IESA DeepTech Hackathon

Idea Submission Template

FILE NAMING CONVENTION

KLE CoreX_PS01

Team Details

Team Name: Enter Team Name Here...

SR. NO	ROLE	NAME	ACADEMIC YEAR
1	**Team Leader**	*Neela Fakkirgoudar*	*2nd year (4th sem)*

Idea Submission Template

COLLEGE NAME

KLE Technological University, Hubballli

TEAM LEADER CONTACT NUMBER

8217046526

TEAM LEADER EMAIL ADDRESS

neelamfakkirgoudra@gmail.com

Problem Statement Addressed

Edge AI–Based Defect Classification System for Semiconductor Wafer/ Die Images

DESCRIPTION / DETAILS

Semiconductor manufacturing produces large volumes of wafer and die inspection images. Defects in these images can reduce yield and cause failures, but traditional centralized/manual inspection creates latency, bandwidth bottlenecks, and high infrastructure cost, making it difficult to scale to real-time production needs. Our task is to develop an Edge AI–based solution capable of detecting and classifying defects in semiconductor wafer and die images using AI/ML techniques. The design must carefully balance accuracy, latency, computational efficiency, and ease of deployment on edge based hardware.

Idea Description - Describe your Idea/Solution/Prototype

KEY CONCEPT & APPROACH

The core idea is to develop a lightweight Edge-AI-based system for automatic classification of semiconductor wafer defects.
The approach uses transfer learning with a compact deep learning model (MobileNetV2) trained on wafer image data to identify and classify defects into predefined categories.
The model is optimized for low memory footprint and fast inference, enabling deployment on resource-constrained edge devices for real-time inspection.

SOLUTION OVERVIEW

The proposed solution automates wafer inspection by replacing manual visual analysis with an AI-driven classification system. Wafer images are preprocessed and passed through a trained convolutional neural network that classifies each image as clean or defective.
The trained model is exported to ONNX format for portability and edge deployment, allowing fast and reliable defect detection in semiconductor manufacturing environments while reducing inspection time, human error, and operational cost.

IESA INDIA ELECTRONICS & SEMICONDUCTOR ASSOCIATION India's ESDM Leadership and Knowledge Partner

Proposed Solution – Describe your Idea/Solution/Prototype

SOLUTION DETAILS

The proposed solution implements an end-to-end Edge-AI pipeline for automated wafer defect classification.

The methodology begins with preprocessing wafer images, including resizing, normalization, and label encoding. A lightweight convolutional neural network based on MobileNetV2 is fine-tuned using transfer learning to efficiently learn defect patterns from the dataset while maintaining low computational complexity.

The model is trained and evaluated using Python and TensorFlow/Keras, with performance measured using accuracy, precision, recall, and confusion matrix analysis. After training, the model is exported to ONNX format to ensure hardware-agnostic deployment and compatibility with edge inference engines.

This solution enables real-time defect detection on edge devices, minimizes dependency on cloud infrastructure, and supports scalable deployment in semiconductor manufacturing environments.

IESA® INDIA ELECTRONICS & SEMICONDUCTOR ASSOCIATION
India's ESDM Leadership and Knowledge Partner

Innovation and Uniqueness

KEY INNOVATION

The core innovation of this solution lies in deploying a lightweight deep learning–based wafer defect detection model directly on edge devices. By combining transfer learning with ONNX model optimization, the system enables fast and reliable inference without dependence on cloud connectivity. This approach reduces latency and computational overhead while maintaining acceptable classification performance for real-world manufacturing environments.

COMPETITIVE ADVANTAGE

Compared to traditional inspection systems and cloud-based AI solutions, this approach offers lower operational cost, reduced inference latency, and improved data privacy. The edge-deployable ONNX model allows easy integration across different hardware platforms, making the solution scalable, efficient, and suitable for continuous industrial deployment.

Impact and Benefits

Primary Impact

The proposed solution significantly improves semiconductor wafer inspection by enabling fast, automated defect detection directly at the edge. This reduces manual inspection effort, minimizes human error, and allows early identification of defective wafers, leading to better yield and improved production efficiency.

Quantifiable Outcomes

- Reduction in inspection time by up to 40–50%
- Faster inference with low-latency edge deployment
- Reduced dependency on high-end cloud infrastructure
- Improved consistency and reliability in defect classification

Technology & Feasibility/Methodology Used

IMPLEMENTATION STRATEGY

The system is implemented using a deep learning–based image classification pipeline optimized for edge deployment. Wafer images are preprocessed and resized, then passed through a lightweight CNN model (MobileNetV2) trained using transfer learning. The trained model is exported to ONNX format for efficient inference on CPU-based edge devices. The solution is designed to be modular, scalable, and deployable without reliance on cloud infrastructure.

Software Architecture

- Image preprocessing using PyTorch and torchvision
- Deep learning model training using MobileNetV2
- Model evaluation using accuracy, precision, recall, and confusion matrix
- ONNX model export for cross-platform inference
- Python-based inference pipeline for edge deployment

Hardware Components

- CPU-based edge systems (industrial PCs)
- Optional GPU for training phase only
- Compatible with low-power devices such as Intel-based edge nodes

Development Tools

- Python
- PyTorch
- ONNX Runtime
- NumPy, OpenCV
- GitHub for version control

GitHub & Video Link

GitHub Repository

https://github.com/neelamfakkirgoudra/wafer-defect-edge-ai

Research and References

Research Background & Methodology

The proposed idea is grounded in convolutional neural network (CNN) theory, which is well established for image-based pattern recognition tasks. CNNs automatically learn hierarchical visual features such as edges, textures, and defect patterns from high-resolution SEM wafer images. Transfer learning using pre-trained lightweight models enables effective feature extraction even with limited datasets while reducing computational complexity. The use of ONNX-based model optimization supports efficient inference on edge devices, aligning the solution with principles of low-latency, resource-efficient Edge AI systems widely adopted in industrial inspection applications.

References & Citations

List key papers, articles, or data sources.

Research Gate

IEEE Papers

SEM wafer images were collected and curated from publicly available academic datasets and open research sources for educational and research purposes.