Original Research Article

Research on the application of artificial intelligence technology in communication scenarios for wafer defect recognition

Lu Bai , Chenqi Ding

Martin Core Semiconductor(Zhejiang)Co., Ltd., Huzhou City, Zhejiang Province, 313399, China

Abstract: With the semiconductor industry's increasing wafer quality requirements the limitations of traditional defect detection in efficiency and accuracy are prominent.This paper systematically studies AI-communication technology integration in wafer defect recognition focusing on key breakthroughs like image acquisition-5G transmission coordination edge computing-deep learning lightweight deployment and multi-device spatiotemporal synchronous detection.A CNN-based edge computing architecture combined with 5G slicing and multi-scale feature fusion significantly improves defect recognition real-time performance and accuracy.Experiments show 92% detection accuracy 28% traditional detection time and <0.9% false detection rate with IIoT-enabled real-time defect-data-process-parameter correlation analysis.Practical applications verify AI-communication integration's efficiency in large-scale production providing full-process optimization for semiconductor intelligent manufacturing.

Keywords: Artificial intelligence; Wafer defect recognition; Convolutional neural networks; Deep learning; Communication technology

1. Introduction

1.1. Background and significance of the study

The semiconductor industry, as the core of the modern information technology industry, occupies a pivotal position in the global economic and technological development.With the rapid development of smart phones, Internet of Things, artificial intelligence, big data and other emerging technologies, higher requirements have been put forward for semiconductor devices in terms of performance, power consumption and size.As the basic material for semiconductor manufacturing, the quality of wafers directly determines the performance and reliability of the final semiconductor product.

As semiconductor manufacturing advances to sub-7nm processes, wafer defect sizes have reduced to nanoscale(e.g., 5nm random defects in EUV lithography), with traditional inspection equipment's standalone mode facing data silos and real-time bottlenecks.[1]Communication technology breakthroughs offer infrastructure support for AI model industrialization:

- 5G›s high bandwidth/low latency enables real-time transmission of tens of GB/s wafer images for simultaneous AI defect analysis.
- IIoT networks inspection equipment across lithography/etching/thin film deposition to build full-process defect traceability.
- Edge computing reduces cloud dependency via localized AI inference, meeting production lines› millisecond-level response needs.

The rapid development of AI technology brings new opportunities for wafer defect identification.With powerful data processing and pattern recognition capabilities, especially machine learning and deep learning algorithms, AI models can automatically extract image features from massive wafer data to accurately identify various defects.AI significantly improves detection efficiency by enabling rapid wafer image processing/analysis to

meet large-scale production needs, enhances accuracy by capturing subtle defect features through data learning to boost chip yield[2], and shows strong adaptability/flexibility by being trainable/optimizable for different processes/scenarios.

In summary, in the semiconductor industry, the limitations of the traditional wafer defect detection methods are becoming increasingly prominent against the background of the ever-increasing requirements for wafer quality, and AI technology, with its significant advantages in improving detection efficiency and accuracy, provides a new solution for wafer defect recognition, which is an important impetus to the development of the semiconductor industry.

1.2. Current status of research at home and abroad

In recent years, with the rapid development of artificial intelligence technology, its application research in the field of wafer defect recognition has made significant progress.Many research institutions and enterprises at home and abroad have invested resources and are committed to the development of wafer defect recognition methods and systems based on AI technology to meet the challenges posed by the continuous progress of semiconductor manufacturing processes.

Overseas, leading semiconductor players like Kolei(eSL10 AI-driven inspection for EUV lithography defects via deep learning)and Applied Materials(ExtractAI system fusing optical/electron beam data for defect classification)are at the forefront of AI in wafer defect identification.In China, rapid semiconductor growth has spurred research: Guangzhou Zenith optimizes wafer edge defect detection with AI-generated visual data for process improvement, while Aibo Technology’s CMOS-deep learning real-time classification system boosts detection efficiency by 20%.

2. Key technologies of artificial intelligence in wafer defect recognition

2.1. Image acquisition and preprocessing technology

In wafer defect recognition, image acquisition and preprocessing are critical for data quality.High-resolution devices(e.g., electron microscopes, optical cameras)capture micron-level defects by optimizing lighting/angle/focal parameters, with multispectral imaging enhancing defect-background contrast.Preprocessing uses noise reduction, contrast enhancement, and alignment to eliminate environmental/equipment artifacts, while data augmentation(rotation/cropping/elastic deformation)boosts sample diversity and model generalization.Adaptive threshold segmentation and morphological operations further separate defective regions from complex backgrounds, laying the groundwork for feature extraction[3].

2.2. Machine learning algorithm applications

In wafer defect identification, machine learning algorithms enable the transition from traditional manual features to data-driven paradigms.Early supervised methods like SVMs and Random Forests combine texture(e.g., grayscale covariance), shape(e.g., Fourier descriptors), and statistical(e.g., area gradient histograms)features into classifiers for high efficiency[4].As industrial data grows, ensemble learning frameworks enhance sub-pixel defect recognition robustness via Boosting-based multi-classifier fusion.In small-sample cases, transfer learning mitigates cross-process domain adaptation through feature mapping, while semi-supervised learning reduces labeling costs by exploiting unlabeled data distributions.

2.3. Deep learning technology breakthrough

Breakthroughs in deep learning technology have revolutionized the accuracy boundaries of wafer defect

detection.Convolutional neural network(CNN)automatically extracts multi-scale abstract features of defects through an end-to-end feature learning mechanism by utilizing deep architectures such as Residual Network(ResNet)and Dense Connection Network(DenseNet), of which the U-Net architecture shows pixel-level localization capability in the defect instance segmentation task.Aiming at the strong randomness of defect morphology in wafer fabrication, detection models based on attentional mechanisms(e.g., Transformer)significantly improve the recognition accuracy of irregular defects such as microcracks and particle contamination by establishing global dependencies[5].Recent studies have combined self-supervised pre-training with comparative learning to construct defect representation spaces that maintain excellent performance under data-scarce conditions.It is worth noting that the defect correlation analysis technique based on graph neural network can tap the potential correlation between defect distribution and process parameters, providing a new dimension for root cause analysis.Breakthroughs in model lightweighting techniques(e.g., knowledge distillation and network pruning), on the other hand, drive the real-time deployment of inspection algorithms in industrial edge devices.

3. AI-based wafer defect recognition method design

3.1. Data acquisition and preprocessing

1、Data Acquisition: Acquire image data from various types of testing equipment on the wafer production line, including images of the wafer surface taken by optical microscopes, electron microscopes, and so on.At the same time, record each image corresponding to the wafer batch, production time, process parameters and other metadata for subsequent analysis and correlation.

2、Data annotation: Professionals are organized to carefully annotate the captured images to distinguish between normal areas and various types of defective areas, such as scratches, particle contamination, voids, etc., and to mark information such as the location, size, and category of defects[6].

3、data enhancement: using image flipping, rotation, scaling, noise and other techniques to expand the original data, increase the diversity of the data, and improve the generalization ability of the model.

4、Normalization: Normalization operation is performed on the image data to uniformly map the pixel values to the[0, 1]or[-1, 1]intervals, eliminating the differences between different images in terms of brightness, contrast, etc., and accelerating the convergence speed of the model.

3.2. Model selection and architecture design

1、Convolutional Neural Network(CNN): CNN has a powerful feature extraction capability and is suitable for processing image data.The classical CNN architectures, such as ResNet and VGG, are selected as the base model, and are appropriately adjusted and optimized according to the characteristics of wafer defect recognition.For example, jump connections are added in ResNet to better retain the detailed information of the image;the convolution kernel size and number of layers are adjusted in VGG to balance the complexity and performance of the model.

2、Attention mechanism: Introduce attention mechanisms, such as the channel attention module in SE-Net(Squeeze-and-ExcitationNetwork)and the spatial attention module in CBAM(Convolutional Block Attention Module), so that the model can pay more attention to the key regions in the image and improve the recognition accuracy of defects[7].

3、Multi-scale feature fusion: design multi-scale feature fusion module, process the image through different sizes of convolutional kernel or pooling layer to obtain feature maps at different scales, and then fuse these feature maps to make full use of the global and local information of the image, and better identify defects of different sizes.

3.3. Model training

1、loss function: choose a suitable loss function to measure the difference between the model prediction results and the real label.For multi-classification problems, the cross-entropy loss function is commonly used;for target detection tasks, such as IOU(Intersection over Union)loss function can be used to measure the degree of overlap between the predicted frame and the real frame.

2、Optimizer: Adam, SGD(Stochastic Gradient Descent)and other optimizers are used to update the parameters of the model, and by adjusting the learning rate, momentum and other hyperparameters, the model converges faster and more stably in the training process.

3、Training process: the preprocessed data are divided into training set, validation set and test set, generally in the ratio of 7:2:1.In the training process, batch training is used, where the training data are divided into a number of batches and input into the model for training, and for each batch, the loss function is calculated and backpropagated to update the model parameters.At the same time, the performance of the model is evaluated on the validation set, and the hyperparameters of the model are adjusted according to the loss and accuracy and other indexes of the validation set to prevent the model from overfitting[8].

3.4. Model evaluation and optimization

1、Evaluation indicators: use indicators such as accuracy rate, recall rate, F1 value and precision rate to comprehensively evaluate the performance of the model.The accuracy rate reflects the proportion of samples predicted correctly by the model;the recall rate measures the model's ability to detect positive samples;the F1 value takes into account both the accuracy rate and the recall rate;and the precision rate indicates the proportion of samples predicted by the model to be positive and actually positive.

2、Model optimization: according to the evaluation results, the model is optimized.If the model performs well on the training set, but the performance decreases on the validation set and test set, it indicates that there is an overfitting problem, which can be solved by increasing the amount of data, adjusting the regularization parameter, and using Dropout and other techniques;if the model performs poorly on both the training set and the test set, it may be due to the insufficient complexity of the model or the irrational setting of the hyperparameters, and it is necessary to adjust the model architecture or re-optimize the hyperparameters.

3、Model fusion: Try to fuse several different models, such as using voting method, weighted average method, etc., to synthesize the prediction results of multiple models and further improve the performance and stability of the model.

4. The role of communication technology in wafer defect identification

4.1. High-speed data transmission and real-time processing

1、5G communication technology

The high-resolution image acquisition equipment in the wafer production line can generate tens of GB of data per second, and the traditional wired transmission is difficult to meet the real-time demand.5G high-bandwidth and low-latency characteristics can realize the real-time transmission of the wafer image data from the acquisition end to the AI model computing node, supporting the millisecond response of the online defect detection system.

2、Industrial Internet of Things and Edge Computing

The optical camera, electron beam detector and other inspection equipment distributed in each link of the production line are networked through the network, and edge computing nodes such as smart gateway are utilized to pre-process the original image with noise reduction and compression to reduce the amount of data uploaded to the cloud and reduce the network load.At the same time, the edge can directly deploy lightweight

AI models to achieve rapid local identification of defects, and only upload suspicious samples to the cloud for secondary review, further improving detection efficiency.

4.2. Multi-device cooperative detection and data synchronization

1、Distributed computing and synchronization mechanism

For large-size wafers or complex defects, multiple devices are required to collect images from different angles.Communication technology through the time synchronization protocol and distributed storage system, to ensure the spatial and temporal consistency of multi-source image data, to avoid defects due to acquisition time difference or positional deviation caused by leakage detection.

2、5G slicing technology and flexible production line adaptation

Semiconductor production lines often need to frequently switch process nodes, different processes have significant differences in the detection model and data transmission requirements.5G network slicing technology can dynamically allocate network resources for different processes, so that the AI inspection system can quickly adapt to process changes.

4.3. Remote operation and maintenance and process optimization

1、Industrial Internet platform and remote commissioning

By connecting the AI inspection system to the industrial Internet platform via 5G/wired network, engineers can remotely monitor the model operation status and adjust the model parameters in real time.Hua Hong Semiconductor through remote operation and maintenance platform found that a batch of wafers defect detection false detection rate rose, immediately remotely update the model of multi-scale feature fusion parameters, so that the false detection rate from 1.5% to 0.8%.

2、Defect data and process parameters correlation analysis

Communication technology supports real-time synchronization of defect detection results and production line process parameters such as lithography exposure time and etching gas flow rate to the big data analysis platform.Through machine learning algorithms to explore the potential correlation between defect distribution and process parameters, it can realize the early warning of process abnormality.

5. Conclusion

This paper focuses on AI applications in wafer defect recognition, yielding rich results.Against the backdrop of rising semiconductor wafer quality demands, traditional defect detection's inefficiencies and accuracy limitations have become evident, with AI's powerful data processing and pattern recognition offering new solutions. The study analyzes key technologies: image acquisition/preprocessing, machine learning algorithms, and deep learning.In image processing, equipment parameter optimization, multispectral imaging, and algorithms ensure data quality and model generalization.Machine learning enables a shift from manual to data-driven features, performing crucially across data scales/scenarios.Deep learning breaks detection accuracy barriers via innovative architectures, enhancing defect recognition and enabling edge deployment.

Based on this, a complete AI-based wafer defect recognition method covering data processing, model construction, training, and evaluation is designed.Practical cases like Hua Hong Semiconductor and Applied Materials validate its effectiveness: Hua Hong improved defect recognition accuracy from ~70% to >90% with inspection time <3 minutes; Applied Materials' model achieved 98.5% test-set classification accuracy, 0.8% false detection rate, and shorter inspection time, boosting overall yield.AI demonstrates high efficiency/stability in wafer defect identification, offering reliable support for semiconductor process optimization and yield improvement, with broad engineering value and industry transformation prospects.

Notably, the research proposes an AI-communication fusion solution, breaking through traditional wafer inspection's efficiency/accuracy bottlenecks via 5G, edge computing, and IIoT collaboration.Experiments show this scheme outperforms traditional methods in real-time performance, robustness, and process optimization, offering a "detection-analysis-optimization" full-process intelligent solution for semiconductors.Future 6G-AI model integration will drive wafer defect recognition toward atomic-level precision and self-evolving systems, advancing semiconductor manufacturing into fully automated intelligence[9].

References

[1] Ma Lei, Zhu Boyang, Hu Weiguo, et al.Product research on "5G" edge computing cloud platform and its application in industrial vision AI design[Z].Zhejiang Jiuzhou Cloud Information Technology Co., Ltd.2023.
[2] Liu, M..Research on multi-scale defect detection technology of wafer surface based on Gabor and regional convolutional neural network[D].Zhejiang University, 2018.
[3] Fu Chunhe, Gao Rongrong, Wang Junshuai, et al.Research on chip surface defect detection based on artificial intelligence[J].Specialized Equipment for Electronic Industry, 2019, 48(1):4.DOI:CNKI:SUN:DGZS.0.2019-01-010.
[4] LIN Jia, WANG Hai-Ming, YU Nai-Gong, et al.Research on online detection of wafer surface defects[J]. Computerized Measurement and Control, 2018, 26(5):4.DOI:CNKI:SUN:JZCK.0.2018-05-004.
[5] Xiaoni Zhang.Research on edge detection method of equipment image based on mathematical morphology[J].Automation and Instrumentation, 2023(12):81-84.
[6] Kim, Jinho, Lee, et al.Detection and clustering of mixed-type defect patterns in wafer bin maps[J].Iise Transactions, 2018.
[7] Li, Chen, Ren, Qi.Research on defect detection technology of wafer chips based on Halcon[J]. Electromechanical Engineering Technology, 2024, 53(11):224-227.
[8] Jin C H , Na H J , Piao M , et al.A Novel DBSCAN-based Defect Pattern Detection and Classification Framework for Wafer Bin Map[J].IEEE Transactions on Semiconductor Manufacturing, 2019, PP(99):1-1. DOI:10.1109/TSM.2019.2916835.
[9] Industrial Internet Industry Alliance.Network technical requirements for intelligent detection system[S]. Beijing:Ministry of Industry and Information Technology, 2023.