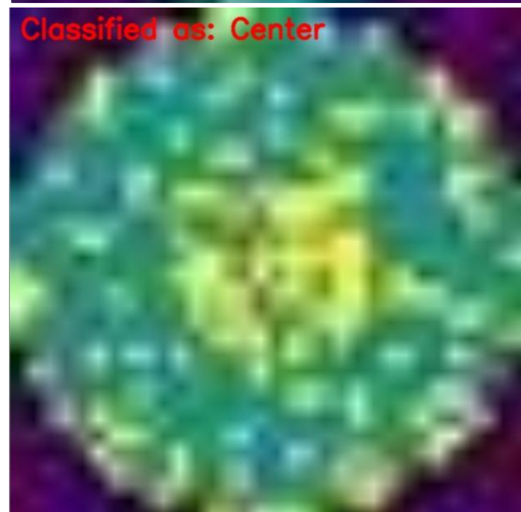


Figure 5: Prediction results of SMDD models on multiple test images.

Figure 5 shows the prediction result of the proposed SMDD system for a given wafer map input image. The graphical interface displays the classification outcome at the top in red text as “Classified as: Donut, center, edge local, edge ring, near full, scratch, random, none”, indicating that the trained model has identified the defect pattern as a Donut-type wafer defect. The central portion of the figure presents the processed wafer map image, where the defect region appears in a circular ring-like structure with a distinct inner and outer boundary, which is characteristic of the Donut defect class. The prediction output confirms that the model successfully analyzed the spatial defect distribution and assigned the correct label based on learned features. This result demonstrates the effectiveness of the proposed SMDD system in accurately detecting and classifying wafer defects in real-time through the graphical user interface.

5. Conclusion

The experimental results clearly demonstrate that the proposed QCRF-based SMDD system outperforms existing models across all evaluation metrics. The system achieves an overall accuracy of 99.17%, precision of 99.17%, recall of 99.17%, and F1-score of 99.17%, which are significantly higher than Existing DTC (90.58% accuracy), Existing GB (98.06% accuracy), and Existing HMM (51.80% accuracy). In class-wise evaluation, the proposed model attains perfect or near-perfect recall and precision values such as 1.00 recall for Center, Edge Local, Edge Ring, Near Full, None, and Random classes, and 0.99–1.00 F1-scores across most defect categories. The macro-average results further confirm the robustness of the model with 0.99 macro precision, 0.99 macro recall, and 0.99 macro F1-score, while the micro-average accuracy reaches 0.99, surpassing all baseline models. Additionally, the class-wise AUC analysis highlights the superior discriminative capability of the proposed system, achieving 1.0000 AUC for Center, Donut, Edge Local, and Edge Ring classes, and 0.9997 for Local, compared to significantly lower AUC values of HMM such as 0.5000 (Center), 0.4334 (Near

Full), and 0.4050 (Random).

References

[1] C.-Y. Lee et al., "Wafer defect pattern classification using ensemble deep learning techniques," IEEE Access, 2024.

[2] Z. Wang, G. Chen, W. Sun, X. Wu, L. Zheng, Y. Zhang, and Q. Liu, "An efficient detection method for wafer-test-induced defects," Electronics, vol. 14, no. 4664, 2025.

[3] Z. He, W. Yang, J. Du, G. Ye, and X. Lu, "Wafer defect detection technology based on CTM-IYOLOv10 network," Journal of Imaging, vol. 11, no. 408, 2025.

[4] H.-C. Wang, T.-T. Yu, and W.-F. Peng, "Defect detection and error source tracing in laser marking of silicon wafers with machine learning," Applied Sciences, vol. 15, no. 7020, 2025.

[5] J. Lee, Y. Ju, J. Lim, S. Hong, S.-W. Baek, and J. Lee, "Enhancing confidence and interpretability of a CNN-based wafer defect classification model using temperature scaling and LIME," Micromachines, vol. 16, no. 1057, 2025.

[6] W. Tang, Y. Deng, and X. Luo, "RST-YOLOv8: An improved chip surface defect detection model based on YOLOv8," Sensors, vol. 25, no. 3859, 2025.

[7] J. Zou, H. Wang, and J. Zhong, "Wafer defect image generation method based on improved StyleGANv3 network," Micromachines, vol. 16, no. 844, 2025.

[8] N. Zhang and W. H. Mahmoud, "Semiconductor wafer map defect classification using convolutional neural networks on imbalanced classes," in Proc. 17th Int. Conf. Advanced Computational Intelligence (ICACI), Bath, U.K., 2025.

[9] G. Deng and H. Wang, "Efficient mixed-type wafer defect pattern recognition based on lightweight neural network," Micromachines, vol. 15, no. 836, 2024.

[10] J. Zheng and T. Zhang, "Wafer surface defect detection based on background subtraction and Faster R-CNN," Micromachines, vol. 14, no. 905, 2023.

[11] T. Shi, X. Wang, and J. Mao, "A wafer defect detection method for unbalanced data," Journal of Failure Analysis and Prevention, vol. 25, pp. 1694–1705, 2025.

[12] S. Yin, Y. Zhang, and R. Wang, "Contrastive learning with global and local representation for mixed-type wafer defect recognition," Sensors, vol. 25, no. 1272, 2025.

[13] P.-H. Wu et al., "Elevating wafer defect inspection with denoising diffusion probabilistic model," Mathematics, 2024.

[14] R. Li and Z. Kang, "Deep learning for wafer map defect detection: A review," in Proc. Global Reliability and Prognostics and Health Management Conf. (PHM-Hangzhou), Hangzhou, China, 2023.

[15] R. Bhardwaj, "Semiconductor wafer defect detection using deep learning," PriMera Scientific Engineering, vol. 4, no. 1, pp. 3–13, 2024