IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 26 May 2024, accepted 28 June 2024, date of publication 3 July 2024, date of current version 15 July 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3422616

RESEARCH ARTICLE

Hierarchical Attention Module-Based Hotspot Detection in Wafer Fabrication Using Convolutional Neural Network Model

MOBEEN SHAHROZ[1], MUDASIR ALI[2], ALISHBA TAHIR[1], HENRY FABIAN GONGORA[3,4,5], CARLOS UC RIOS[3,4,6], MD ABDUS SAMAD[7], (Member, IEEE), AND IMRAN ASHRAF[7], (Member, IEEE)

[1]Department of Artificial Intelligence, The Islamia University of Bahawalpur, Bahawalpur, Punjab 63100, Pakistan
[2]Department of Computer Science, The Islamia University of Bahawalpur, Bahawalpur, Punjab 63100, Pakistan
[3]Universidad Europea del Atlántico, 39011 Santander, Spain
[4]Universidad Internacional Iberoamericana, Campeche 24560, Mexico
[5]Universidad de La Romana, La Romana, Dominican Republic
[6]Universidad Internacional Iberoamericana, Arecibo, PR 00613, USA
[7]Department of Information and Communication Engineering, Yeungnam University, Gyeongsan-si 38541, Republic of Korea

Corresponding authors: Imran Ashraf (imranashraf@ynu.ac.kr) and Md Abdus Samad (masamad@yu.ac.kr)

This work was supported by European University of Atlantic, Spain.

ABSTRACT Wafer mappings (WM) help diagnose low-yield issues in semiconductor production by offering vital information about process anomalies. As integrated circuits continue to grow in complexity, doing efficient yield analyses is becoming more essential but also more difficult. Semiconductor manufacturers require constant attention to reliability and efficiency. Using the capabilities of convolutional neural network (CNN) models improved by hierarchical attention module (HAM), wafer hotspot detection is achieved throughout the fabrication process. In an effort to achieve accurate hotspot detection, this study examines a variety of model combinations, including CNN, CNN+long short-term memory (LSTM) LSTM, CNN+Autoencoder, CNN+artificial neural network (ANN), LSTM+HAM, Autoencoder+HAM, ANN+HAM, and CNN+HAM. Data augmentation strategies are utilized to enhance the model's resilience by optimizing its performance on a variety of datasets. Experimental results indicate a superior performance of 94.58% accuracy using the CNN+HAM model. K-fold cross-validation results using 3, 5, 7, and 10 folds indicate mean accuracy of 94.66%, 94.67%, 94.66%, and 94.66%, for the proposed approach, respectively. The proposed model performs better than recent existing works on wafer hotspot detection. Performance comparison with existing models further validates its robustness and performance.

INDEX TERMS Wafer hotspot detection, hierarchical attention module, autoencoder, data augmentation, hybrid attention module, deep learning, image classification, convolutional neural networks.

I. INTRODUCTION

The silicon wafers are crucial parts of the manufacturing process and are used in the creation of semiconductor devices. Lithography is used to produce the complex patterns needed for semiconductor operations. The chip's ability to work highly depends on these patterns [1]. On the other hand, a variety of manufacturing-related variables, including process parameters and ambient conditions, may cause surface defects on the wafer. Defects on the wafer surface are a serious problem, as they can drastically reduce the yield of wafer fabrication. Accurately identifying and fixing these flaws is very important for a number of reasons. In the first place, it expedites the early detection of flaws in the

The associate editor coordinating the review of this manuscript and approving it for publication was Mehrdad Saif.

production process, enabling prompt remedial action thereby improving the quality of wafers. Second, it makes it easier to change manufacturing settings to increase the effectiveness of production [2]. Thirdly, it is essential to reduce scrap rates, which minimizes wasteful material use and manufacturing expenses.

In the semiconductor industry, wafers and silicon are closely related components, yet they have diverse functions and unique qualities. A wafer is a thin, disc-shaped substrate that is mainly composed of silicon, however, it can also be formed of other materials. It serves as the manufacturing canvas for semiconductor devices [3]. Wafers are carefully cut, polished, and ready to be used in the fabrication of microchips and integrated circuits. Conversely, silicon describes the particular substance that is employed in the production of semiconductors. It is a crystalline material with superior electrical characteristics, which makes it the best option for electronic component creation. In essence, wafers serve as the structural foundation for silicon-based semiconductor devices. They are designed for certain uses and are available in a range of sizes [4]. Contrarily, silicon is a raw material that must go through a number of steps in order to be transformed into a useful semiconductor component. To provide the required electrical qualities, the silicon is grown cut, and doped by these procedures. Wafers are essentially a type of refined silicon utilized in the production of integrated circuits. The semiconductor industry relies heavily on silicon and wafers, with silicon providing the building blocks and wafers acting as the canvas for the sophisticated electronic gadgets that run our contemporary society [5].

Wafers are thin substrates that are flat and shaped like discs and are usually constructed of silicon or another semiconductor material. As the raw material used to create integrated circuits, microchips, and other semiconductor devices, they are essential to the field of electronics. These wafers are the building blocks of contemporary electronic components because they are finely constructed, polished to almost flawlessness, and available in sizes ranging from a few inches to greater diameters [6]. Wafers are manufactured by growing, slicing, and polishing semiconductor material to the appropriate thickness and surface quality. These wafers serve as a canvas on which elaborate designs are made using etching and photolithography techniques. Electronic circuits are built on these patterns, which consist of transistors, capacitors, and interconnects. Microelectronics also relies heavily on wafers; the tiniest and most sophisticated chips are made on wafers as tiny as 300 mm in diameter. Because of their extraordinary electrical qualities and capacity to house intricate integrated circuits, wafers are essential to the semiconductor industry as they allow for the miniaturization of electronic devices [7]. These wafer-based chips are the foundation of many different technologies, including computers, cell phones, medical equipment, and automobile systems. Their significance is only going to increase with the ongoing need for electronics that are increasingly compact, potent, and energy-efficient. In wafers lay the groundwork for the technical breakthroughs that have shaped our contemporary world, making them the unsung heroes of the digital age [8].

In wafer fabrication [1], hotspots refer to purposely generated localized defects or anomalies on a semiconductor wafer [2]. These hotspots are strategically created to serve various purposes, such as testing equipment sensitivity, evaluating process variations, or validating defect detection algorithms. The procedure for constructing hotspots begins with meticulous planning, where specific areas or patterns on the wafer are identified for hotspot generation [3]. A mask is then designed to outline this pattern, which guides the photolithography process to expose a photoresist layer on the wafer surface. Subsequently, etching techniques are employed to selectively remove material from the exposed areas, forming the desired hotspot pattern [6]. Despite the controlled nature of this process, hotspots are considered defective because they can compromise the integrity and functionality of semiconductor devices. Hotspots may lead to electrical shorts, reduced device performance, or complete device failure [4]. These defects can arise due to various factors, including process variations, material impurities, or incomplete etching. Consequently, wafers with hotspots are typically rejected during quality control inspections to ensure that only defect-free semiconductor devices are delivered to customers. Furthermore, hotspots can increase manufacturing costs and decrease overall yield, emphasizing the importance of minimizing their occurrence through rigorous process optimization and quality assurance measures [7].

Inadequate to satisfy the demands of contemporary industrialized goods, early identification of wafer hotspots is frequently carried out manually by skilled inspectors. This approach has drawbacks, including low efficiency, poor accuracy, high expense, and strong subjectivity. In the realm of wafer inspection, machine vision-based hotspot detection techniques have currently supplanted manual inspection [9]. Deep learning vision-based fault detection techniques frequently rely on laborious, human feature extraction. The shortcomings of feature representation and extraction, data preprocessing, and model learning procedures have been overcome by the development of computer vision-based detection techniques, particularly the introduction of neural networks like convolutional neural networks. With their quick development and widespread use in the field of hotspot detection in semiconductor wafers, neural networks have gained popularity due to their high efficiency, accuracy, low cost, and great objectivity [10]. In this regard, the following contributions are made

- This study proposes a novel approach for wafer hotspot detection that utilizes a hierarchical attention module (HAM) combined with a customized convolutional neural network (CNN). The HAM is introduced to enhance detection accuracy. In addition, data augmentation is also used to improve the resilience and robustness of the proposed approach.

- Deep learning models are also used in this study such as convolutional neural networks (CNN), artificial neural networks (ANN), and long short-term memory (LSTM). In addition, the combination of deep learning models are also utilized like CNN+LSTM, CNN+Autoencoder, CNN+ANN, LSTM+HAM, and Autoencoder+HAM. The approach enhances the training dataset's richness and generalization across real-world wafer images.
- The dataset consists of nine classes of wafer Map images to capture spatial and temporal dependencies. Experiments involve extensive investigation of stand-alone and combined models, k-fold cross-validation, and performance comparison with previous studies. The proposed CNN+HAM model outperforms other deep learning models adopted in experiments to enhance semiconductor manufacturing quality control processes with enhanced robustness and accuracy.

This study is separated into several subsequent sections: A survey of the literature on recent research and technological advancements is included in Section II. Section III entails providing an overview of the study strategy, including the data gathering strategies, data analysis approaches, and proposed approach for hotspot detection. Results and a discussion of the suggested approach are presented in Section IV. The conclusion in Section V talks about the conclusion and future research directions.

II. RELATED WORK

A fabrication technique that makes it possible to produce ultralow-loss, high-confinement, anomalous-GVD Si3N4 PIC at high yield and repeatable wafer scale. When compared to previously published Si3N4 manufacturing procedures, this optimized process which uses standard CMOS foundry techniques has several benefits [11]. One meter-long spiral waveguide with 2.4 dB m1 loss may be produced using dies that are only 5 mm2 in size because of the process's high yield and ability to work over broad regions [12]. The study reveals that the inherent absorption-limited Q factor of Si3N4 microresonators can exceed 2 × 108, highlighting the importance of cleaning in microelectronic integrated circuit production. The whole cleaning procedure took place in a spotless room. During cleaning, ultrasonic agitation mainly eliminates particles [13]. The technique for measuring and identifying fractures along wafer edges using dark-field infrared scattering imaging is used, which allows edge cracks to be detected at the micron scale. The study evaluates new technology expenses and costs for next-generation manufacturing facilities with annual module power capabilities ranging from a few hundred megawatts to one gigawatt [14]. Certain vocabulary is related to layout hotspot identification. Designed layout patterns are transferred onto silicon wafers using a highly variable lithographic technique. This study introduces a high-dimensional feature extraction approach using machine learning and convolutional neural networks, and a biassed learning technique to improve hotspot identification accuracy and reduce false alarm penalties [15]. The goals of the hotspot identification process are to minimize runtime, prevent inaccurate predictions for non-hotspot clips, and find as many actual hotspots as feasible. The author's proposed model achieves an average accuracy of 94.5% with a false-alarm count of 33155.4, which is higher than previous efforts [16].

For the purpose of semiconductor process monitoring, multivariate sensor data were evaluated using CNN-based fault classification and fault diagnosis. The convolutional layer of the conventional CNN was changed in the suggested FDCCNN model to take into account the structural properties of the data, which improved classification performance and training speed [17]. Sensor data with varying lengths is produced by semiconductor fabrication procedures.CNNs and a self-attention mechanism are used in the author's suggested model to encode variable-length signals into fixed-size vectors [18]. Resolution enhancement methods like SRAFs, co-optimization, and OPC improve layout pattern printability. Lithography simulation time is long, so creating high-accuracy hotspot identification algorithms is crucial to reduce turnaround time [19].

Identifying hotspots for lithography at the IC design stage is a crucial step in obtaining high yields at modern technological nodes. to suggest a CNN architecture that does not need pooling and works well with post-OPC mask pictures [20]. In EDSE, layout patterns are generated randomly yet realistically by utilizing just the fundamental design guidelines provided by the PDK. There are currently commercial CAD tools available that can do this work [21]. Based on 10%-50% of training data, the author's experimental results show that these approaches can produce 2.9-4.5% greater accuracy at the same false alarm levels as the state-of-the-art work [22], [23], [24]. Wafer map failure pattern recognition (WMFPR) has been the subject to research analysis, however, the majority of researches employed raw wafer maps as the input data for their classification algorithms [25]. The success of the suggested WMFPR and WMSR also depends on this decreased representation. WMFPR's accuracy for the test set (118 595 wafer mappings) was 94.63% [26].

The influence of lithography hotspots on manufacturing yield is significant. It is now a crucial issue to identify these prohibited pattern topologies at the early physical design or physical confirmation stages. Provide a very efficient hotspot identification method based on the study's PCA SVM classifier. Using the Propsed technique, which correctly identifies more than 80% of the hotspots on all testing layouts, maximizes accuracy while minimizing false alarms [27]. High accuracy and data learning algorithms are made possible by pattern matching techniques, which also offer great adaptability to new lithographic regulations and procedures [28]. Three- and two-tier functioning 3-D circuits were exhibited, along with the development of the 3-D technology and design guidelines [29]. The method allows the creation of 3-D circuits using various technologies and materials, as well as the unlimited placement of dense-vertical connections

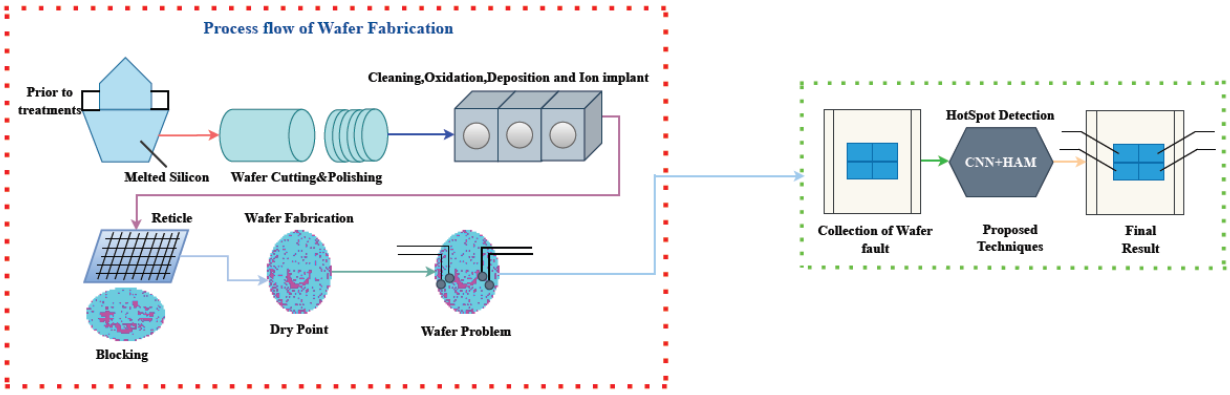

FIGURE 1. Workflow of hotspot detection using proposed approach in wafer manufacturing process.

between levels. For image applications, where a device with intricate pixel circuitry can achieve a 100% fill factor, the technology's benefits are clear [30]. The Defense Advanced Research Projects Agency (DARPA)-sponsored multichip programmer of today focuses on applying 3-D technology to mixed-signal applications [31]. The circuit model, consisting of lumped circuit RLCG parts, was used for eye-diagram and TDR/TDT simulations, showing slower signal rise time but maintaining signal integrity [32].

Four thousand polycrystalline silicon solar cells were analyzed in order to evaluate photovoltaic (PV) microcracks [33]. The cracks were inspected using electron microscopy, which made it easier to find the cracks by capturing images from the Back Scratched Electron Diffraction (BSED) and Everhart-Thornley Detector (ETD) [34]. It was discovered that the microcracks ranged in size from 50 mm to a maximum of 4 mm [35]. A new method for extracting features via layer patterns, which have far less discriminative information than metal layer patterns, is provided by the deep layout metric learning mechanism [36]. Additionally, a new via-layer benchmark suite has been employed for thorough verification in order to assess the actual performance of hotspot detectors [37]. The authors analyzed hotspot stress on two GG modules, revealing high temperatures leading to cell failure, fractures, and broken fingers in the restressed module, causing a significant 8.2% Pmax loss [38].

The behavior of substantial damage has been analyzed under the surface on a 300 mm diameter, 6 μm thick monocrystalline silicon wafer caused by ultrafine dry polishing. Dry polishing improves the strength and homogeneity of the ultra-thin wafer by eliminating micro-cracks and high-pressure phases. Polishing merely created tension in the upper, thinner surface [39]. Dry polishing does, nonetheless, also increase the risk of electrical dependability. In the meanwhile, the results shed light on the dependability of sophisticated electronic packaging [40]. Wing-out wafer-level package technology holds potential in radar, communication test, and measurement applications due to its unique blend of technological and business benefits. It can meet RF standards with FOWLP, offering cost and performance options, and co-designing the chip and package [41].

A machine learning-based, non-invasive technique for diagnosing faults in PV modules was put forth. Through the automated identification and precise categorization of non-uniformity in PV modules, the suggested technique improves the intelligence of defect diagnosis. The author's suggested methods have a low computing cost and a 94.10% accuracy rate [42]. The Naive Bayes (nBayes) classifier, a machine learning system, is trained to identify the categorized hotspots. A 42.24-kWp PV system is used for the experimental findings, which show that a mean identification rate of about 94.1% is obtained for the set of 375 samples [43].

A. GAP ANALYSIS

Developments in photovoltaic and semiconductor technologies have enhanced system dependability and problem identification. Edge fracture identification, signal-to-noise ratios, and pattern printability for nanoscale circuits are being improved by high SNR micro-crack detection, sophisticated optical fault detection techniques, machine learning models like CNNs, and resolution improvement technologies in lithography. Deep learning techniques in photovoltaics have demonstrated fault detection accuracy of around 94% to 94.5%; additional efficiency improvements are anticipated with quantum computing. High-performance optical systems are made possible by new manufacturing processes for Si3N4 photonic integrated circuits, while solar energy dependability is increased by knowledge of microcracks in polycrystalline silicon solar cells. In order to improve accuracy and computational efficiency for hotspot identification in wafer production, this research presents a hierarchical attention module combined with a convolutional neural network (CNN). This strategy outperforms conventional techniques and current deep learning models by enabling the machine to concentrate on important characteristics at various sizes. This method exhibits practical applicability for real-time industrial applications by minimizing false positives and negatives.

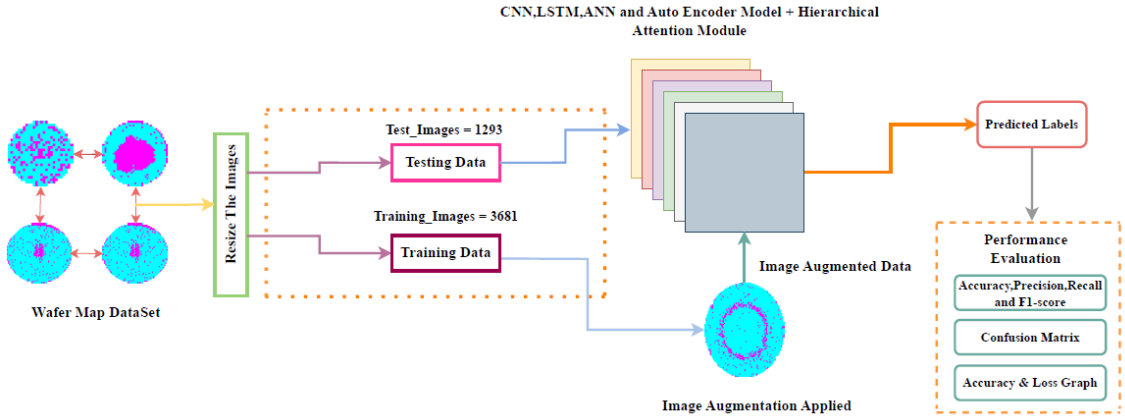

FIGURE 2. Proposed methodological architecture for wafer hotspot detection using deep learning.

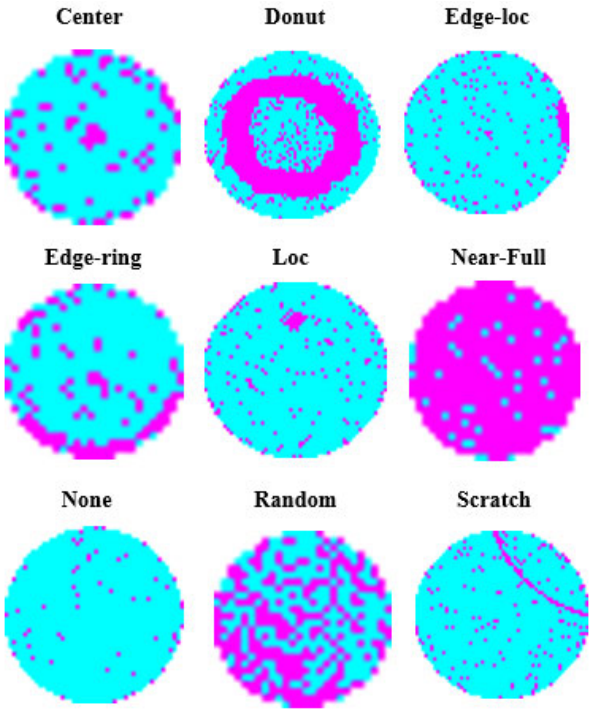

FIGURE 3. Typical wafer map defect patterns in the wafer map dataset.

The results demonstrate the promise of cutting-edge machine learning approaches for enhancing production procedures while also setting a new standard in the semiconductor manufacturing industry. This research provides the groundwork for further research and technical developments in the field.

III. METHODOLOGY

This study proposes a novel approach for hotspot detection in wafers comprising CNN and HAM. The workflow of the proposed approach is presented in Figure 1. In the proposed strategy, initially, it is necessary to ensure that the wafer map pictures and defect image data from the analysis are included in the imported datasets. Second, image augmentation is used to improve model performance and enhance datasets. CNN, ANN, LSTM, and Autoencoders are deep learning models that have been developed. Task-dependent feature selectors have not been explicitly created. Given the high feature studies at each layer in deep learning models, following the completion of the model compilation, the model is achieved by the use of 3681 wafer map images, and 1293 are trained through model testing. In this dataset, multi-classification has been applied.

TABLE 1. Statistics of dataset.

Classes	Training Image	Testing Image
Scratch	409	150
Center	409	165
Edge-ring	409	165
Donut	409	139
Random	409	150
Edge-loc	409	165
None	409	140
Near-Full	409	54
Loc	409	165
Total	3681	1293

To enhance the model's functionality even further, a HAM is incorporated. This approach improves the accuracy and efficacy of the defect identification process by allowing the model to focus on certain regions or characteristics within the wafer map images. Research is well-equipped to precisely identify faults on semiconductor wafers with the combination of models, making a valuable contribution to the progress of quality control in semiconductor production. Model evolution and outcomes are carried out once the model has been trained using wafer map images. The results and testing of the suggested evolution model are carried out. The model is now prepared to extract the features from the wafer map. The architecture of the proposed approach is illustrated in Figure 2.

A. DATASET DESCRIPTION

The wafer map (WM) datasets used in this study are freely accessible on Kaggle. There are nine classes in it. The distribution of the dataset is balanced in terms of the number of categories. 'Center', 'Donu't, 'Edge-loc', 'Edge-ring',

'Loc', 'Near-Full', 'None', 'Random', and 'Scratch' are the names of the classes. We utilized the research dataset, which was closely examined, to identify hotspots. Process problems and wafer map patterns from the CP yield wafer acceptance test might occur in semiconductors. Particles assist engineers in finding clues. However, there is a problem with categorizing the wafer map patterns without an operator's help. This topic has been the subject of other studies, and this study presents the outcomes of applying deep learning. Examples of common patterns are shown in Figure 3, and each one contains information about a distinct process failure. The dataset's statistic is displayed in Table 1.

B. IMAGE AUGMENTATION

The primary goal of image augmentation is to artificially expand the size of the dataset using a variety of transformations to the original images. By exposing the model to several iterations of the source data during training, image augmentation aids the neural network learn invariant features and reduces overfitting. For instance, rotating an image can help the model recognize objects from different orientations, and flipping horizontally can simulate variations in the scene. Through these diverse augmentations, the model becomes more resilient to variations in real-world data, leading to improved performance when faced with unseen or distorted images during inference. Ultimately, the incorporation of image augmentation techniques in hotspot detection significantly contributes to the optimization of semiconductor manufacturing processes, resulting in improved product quality and higher yields.

C. DEEP LEARNING MODELS

Deep learning has a transformative influence on the area of wafer hotspot detection within semiconductor manufacturing. Utilizing advanced techniques, including CNN, ANN, LSTM, and autoencoder, has revolutionized the ability to precisely identify hotspots on semiconductor wafers. Deep learning models excel in their capacity to scrutinize intricate patterns and defects within wafer images, making them a highly effective asset in the realm of quality control. The capability to acquire knowledge and adapt from extensive datasets significantly enhances the precision and efficiency of hotspot detection, thereby leading to substantial enhancements in manufacturing processes and the overall quality of products. This technology has evolved into an indispensable asset for semiconductor fabrication, as it ensures the early detection of potential issues, resulting in increased yields and the production of defect-free products.

A CNN is a class of deep neural networks specifically designed for image recognition and processing tasks. Its architecture is inspired by the human visual system, employing layers of learnable filters or kernels to extract hierarchical features from input images. Filters go across the input images in the convolutional layers, convolving them to identify patterns like edges and textures. Pooling layers down sample feature maps and helps to reduce complexity. To create predictions, the fully linked layers merge the acquired characteristics. This hierarchical and localized feature learning enables CNNs to achieve remarkable accuracy in computer vision applications. This technique exploited the CNN model and had wafer hotspot detection. The output of a convolutional layer with ReLU activation is given by

$$C(i,j) = \text{ReLU}\left(\sum_m \sum_n I(i+m, j+n) \cdot F(m,n) + b\right) \tag{1}$$

An ANN is a computational model that is used to do tasks like pattern recognition, classification, and regression. It is inspired by the structure and operation of the human brain. A Wafer hotspot detection has been employed in this component. An ANN is made of the input layer, hidden and output layer with linked nodes arranged into layers. The first layer of the system receives the data, which is then processed and modified as it moves through the hidden levels using a sequence of weighted connections and activation functions. The weights are iteratively adjusted during training, guided by a learning algorithm, to reduce the discrepancy between the real and anticipated outputs. The final layer produces the network's output. The output of a fully connected layer with activation is given by

$$y = \text{Activation}\left(\sum_{i=1}^{n} w_i \cdot x_i + b\right) \tag{2}$$

where y is the output, $cdot$ is the activation, w_i are the weights, b is the bias term and n is the number of input features.

LSTM was created to solve the vanishing gradient issue and make it possible to learn long-term relationships in sequential data more successfully. Time series prediction and natural language processing are two applications where LSTMs excel. Unlike conventional RNNs, LSTMs have gating mechanisms and memory cells that enable them to selectively remember or forget information over long sequences. LSTM has a memory cell, input gate, forget gate, and output gate. It stores information over time, regulates information flow, and learns to adapt its gate parameters through backpropagation. This architecture makes LSTMs effective in handling long-term dependencies, addressing the limitations of standard RNNs, and making them powerful tools for sequential information processing. The LSTM model has been deployed in this component to locate wafer hotspots.

A kind of ANN type called an autoencoder is made for unsupervised learning. and dimensionality reduction. The fundamental idea behind an auto-encoder is to encode input data into a lower-dimensional representation and then reconstruct the input from this encoded representation. During training, the network learns to minimize the difference between the input and the reconstructed output. The bottleneck layer in the middle of the auto-encoder serves as a compressed representation of the input data. Autoencoders find applications in tasks such as data denoising, anomaly

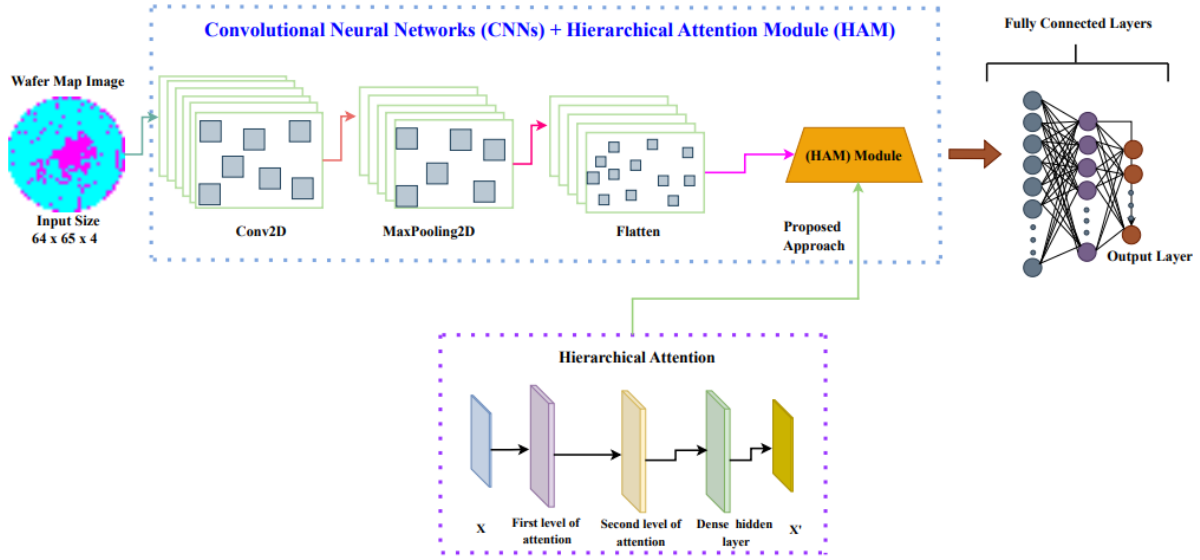

FIGURE 4. Hybridization of proposed model based on CNN and HAM.

TABLE 2. Hyperparameters of all the models used in this research.

Models	Layers	Kernal_size	Activation Function	Epoch
CNN	8	(4,4)	relu,softmax	30
CNN+ANN	8	(3,3)	relu,softmax	30
CNN+LSTM	11	(3,3)	relu,softmax	30
CNN+Auto encoder	9	(3,3)	relu,softmax,sigmid	30
CNN+HAM	13	(4,4)	relu,softmax	30
ANN+HAM	14	(4,4)	relu,softmax	30
LSTM+HAM	9	-	softmax	30
Autoencoder+ HAM	17	-	relu,sigmid,softmax	30

detection, and feature learning. By forcing the model to capture essential features in a reduced-dimensional space, autoencoders can effectively learn meaningful representations of complex data, making them valuable tools in various domains of machine learning. This component uses the auto-encoder approach to implement wafer hotspot detection. Autoencoder model, the reconstruction of the input $\hat{x}\hat{f}$rom the encoded representation z can be represented by the following equation:

$$\hat{X} = \text{Decoder}(\text{Encoder}(X)) \tag{3}$$

While deep learning models need large amounts of data for training, data gathering, and labeling is not an easy task. Recent efforts on surrogate modeling can potentially help in this regard. For example, [44] combined surrogate modeling and physics-informed ML to build a neural network that can perform better even while small datasets are used. Physic-based prior information can be incorporated without calculating derivates in the loss function. Similarly, the concerns raised about the computational complexity of deep learning models can be mitigated using the Gaussian process (GP), as evaluated by [45]. The study designed a GP-based framework to quantify prediction-related uncertainties, thereby reducing additional evaluation of deep learning models. Another similar endeavor is [46] where GP is combined with an autoencoder for multi-task spatiotemporal regression to reduce computational complexity.

1) HIERARCHICAL ATTENTION MODULE

The HAM is a sophisticated neural network architecture designed to enhance the model's ability to focus on relevant information at different levels of abstraction within a given input sequence. Operating on the principle of hierarchical attention mechanisms, HAM integrates multiple layers of attention to selectively weigh the importance of various parts of the input data. This enables the model to discern and prioritize features across different hierarchical levels, capturing both local and global dependencies. The working principle of HAM involves the application of attention mechanisms at multiple hierarchical levels, allowing the model to give various input sequence parts differing levels of priority. By dynamically adjusting attention weights, the module effectively attends to salient features and suppresses irrelevant information, facilitating improved performance in tasks such as natural language processing, image recognition, and sequence generation. This component uses a variety of models, including CNN, ANN, LSTM, and autoencoder, in conjunction with the HAM module model to identify wafer hotspots. The architecture of CNN and HAM models are shown in Figure 4 and the hyperparameters of all models are

given in Table 2. The 'categorical cross entropy' loss is used for all models. Similarly, the 'Adam' optimizer is used along with the 'accuracy' as the evaluation metric.

2) EVALUATION PARAMETERS

In wafer hotspot detection, the assessment of detection algorithm effectiveness relies heavily on key evaluation parameters. Accuracy, precision, recall, the F1 score, and the confusion matrix are integral to measuring the performance of these algorithms. Accuracy serves as an indicator of overall correctness and is calculated using correct predictions (true positive (TP), and true negative (TN)), and total predictions. Precision plays a crucial role in minimizing false alarms, assessing the proportion of true hotspots among the predicted ones through the consideration of TP and false positives (FP). On the flip side, recall evaluates the algorithm's proficiency in identifying actual hotspots using TP and false negatives (FN). The F1 score considers precision and recall, providing a comprehensive measure of the algorithm's overall effectiveness. To gain deeper insights into strengths and weaknesses, the confusion matrix breaks down TP, TN, FP, and FN. Altogether, these evaluation parameters collaboratively contribute to refining the accuracy and reliability of wafer hotspot detection systems. Here are the equations and formulas for accuracy, precision, recall, and F1 score.

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{4}$$

$$Precision = \frac{(TP)}{(TP + FP)} \tag{5}$$

$$Recall = \frac{(TP)}{(TP + FP)} \tag{6}$$

$$F1\ score = \frac{2 * (Precision * Recall)}{(Precision + Recall)} \tag{7}$$

IV. RESULTS AND DISCUSSION

Deep learning models' capacity to precisely analyze and interpret data in order to provide predictions determines how well they work. Several criteria are used to assess this performance, including recall, accuracy, precision, and the F1 score.

Convolutional Neural Networks: CNNs demonstrate impressive effectiveness in addressing wafer map hotspot classification tasks, employing evaluation metrics such as accuracy, precision, recall, and the F1 score. These networks acquire insights into patterns and features from wafer map datasets and are evaluated using separate test datasets. Precision computes the percentage of properly predicted positive cases, whereas accuracy evaluates the ratio of correctly categorized instances to the overall sample size. Accuracy, precision, recall, and F1 scores are increased in CNNs by optimizing their weights and biases. CNNs are well known for their effectiveness and may be used for a wide range of classification tasks, especially when dealing with picture data. The CNN algorithm showed

TABLE 3. Performance of convolutional neural network.

Epochs	Accuracy	Precision	Recall	F1-Score
5	0.669	0.7145	0.669	0.684
10	0.8484	0.8532	0.8484	0.8475
15	0.8716	0.8735	0.8716	0.8714
20	0.819	0.8291	0.819	0.8131
25	0.8948	0.8954	0.8948	0.8945
30	0.8894	0.9011	0.8894	0.8894

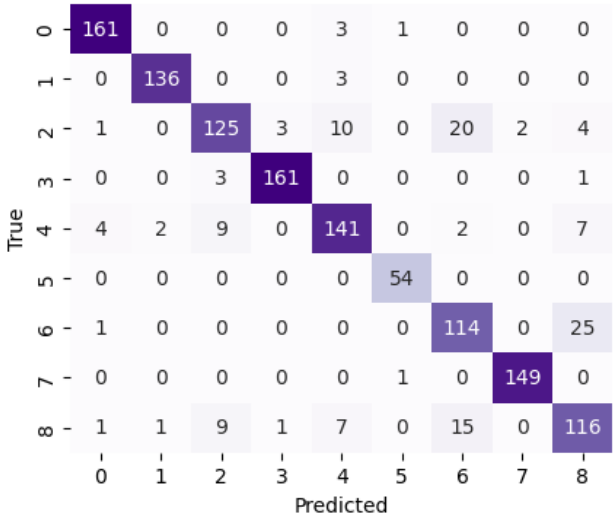

FIGURE 5. Confusion matrix for CNN model.

impressive performance recall is 89.48%, accuracy is 89.48%, precision is 89.54%, and F1 score is 89.45%. As seen in Table 3, these findings demonstrate the model's capacity to correctly forecast outcomes and successfully capture positive examples.

Figure 5 presents the CNN model's predictions using testing data in the form of the confusion matrix. The confusion matrix shows results in a cross-section of predicted and actual classes such as 0 to 8. The X-axis shows predicted results and the y-axis shows actual classes. The diagonal of the confusion matrix shows the true prediction performed by such as 0 classes in true classes and predicted classes show 1157 correct predictions out of 1293 while 136 predictions are incorrect.

A. HYBRID MODEL OF CNN AND LSTM

CNNs are highly effective in classifying wafer map hotspots. A novel approach combines CNN with LSTM to capture spatial patterns and temporal dependencies in wafer map hotspot data. The CNN component excels in feature extraction, while the LSTM model models sequential data. The integrated model, combining two powerful neural network architectures, enhances classification accuracy, precision, recall, and F1 scores in wafer map data, making it promising for various spatial and temporal classification scenarios. The CNN+LSTM algorithm performed admirably, with an accuracy of 90.02%, a precision of 90.53%, a recall of 90.02%, and an F1 score of 90.12%. These results show that the model can make accurate predictions, as shown in Table 4.

TABLE 4. Result for hybrid CNN+LSTM model.

Epochs	Accuracy	Precision	Recall	F1-Score
5	0.5329	0.557	0.5329	0.5239
10	0.7022	0.705	0.7022	0.6992
15	0.8112	0.8242	0.8112	0.8146
20	0.8484	0.853	0.8484	0.8495
25	0.8995	0.9014	0.8995	0.8992
30	0.9002	0.9053	0.9002	0.9012

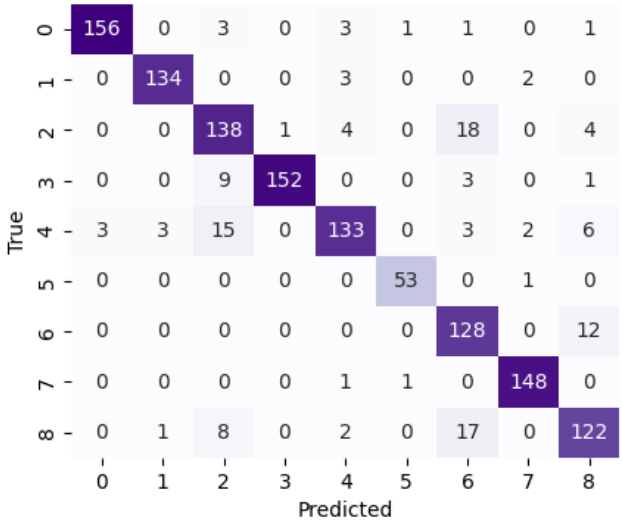

FIGURE 6. Confusion matrix for CNN+LSTM model.

TABLE 5. Result for CNN+Autoencoder model.

Epochs	Accuracy	Precision	Recall	F1-Score
5	0.6156	0.6579	0.6156	0.6111
10	0.7703	0.7762	0.7703	0.7645
15	0.8051	0.8125	0.8051	0.8019
20	0.7881	0.7908	0.7881	0.7865
25	0.8917	0.8941	0.8917	0.8919
30	0.9157	0.9171	0.9157	0.9156

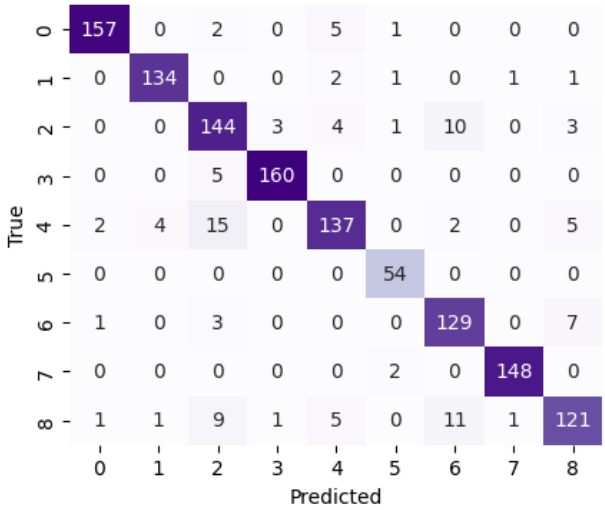

FIGURE 7. Confusion matrix for CNN+Autoencoder model.

The confusion matrix for the CNN+LSTM model is shown in Figure 6. Results are displayed in the confusion matrix as a cross-section of actual and anticipated classes, ranging from 0 to 8. Results show that the CNN+LSTM model has 129 wrong predictions out of 1293 total predictions while 1164 predictions are correct. These numbers are better than only using the CNN model for prediction.

B. RESULTS FOR HYBRID CNN+AUTOENCODER MODEL

CNN has indeed showcased remarkable effectiveness when it comes to addressing wafer map hotspot classification tasks. To further enhance the capabilities of these models, a cutting-edge approach involves the incorporation of CNNs with autoencoders. This hybrid model combines the strengths of CNNs in feature extraction from wafer map images with the unsupervised learning abilities of autoencoders. Autoencoders are particularly adept at capturing the underlying structure and reducing the dimensionality of complex data. When merged with CNNs, this integrated model becomes a powerful tool for hotspot classification tasks. By jointly optimizing their weights and biases, the integrated CNNs and Auto encoder's model achieves heightened levels of accuracy, precision, recall, and F1 scores in wafer map hotspot classification. With an accuracy of 91.57%, precision of 91.71%, recall of 91.57%, and F1 score of 91.56%, the CNN+Auto encoder algorithm performed wonderfully. As seen in Table 5, these findings demonstrate the model's ability to correctly collect positive instances and forecast outcomes.

A confusion matrix representing all of the predictions produced by the CNN+Auto encoder model using testing data is displayed in Figure 7. The confusion matrix shows the results as a cross-section of the expected and actual classes, with values ranging from 0 to 8. The predicted results are shown on the x-axis, while the actual classes are shown on the y-axis. On the diagonal of the confusion matrix, for instance, is the true prediction provided by each class. Results indicate a total of 1184 correct predictions which is higher than CNN alone, as well as, a hybrid of CNN with LSTM. The hybrid CNN+Autoencoder makes only 109 wrong predictions out of 1293 total predictions.

C. RESULTS FOR CNN+ANN MODEL

In fact, CNNs have proven to be remarkably successful in wafer map hotspot classification tasks. In a strategy to advance these models' capabilities, CNNs and ANNs are combined in a forward-looking manner. The combined strengths of ANNs, which provide a wider range of data processing capabilities, and CNNs, which excel at extracting spatial features from pictures, are utilized in this integrated model. The model gets a thorough grasp of the underlying patterns and relationships in the wafer map data by integrating several neural network architectures. The CNN and ANN components are taken into consideration throughout the

TABLE 6. Result for CNN+ANN model.

Epochs	Accuracy	Precision	Recall	F1-Score
5	0.7417	0.7989	0.7417	0.7359
10	0.8059	0.8274	0.8059	0.8108
15	0.9002	0.9057	0.9002	0.9012
20	0.9365	0.9367	0.9365	0.9364
25	0.9041	0.9089	0.9041	0.9043
30	0.928	0.932	0.928	0.9285

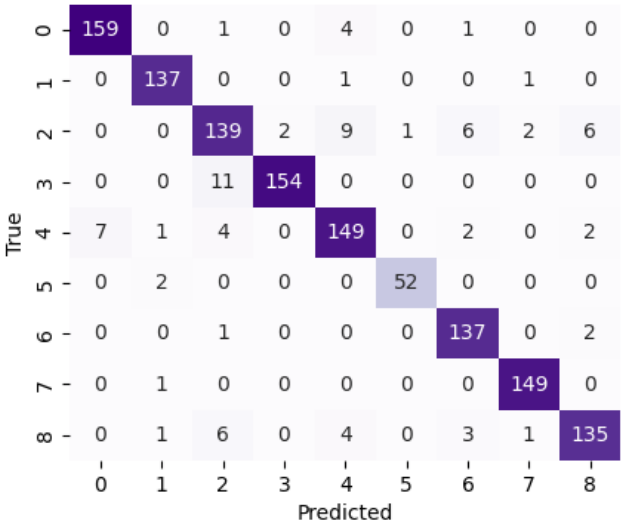

FIGURE 8. Confusion matrix for CNN+ANN model.

TABLE 7. Result for LSTM+HAM model.

Epochs	Accuracy	Precision	Recall	F1-Score
5	0.5282	0.5322	0.5282	0.514
10	0.6744	0.6882	0.6744	0.6776
15	0.7687	0.7744	0.7687	0.7577
20	0.8043	0.8149	0.8043	0.7991
25	0.8522	0.8549	0.8522	0.852
30	0.8917	0.8943	0.8917	0.8919

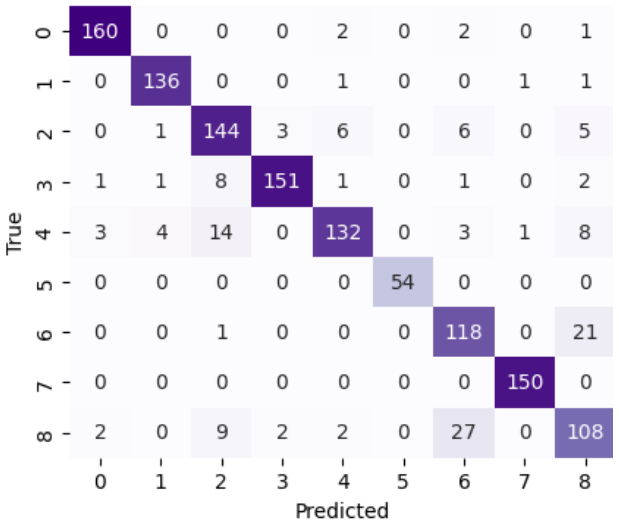

FIGURE 9. Confusion matrix for LSTM+HAM model.

iterative optimization process, which results in an integrated model that achieves high levels of recall, accuracy, precision, and F1 scores in wafer map hotspot classification tasks. The CNN+ANN algorithm has excellent performance, with an accuracy of 93.65%, precision of 93.67%, recall of 93.65%, and F1 score of 93.64%. These results show that the model can accurately gather positive examples and predict outcomes, as shown in Table 6.

Figure 8 shows the confusion matrix that summarises all of the predictions made by the CNN+ANN model using testing data. With values ranging from 0 to 8, the confusion matrix displays the findings as a cross-section of the actual and predicted classes. The model shows a superior performance than CNN, CNN+LSTM, and CNN+Autoencoder by making 1211 correct predictions while only 82 predictions are wrong. It shows the model's capability to accurately predict various classes of wafer map hotspots.

D. EXPERIMENTAL RESULTS FOR LSTM+HAM

To complement the prowess in wafer map hotspot classification, LSTM and the integration of the HAM offer a promising approach. LSTMs are a kind of recurrent neural network intended to identify patterns and sequential relationships in data. When combined with HAM, which provides a mechanism for the model to focus on the most relevant information at different hierarchical levels, the resulting model becomes well-suited for analyzing complex wafer map datasets. LSTMs process sequential wafer map data, considering manufacturing process temporal dependencies. HAM allows selective attention to areas of interest, enhancing hotspot detection accuracy. Hierarchical attention structure helps discern critical patterns at multiple scales. High accuracy, precision, recall, and F1 scores make them valuable for image data classification tasks. the LSTM+HAM algorithm performs quite well as accuracy of 89.17%, precision of 89.43%, recall of 89.17%, and F1 score of 89.19%. As seen in Table 7, these findings demonstrate that the model can reliably collect good instances and forecast results.

A confusion matrix encapsulating all of the predictions generated by the LSTM+HAM model using testing data is displayed in Figure 9. The confusion matrix shows the performance of a hybrid model in correctly predicting the samples for each class. Results indicate the LSTM+HAM models make 1153 correct predictions while 140 predictions are wrong. This performance is poor compared to CNN, CNN+LSTM, CNN+Autoencoder, and CNN+ANN models.

E. RESULTS FOR AUTOENCODERS+HAM MODEL

Autoencoders excel at uncovering latent patterns and features within data, which is particularly valuable in the context of wafer map analysis. By employing HAM, this integrated model can selectively emphasize crucial regions of the wafer maps, thereby elevating the precision and accuracy of hotspot

TABLE 8. Result for Auto encoder+HAM model.

Epochs	Accuracy	Precision	Recall	F1-Score
5	0.6419	0.6315	0.6419	0.624
10	0.8399	0.8508	0.8399	0.84
15	0.8012	0.8144	0.8012	0.8006
20	0.8925	0.8981	0.8925	0.8941
25	0.9226	0.9271	0.9226	0.9227
30	0.9187	0.9242	0.9187	0.9191

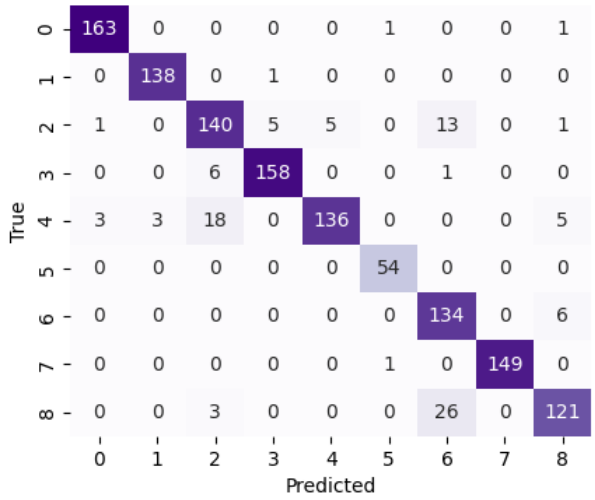

FIGURE 10. Confusion matrix for Auto encoder+HAM model.

TABLE 9. Result for ANN+HAM model.

Epochs	Accuracy	Precision	Recall	F1-Score
5	0.7332	0.7371	0.7332	0.7217
10	0.8074	0.8127	0.8074	0.8065
15	0.8685	0.8728	0.8685	0.8682
20	0.9134	0.9146	0.9134	0.9138
25	0.9288	0.9298	0.9288	0.9289
30	0.9451	0.9481	0.9451	0.9451

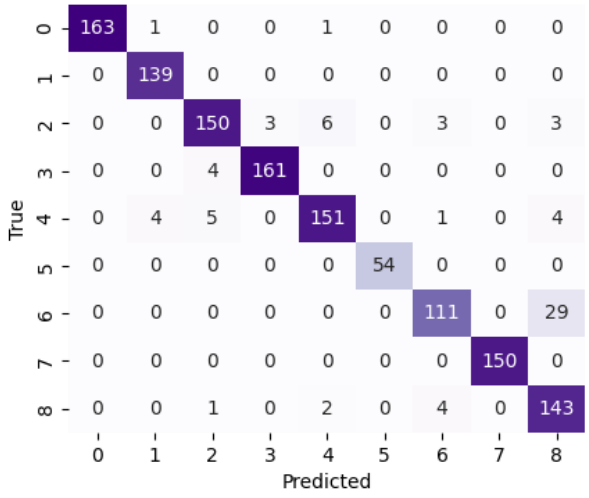

FIGURE 11. Confusion matrix for ANN+HAM model.

detection. The hierarchical nature of the attention module enables the model to discern critical patterns across multiple scales, which is essential for situations where hotspot features may manifest at various levels of granularity within the data. His holistic solution enhances the precision, recall, and F1 scores, establishing its efficacy in addressing wafer map hotspot classification and solidifying its applicability in image data classification tasks. The Auto encoder+HAM algorithm works rather well, with an accuracy of 92.26%, precision of 92.71%, recall of 92.26%, and F1 score of 92.27%. These results show that the model can consistently gather excellent examples and estimate outcomes, as seen in Table 8.

Figure 10 shows a confusion matrix that contains all of the predictions made by the Auto encoder+HAM model utilizing testing data. With values ranging from 0 to 8, the confusion matrix displays the findings as a cross-sectional of the actual and predicted classes. Results regarding correct and wrong predictions by the hybrid Autoencoder+HAM model indicate that 1193 samples are predicted correctly out of 1293 samples while 100 predictions are wrong. The performance of the Autoencoder+HAM model is better than LSTM+HAM, however, poor than other hybrid models implemented in this study.

F. HYBRID MODEL ANN+HAM

To further bolster the capabilities of wafer map hotspot classification, the inclusion of ANN in conjunction with the HAM provides a promising alternative. ANNs are a useful addition to the toolset for wafer map analysis because of their reputation for being able to recognize intricate patterns and correlations within data. When paired with HAM, this integrated model becomes proficient in focusing on critical areas within the wafer maps, thereby enhancing precision and accuracy in hotspot detection. The hierarchical structure of the attention mechanism empowers the model to discern crucial patterns across multiple scales, a crucial feature in scenarios where hotspot characteristics may manifest at varying levels of granularity within the dataset. This holistic solution not only contributes to higher precision, recall, and F1 scores but also solidifies its efficacy in addressing wafer map hotspot classification. The ANN+HAM algorithm performs rather well with an accuracy of 94.51%, precision of 94.81%, recall of 94.51%, and F1 score of 94.51%. As Table 9 demonstrates, these findings demonstrate that the model can reliably collect top-notch instances and estimate results.

A confusion matrix including every prediction the ANN+HAM model produced using testing data is displayed in Figure 11. The confusion matrix shows the results as a cross-sectional of the actual and projected classes, with values ranging from 0 to 8. The model is able to make 1222 predictions showing superior performance compared to other models used in this study. The model has only 71 wrong predictions.

G. EXPERIMENTAL RESULTS FOR CNN+HAM HYBRID MODEL

The inclusion of CNN in conjunction with the HAM presents a compelling alternative. CNNs are renowned for their

TABLE 10. Result for CNN+HAM model.

Epochs	Accuracy	Precision	Recall	F1-Score
5	0.7548	0.7629	0.7548	0.7554
10	0.8484	0.8461	0.8484	0.846
15	0.8832	0.8885	0.8832	0.8832
20	0.9157	0.916	0.9157	0.9153
25	0.9397	0.9407	0.9397	0.9397
30	0.9458	0.948	0.9458	0.9459

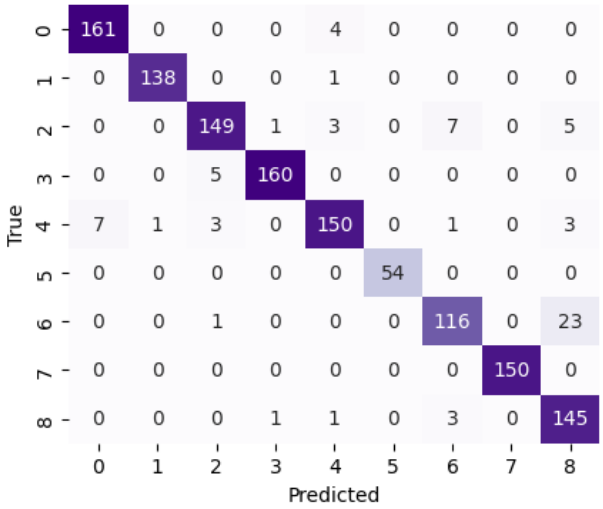

FIGURE 12. Confusion matrix for CNN+HAM model.

prowess in discerning intricate patterns and features within data, and their application within the context of wafer map analysis holds significant promise. The hierarchical structure of the attention mechanism equips the model with the ability to identify critical patterns across various scales, a fundamental requirement in scenarios where hotspot features may manifest at different levels of granularity within the dataset. When combined with HAM, this integrated model becomes adept at directing its attention toward crucial regions of the wafer maps, resulting in an elevation of precision and accuracy in hotspot detection. The CNN+HAM algorithm works rather well, with an accuracy of 94.58%, precision of 94.48%, recall of 94.58%, and F1 score of 94.59%. These results show that the model can consistently gather excellent cases and predict outcomes, as Table 10 illustrates.

Figure 12 provides a confusion matrix with all of the predictions the CNN+HAM model made using testing data. With values ranging from 0 to 8, the confusion matrix displays the findings as a cross-sectional of the actual and predicted classes. The CNN+HAM hybrid outperforms all other models implemented in this study with 1223 correct predictions with only 70 wrong predictions for nine classes of wafer map hotspots.

H. RESULTS OF K-FOLD CROSS-VALIDATION

The proposed approach has to go through k-fold cross-validation in order to validate the outcome. The findings for k = 3,5,7, and 10 are shown in Tables 11, 12, 13, and 14,

TABLE 11. Model performance with k = 3.

Fold	Accuracy	Precision	Recall	F1-Score
1	0.9397	0.9429	0.9397	0.9403
2	0.9490	0.9490	0.9490	0.9488
3	0.9512	0.9515	0.9513	0.9512
Mean	0.9466	0.9478	0.9467	0.9468

TABLE 12. Model performance with k = 5.

Fold	Accuracy	Precision	Recall	F1-Score
1	0.9537	0.9566	0.9537	0.9541
2	0.9189	0.9217	0.9189	0.9195
3	0.9537	0.9519	0.9537	0.9536
4	0.9729	0.9731	0.9729	0.9728
5	0.9341	0.9348	0.9466	0.9468
Mean	0.9467	0.9476	0.9492	0.9494

TABLE 13. Model performance with k = 7.

Fold	Accuracy	Precision	Recall	F1-Score
1	0.9568	0.9622	0.9568	0.9576
2	0.9351	0.9382	0.9351	0.9352
3	0.9297	0.9306	0.9297	0.9297
4	0.9405	0.9421	0.9405	0.9403
5	0.9784	0.9784	0.9784	0.9783
6	0.9673	0.9707	0.9673	0.9681
7	0.9185	0.9208	0.9185	0.9184
Mean	0.9466	0.9490	0.9466	0.9468

TABLE 14. Model performance with k = 10.

Fold	Accuracy	Precision	Recall	F1-Score
1	0.9538	0.9607	0.9538	0.9544
2	0.9538	0.9561	0.9538	0.9537
3	0.9231	0.9228	0.9231	0.9240
4	0.9147	0.9157	0.9147	0.9144
5	0.9379	0.9458	0.9379	0.9374
6	0.9690	0.9705	0.9689	0.9690
7	0.9767	0.9771	0.9767	0.9767
8	0.9690	0.9713	0.9690	0.9694
9	0.9379	0.9418	0.9379	0.9381
10	0.9302	0.9328	0.9302	0.9297
Mean	0.9466	0.9495	0.9466	0.9467

respectively. The results demonstrated in Table 13 show that the proposed model performs much better with k = 5 and k = 7 in K-fold cross-validation. Overall, the performance of the proposed model is similar for all values of k, with marginal differences.

I. COMPARATIVE ANALYSIS OF ALL MODELS

All of the models employed in this study have results displayed in Table 15. For accuracy, precision, recall, and F1 scores, results are given. The proposed model performs better than the other models, which are all optimized to obtain the best results with the provided dataset: CNN, CNN+LSTM, CNN+Auto encoder, CNN+ANN, LSTM+HAM, Autoencoder+HAM, ANN+HAM, and CNN+HAM. Still, the performance of the hybrid models is better than all the other models now in use. Of all the models employed in this study, the best results are obtained using CNN+HAM with a 0.9458 accuracy score. The CNN+HAM

TABLE 15. Results of deep learning models using hyperparameter tuning.

Models	Epoch	Accuracy	Precision	Recall	F1-Score
CNN	25	0.8948	0.8954	0.8948	0.8945
CNN+LSTM	30	0.9002	0.9053	0.9002	0.9012
CNN+Auto encoder	30	0.9157	0.9171	0.9157	0.9156
CNN+ANN	20	0.9365	0.9367	0.9365	0.9364
RNN+HAM	30	0.8685	0.8682	0.8685	0.8681
GRU+HAM	30	0.8955	0.8987	0.8955	0.8933
LSTM+HAM	30	0.8917	0.8943	0.8917	0.8919
Autoencoder+HAM	25	0.9226	0.9271	0.9226	0.9227
ANN+HAM	30	0.9451	0.9481	0.9451	0.9451
CNN+HAM	30	0.9458	0.948	0.9458	0.9459

TABLE 16. Comparative analysis with existing literature.

Reference	Year	Technique	Result
[47]	2020	CNN-FC, CNN-GAP	Accuracy = 90.9%, 93.6%
[48]	2018	SENet	Accuracy=93.55%
[49]	2019	MobileNetV3	Accuracy=92.09%
[50]	2019	CNN	Accuracy = 93.25%
[51]	2020	CNN	Accuracy = 94.00%
[52]	2020	CNN	Accuracy = 92.13%
[53]	2020	CNN	Accuracy = 90.44%
[54]	2019	CNN	Accuracy = 93.00%
[55]	2019	CNN	Accuracy = 89.5%
[56]	2019	CNN	Accuracy = 91.2%
[57]	2019	SCSDAE	Accuracy = 94.39%
[58]	2020	SVAE	Accuracy = 90.9%
[59]	2018	CNN	Accuracy = 91.00%, Precision = 94.9%, Recall = 94.5%
[60]	2021	DCNN	Accuracy = 93.75%, Precision = 93.81%, Recall = 93.79%, F1-score = 93.76%
Proposed	2023	CNN+HAM	Accuracy = 94.58%, Precision = 94.8%, Recall = 94.58%, F1-score = 94.59%
K-fold	2023	CNN+HAM with K-fold	Accuracy = 97.84%, Precision = 97.84%, Recall = 97.84%, F1-score = 97.83%

model also has superior results concerning precision, recall, and F1 score showing its potential for accurate hotspot detection in wafer maps.

J. LIMITATIONS

The proposed approach shows promising results and performs well compared to existing approaches on wafer hotspot detection, however, it has several limitations. First, for generalizability, further experiments are needed on additional datasets. Second, a total of nine types of wafer defects have been considered in this study, adding more would broaden the scope of the model's applicability. Third, deep learning models are complex, require large amounts of training data, and have higher computational complexity. These aspects can be further investigated to improve their performance. Lastly, the use of transfer learning can be investigated in the context of wafer hotspot detection in the future.

K. PERFORMANCE COMPARISON WITH EXISTING APPROACHES

A comparative study is conducted between several models from the existing body of literature that used various machine and deep learning algorithms to detect defects in semiconductors. For example, [47] uses CNN-fully connected (CC-FC) and CNN with global average pooling (CNN-GAP) models and reports a 90.9% and 93.6% accuracy for FC and GAP, respectively. Similarly, other studies [48], [49] have utilized SENet, MobileNet, etc. for the same purpose. Table 16 shows a performance comparison of the current study with these studies indicating a better performance of the proposed model. In addition, we have considered other CNN variants for image classification for comparison and implemented them on the dataset used in the current study to carry out an extensive investigation. The results from those studies also indicate that the CNN+HAM provides better results.

V. CONCLUSION

In the realm of semiconductor fabrication, the quest for efficient and reliable methods to detect wafer hotspots during the production process has led to the integration of advanced artificial intelligence techniques. Leveraging CNNs in combination with various attention mechanisms, such as the hierarchical attention module, this approach explores an array of models to optimize hotspot detection. The models considered include CNN, CNN+LSTM, CNN+Autoencoder, CNN+ANN, LSTM+HAM, Autoencoder+HAM, ANN+HAM and CNN+HAM. Among these, the CNN+HAM model emerged as the frontrunner, showcasing the highest accuracy of 94.58%. According to the experimental results, CNN+HAM performed better than expected, with a k-fold cross-validation accuracy score of 97.84% for values of k that are 3,5, 7, and 10, respectively. This result underscores the effectiveness of the HAM's attention mechanism in enhancing the CNN's

ability to detect and prioritize hotspots, thereby providing a robust and accurate solution for raising the standard and effectiveness of the procedure used in the manufacture of semiconductors. To encourage future developments in flaw identification so that engineers and researchers may verify and test against new designs and feature sizes, much thought must be given to creating a repository with actual flaws. Combining ongoing innovation, expansion, and development shows a lot of potential for reliable and effective defect identification. The model's robustness and generalizability should be improved by adding data from different wafer production conditions. Investigating complex attention methods can enhance the model's focus on important wafer areas. Integrating the model with real-time monitoring systems for hotspot identification and remedial measures is also important. Combining CNN-based methods with reinforcement learning can increase detection accuracy in dynamic manufacturing situations through adaptive learning.

ACKNOWLEDGMENT

(Mobeen Shahroz, Mudasir Ali, and Alishba Tahir are co-first authors.)

REFERENCES

[1] J. Ma, T. Zhang, C. Yang, Y. Cao, L. Xie, H. Tian, and X. Li, "Review of wafer surface defect detection methods," *Electronics*, vol. 12, no. 8, p. 1787, Apr. 2023.

[2] 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.

[3] J. Zhu, J. Liu, T. Xu, S. Yuan, Z. Zhang, H. Jiang, H. Gu, R. Zhou, and S. Liu, "Optical wafer defect inspection at the 10 nm technology node and beyond," *Int. J. Extreme Manuf.*, vol. 4, no. 3, Sep. 2022, Art. no. 032001.

[4] P. Cui and J. Wang, "Out-of-distribution (OOD) detection based on deep learning: A review," *Electronics*, vol. 11, no. 21, p. 3500, Oct. 2022.

[5] T. N. Theis and H.-S. P. Wong, "The end of Moore's law: A new beginning for information technology," *Comput. Sci. Eng.*, vol. 19, no. 2, pp. 41–50, Mar. 2017.

[6] 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.

[7] Y.-F. Yang and M. Sun, "Hybrid quantum-classical machine learning for lithography hotspot detection," in *Proc. 33rd Annu. SEMI Adv. Semiconductor Manuf. Conf. (ASMC)*, May 2022, pp. 1–6.

[8] J. Hwang and H. Kim, "Variational deep clustering of wafer map patterns," *IEEE Trans. Semicond. Manuf.*, vol. 33, no. 3, pp. 466–475, Aug. 2020.

[9] P. Zeng, M. Zheng, H. Chen, G. Chen, Z. Shu, L. Chen, H. Liang, Y. Zhou, Q. Zhao, and H. Duan, "Wafer-level highly dense metallic nanopillar-enabled high-performance SERS substrates for molecular detection," *Nanomaterials*, vol. 13, no. 11, p. 1733, May 2023.

[10] S.-H. Chen, C.-H. Kang, and D.-B. Perng, "Detecting and measuring defects in wafer die using GAN and YOLOv3," *Appl. Sci.*, vol. 10, no. 23, p. 8725, Dec. 2020.

[11] J. Liu, G. Huang, R. N. Wang, J. He, A. S. Raja, T. Liu, N. J. Engelsen, and T. J. Kippenberg, "High-yield, wafer-scale fabrication of ultralow-loss, dispersion-engineered silicon nitride photonic circuits," *Nature Commun.*, vol. 12, no. 1, p. 2236, Apr. 2021.

[12] I. Teerlinck, P. W. Mertens, H. F. Schmidt, M. Meuris, and M. M. Heyns, "Impact of the electrochemical properties of silicon wafer surfaces on copper outplating from HF solutions," *J. Electrochem. Soc.*, vol. 143, no. 10, pp. 3323–3327, Oct. 1996.

[13] B. Bera, "Silicon wafer cleaning: A fundamental and critical step in semiconductor fabrication process," *Int. J. Appl. Nanotechnol.*, vol. 5, no. 1, pp. 8–13, 2019.

[14] S. Wieghold, A. E. Morishige, L. Meyer, T. Buonassisi, and E. M. Sachs, "Crack detection in crystalline silicon solar cells using dark-field imaging," *Energy Proc.*, vol. 124, pp. 526–531, Sep. 2017.

[15] H. Yang, J. Su, Y. Zou, Y. Ma, B. Yu, and E. F. Y. Young, "Layout hotspot detection with feature tensor generation and deep biased learning," in *Proc. 54th Annu. Design Autom. Conf.*, Jun. 2017, pp. 1–6.

[16] W.-Y. Wen, J.-C. Li, S.-Y. Lin, J.-Y. Chen, and S.-C. Chang, "A fuzzy-matching model with grid reduction for lithography hotspot detection," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 33, no. 11, pp. 1671–1680, Nov. 2014.

[17] K. B. Lee, S. Cheon, and C. O. Kim, "A convolutional neural network for fault classification and diagnosis in semiconductor manufacturing processes," *IEEE Trans. Semicond. Manuf.*, vol. 30, no. 2, pp. 135–142, May 2017.

[18] E. Kim, S. Cho, B. Lee, and M. Cho, "Fault detection and diagnosis using self-attentive convolutional neural networks for variable-length sensor data in semiconductor manufacturing," *IEEE Trans. Semicond. Manuf.*, vol. 32, no. 3, pp. 302–309, Aug. 2019.

[19] Y. Lin, X. Xu, J. Ou, and D. Z. Pan, "Machine learning for mask/wafer hotspot detection and mask synthesis," *Proc. SPIE*, vol. 10451, pp. 72–84, Oct. 2017.

[20] H. Yang, Y. Lin, B. Yu, and E. F. Y. Young, "Lithography hotspot detection: From shallow to deep learning," in *Proc. 30th IEEE Int. Syst.-on-Chip Conf. (SOCC)*, Sep. 2017, pp. 233–238.

[21] G. R. Reddy, K. Madkour, and Y. Makris, "Machine learning-based hotspot detection: Fallacies, pitfalls and marching orders," in *Proc. IEEE/ACM Int. Conf. Comput.-Aided Design (ICCAD)*, Nov. 2019, pp. 1–8.

[22] Y. Chen, Y. Lin, T. Gai, Y. Su, Y. Wei, and D. Z. Pan, "Semi-supervised hotspot detection with self-paced multi-task learning," in *Proc. 24th Asia South Pacific Design Autom. Conf.*, Jan. 2019, pp. 420–425.

[23] J. K. Kaldellis, M. Kapsali, and K. A. Kavadias, "Temperature and wind speed impact on the efficiency of PV installations. Experience obtained from outdoor measurements in Greece," *Renew. Energy*, vol. 66, pp. 612–624, Jun. 2014.

[24] V. Sharma and S. S. Chandel, "A novel study for determining early life degradation of multi-crystalline-silicon photovoltaic modules observed in western Himalayan Indian climatic conditions," *Sol. Energy*, vol. 134, pp. 32–44, Sep. 2016.

[25] M. Santhakumari and N. Sagar, "A review of the environmental factors degrading the performance of silicon wafer-based photovoltaic modules: Failure detection methods and essential mitigation techniques," *Renew. Sustain. Energy Rev.*, vol. 110, pp. 83–100, Aug. 2019.

[26] M.-J. Wu, J. R. Jang, and J.-L. Chen, "Wafer map failure pattern recognition and similarity ranking for large-scale data sets," *IEEE Trans. Semicond. Manuf.*, vol. 28, no. 1, pp. 1–12, Feb. 2015.

[27] B. Yu, J.-R. Gao, D. Ding, X. Zeng, and D. Z. Pan, "Accurate lithography hotspot detection based on principal component analysis-support vector machine classifier with hierarchical data clustering," *J. Micro/Nanolithography, MEMS, MOEMS*, vol. 14, no. 1, Nov. 2014, Art. no. 011003.

[28] N. S. Beattie, R. S. Moir, C. Chacko, G. Buffoni, S. H. Roberts, and N. M. Pearsall, "Understanding the effects of sand and dust accumulation on photovoltaic modules," *Renew. Energy*, vol. 48, pp. 448–452, Dec. 2012.

[29] MITLL Low-Power FDSOI, CMOS, "MITLL low-power FDSOI CMOS process design guide," MIT Lincoln Lab., Lexington, MA, USA, Version: 3D01, 2006.

[30] J. A. Burns, B. F. Aull, C. K. Chen, C.-L. Chen, C. L. Keast, J. M. Knecht, V. Suntharalingam, K. Warner, P. W. Wyatt, and D.-R. Yost, "A wafer-scale 3-D circuit integration technology," *IEEE Trans. Electron Devices*, vol. 53, no. 10, pp. 2507–2516, Oct. 2006.

[31] M. W. Akram, G. Li, Y. Jin, X. Chen, C. Zhu, and A. Ahmad, "Automatic detection of photovoltaic module defects in infrared images with isolated and develop-model transfer deep learning," *Sol. Energy*, vol. 198, pp. 175–186, Mar. 2020.

[32] C. Ryu, J. Lee, H. Lee, K. Lee, T. Oh, and J. Kim, "High frequency electrical model of through wafer via for 3-D stacked chip packaging," in *Proc. 1st Electron. Systemintegr. Technol. Conf.*, Sep. 2006, pp. 215–220.

[33] W. Yuan, Y. Lu, M. Li, B. Pan, Y. Gao, Y. Tian, Z.-Q. Li, L. Ji, Y. Huang, H. Chen, Y. Yao, and S. Park, "Machine learning hotspot prediction significantly improve capture rate on wafer," in *Proc. Int. Workshop Adv. Patterning Solutions (IWAPS)*, Nov. 2020, pp. 1–4.

[34] E. Sovetkin, E. J. Achterberg, T. Weber, and B. E. Pieters, "Encoder–decoder semantic segmentation models for electroluminescence images of thin-film photovoltaic modules," *IEEE J. Photovolt.*, vol. 11, no. 2, pp. 444–452, Mar. 2021.

[35] Y.-F. Yang and M. Sun, "Semiconductor defect detection by hybrid classical-quantum deep learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 2313–2322.

[36] M. Dhimish, "Micro cracks distribution and power degradation of polycrystalline solar cells wafer: Observations constructed from the analysis of 4000 samples," *Renew. Energy*, vol. 145, pp. 466–477, Jan. 2020.

[37] H. Geng, H. Yang, L. Zhang, J. Miao, F. Yang, X. Zeng, and B. Yu, "Hotspot detection via attention-based deep layout metric learning," in *Proc. IEEE/ACM Int. Conf. Comput. Aided Design (ICCAD)*, Nov. 2020, pp. 1–8.

[38] M. Afridi, A. Kumar, F. I. Mahmood, and G. TamizhMani, "Comparative analysis of hotspot stress endurance in pristine and thermal cycled prestressed glass–glass photovoltaic modules," *Sustainability*, vol. 15, no. 16, p. 12131, Aug. 2023.

[39] A. Bouraiou, M. Hamouda, A. Chaker, M. Mostefaoui, S. Lachtar, M. Sadok, N. Boutasseta, M. Othmani, and A. Issam, "Analysis and evaluation of the impact of climatic conditions on the photovoltaic modules performance in the desert environment," *Energy Convers. Manage.*, vol. 106, pp. 1345–1355, Dec. 2015.

[40] X. Zhang, C. Yang, Y. Zhang, A. Hu, M. Li, L. Gao, H. Ling, and T. Hang, "Sub-surface damage of ultra-thin monocrystalline silicon wafer induced by dry polishing," *Electron. Mater. Lett.*, vol. 16, no. 4, pp. 355–362, Jul. 2020.

[41] T. Braun, T. D. Nguyen, S. Voges, M. Wöhrmann, R. Gernhardt, K.-F. Becker, I. Ndip, D. Freimund, M. Schneider-Ramelow, K.-D. Lang, D. Schwantuschke, E. Ture, M. Pretl, and S. Engels, "Fan-out wafer level packaging of GaN components for RF applications," in *Proc. IEEE 70th Electron. Compon. Technol. Conf. (ECTC)*, Jun. 2020, pp. 7–13.

[42] J. Kurnik, M. Jankovec, K. Brecl, and M. Topic, "Outdoor testing of PV module temperature and performance under different mounting and operational conditions," *Sol. Energy Mater. Sol. Cells*, vol. 95, no. 1, pp. 373–376, Jan. 2011.

[43] K. A. K. Niazi, W. Akhtar, H. A. Khan, Y. Yang, and S. Athar, "Hotspot diagnosis for solar photovoltaic modules using a Naive Bayes classifier," *Sol. Energy*, vol. 190, pp. 34–43, Sep. 2019.

[44] J. Donnelly, A. Daneshkhah, and S. Abolfathi, "Physics-informed neural networks as surrogate models of hydrodynamic simulators," *Sci. Total Environ.*, vol. 912, Feb. 2024, Art. no. 168814.

[45] J. Donnelly, S. Abolfathi, J. Pearson, O. Chatrabgoun, and A. Daneshkhah, "Gaussian process emulation of spatio-temporal outputs of a 2D inland flood model," *Water Res.*, vol. 225, Oct. 2022, Art. no. 119100.

[46] J. Donnelly, A. Daneshkhah, and S. Abolfathi, "Forecasting global climate drivers using Gaussian processes and convolutional autoencoders," *Eng. Appl. Artif. Intell.*, vol. 128, Feb. 2024, Art. no. 107536.

[47] D. V. Patel, R. Bonam, and A. A. Oberai, "Deep learning-based detection, classification, and localization of defects in semiconductor processes," *J. Micro/Nanolithography, MEMS, MOEMS*, vol. 19, no. 2, p. 1, Apr. 2020.

[48] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[49] A. Howard, M. Sandler, B. Chen, W. Wang, L.-C. Chen, M. Tan, G. Chu, V. Vasudevan, Y. Zhu, R. Pang, H. Adam, and Q. Le, "Searching for MobileNetV3," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1314–1324.

[50] N. Yu, Q. Xu, and H. Wang, "Wafer defect pattern recognition and analysis based on convolutional neural network," *IEEE Trans. Semicond. Manuf.*, vol. 32, no. 4, pp. 566–573, Nov. 2019.

[51] M. B. Alawieh, D. Boning, and D. Z. Pan, "Wafer map defect patterns classification using deep selective learning," in *Proc. 57th ACM/IEEE Design Autom. Conf. (DAC)*, Jul. 2020, pp. 1–6.

[52] S. Kang, "Rotation-invariant wafer map pattern classification with convolutional neural networks," *IEEE Access*, vol. 8, pp. 170650–170658, 2020.

[53] U. Batool, M. I. Shapiai, H. Fauzi, and J. X. Fong, "Convolutional neural network for imbalanced data classification of silicon wafer defects," in *Proc. 16th IEEE Int. Colloq. Signal Process. Appl. (CSPA)*, Feb. 2020, pp. 230–235.

[54] V. S. Ajna and N. George, "Detection of hotspots in layout patterns using deep learning," in *Proc. 10th Int. Conf. Comput., Commun. Netw. Technol. (ICCCNT)*, Jul. 2019, pp. 1–6.

[55] J. Yu, "Enhanced stacked denoising autoencoder-based feature learning for recognition of wafer map defects," *IEEE Trans. Semicond. Manuf.*, vol. 32, no. 4, pp. 613–624, Nov. 2019.

[56] 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.

[57] J. Yu, X. Zheng, and J. Liu, "Stacked convolutional sparse denoising auto-encoder for identification of defect patterns in semiconductor wafer map," *Comput. Ind.*, vol. 109, pp. 121–133, Aug. 2019.

[58] Y. Kong and D. Ni, "A semi-supervised and incremental modeling framework for wafer map classification," *IEEE Trans. Semicond. Manuf.*, vol. 33, no. 1, pp. 62–71, Feb. 2020.

[59] 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.

[60] H. Zheng, S. W. A. Sherazi, S. H. Son, and J. Y. Lee, "A deep convolutional neural network-based multi-class image classification for automatic wafer map failure recognition in semiconductor manufacturing," *Appl. Sci.*, vol. 11, no. 20, p. 9769, Oct. 2021.

MOBEEN SHAHROZ received the M.C.S. degree and the M.S. degree in computer science from the Department of Computer Science, Khwaja Fareed University of Engineering and Information Technology (KFUEIT), Rahim Yar Khan, Pakistan, in 2018 and 2020, respectively. He is currently pursuing the Ph.D. degree with the Department of Artificial Intelligence, Faculty of Computing, The Islamia University of Bahawalpur (IUB), Bahawalpur, Pakistan. He worked as a Research Assistant with the Fareed Computing and Research Center, KFUEIT University, and the Sir Sadiq Research and Computing Center, IUB University. He is working as a Lecturer with the Department of Artificial Intelligence, IUB University. His current research interests include the Internet of Things (IoT), electronic appliances, medical image classification, bioinformatics, DNA next-generation sequencing, natural language processing (NLP), text classification, and image classification.

MUDASIR ALI received the B.S.C.S. degree from the Department of Computer Science, Institute of Southern Punjab, Multan, in 2021, and the M.S. degree in computer science from the Department of Computer Science, The Islamia University of Bahawalpur, Bahawalpur, Pakistan, in 2024, where he is currently pursuing the Ph.D. degree in computer science. His current research interests include the Internet of Things (IoT), artificial intelligence, data mining, cyber security, machine learning, deep learning, and image classification.

ALISHBA TAHIR received the Bachelor of Science in Software Engineering (BSSE) degree from the Department of Computer Engineering, Khwaja Fareed University of Engineering and Information Technology (KFUEIT) in 2022. She is currently pursuing the Master of Science in Artificial Intelligence (MSAI) degree from the Department of Artificial Intelligence, The Islamia University of Bahawalpur, Bahawalpur, Pakistan. Her current research interests include natural language processing (NLP), text mining, image classification, medical diagnosis, DNA sequencing, cyber security, machine learning, and deep learning.

HENRY FABIAN GONGORA was born in Veracruz, Mexico. He received the B.S. degree in electronics engineering from the Technological Institute of Veracruz, in 2009, and the M.Eng. and Ph.D. degrees in electrical engineering from the National Autonomous University of Mexico (UNAM), Mexico City, in 2014 and 2018, respectively. Since 2019, he has been a Professor with the School of Engineering, Autonomous University of Campeche (UAC). His current research interests include frequency-selective surfaces, metasurfaces, antenna arrays, and microwave circuits.

CARLOS UC RIOS is currently working as a Professor with Universidad Europea del Atlántico, Spain. He is also affiliated with Universidad Internacional Iberoamericana, Campeche, Mexico, and Universidad Internacional Iberoamericana, Arecibo, PR, USA.

MD ABDUS SAMAD (Member, IEEE) received the Ph.D. degree in information and communication engineering from Chosun University, South Korea. He worked as an Assistant Professor with the Department of Electronics and Telecommunication Engineering, International Islamic University Chittagong, Chattogram, Bangladesh, from 2013 to 2017. He has been working as a Research Professor with the Department of Information and Communication Engineering, Yeungnam University, South Korea. His research interests include signal processing, antenna design, electromagnetic wave propagation, applications of artificial neural networks, and millimeter-wave propagation by interference and atmospheric causes for 5G and beyond wireless networks. He received the Prestigious Korean Government Scholarship (GKS) for his Ph.D. study.

IMRAN ASHRAF (Member, IEEE) received the M.S. degree (Hons.) in computer science from Blekinge Institute of Technology, Karlskrona, Sweden, in 2010, and the Ph.D. degree in information and communication engineering from Yeungnam University, Gyeongsan-si, South Korea, in 2019. He has worked as a Postdoctoral Fellow with Yeungnam University. He is currently working as an Assistant Professor with the Information and Communication Engineering Department, Yeungnam University. His research interests include positioning using next-generation networks, communication in 5G and beyond, location-based services in wireless communication, smart sensors (LIDAR) for smart cars, and data analytics.