IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 11 September 2023, accepted 21 September 2023, date of publication 2 October 2023,
date of current version 10 October 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3321025

RESEARCH ARTICLE

An Attention-Augmented Convolutional Neural Network With Focal Loss for Mixed-Type Wafer Defect Classification

UZMA BATOOL[1], MOHD IBRAHIM SHAPIAI[1], (Member, IEEE), SALAMA A. MOSTAFA[2], AND MOHD ZAMRI IBRAHIM[3]

[1]Centre for Artificial Intelligence and Robotics iKohza, Malaysia–Japan International Institute of Technology, Universiti Teknologi Malaysia, Kuala Lumpur 54100, Malaysia
[2]Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Parit Raja, Johor 86400, Malaysia
[3]Faculty of Electrical and Electronics Engineering Technology, Universiti Malaysia Pahang, Pekan, Pahang 26600, Malaysia

Corresponding authors: Mohd Ibrahim Shapiai (md_ibrahim83@utm.my) and Salama A. Mostafa (salama@uthm.edu.my)

This research was supported in part by the University Fundamental Research Grant scheme RDU200318 awarded by Universiti Malaysia Pahang Al-Sultan Abdullah (UMPSA) Malaysia; and in part by Universiti Tun Hussein Onn Malaysia (UTHM) through Tier 1 vot. Q166.

ABSTRACT Silicon wafer defect classification is crucial for improving fabrication and chip production. Although deep learning methods have been successful in single-defect wafer classification, the increasing complexity of the fabrication process has introduced the challenge of multiple defects on wafers, which requires more robust feature learning and classification techniques. Attention mechanisms have been used to enhance feature learning for multiple wafer defects. However, they have limited use in a few mixed-type defect categories, and their performance declines as the number of mixed patterns increases. This work proposes an attention-augmented convolutional neural networks (A2CNN) model for enhanced discriminative feature learning of complex defects. The A2CNN model emphasizes the features in the channel and spatial dimensions. Additionally, the model adopts the focal loss function to reduce misclassification and a global average pooling layer to enhance the network's generalization by reducing overfitting. The A2CNN model is evaluated on the MixedWM38 wafer defect dataset using 10-fold cross-validation. It achieves impressive results, with accuracy, precision, recall, and F1-score reported as 98.66%, 99.0%, 98.55%, and 98.82% respectively. Compared to existing works, the A2CNN model performs better by effectively learning valuable information for complex mixed-type wafer defects.

INDEX TERMS Anomaly detection, deep learning, mixed-type defects, multi-defect classification, pattern recognition, channel attention, spatial attention, wafer map classification, semiconductor fabrication defects.

I. INTRODUCTION

The demand for silicon chips is increasing because of their use in various applications, including emerging technologies such as the Internet of Things (IoT), autonomous vehicles, and 5G/6G networks. To meet this demand, semiconductor companies target enhanced production by addressing yield-limiting factors, such as wafer defects. A significant cause of wafer fabrication defects is flawed manufacturing equipment and processes [1], [2]. In chip manufacturing, integrated circuits (ICs) are built on a single-crystal silicon wafer through hundreds of steps in front-end (wafer fabrication and probe) and back-end (assembly and test) operations [3]. Front-end operations may cause wafer defects, risking circuit failure. Defective chips are discarded, reducing the overall chip production. Minimizing defects will prevent the waste of flawed chips, resulting in increased production [4], [5].

The associate editor coordinating the review of this manuscript and approving it for publication was Byung-Gyu Kim.

Wafer defects are identified during the post-fabrication wafer probe or die-sorting phase. Integrated circuits undergo a circuit probe (CP) test to generate a wafer map highlighting the failed circuits [6]. This map forms the foundation of wafer bin maps (WBMs) that are color-coded to distinguish faulty and intact dies/ICs. WBMs are used for defect monitoring, analysis, and root-cause assessment.

Wafer defect monitoring approaches include manual inspection by domain experts [7] and automated visual inspection (AVI) using computer-aided methods [8]. Manual inspections are costly, inefficient, and inaccurate, pushing for more suitable solutions for shrinking chip sizes. AVI, specifically through machine learning, is a preferred, efficient, and reliable choice. However, machine learning methods require feature engineering and struggle with large-scale, low-quality, noisy data. With its automatic feature acquisition ability, deep learning methods, such as convolutional neural networks (CNNs), excel in AVI for large, noisy, and incomplete datasets. In industrial defect inspections [9], deep learning can be used for defect classification, localization, and segmentation [10]:

A. DEFECT CLASSIFICATION

Defect classification identifies image-level defects using supervised or unsupervised learning to determine defect presence (anomaly detection) [11] or specify the exact defect label. Classification ensures early defect detection for swift product quality assessment and is ideal when defect-type understanding suffices for further actions. However, it does not offer intricate defect orientation information and faces challenges with images containing multiple non-uniformly shaped defects.

B. DEFECT LOCALIZATION

Defect localization determines the precise location of the defects within an image and marks the defect category. It is typically carried out by object detection methods [12] grouped into two-stage and one-stage object detectors: Two-stage detectors first generate candidate boxes through methods such as selective search or region proposal networks (RPNs) and then apply classification. Examples include R-CNN [13], Fast R-CNN [14], Faster R-CNN [15], and spatial pyramid pooling (SPP-net) [16]. They offer high detection accuracy due to their refined two-step process, adeptly handling size variations and complex defect scenarios. However, they are slower, with more intricate architectures and resource-demanding training pipelines. One-stage detectors, exemplified by models such as You Only Look Once (YOLO) [17] and Single Shot MultiBox Detector (SSD) [18], perform simultaneous predictions of object bounding boxes and classification in a single step. This approach yields faster inference speeds and simpler architectures, making it suitable for speed-critical real-time applications. However, their localization accuracy might be lower than two-stage methods, particularly for smaller or densely packed defect instances.

On the whole, defect localization furnishes precise spatial information for corrective measures. However, it necessitates annotated training data, which is labor-intensive and time-consuming. Moreover, it is unsuitable for defects in close proximity or those that overlap without distinct boundaries.

C. DEFECT SEGMENTATION

Defect segmentation [19], [20] provides intricate insights into defect attributes; size, shape, and precise extent via pixel-level outlining. It is effective for capturing irregular or non-uniform defects that lack predefined shapes. However, training and deploying segmentation models can impose substantial computational demands and require annotated pixel masks for training and assessment. Semantic segmentation and instance segmentation are prevalent strategies. In semantic segmentation, each pixel within an image is categorized as part of the defect class or background. FCN [21], U-Net [22], and SegNet [23] are commonly employed architectures for semantic segmentation. Instance segmentation identifies defect regions and distinguishes between distinct instances of the same defect class. Example architectures include Mask R-CNN [24] and Mask-refined R-CNN [25].

Several studies have demonstrated the effectiveness of deep learning for silicon wafer defects [26], [27] with a focus on single-type defects [28], [29], [30], [31], [32]. However, there is a need to investigate more mixed-type defects, which are becoming common with the increasing complexity of fabrication processes. The current deep networks perform well for wafers with few simple defects. However, their accuracy decreases with the defect intricacy and the number of defects, for example, three and four defects on a wafer. This is because the CNN-based models suffer spatial information loss and necessitate refinement mechanisms to improve their performance for closely located multiple smaller defects with similar features [33], [34].

D. CONTRIBUTIONS

This work proposes a CNN-based model for mixed-type wafer defect classification, which maintains accuracy despite increasing defect complexity and count. To augment CNN's ability for minute defects and discriminative learning of similar classes, an attention module comprising spatial and channel attention is integrated, ensuring a more comprehensive feature set. The attention module emphasizes the most pertinent and informative features for enhanced detection accuracy. The focal loss function is employed instead of the cross-entropy function to minimize false negatives. The misclassification rate is reduced by emphasizing harder examples of each class. The global average pooling (GAP) layer is adopted as a regularization method, improving generalization and mitigating overfitting. All these measures ensure that the model is efficient and accurate in handling complex multi-pattern defects commonly found in wafer maps. The proposed model is thoroughly evaluated on the MixedWM38 dataset, which encompasses 38 different defect

classes. Comparative analysis with previous models is conducted to showcase the efficacy of the proposed model in accurately detecting and classifying complex multi-defect patterns. The main contributions of this work are summarized as follows:

- A novel CNN-based model is proposed that integrates a convolutional (Conv) module with an attention module to enhance the accuracy of mixed-type wafer defect classification. The Conv module extracts feature maps, which are further refined by incorporating channel and spatial information. This augmentation results in a more representative feature set, particularly beneficial for identifying small-sized defects that resemble one another. The model is named the attention-augmented convolutional neural network (A2CNN).
- The focal loss function is used instead of categorical cross-entropy to reduce false negatives and enhance the model's ability to correctly classify challenging examples in each class.
- A global average pooling layer is adopted as a regularization technique to enhance generalization capabilities, prevent overfitting, and reduce the model's parameters and training time.
- An operative evaluation of the A2CNN on the MixedWM38 dataset is performed, and the performance is compared against notable existing models through accuracy, precision, and recall.

The rest of the paper is arranged as follows. Section II reviews the related literature on mixed-type defects, attention mechanisms, focal loss function, and GAP. Section III describes the dataset, preprocessing and distribution procedures, proposed model architecture, and performance measures. Section IV presents the results, comparison, and analysis. Finally, Section V concludes the work and presents the findings.

II. RELATED WORK

This section examines literature concerning deep learning for mixed-type wafer defect classification. Additionally, it delves into attention mechanisms, focal loss function, and GAP for advanced deep-learning models.

A. DEEP LEARNING FOR MIXED-TYPE WAFER MAP DEFECT CLASSIFICATION

To identify mixed-type wafer defects, Tello et al. [35] applied a splitter to differentiate single and mixed patterns. Then, a randomized general regression network is used to classify single defects, and a CNN (Convolutional Neural Networks) for mixed patterns. Kyeong and Kim [36] proposed an ensemble of four CNNs trained on basic single defects to identify their appearance in mixed-type defects. However, training multiple CNNs is computationally expensive and time-consuming, particularly with many single defect types. Kong and Ni [37] focused on overlapped and non-overlapped multi-pattern wafers. A seed-filling method was employed for non-overlapped pattern segmentation followed by CNN classification. Overlapped patterns were classified through pattern-matching. The subsequent work [38] introduced pattern segmentation via a UNet-based boundary detection and pattern unwrapping before CNN-based classification. Further, quantitative analysis by pattern remapping and impact calculation is conducted. These approaches involve several intermediate steps, wherein the outcome of each step relies on the proper working of the previous ones.

Byun and Baek [39] adopted pre-trained CNNs with convolutional auto-encoders as weight initializers. Separating the single and mixed-type defects first and then classifying the exact pattern. The model is trained on the single defect dataset and tested on a few synthetic mixed-type defects with only two patterns. The accuracy for the single patterns is sufficient but dissatisfactory for the mixed patterns. Lee and Kim [40] presented a semi-supervised convolutional generative model for labeled and non-labeled data. Zhuang et al. [41] employed a real and private dataset of six individual types to train an ensemble of deep belief networks. Shin et al. [42] produced mixed-type wafer maps through data augmentation, combining WM-811K single defect wafer maps for CNN training. These studies were based on limited data collected from own sources or generated synthetically, unavailable in the public domain for model benchmarking.

Wang et al. [43] introduced MixedWM38, a mixed-type wafer defect dataset and employed a deformable convolutional network (DC-Net) for classification. DC-Net integrated positional information and displacement adjustment through a deformable convolutional kernel to improve CNN-based classification. In another study, a multi-scale information fusion transformer (MSF-Trans) [44] was proposed in which a multi-head self-attention mechanism assigned weights to feature values for an optimized feature set learning. Wei and Wang [45] proposed a multi-resolution wavelet-integrated attention network (MRWA-Net) expanding the learning of frequency components from the wavelet domain. In [46], WaferSegClassNet (WSCN), an end-to-end encoder-decoder-based defect segmentation network, was proposed for MixedWM38. WSCN employed N-pair contrastive loss, BCE-Dice loss (binary cross-entropy + Dice loss), and categorical cross-entropy loss for encoding, segmentation, and classification, respectively. A U-Net with residual attention block was proposed for segmentation in [47]. Segmentation masks were generated by the defect masking technique, but the model was evaluated only for a few classes of MixedWM38. These studies demonstrated good accuracy for single and a few simple multi-defect wafers but could not cater to many small defects with complex patterns.

In summary, the previous deep learning methods for mixed-type wafer recognition are based on resource-intensive ensemble networks, methods with multiple intermediate steps, and pre-trained networks adopted from other domains through transfer learning. They have only been evaluated on smaller datasets with fewer defect types and exhibit good results for single and two mixed defects but need to be optimized for wafers with three and four mixed patterns.

In contrast, the current work proposes A2CNN, a custom-made CNN, which combines channel and spatial information into the feature maps for more comprehensive defect features. The focal loss function addresses the classification of challenging class samples, and a GAP layer prevents overfitting and increases generalization. The proposed model offers a one-shot solution for the diverse MixedWM38 patterns.

B. ATTENTION MECHANISM

Attention mechanisms originally introduced for machine translation [48] have a transformative impact in various domains, enabling models to focus on specific elements within input data [49]. The concept of attention is based on human perception. From the available information, human beings select the relevant part for processing, ignoring the remaining, thus greatly improving the efficiency and accuracy of perceptual information processing [50]. Its application in disaster assessment is a difference-aware attention network (D2ANet) [51] for building localization and change detection using dual-temporal satellite imagery. The Difference-Aware Attention (D2A) block integrates the Dual-Temporal Aggregation (DTA) and Difference-Aware (DA) modules to investigate multi-level transformations within pre- and post-disaster features. The DTA module selectively amplifies change-sensitive channels to extract global change information, while the DA module is a localized attention block that captures correlations across positions and channels to discern various damage scales.

Attention mechanisms have substantially shaped object classification and detection tasks, effectively merging with machine vision applications for enhanced defect identification. An edge and multi-scale reverse attention network (EMRA-Net) is proposed in [52] for tiny and low-contrast surface defect detection. For improved detection accuracy, global dynamic convolution features, global multi-scale fusion features, and local pyramid edge features were obtained and fused by self-learning weight and spatial and channel attention modules. The Defect-Aux-Net [53], built upon multi-task learning with an attention mechanism, undertakes the concurrent tasks of classification, segmentation, and detection of industrial surface defects. The feature pyramid network (FPN) and ResNet-50 inspired the network architecture. By incorporating positional and channel attention, the network enhances the learning of subtle defects, enriching feature maps with global context. The dynamic weighting of each task's loss collectively serves to refine the network's overall loss function. Subsequently boosting accuracy and efficiency across all tasks.

Two renowned architectures which utilize channel-spatial attention mechanisms are the bottleneck attention module (BAM) [54] and the convolutional block attention module (CBAM) [55]. BAM adopts a bottleneck structure, extracting channel and spatial attention maps via the global average pooling of feature maps. This enables the module to concentrate on fine-grained features and global context. In contrast, CBAM sequentially integrates channel and spatial attention, capturing inter-channel relationships and spatial dependencies. CBAM employs MaxPooling and AvgPooling to compute feature map statistics, leading to enriched feature representation and enhanced contextual comprehension.

C. FOCAL LOSS FUNCTION

The focal loss function, first applied in RetinaNet [56], has proven valuable for object detection in dense scenes and several other studies [57], [58]. Built upon the cross-entropy loss, it addresses foreground-background class imbalance and effectively boosts performance of minority classes for class imbalanced data. The model enhances object detection or classification tasks focusing on challenging cases by down-weighting easy instances and amplifying hard ones during training.

D. GLOBAL AVERAGE POOLING (GAP)

GAP is a widely employed technique in deep learning to enhance feature extraction and boost classification accuracy [59]. It serves as a means to mitigate overfitting by incorporating spatial information regularization. GAP replaces the spatial dimensions of feature maps with a single value per channel. It offers improved generalization, reduced overfitting, quicker convergence, and robustness to image translation and scaling.

III. MATERIALS AND METHODS

This section provides the design and development details of the A2CNN model. The whole process is illustrated in Fig. 1. It includes two main phases:

- Data Preprocessing and Distribution: This encompasses the methods applied to the initial wafer map dataset, rendering it suitable for CNN compatibility and partitioning it into training, validation, and testing subsets.
- A2CNN Development: This entails model design, iterative training, and validation on the designated dataset

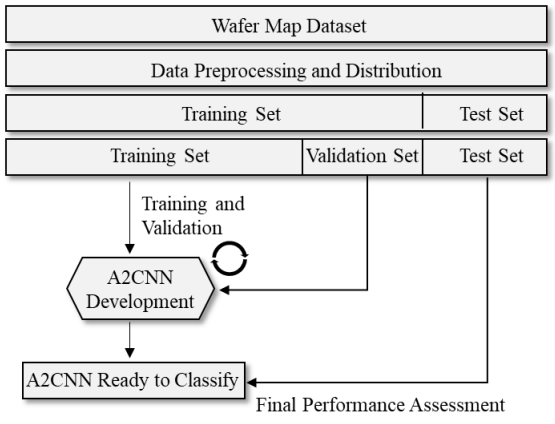


FIGURE 1. Block diagram of the A2CNN development process; the model was trained and validated iteratively on the preprocessed data and evaluated on the test set for performance assessment.

portions and evaluating its performance on the test set to produce an operational model for wafer map classification.

A. DATASET

MixedWM38 is a public domain wafer map dataset of single and mixed-type defects. Comprised of real and artificially generated 38,015 wafer maps in 38 defect classes, divided into four groups: Group_1 consists of nine single-type defect classes, encompassing a normal class and eight distinct classes featuring a solitary defect pattern. The remaining three groups combine fundamental defect types in configurations of two, three, and four defect patterns. Group_2 comprises 13 classes with two mixed-type defects, Group_3 presents 12 classes with three mixed-type defects, and Group_4 entails four classes featuring four mixed-type defects. Class sizes vary; C7 with 866, C9 with 149, C24 with 2,000, and the remaining 35 classes have 1,000 wafer maps in each. Wafer maps have consistent dimensions of 52×52 pixels. A pixel is assigned a value of 0, 1, or 2; 0 signifies background/non-wafer pixels, 1 represents qualified/good dies, and 2 denotes flawed/defective dies. Image samples in the four class groups are illustrated in Fig. 2.

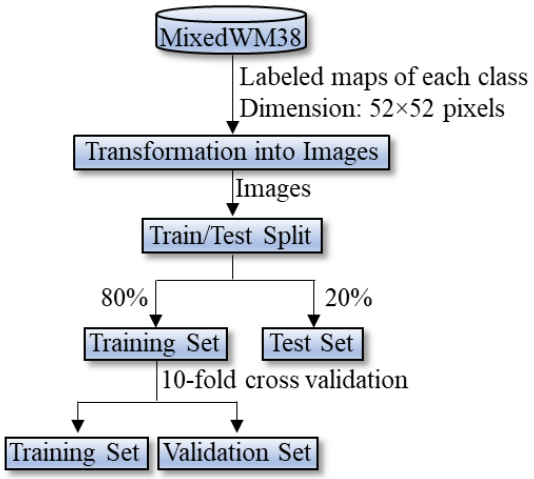


FIGURE 2. Data preprocessing steps: wafer maps were transformed into images and divided into training, validation, and test sets.

B. DATA PREPROCESSING AND DISTRIBUTION

In data preprocessing, the wafer maps were transformed into grayscale images by mapping pixel values from [2, 0, 1] to [0, 127, 255] and subsequent conversion into images. These images were split into training and test sets following an 80/20 ratio, maintaining a balanced representation across classes. For 10-fold cross-validation, the training set was divided into 10 distinct non-overlapping training and validation subsets. This iterative process led to ten datasets by combining nine folds for training while utilizing the remaining one for validation. Data preprocessing steps are illustrated in Fig. 3, and the distribution of class samples across sets is outlined in Table 1.

TABLE 1. Division of class samples into training, validation, and test sets.

	Class	Total Samples	Training Set (80%) divided into 10 folds of training and validation sets	Test Set (20%)
1	C7 (R)	866	693 (69*9 + 72): 9 folds with 62 and 10-th fold with 72 samples	173
2	C9 (NF)	149	119 (11*9 + 20): 9 folds with 11 and 10-th fold with 20 samples	30
3	C24 (C+EL+S)	2,000	1,600 (160*10): 10 folds with 160 samples each	400
4	All Others	1,000	800 (80*10): 10 folds with 80 samples each	200
	Total	38,015	30,412	7,603

C. THE ATTENTION-AUGMENTED CONVOLUTIONAL NEURAL NETWORK (A2CNN)

The A2CNN architecture is illustrated in Fig. 4 with an elaborated design of the modules and blocks inside the modules. The Conv module extracts feature maps from the input image, which are passed to the Attention module for more contextual information on salient, class-specific features. The attention blocks within this module blend feature maps across the channel and spatial dimensions. The refined features obtained from each attention block are aggregated and passed to the Classification module, where the final decision on the image label is made.

The Conv module consists of three Conv blocks, each having a Conv layer with Batch Norm (batch normalization), and ReLU (Rectified Linear Unit). Batch Norm expedites training, mitigates internal covariate shifts, and aids in avoiding local minima [60], [61]. Also, zero padding is employed in convolutions to retain information on defect patterns at the wafer edges. Maxpooling layers in the last two Conv blocks downsize the feature maps.

The Attention module comprises eight Attention blocks, each having a Conv block with channel and spatial attention layers, added sequentially based on the CBAM architecture. This module enhances the significant features extracted by the Conv module, emphasizing the most pertinent ones. This enhancement is achieved by amalgamating channel and spatial attention maps in the feature maps, generating intermediate channel-refined and fully refined attention maps, as summarized in (1) and (2):

$$F' = \mathrm{M_c}\,(F) \otimes F \tag{1}$$

$$F'' = \mathrm{M_s}\left(F'\right) \otimes F' \tag{2}$$

Here, F and F' denote feature maps and intermediate channel refined attention maps, while F'' represents fully refined attention maps. M_c, M_s and $\otimes$ represent channel attention maps, spatial attention maps and element-wise multiplication. M_c and M_s can be expressed by (3) and (4) as:

$$\mathrm{M_c}(F) = \sigma\,(\mathrm{MLP}\,(\mathrm{AvgPool}\,(F)) + \mathrm{MLP}\,(\mathrm{AvgPool}\,(F))) \tag{3}$$

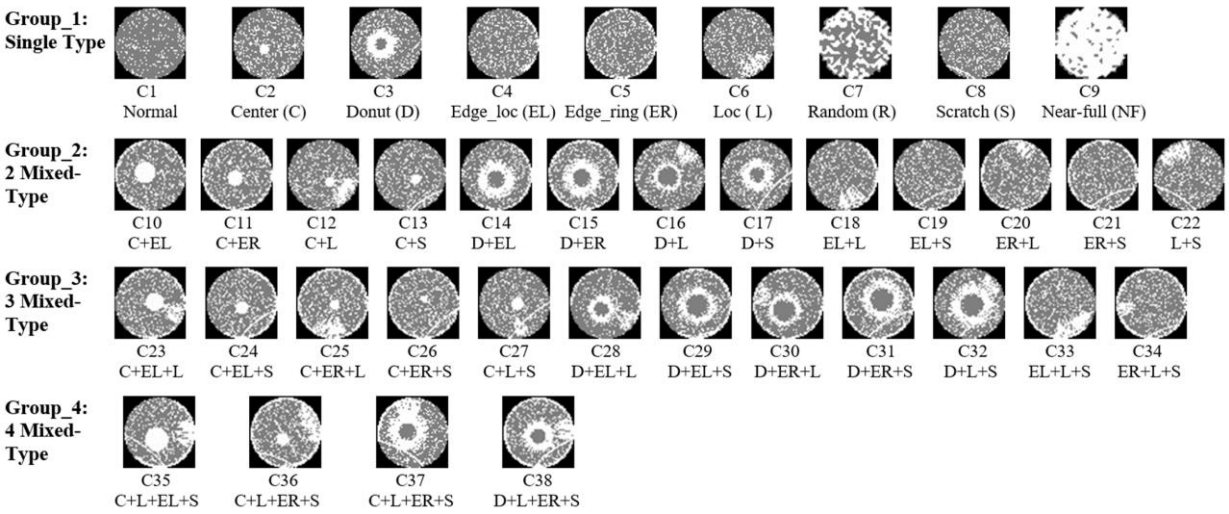


FIGURE 3. MixedWM38 image samples in four class groups based on number of defect patterns on wafer surface.

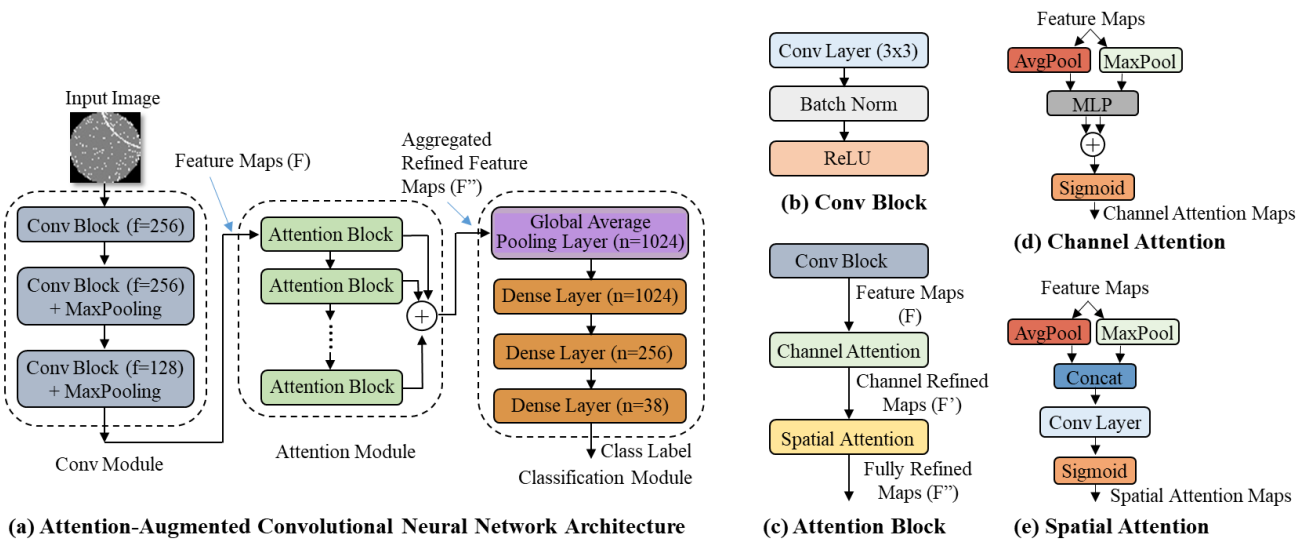


FIGURE 4. A2CNN architecture with the configuration of the sub-modules and blocks: (a) Three main modules, (b)-(c) Conv and attention blocks, (d)-(e) Channel and spatial attention.

Here, MLP (Multi-Layer Perceptron) stands for multi-layer perceptron, while AvgPool(F) and MaxPool(F) refer to the operations of average pooling and max pooling applied to the input feature maps F, respectively.

$$\mathrm{M_s}(F) = \sigma\left(\mathrm{conv}(\mathrm{concat}\left[\mathrm{AvgPool}(F); \mathrm{maxPool}(F)\right])\right) \tag{4}$$

Equation (4) illustrates that the feature maps obtained by concatenating average and maxpooled representations of F underwent a convolution operation before being subjected to a sigmoid function.

The outputs of the Attention blocks were summed and forwarded to the Classification module, which includes a GAP and three dense layers. The GAP layer calculates the average of each feature map, reflecting in a global feature vector with reduced dimensionality that represents the entire input image. This operation is formally defined by (5):

$$\mathrm{GAP}(C_i) = 1/N \times \sum x_j \tag{5}$$

Here, C_i represents the value of the i-th channel following the global average pooling, N denotes the total number of spatial locations within the feature map, x_j represents the activation value at the j-th spatial location of the i-th channel. Averaging a feature map transforms it into a confidence map category, enhancing its resilience to spatial translations in the input data. This process, in turn, boosts the network's ability to generalize while concurrently reducing parameters to prevent overfitting.

After the GAP layer, the Classification module contains three dense layers featuring 1024, 256, and 38 nodes. The 38 nodes in the output layer correspond to the 38 classes of the MixedWM38. The activation function employed here is

Softmax, which yields class probabilities between 0 and 1, with the sum of all probabilities equating to 1. The optimization is performed by Adam [62] with a learning rate of 0.001. As for the loss function, categorical focal loss, with default settings, is employed, which can be expressed by (6):

$$\mathrm{FL}\,(p) = -\alpha\,(1-p)^{\gamma}\log(p) \tag{6}$$

Here, α is the balancing factor that adjusts a class's contribution based on its frequency. P represents the predicted probability of the correct class, and γ acts as the focusing parameter, governing the rate at which the loss decreases with an increase in predicted probability. The $(1-p)^{\gamma}$ is the modulation factor that amplifies the contribution of challenging or, conversely, the contribution of well-classified examples.

During training, the network was presented with batches of normalized images augmented by horizontal and vertical flips for increased robustness. However, in validation and testing, only normalized images were provided. The model was subjected to ten tests following a 10-fold cross-validation. Each test involved training the model for 100 epochs using a batch training method with 32 images in each batch. At the end of each epoch, the model was validated and saved only if it performed better on the validation data than the previously saved model. The best-validated model was then evaluated on the test set. Table 2 conveniently encapsulates the configuration hyperparameters employed in the A2CNN model.

TABLE 2. The hyperparameter values of the proposed A2CNN.

Hyperparameters	Values
Input size	52 × 52 × 1
Convolutional layers	11 (3 in Conv and 8 in Attention module) filter size 3×3 and stride 1
Filters in Conv layers	266/256/256/128/128/128/128/128/128/128/128
Activation function	ReLU
Zero padding	Yes
Max pooling layers	2 (filter size 2×2 and stride 1)
Attention type	Channel and Spatial
Nodes in the GAP layer	128
Dense layers	3 (Nodes: 1024, 256, and 38)
Output layer	Softmax activation: 38 nodes for 38 classes
Batch size	32
Optimizer	Adam (learning rate 0.001)
Loss function	Categorical focal loss

D. PERFORMANCE EVALUATION METRICS

The model's performance is assessed by accuracy (7) and (8), precision (9), recall (10), and F1-score (11), which depend on TP (true positive), TN (true negative), FP (false positive), and FN (false native):

$$\text{Accuracy} = (\mathrm{TP}+\mathrm{TN})/(\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}) \tag{7}$$

Equation (7) gives the model's overall accuracy. Here, TP+TN counts to the correct predictions for all classes, and TP+TN+FP+FN is the total prediction made by the classifier (total samples in all classes). Similarly, the accuracy of a single class is given by (8).

$$\text{Class Accuracy} = \mathrm{TP}/P \tag{8}$$

Here, P stands for the positive samples of a particular class, and TP represents the correctly predicted samples of that class.

$$\text{Precision} = \mathrm{TP}/(\mathrm{TP}+\mathrm{FP}) \tag{9}$$

Equation (9) defines precision, TP represents the correctly predicted class samples and TP+FP the total number of positive predictions for the class (including samples not belonging to the class).

Recall or the True Positive Rate (TPR), is defined as:

$$\text{Recall} = \mathrm{TP}/(\mathrm{TP}+\mathrm{FN}) \tag{10}$$

Here, TP is the correctly predicted sample, and TP+FN represents the total number of class instances (including samples predicted as belonging to another class).

The F1-score or F1-measure combines precision and recall by the harmonic mean as given by (11).

$$\begin{aligned}\text{F1}-\text{score} = {} & [(1+\beta^2)\times\text{Recall}\times\text{Precision})]/ \\ & \times\ [\beta^2\times\text{Recall}\times\text{Precision}]\end{aligned} \tag{11}$$

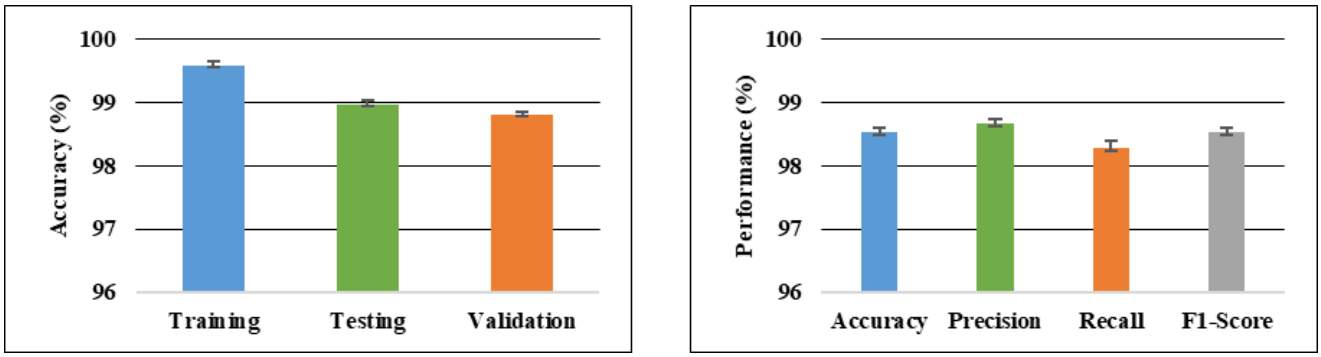
Here, coefficient β adjusts the relative importance of precision versus recall, and its value is 1.0 for the F1-measure.

IV. RESULTS AND DISCUSSION

This section presents experimental results and a detailed analysis of the outcomes. The first part of this section focuses on the performance evaluation of A2CNN, including a comparison with previous studies. The findings are analyzed to determine the approach's effectiveness in achieving accurate classification results. Additionally, an in-depth examination of the model components is conducted to assess their role in improving the classification performance.

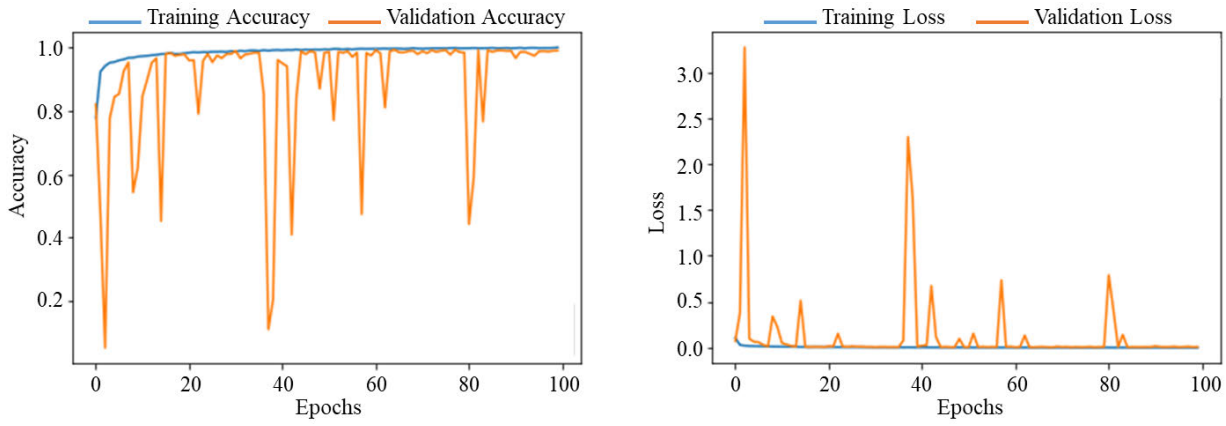
A. EXPERIMENTAL RESULTS

The A2CNN's stability and robustness were ensured through ten repetitions of the experiment, varying training and validation sets while keeping the test set consistent. Comprehensive results of all repetitions in 10-fold across-validation can be seen in Appendix (Tables 4 and 5). Fig. 5 presents the aggregated values of the tests. Fig. 5(a) illustrates the average accuracy of the tests (training 99.59%, std. dev. 0.12; validation 98.97%, std. dev. 0.13; testing 98.81%, std. dev. 0.1). Fig. 5(b) illustrates the average percentage of class accuracy (98.47%, std. dev. 0.18), precision (98.77%, std. dev. 0.14), recall (98.27%, std. dev. 0.20), and F1 score (98.59%, std. dev. 0.14) for the test set. The values were stable across the trest with minimal deviation, which underscores A2CNN's reliability, robustness, and generalization across diverse datasets and unseen data.



(a) The average accuracy of 10 folds (b) The average performance of 10 folds

FIGURE 5. Aggregated results of the 10-fold cross-validation: (a) Average accuracy of the Training, Testing, and Validation in ten repetitions of the experiment, (b) Average Accuracy, Precision, Recall, and F1-Score for test set in ten repetitions.



(a) Training and validation accuracy (b) Training and validation loss

FIGURE 6. The training and validation graphs of the A2CNN: (a) Accuracy graphs, (b) Loss graphs.

Test three produced the highest values for the performance measures, which are compared to prior studies. The training and validation accuracy graphs of this test are displayed in Fig. 6(a). Training accuracy exceeded 90% initially and steadily rose to approximately 99%. Validation accuracy, while subject to occasional dips, maintained its value during training, showcasing A2CNN's rapid convergence and maximum accuracy attainment. Fig. 6(b) reveals nearly zero training and validation loss. The overlap of the curves confirms the avoidance of overfitting and the stability of the model.

Table 3 presents A2CNN's accuracy, precision, and recall, along with comparisons to WSCN [46], MSF-Trans [44], and DC-Net [43]. A2CNN achieved an average classification accuracy of 98.66%, outperforming the other methods (98.34%, 97.17%, and 93.20%). Its average precision (99.00%) and recall (98.55%) also surpassed the others, with precision gains of 1.00%, 1.86%, and 5.00%, and recall gains of 0.37%, 1.84%, and 3.55% with lower standard deviations for these metrics. A2CNN's lower precision bound is 97%, compared to WSCN's 92%, MSF-Trans's 92.78%, and DC-Net's 60%, which suggests fewer false positives for A2CNN. Similarly, A2CNN's recall rate is significant, with only minor recognition errors. Specifically, it exhibits superior recall rates for defects like C14, C20, C27, C33, C37, and C38. C37's 99% recall rate surpasses WSCN by approximately 4%, MSF-Trans by 6.87%, and DC-Net by 10%.

As for the defect groups, A2CNN excelled in single-pattern classes (C1-C8), achieving flawless classification for C1 and C6. However, performance on C9, with only 149 samples, showed room for improvement, possibly due to limited training data. A2CNN consistently outperformed in mixed patterns (C10-C38). Notably, it effectively classified three and four mixed-pattern classes (C23-C38), known for their complexity. For example, WSCN had 99%, MSF-Trans 95.65%, and DC-Net only a 92% precision for C38 compared to the 100% of A2CNN. WSCN, MSF-Trans, and DC-Net struggled more with such patterns, where A2CNN excelled, demonstrating its recognition strength.

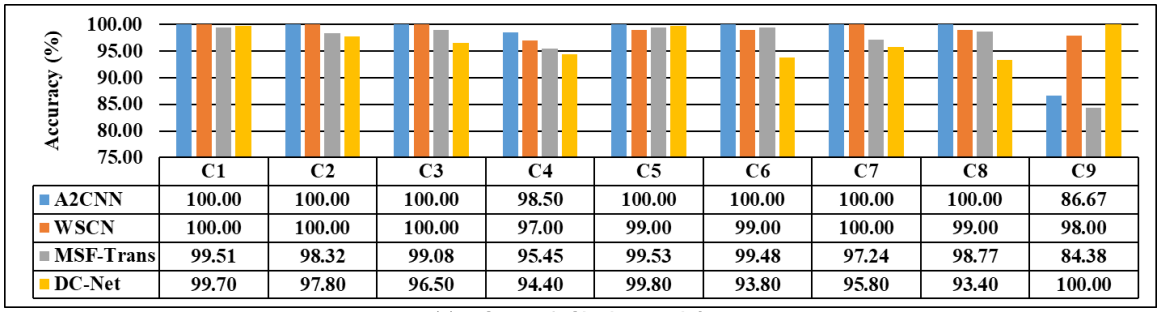
Fig. 7 displays a performance comparison among the models using bar graphs to present accuracy across the four class groups. In Fig. 7(a), A2CNN excels in recognizing single-type defects, except for C9, where it outperforms MSF-Trans but not the other models. Fig. 7(b) shows A2CNN's consistently high accuracy for most classes with two defect patterns, while other models exhibit varying accuracy levels. Fig. 7(c) and (d) highlight A2CNN's consistent superiority in recognizing three and four mixed-type defects, a notable achievement.

TABLE 3. The test set results of the proposed A2CNN and a comparison with WSCN [46], MSF-Trans [44], and DC-Net [43].

	Accuracy (%)				Precision (%)				Recall (%)			
Class	**A2CNN (Proposed)**	**WSCN [46]**	**MSF-Trans [44]**	**DC-Net [43]**	**A2CNN (Proposed)**	**WSCN [46]**	**MSF-Trans [44]**	**DC-Net [43]**	**A2CNN (Proposed)**	**WSCN [46]**	**MSF-Trans [44]**	**DC-Net [43]**
C1	**100.00**	**100.00**	99.51	99.70	**100.00**	**100.00**	99.04	94.00	**100.00**	**100.00**	**100.00**	91.00
C2	**100.00**	**100.00**	98.32	97.80	**99.00**	**99.00**	97.22	93.00	**100.00**	**100.00**	97.77	97.00
C3	**100.00**	**100.00**	99.08	96.50	97.00	96.00	**97.73**	95.00	**100.00**	**100.00**	98.62	93.00
C4	**98.50**	97.00	95.45	94.40	**99.00**	98.00	97.96	96.00	**98.00**	**98.00**	96.97	91.00
C5	**100.00**	99.00	99.53	99.80	**98.00**	97.00	97.72	93.00	**100.00**	99.00	99.53	97.00
C6	**100.00**	99.00	99.48	93.80	**100.00**	99.00	97.95	99.00	**100.00**	99.00	98.96	100.00
C7	**100.00**	**100.00**	97.24	95.80	**98.00**	92.00	95.95	90.00	**100.00**	**100.00**	97.93	94.00
C8	**100.00**	99.00	98.77	93.40	**99.00**	97.00	96.39	60.00	**100.00**	**100.00**	98.77	88.00
C9	86.67	98.00	84.38	**100.00**	**100.00**	**100.00**	96.30	97.00	87.00	98.00	81.25	**93.00**
C10	98.50	98.00	97.55	**99.20**	**98.00**	**98.00**	96.62	94.00	98.00	**99.00**	98.04	94.00
C11	**100.00**	100.00	99.47	97.90	98.00	**99.00**	96.41	92.00	**100.00**	**100.00**	99.47	99.00
C12	**99.50**	99.00	97.57	98.50	**99.00**	**99.00**	97.58	92.00	99.00	**100.00**	98.06	96.00
C13	**100.00**	99.00	99.52	96.70	**99.00**	98.00	96.76	97.00	**100.00**	**100.00**	**100.00**	89.00
C14	99.00	94.00	97.87	**99.30**	99.00	**100.00**	98.91	96.00	**99.00**	95.00	96.81	92.00
C15	**99.00**	**99.00**	97.69	96.10	**99.00**	98.00	96.38	91.00	**99.00**	**99.00**	98.61	98.00
C16	96.50	95.00	95.69	**98.30**	98.00	**100.00**	96.62	94.00	96.00	95.00	95.69	**97.00**
C17	**100.00**	**100.00**	96.94	92.80	**100.00**	96.00	97.96	96.00	**100.00**	**100.00**	97.96	94.00
C18	**99.50**	99.00	96.90	93.90	98.00	**99.00**	96.09	98.00	99.00	**100.00**	97.79	89.00
C19	98.00	97.00	**98.61**	92.30	**100.00**	99.00	99.53	94.00	**98.00**	**98.00**	97.69	91.00
C20	**99.00**	96.00	97.79	94.60	**99.00**	**99.00**	95.68	95.00	**99.00**	96.00	97.79	91.00
C21	**100.00**	**100.00**	**100.00**	90.70	98.00	98.00	**99.51**	96.00	**100.00**	100.00	99.51	92.00
C22	**99.00**	97.00	96.50	90.30	**100.00**	96.00	97.49	98.00	**99.00**	97.00	97.00	88.00
C23	96.00	**97.00**	95.07	88.90	98.00	97.00	98.98	**99.00**	96.00	**97.00**	96.06	96.00
C24	**100.00**	99.00	99.24	89.40	99.00	99.00	**99.24**	92.00	**100.00**	98.00	98.98	**100.00**
C25	**99.00**	97.00	96.76	91.40	**98.00**	**98.00**	92.78	93.00	**99.00**	99.00	97.30	91.00
C26	99.50	**100.00**	98.32	92.50	**99.00**	**99.00**	97.80	97.00	99.00	**100.00**	99.44	97.00
C27	**99.00**	97.00	97.09	90.50	**99.00**	**99.00**	98.50	97.00	**99.00**	97.00	95.63	93.00
C28	**99.00**	97.00	96.39	88.30	**100.00**	95.00	94.90	95.00	**99.00**	97.00	95.88	91.00
C29	**99.00**	96.00	93.93	90.50	**99.00**	97.00	96.68	98.00	**99.00**	97.00	95.33	97.00
C30	**100.00**	**100.00**	96.93	92.30	99.00	94.00	**99.55**	89.00	**100.00**	**100.00**	96.49	**100.00**
C31	**99.00**	98.00	96.53	91.50	**99.00**	98.00	92.45	90.00	**99.00**	**99.00**	97.03	94.00
C32	**97.50**	97.00	96.34	88.30	**99.00**	98.00	96.30	**99.00**	**97.00**	**97.00**	95.29	88.00
C33	**98.00**	96.00	93.33	86.20	**100.00**	**100.00**	98.92	97.00	**98.00**	96.00	93.85	93.00
C34	**98.00**	97.00	97.44	89.00	**100.00**	97.00	96.37	98.00	**98.00**	**98.00**	95.38	94.00
C35	**95.00**	94.00	94.44	87.00	99.00	**100.00**	96.91	96.00	95.00	95.00	94.95	**99.00**
C36	**98.50**	97.00	93.27	90.00	**99.00**	98.00	96.95	**99.00**	**98.00**	**98.00**	91.83	96.00
C37	**99.50**	95.00	92.59	86.00	**100.00**	99.00	97.55	95.00	**99.00**	95.00	92.13	89.00
C38	**99.00**	95.00	97.01	88.00	**100.00**	99.00	95.65	92.00	**99.00**	95.00	95.51	92.00
Avg (Std)	**98.66** (2.29)	98.34 **(1.80)**	97.17 (2.77)	93.20 (4.16)	**99.00 (0.79)**	98.00 (1.73)	97.14 (1.60)	94.00 (6.23)	**98.55** (2.26)	98.18 **(1.74)**	96.71 (3.21)	95.00 (3.56)

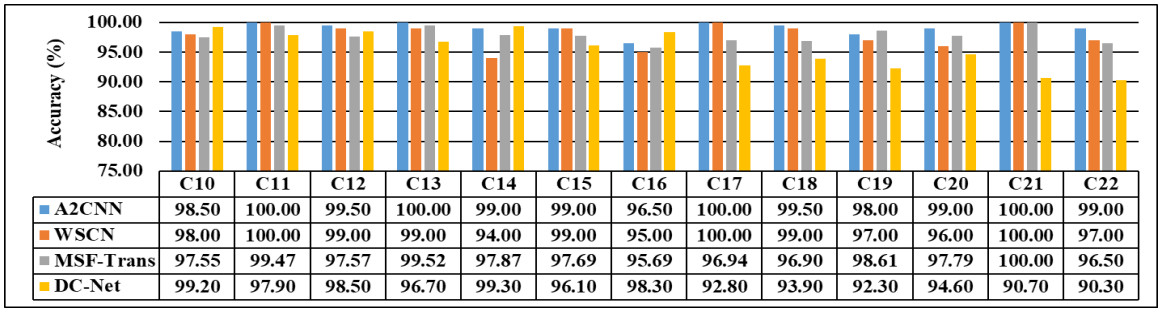
The attention module within A2CNN contributes to this success by facilitating the extraction of class-specific features and reducing irrelevant noise. Combining local information from CNN feature maps with global features in the channel and spatial dimensions results in a more comprehensive and representative feature set.

Further, the GAP computes an average representation of features, providing a more generalized and robust understanding of input features. Additionally, the loss function's focus on hard examples reduces misclassification. All this enhances the model's ability to distinguish between small defects in different classes and improves generalization.



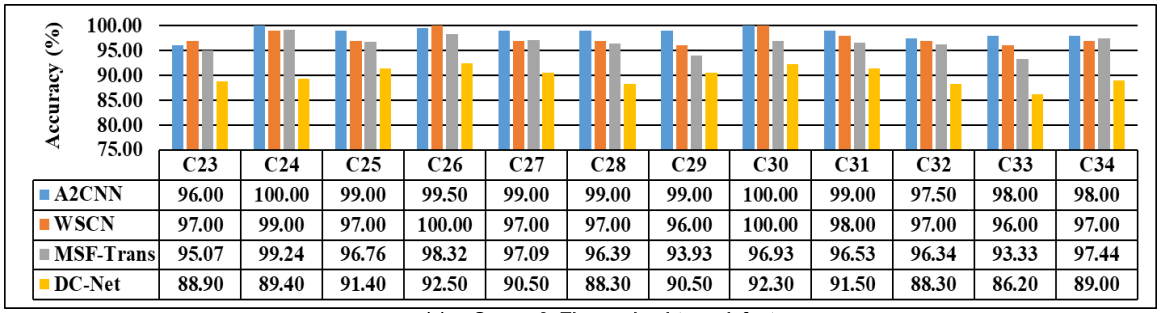
	C1	C2	C3	C4	C5	C6	C7	C8	C9
A2CNN	100.00	100.00	100.00	98.50	100.00	100.00	100.00	100.00	86.67
WSCN	100.00	100.00	100.00	97.00	99.00	99.00	100.00	99.00	98.00
MSF-Trans	99.51	98.32	99.08	95.45	99.53	99.48	97.24	98.77	84.38
DC-Net	99.70	97.80	96.50	94.40	99.80	93.80	95.80	93.40	100.00

(a) Group_1: Single type defects



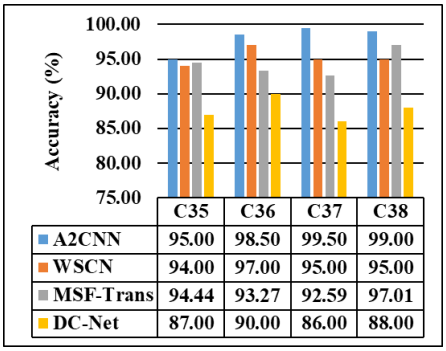
	C10	C11	C12	C13	C14	C15	C16	C17	C18	C19	C20	C21	C22
A2CNN	98.50	100.00	99.50	100.00	99.00	99.00	96.50	100.00	99.50	98.00	99.00	100.00	99.00
WSCN	98.00	100.00	99.00	99.00	94.00	99.00	95.00	100.00	99.00	97.00	96.00	100.00	97.00
MSF-Trans	97.55	99.47	97.57	99.52	97.87	97.69	95.69	96.94	96.90	98.61	97.79	100.00	96.50
DC-Net	99.20	97.90	98.50	96.70	99.30	96.10	98.30	92.80	93.90	92.30	94.60	90.70	90.30

(b) Group_2: Two mixed-type defects



	C23	C24	C25	C26	C27	C28	C29	C30	C31	C32	C33	C34
A2CNN	96.00	100.00	99.00	99.50	99.00	99.00	99.00	100.00	99.00	97.50	98.00	98.00
WSCN	97.00	99.00	97.00	100.00	97.00	97.00	96.00	100.00	98.00	97.00	96.00	97.00
MSF-Trans	95.07	99.24	96.76	98.32	97.09	96.39	93.93	96.93	96.53	96.34	93.33	97.44
DC-Net	88.90	89.40	91.40	92.50	90.50	88.30	90.50	92.30	91.50	88.30	86.20	89.00

(c) Group_3: Three mixed-type defects



	C35	C36	C37	C38
A2CNN	95.00	98.50	99.50	99.00
WSCN	94.00	97.00	95.00	95.00
MSF-Trans	94.44	93.27	92.59	97.01
DC-Net	87.00	90.00	86.00	88.00

(d) Group_4: Four mixed-type defects

FIGURE 7. The comparison of A2CNN, WSCN, MFS-Trans, and DC-Net based on the percentage of class accuracy in four defect groups: (a) Single type defects, (b) Two mixed-type defects, (c) Three mixed-type defects, and (4) Four mixed-type defects.

Please refer to the confusion matrix in Appendix for detailed class-wise results (Fig. 12).

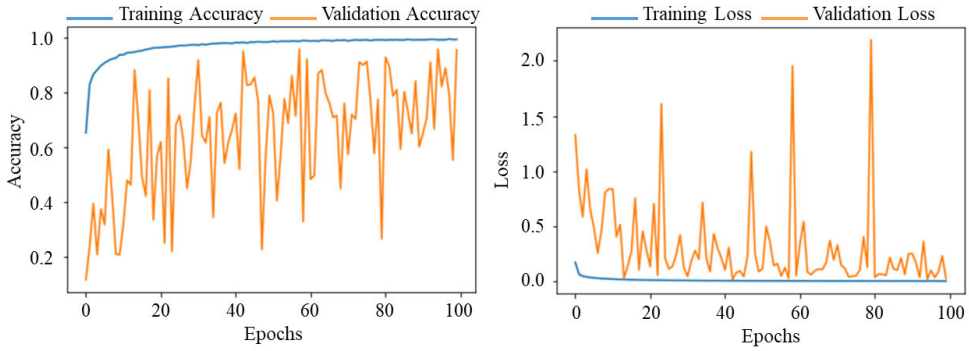
In the ablation study, removing the Attention module while keeping other parameters constant resulted in maximum validation accuracy of 95.89% and training accuracy of 98.73%. Evaluation measures on the test set; accuracy (95.33%), precision (95.74%), recall (95.08%), and F1-score (95.21%), were significantly lower compared to the model with the Attention module. Also, the overfitting was observed in the training and validation graphs as displayed in Fig. 8.

B. ANALYSIS AND DISCUSSION

The overall remarks of this work based on the obtained results are as follows:

1) THE EFFECT OF THE ATTENTION MECHANISM ON FEATURE LEARNING

To investigate the impact of attention on feature learning of complex defects, feature maps from the last convolutional layer of the attention module were extracted for defects in



(a) Training and validation accuracy (b) Training and validation loss

FIGURE 8. The ablation experiment training and validation graphs, removing the Attention module results in overfitting: (a) Accuracy graphs, (b) Loss graphs.

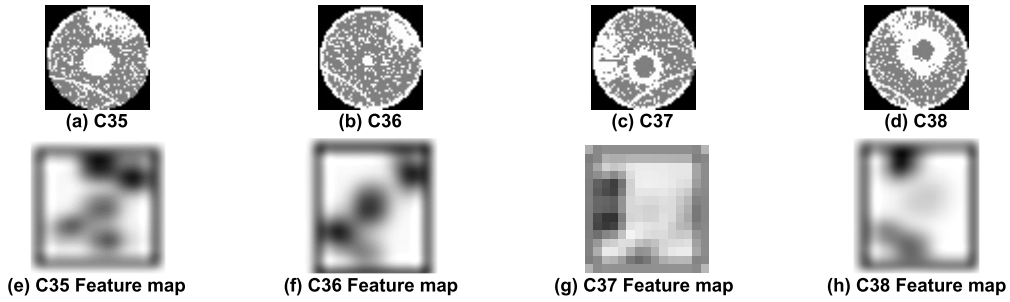


FIGURE 9. The class samples from C35-38 and corresponding feature maps: (a) -(d) C35-C38 class samples, (e)-(h) Feature maps of C35-C38.

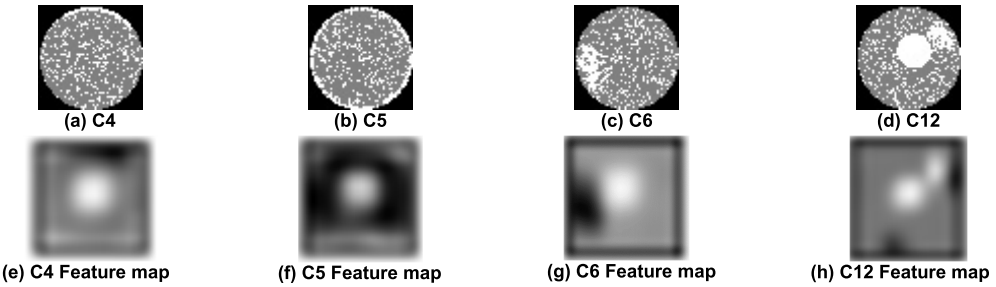


FIGURE 10. The defect types with the pattern on wafer edges: (a) C4, (b) C5, (c) C6, (d) C12, (e) C4 Feature map, (f) C5 Feature map, (g) C6 Feature map, (h) C12 Feature map.

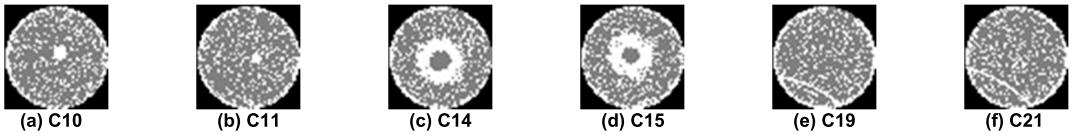


FIGURE 11. The difficult class examples: (a) C10 (C+EL) resembles (b) C11 (C+ER), (c) C14 (D+EL) resembles (d) C15 (D+ER), and (e) C19 (EL+S) resembles (f) C21 (ER+S).

Group_4. These maps demonstrate the network's sensitivity to the defect area, with varying intensities of grey indicating the strength of the response for a particular defect type. Fig. 9 (a)-(d) displays the class images of C35-38, while Fig. 9 (e)-(h) shows the corresponding feature maps. The emphasis on the defect region in the feature maps demonstrates the role of the attention mechanism in highlighting meaningful features, thus improving the model's ability to distinguish between complex defect types.

2) THE EFFECT ON DEFECT PATTERNS ON WAFER MAP EDGES

To evaluate A2CNN's performance for defect patterns on the wafer edges, some example images and feature maps

TABLE 4. The training, validation, and test accuracy and loss for 10-fold cross-validation.

	Training Accuracy	Training Loss	Validation Accuracy	Validation Loss	Test Accuracy (7)	Test Error (1-Test accuracy)
Test 1	99.66	8.00E-02	98.91	0.93	98.74	1.26
Test 2	99.58	1.10E-01	98.78	0.79	98.91	1.09
Test 3	99.53	1.30E-01	99.15	0.44	**98.96**	1.04
Test 4	99.75	6.00E-02	99.21	0.65	98.87	1.13
Test 5	99.37	1.90E-01	98.88	0.50	98.63	1.37
Test 6	99.65	9.00E-02	98.88	0.86	98.80	1.20
Test 7	99.43	1.60E-01	98.88	0.59	98.79	1.21
Test 8	99.51	1.30E-01	98.88	0.61	98.68	1.32
Test 9	99.64	1.10E-01	99.01	0.63	98.86	1.14
Test 10	99.74	6.00E-02	99.08	0.86	98.82	1.18
Avg (Std)	99.59 (0.12)	0.11 (0.04)	98.97 (0.13)	0.69 (0.16)	98.81 (0.10)	1.20 (0.10)

TABLE 5. The test performance of 10-fold cross-validation.

	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Test 1	98.52	98.66	98.29	98.53
Test 2	98.62	98.84	98.42	98.71
Test 3	98.66	99.00	98.55	98.82
Test 4	98.42	98.89	98.21	98.68
Test 5	98.27	98.55	97.97	98.37
Test 6	98.65	98.63	98.42	98.61
Test 7	98.64	98.66	98.45	98.68
Test 8	98.32	98.82	98.13	98.47
Test 9	98.11	98.92	97.97	98.39
Test 10	98.52	98.74	98.34	98.66
Avg (Std)	98.47 (0.18)	98.77 (0.14)	98.27 (0.20)	98.59 (0.14)

from the network's last layer are presented in Fig. 10. Thirty defect types have such patterns, including three in single defects (C4-C6), 11 in two mixed types (C10-C12, C14-C16, C18-C22), and all three and four mixed types (C23-C38). Since these patterns comprise a large proportion of the dataset, achieving high precision for them indicates an overall good performance. A2CNN achieved relatively higher values for the three basic patterns and, consequently, for all types that appear in combination. A2CNN achieved good recognition rates for these patterns due to the zero-padding applied in convolutional layers, which retains information at the edges.

3) THE EFFECT OF FOCAL LOSS AND GLOBAL AVERAGE POOLING

A2CNN has shown improved recognition rates for difficult class samples that resemble other classes. For instance, the defect patterns in Fig. 11 (a), (c), and (e) resemble those in Fig. 11 (b), (d), and (f), respectively. A2CNN correctly classified them because of the focal loss function, which assigns more weight to the difficult class samples.

Difficult-to-classify samples contribute most to the misclassification rate. The focal loss function made the model focus on such samples while giving less weightage to well-classified samples. Furthermore, the GAP layer improved the classification accuracy by regularizing the spatial information of the feature maps, which resulted in improved generalization. These results highlight that refinement mechanisms in the A2CNN architecture have effectively contributed to its overall performance in facilitating the recognition of complex multi-defect patterns.

4) LIMITATIONS AND POSSIBLE IMPROVEMENTS

The proposed A2CNN model is designed to classify fabrication defects using only wafer map images, providing the class label as the classification result. However, it does not provide information on the attributes of the defect, such as the number of patterns on a wafer, their size, and spatial orientation. Additionally, it does not relate the defect patterns to their cause of generation, typically done by a defect inspection system. The model is evaluated on a single dataset, and its generalizability to other datasets needs further investigation.

V. CONCLUSION

This study proposed the A2CNN model, an attention-augmented CNN, for complex multi-pattern wafer map defect classification. The model maintains classification accuracy as the defect patterns increase on the wafer surface. The architecture leverages multiple channels and spatial attention layers to focus on critical areas of the defect image and extract fine discriminative features for improved defect classification. The focal loss function further refines the results, producing good precision for difficult class samples. Additionally, a GAP layer reduces the number of trainable parameters, improving network generalization and reducing overfitting. The experimental results demonstrate that A2CNN outperforms existing works, particularly in accuracy, precision, and recall. Specifically achieving 98.66%, 99.0%, and 98.55%, respectively, compared to lower values in existing works. The proposed model exhibits high performance on easy defect categories and effectively handles complex mixed-type defects, particularly those with three and four mixed-type defect patterns. This was accomplished by accurately capturing essential class features for predicting the class label. While the A2CNN has shown promising results in detecting multi-pattern defects in a single dataset, further investigations are required to evaluate its performance on more datasets. Additionally, the current model outputs only

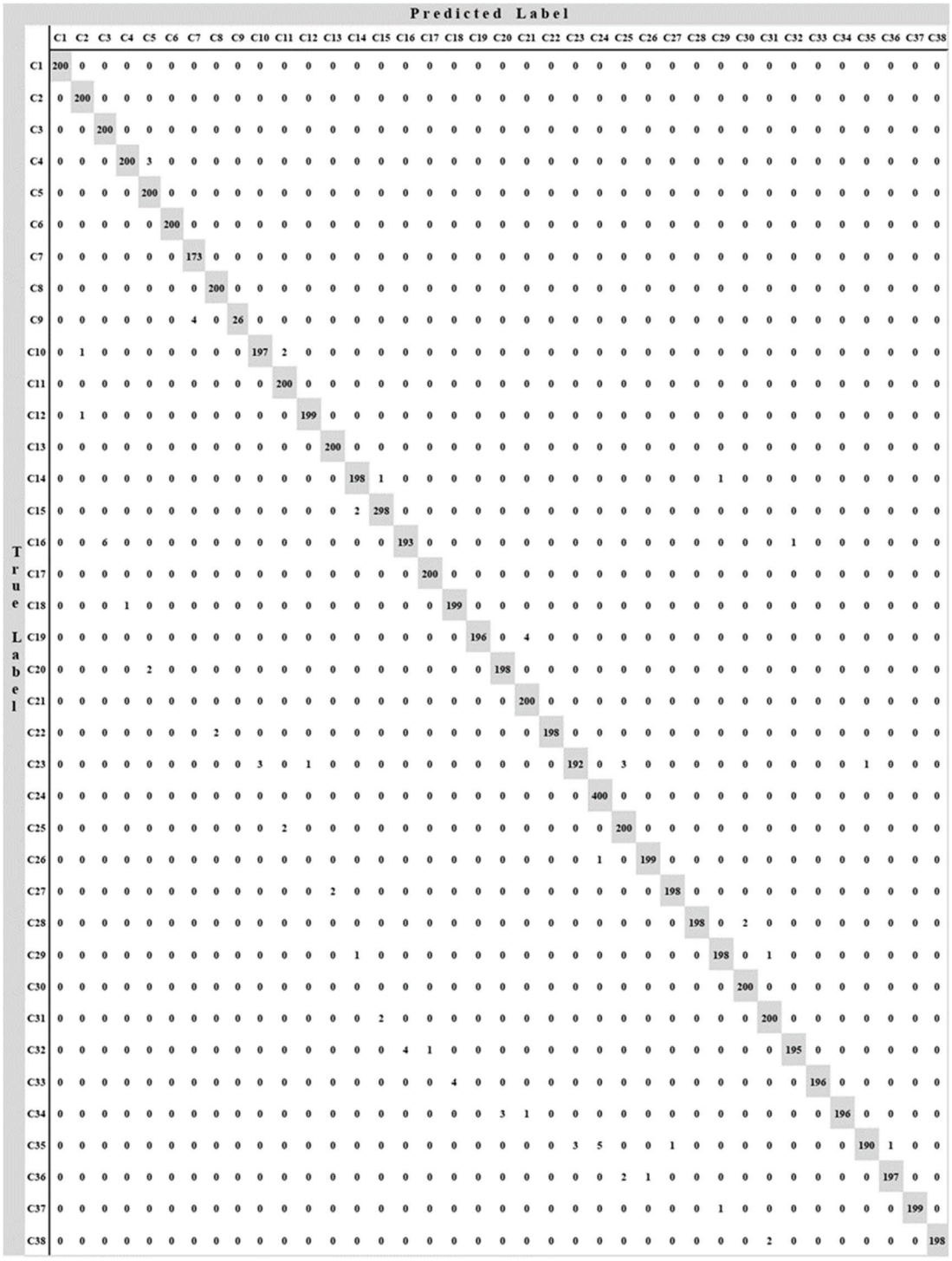


FIGURE 12. Confusion matrix of test set illustrating the number of class samples.

the defect label without further information regarding the root cause analysis, hindering its usefulness in real-world applications. Future work will focus on addressing these limitations. Efforts will be made to provide more comprehensive information with defect segmentation and faster response time for more effective root cause analysis and efficient problem-solving in real-world applications.

APPENDIX

See Table 4, Table 5, and Fig. 12.

REFERENCES

[1] L. Milor, "A survey of yield modeling and yield enhancement methods," *IEEE Trans. Semicond. Manuf.*, vol. 26, no. 2, pp. 196–213, May 2013, doi: 10.1109/TSM.2013.2243766.

[2] C. Liu and C. Chien, "An intelligent system for wafer bin map defect diagnosis: An empirical study for semiconductor manufacturing," *Eng. Appl. Artif. Intell.*, vol. 26, pp. 1479–1486, Feb. 2013, Accessed: Feb. 9, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0952197612003041

[3] L. Mönch, R. Uzsoy, and J. W. Fowler, "A survey of semiconductor supply chain models Part I: Semiconductor supply chains, strategic network design, and supply chain simulation," *Int. J. Prod. Res.*, vol. 56, no. 13, pp. 4524–4545, Jul. 2018, doi: 10.1080/00207543.2017.1401233.

[4] T. Yuan, W. Kuo, and S. J. Bae, "Detection of spatial defect patterns generated in semiconductor fabrication processes," *IEEE Trans. Semicond. Manuf.*, vol. 24, no. 3, pp. 392–403, Aug. 2011, doi: 10.1109/TSM.2011.2154870.

[5] J.-S. Kim, S.-J. Jang, T.-W. Kim, H.-J. Lee, and J.-B. Lee, "A productivity-oriented wafer map optimization using yield model based on machine learning," *IEEE Trans. Semicond. Manuf.*, vol. 32, no. 1, pp. 39–47, Feb. 2019, doi: 10.1109/TSM.2018.2870253.

[6] L. Mönch, J. W. Fowler, and S. J. Mason, *Production Planning and Control for Semiconductor Wafer Fabrication Facilities: Modeling, Analysis, and Systems* (Operations Research/Computer Science Interfaces Series), vol. 52. New York, NY, USA: Springer, 2012, doi: 10.1007/978-1-4614-4472-5.

[7] C.-Y. Hsu, W.-J. Chen, and J.-C. Chien, "Similarity matching of wafer bin maps for manufacturing intelligence to empower industry 3.5 for semiconductor manufacturing," *Comput. Ind. Eng.*, vol. 142, Apr. 2020, Art. no. 106358, doi: 10.1016/j.cie.2020.106358.

[8] S.-H. Huang and Y.-C. Pan, "Automated visual inspection in the semiconductor industry: A survey," *Comput. Ind.*, vol. 66, pp. 1–10, Jan. 2015, doi: 10.1016/j.compind.2014.10.006.

[9] X. Zheng, S. Zheng, Y. Kong, and J. Chen, "Recent advances in surface defect inspection of industrial products using deep learning techniques," *Int. J. Adv. Manuf. Technol.*, vol. 113, nos. 1–2, pp. 35–58, Mar. 2021, doi: 10.1007/s00170-021-06592-8.

[10] Z. Ren, F. Fang, N. Yan, and Y. Wu, "State of the art in defect detection based on machine vision," *Int. J. Precis. Eng. Manufacturing-Green Technol.*, vol. 9, no. 2, pp. 661–691, Mar. 2022, doi: 10.1007/s40684-021-00343-6.

[11] P. M. Bhatt, R. K. Malhan, P. Rajendran, B. C. Shah, S. Thakar, Y. J. Yoon, and S. K. Gupta, "Image-based surface defect detection using deep learning: A review," *J. Comput. Inf. Sci. Eng.*, vol. 21, no. 4, Aug. 2021, doi: 10.1115/1.4049535.

[12] J. Chu, Z. Guo, and L. Leng, "Object detection based on multi-layer convolution feature fusion and online hard example mining," *IEEE Access*, vol. 6, pp. 19959–19967, 2018, doi: 10.1109/ACCESS.2018.2815149.

[13] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," 2013, *arXiv:1311.2524*.

[14] R. Girshick, "Fast R-CNN," 2015, *arXiv:1504.08083*.

[15] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," 2015, *arXiv:1506.01497*.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," in *Proc. Eur. Conf. Comput. Vis.*, Jun. 2014, pp. 346–361, doi: 10.1007/978-3-319-10578-9_23.

[17] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," 2015, *arXiv:1506.02640*.

[18] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, and C.-Y. Fu, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, Dec. 2015, pp. 21–37, doi: 10.1007/978-3-319-46448-0_2.

[19] T. Liu, Z. He, Z. Lin, G.-Z. Cao, W. Su, and S. Xie, "An adaptive image segmentation network for surface defect detection," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Dec. 29, 2022, doi: 10.1109/TNNLS.2022.3230426.

[20] J. Zhang, R. Ding, M. Ban, and T. Guo, "FDSNET: An accurate real-time surface defect segmentation network," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.* Piscataway, NJ, USA: Institute of Electrical and Electronics Engineers, Jul. 2022, pp. 3803–3807, doi: 10.1109/ICASSP43922.2022.9747311.

[21] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," 2014, *arXiv:1411.4038*.

[22] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," 2015, *arXiv:1505.04597*.

[23] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder–decoder architecture for image segmentation," 2015, *arXiv:1511.00561*.

[24] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," 2017, *arXiv:1703.06870*.

[25] Y. Zhang, J. Chu, L. Leng, and J. Miao, "Mask-refined R-CNN: A network for refining object details in instance segmentation," *Sensors*, vol. 20, no. 4, p. 1010, Feb. 2020, doi: 10.3390/s20041010.

[26] U. Batool, M. I. Shapiai, M. Tahir, Z. H. Ismail, N. J. Zakaria, and A. Elfakharany, "A systematic review of deep learning for silicon wafer defect recognition," *IEEE Access*, vol. 9, pp. 116572–116593, 2021, doi: 10.1109/ACCESS.2021.3106171.

[27] T. Kim and K. Behdinan, "Advances in machine learning and deep learning applications towards wafer map defect recognition and classification: A review," *J. Intell. Manuf.*, vol. 34, no. 8, pp. 3215–3247, Dec. 2023, doi: 10.1007/s10845-022-01994-1.

[28] M. Saqlain, Q. Abbas, and J. Y. Lee, "A deep convolutional neural network for wafer defect identification on an imbalanced dataset in semiconductor manufacturing processes," *IEEE Trans. Semicond. Manuf.*, vol. 33, no. 3, pp. 436–444, Aug. 2020, doi: 10.1109/TSM.2020.2994357.

[29] U. Batool, M. I. Shapiai, N. Ismail, H. Fauzi, and S. Salleh, "Oversampling based on data augmentation in convolutional neural network for silicon wafer defect classification," in *Frontiers in Artificial Intelligence and Applications*. Clifton, VA, USA: IOS Press BV, Sep. 2020, pp. 3–12, doi: 10.3233/FAIA200547.

[30] P.-C. Shen and C.-Y. Lee, "Wafer bin map recognition with autoencoder-based data augmentation in semiconductor assembly process," *IEEE Trans. Semicond. Manuf.*, vol. 35, no. 2, pp. 198–209, May 2022, doi: 10.1109/TSM.2022.3146266.

[31] S. Chen, Y. Zhang, X. Hou, Y. Shang, and P. Yang, "Wafer map failure pattern recognition based on deep convolutional neural network," *Exp. Syst. Appl.*, vol. 209, Dec. 2022, Art. no. 118254, doi: 10.1016/j.eswa.2022.118254.

[32] S. Misra, D. Kim, J. Kim, W. Shin, and C. Kim, "A voting-based ensemble feature network for semiconductor wafer defect classification," *Sci. Rep.*, vol. 12, no. 1, p. 16254, Sep. 2022, doi: 10.1038/s41598-022-20630-9.

[33] S. Qi, J. Yang, and Z. Zhong, "A review on industrial surface defect detection based on deep learning technology," in *Proc. ACM Int. Conf. Proc. Ser.* New York, NY, USA: Association for Computing Machinery, Sep. 2020, pp. 24–30, doi: 10.1145/3426826.3426832.

[34] E. Elizar, M. A. Zulkifley, R. Muharar, M. H. M. Zaman, and S. M. Mustaza, "A review on multiscale-deep-learning applications," *Sensors*, vol. 22, no. 19, p. 7384, Sep. 2022, doi: 10.3390/s22197384.

[35] G. Tello, O. Y. Al-Jarrah, P. D. Yoo, Y. Al-Hammadi, S. Muhaidat, and U. Lee, "Deep-structured machine learning model for the recognition of mixed-defect patterns in semiconductor fabrication processes," *IEEE Trans. Semicond. Manuf.*, vol. 31, no. 2, pp. 315–322, May 2018, doi: 10.1109/TSM.2018.2825482.

[36] K. Kyeong and H. Kim, "Classification of mixed-type defect patterns in wafer bin maps using convolutional neural networks," *IEEE Trans. Semicond. Manuf.*, vol. 31, no. 3, pp. 395–402, Aug. 2018, doi: 10.1109/TSM.2018.2841416.

[37] Y. Kong and D. Ni, "Recognition and location of mixed-type patterns in wafer bin maps," in *Proc. IEEE Int. Conf. Smart Manuf., Ind. Logistics Eng. (SMILE)*, Apr. 2019, pp. 4–8, doi: 10.1109/SMILE45626.2019.8965309.

[38] Y. Kong and D. Ni, "Qualitative and quantitative analysis of multi-pattern wafer bin maps," *IEEE Trans. Semicond. Manuf.*, vol. 33, no. 4, pp. 578–586, Nov. 2020, doi: 10.1109/TSM.2020.3022431.

[39] Y. Byun and J. G. Baek, "Mixed pattern recognition methodology on wafer maps with pre-trained convolutional neural networks," *Proc. 12th Int. Conf. Agents Artif. Intell.*, vol. 2, 2020, pp. 974–979, doi: 10.5220/0009177909740979.

[40] H. Lee and H. Kim, "Semi-supervised multi-label learning for classification of wafer bin maps with mixed-type defect patterns," *IEEE Trans. Semicond. Manuf.*, vol. 33, no. 4, pp. 653–662, Nov. 2020, doi: 10.1109/TSM.2020.3027431.

[41] J. Zhuang, G. Mao, Y. Wang, X. Chen, and Z. Wei, "A neural-network approach to better diagnosis of defect pattern in wafer bin map," in *Proc. China Semicond. Technol. Int. Conf. (CSTIC)*, Jun. 2020, pp. 1–3, doi: 10.1109/CSTIC49141.2020.9282438.

[42] W. Shin, H. Kahng, and S. B. Kim, "Mixup-based classification of mixed-type defect patterns in wafer bin maps," *Comput. Ind. Eng.*, vol. 167, May 2022, Art. no. 107996, doi: 10.1016/j.cie.2022.107996.

[43] J. Wang, C. Xu, Z. Yang, J. Zhang, and X. Li, "Deformable convolutional networks for efficient mixed-type wafer defect pattern recognition," *IEEE Trans. Semicond. Manuf.*, vol. 33, no. 4, pp. 587–596, Nov. 2020, doi: 10.1109/TSM.2020.3020985.

[44] Y. Wei and H. Wang, "Mixed-type wafer defect recognition with multi-scale information fusion transformer," *IEEE Trans. Semicond. Manuf.*, vol. 35, no. 2, pp. 341–352, May 2022, doi: 10.1109/tsm.2022.3156583.

[45] Y. Wei and H. Wang, "Wavelet integrated attention network with multi-resolution frequency learning for mixed-type wafer defect recognition," *Eng. Appl. Artif. Intell.*, vol. 121, May 2023, Art. no. 105975, doi: 10.1016/j.engappai.2023.105975.

[46] S. Nag, D. Makwana, S. C. Teja, S. Mittal, and C. K. Mohan, "WaferSegClassNet—A light-weight network for classification and segmentation of semiconductor wafer defects," *Comput. Ind.*, vol. 142, Nov. 2022, Art. no. 103720, doi: 10.1016/j.compind.2022.103720.

[47] J. Cha and J. Jeong, "Improved U-Net with residual attention block for mixed-defect wafer maps," *Appl. Sci.*, vol. 12, no. 4, p. 2209, Feb. 2022, doi: 10.3390/app12042209.

[48] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*.

[49] S. Chaudhari, V. Mithal, G. Polatkan, and R. Ramanath, "An attentive survey of attention models," *ACM Trans. Intell. Syst. Technol.*, vol. 12, no. 5, pp. 1–32, Oct. 2021, doi: 10.1145/3465055.

[50] Z. Niu, G. Zhong, and H. Yu, "A review on the attention mechanism of deep learning," *Neurocomputing*, vol. 452, pp. 48–62, Sep. 2021, doi: 10.1016/j.neucom.2021.03.091.

[51] J. Mei, Y.-B. Zheng, and M.-M. Cheng, "D2ANet: Difference-aware attention network for multi-level change detection from satellite imagery," *Comput. Vis. Media*, vol. 9, no. 3, pp. 563–579, Sep. 2023, doi: 10.1007/s41095-022-0325-1.

[52] Q. Lin, J. Zhou, Q. Ma, Y. Ma, L. Kang, and J. Wang, "EMRA-net: A pixel-wise network fusing local and global features for tiny and low-contrast surface defect detection," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–14, 2022, doi: 10.1109/TIM.2022.3151926.

[53] V. Sampath, I. Maurtua, J. J. A. Martín, A. Rivera, J. Molina, and A. Gutierrez, "Attention-guided multitask learning for surface defect identification," *IEEE Trans. Ind. Informat.*, vol. 19, no. 9, pp. 9713–9721, Sep. 2023, doi: 10.1109/TII.2023.3234030.

[54] J. Park, S. Woo, J.-Y. Lee, and I. So Kweon, "BAM: Bottleneck attention module," 2018, *arXiv:1807.06514*.

[55] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon, "CBAM: Convolutional block attention module," 2018, *arXiv:1807.06521*.

[56] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," 2017, *arXiv:1708.02002*.

[57] J. Mukhoti, V. Kulharia, A. Sanyal, S. Golodetz, P. H. S. Torr, and P. K. Dokania, "Calibrating deep neural networks using focal loss," in *Proc. 34th Conf. Neural Inf. Process. Syst.*, 2020, pp. 1–12. [Online]. Available: https://github.com/torrvision/

[58] T. F. Romdhane, H. Alhichri, R. Ouni, and M. Atri, "Electrocardiogram heartbeat classification based on a deep convolutional neural network and focal loss," *Comput. Biol. Med.*, vol. 123, Aug. 2020, Art. no. 103866, doi: 10.1016/J.COMPBIOMED.2020.103866.

[59] M. Lin, Q. Chen, and S. Yan, "Network in network," 2023, *arXiv.1312.4400*.

[60] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. 32nd Int. Conf. Mach. Learn. (ICML)*, vol. 1, 2015, pp. 448–456.

[61] J. Bjorck, C. Gomes, B. Selman, and K. Q. Weinberger, "Understanding batch normalization," in *Proc. 32nd Conference on Neural Inf. Process. Syst. (NeurIPS)*, Montréal, QC, Canada, 2018, pp. 1–12. [Online]. Available: https://www.google.ca/search?q=Understanding+Batch+Normalization&oq=Understanding+Batch+Normalization&aqs=chrome.69i57j69i61j69i60j0.588j0j7&client=ubuntu&sourceid=chrome&ie=UTF-8

[62] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, 2015, pp. 1–15.

UZMA BATOOL received the M.S. degree in computer science from the National University of Computer and Emerging Sciences, Pakistan, in 2008, and the Ph.D. degree from the Malaysia-Japan International Institute of Technology (MJIIT), Universiti Teknologi Malaysia (UTM), in 2022. She is an active researcher focusing on pattern recognition using deep learning algorithms, data science, and machine learning.

MOHD IBRAHIM SHAPIAI (Member, IEEE) received the M.Eng. degree from the University of York, U.K., in 2007, and the Ph.D. degree in machine learning from Universiti Teknologi Malaysia, in 2013. From March 2010 to April 2010, he was a Visiting Researcher with the Graduate School of Information, Production and Systems, Waseda University, Japan, under the supervision of Dr. Junzo Watada; and the Faculty of Engineering, Leeds University, U.K., under the supervision of Dr. Vassili Toropov, from June 2012 to July 2012. He is currently a Senior Lecturer with Universiti Teknologi Malaysia and a researcher with the Center of Artificial Intelligence and Robotics (CAIRO). His current research interests include artificial intelligence, machine learning, brain–computer interface, and swarm intelligence. He has been named a Certified NVIDIA Deep Learning Instructor.

SALAMA A. MOSTAFA received the B.Sc. degree in computer science from the University of Mosul, Iraq, in 2003, and the M.Sc. and Ph.D. degrees in information and communication technology from Universiti Tenaga Nasional (UNITEN), Malaysia, in 2011 and 2016, respectively. He is currently the Head of the Center of Intelligent and Autonomous Systems (CIAS), Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia (UTHM). He has produced more than 200 papers in journals, book chapters, conferences, and tutorials. He has completed 14 industrial projects and 23 research projects. His current research interests include autonomous agents, adjustable autonomy, human–computer collaboration, machine learning, optimization, and computer vision.

MOHD ZAMRI IBRAHIM received the B.Eng. and M.Eng. degrees from Universiti Teknologi Malaysia, Malaysia, and the Ph.D. degree from Loughborough University, U.K. He is currently a Senior Lecturer with the Faculty of Electrical and Electronics Engineering, University Malaysia Pahang, Malaysia. His current research interests include computer vision, plasma science, embedded system programming, brain–computer interaction, image processing, intelligent systems, and speech recognition.