Edge-AI-Based Defect Classification System for Semiconductor Images

Kamalesh E
Ravi Prasath N K
Government College of Technology, Coimbatore
kama.71772314122@gct.ac.in

Abstract

Semiconductor manufacturing requires rapid and reliable defect inspection to ensure high yield and process stability. Conventional inspection systems rely on cloud-based processing or manual review, which introduces latency and limits real-time responsiveness. This report presents an Edge-AI-based defect classification system that performs real-time inference directly at the manufacturing edge. The proposed approach integrates a decentralized intelligence architecture, an eight-class defect classification model, and a hardware-aware deployment pipeline optimized for NXP i.MX RT series devices using the NXP eIQ platform.

1 Introduction

The increasing complexity of semiconductor fabrication processes necessitates fast and accurate defect detection mechanisms. Traditional inspection pipelines often depend on centralized cloud infrastructure or manual inspection, leading to increased latency and reduced scalability. As fabrication throughput continues to rise, these approaches become insufficient for real-time decision-making.

Edge Artificial Intelligence (Edge AI) addresses these challenges by enabling inference directly at the data source. By shifting intelligence from the cloud to the manufacturing edge, defect detection can be performed with minimal delay while reducing network dependency. This report explores the feasibility and implementation of an Edge-AI-based defect classification system designed for embedded deployment in semiconductor manufacturing environments.

2 Problem Statement

Existing semiconductor inspection systems face several limitations:

- High inference latency due to cloud-based processing.
- Dependence on continuous network connectivity.
- Limited suitability for low-power embedded hardware.
- Reduced scalability across multiple inspection points.

The objective of this work is to develop a low-latency, scalable, and embedded-ready defect classification solution capable of operating in real-time manufacturing environments.

3 Proposed Solution

The proposed system introduces a decentralized intelligence architecture that enables real-time defect classification directly at the manufacturing edge. Inspection images are captured from the fabrication setup and processed locally using an optimized AI model.

The solution is built upon the following key principles:

- Edge-level image processing to minimize latency.
- High-fidelity multi-class defect classification.
- Hardware-aware model optimization for embedded platforms.
- Seamless software-to-hardware deployment.

4 System Architecture

4.1 Edge AI Processing

Defect detection is performed directly at the manufacturing edge, eliminating the need for cloud-based inference. This approach ensures real-time responsiveness and uninterrupted operation regardless of network availability.

4.2 Defect Classification Model

An eight-class classification model is employed to distinguish clean samples from various structural defect categories such as bridges, cracks, and line edge roughness. This enables consistent and reliable defect identification.

4.3 Embedded Hardware Feasibility

The AI models are optimized to operate within the power and memory constraints of NXP i.MX RT series microcontrollers. This hardware-aware design ensures efficient execution while maintaining classification accuracy.

4.4 Deployment Pipeline

Models are developed using Python-based deep learning frameworks and deployed using the NXP eIQ toolchain. This Python-to-silicon pipeline converts trained models into deployable binaries suitable for embedded execution.

5 Technology and Feasibility

The proposed system is technologically feasible due to its reliance on proven Edge AI methodologies and commercially available embedded platforms. By minimizing computational overhead and eliminating cloud dependency, the solution is well-suited for industrial environments. Furthermore, the modular architecture enables easy replication across multiple inspection points, supporting scalable deployment.

6 Conclusion

This report presents a practical and scalable Edge-AI-based defect classification system for semiconductor manufacturing. By shifting intelligence from centralized cloud infrastructure to the manufacturing edge, the proposed approach enables real-time defect detection, reduces latency, and improves system scalability. Future work will focus on expanding the dataset, refining model accuracy, and validating performance in real fabrication environments.

References

NXP Semiconductors (2024) *eIQ Machine Learning Software*. Available at: `https://www.nxp.com/eiq` (Accessed: 2024).

ONNX (2024) *Open Neural Network Exchange*. Available at: `https://onnx.ai` (Accessed: 2024).

Shi, W., Cao, J., Zhang, Q., Li, Y. and Xu, L. (2016) 'Edge computing: Vision and challenges', *IEEE Internet of Things Journal*, 3(5), pp. 637–646.