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

Team Details

Team Name: TRIT

SR. NO	ROLE	NAME	ACADEMIC YEAR
1	**Team Leader**	Baalekshan E	4
2	Member 1	Anbu Saran G	4
3	Member 2	*{Enter Name}*	*{Enter Year}*
4	Member 3	*{Enter Name}*	*{Enter Year}*

A team can have up to 4 members including the team leader. Add rows if necessary.

COLLEGE NAME

RAMCO INSTITUTE OF TECHNOLOGY

TEAM LEADER CONTACT NUMBER

+91 9585668183

TEAM LEADER EMAIL ADDRESS

baalekshan@gmail.com

Problem Statement

DESCRIPTION / DETAILS

Semiconductor fabrication processes generate massive volumes of wafer and die inspection images, where even microscopic defects can severely impact yield, reliability, and performance. Traditional centralized or manual inspection approaches face challenges such as high latency, excessive bandwidth usage, costly infrastructure, and limited scalability for real-time manufacturing. To overcome these limitations, there is a need for an intelligent inspection system that performs defect analysis closer to the data source using edge computing.

This project focuses on developing an Edge AI–based defect classification system that:

- Detects and classifies defects in semiconductor wafer and die images.
- Operates with low latency and reduced bandwidth dependency.
- Uses lightweight AI/ML models suitable for edge hardware constraints.
- Supports real-time, high-volume inspection workflows.
- Enables easy deployment on edge platforms such as NXP eIQ.

Idea Description

KEY CONCEPT & APPROACH

The core idea of this project is to implement an **Edge AI–based defect inspection system** for semiconductor wafer and die images, where defect detection and classification are performed directly at the edge. A **lightweight YOLO-based deep learning model** is trained to identify multiple microscopic defect types from inspection images and optimized for low-latency inference. By processing data close to the source, the approach minimizes dependency on centralized infrastructure while remaining suitable for deployment on edge platforms such as **NXP eIQ**.

SOLUTION OVERVIEW

The proposed solution addresses the limitations of traditional inspection systems by enabling **real-time, low-latency defect analysis** at the manufacturing edge. Local inference significantly reduces bandwidth usage by transmitting only defect information instead of raw images, while the compact model architecture ensures compatibility with resource-constrained edge hardware. This design supports high-volume inspection workflows, improves scalability, and enables faster decision-making in semiconductor fabrication, directly enhancing yield and process reliability.

Proposed Solution –Edge AI–Based Semiconductor Defect Inspection

SOLUTION DETAILS

Due to the unavailability of real SEM defect images in the public domain, a **custom synthetic data generation pipeline** was developed to closely mimic real semiconductor inspection imagery. Instead of relying solely on AI-generated diffusion data, backend **VLSI GDS layout files** were used as a base, and advanced image processing techniques were applied to transform smooth layout crops into **SEM-like images**. Additional realistic defects were manually introduced to improve dataset diversity and balance.

Key Steps and Implementation

- Converted GDS layout crops into SEM-like images using grayscale conversion, Gaussian and Poisson noise, and edge roughness/LER modeling.
- Manually created defects (Open, Short, Crack, LER, Protrusion) using Paint and Photoshop; supplemented with Zenodo and DeepPCB datasets.
- Annotated all images using Label Studio and trained a **YOLO-based detection model** at 256×256 resolution on an NVIDIA RTX 5050 GPU.
- Achieved ~90% classification accuracy, ~92% detection correctness with a compact **~5.9 MB model.**
- Exported the optimized model to ONNX for **edge deployment** on platforms such as **NXP i.MX RT1170 using NXP eIQ.**

Innovation and Uniqueness

KEY INNOVATION

- **SEM-like synthetic data generation:** Generated realistic SEM-style images from **VLSI GDS layout crops** using grayscale conversion, noise injection, and edge/line edge roughness modeling.
- **Hybrid dataset strategy:** Combined synthetic images, manually created defect patterns, and limited public datasets (Zenodo, DeepPCB) to improve robustness and coverage of rare defects.
- **YOLO-based detection approach:** Used object detection instead of pure classification to enable **both defect localization and classification**.
- **Defect-aware annotation design:** Designed class definitions and annotations to handle **visually overlapping defects** such as crack and open.
- **Edge-optimized lightweight model:** Achieved a compact **~5.9 MB model** optimized for **low-latency edge deployment**.

COMPETITIVE ADVANTAGE

- **SEM-realistic training data:** Trained on a custom **SEM-like synthetic dataset** derived from VLSI GDS layouts, enabling learning without dependence on confidential fab data.
- **Compact edge-ready model:** Lightweight YOLO-based detector with a **~5.9 MB size (without quantization)**, suitable for memory-constrained edge devices.
- **Detection with localization:** Performs **both defect detection and localization**, unlike classification-only approaches, improving inspection reliability.
- **Low-latency edge inference:** Designed for real-time operation, supporting **inline inspection workflows** in semiconductor manufacturing.
- **Portable and scalable deployment:** Exported in **ONNX format**, enabling plug-and-play deployment on edge platforms such as **NXP eIQ**.

Impact and Benefits

PRIMARY IMPACT

- **Edge-level real-time defect detection**, enabling faster inspection decisions and improved manufacturing yield.
- **Immediate identification of critical defects** (e.g., crack vs open) during wafer and die inspection.
- **Reduced inspection latency** by eliminating dependency on centralized processing pipelines.
- **Lower infrastructure and operational costs** through local inference and reduced data movement.
- **Scalable inspection framework** capable of supporting high-volume semiconductor manufacturing without performance bottlenecks.

QUANTIFIABLE OUTCOMES

- **Reduction in data transfer bandwidth** by performing inspection directly at the edge.
- **Faster inspection turnaround time** compared to centralized or manual inspection workflows.
- **~5.9 MB model size (without quantization)** enabling deployment on memory-constrained edge hardware.
- **Low Inference latency per image** on edge-accelerated platforms.
- **~90% defect classification accuracy** with **~92% detection correctness** for critical defects such as cracks, opens, and shorts.
- **Reliable differentiation of visually similar defects** (e.g., crack vs open) within the same inspection image.

Technology & Feasibility/Methodology Used

TECHNOLOGY & FEASIBILITY

- **Lightweight detection model:** YOLOv8-Nano selected for **simultaneous defect detection and classification** with a low memory footprint suitable for edge inference.
- **Edge deployment pipeline:** Model trained using PyTorch and exported to **ONNX**, enabling compatibility with edge runtimes such as **NXP eIQ**.
- **SEM-realistic synthetic dataset:** Custom dataset generated from **VLSI GDS layout crops** and enhanced using image processing to mimic SEM characteristics, reducing dependency on real fab data.
- **Edge-ready hardware support:** Designed to operate on **resource-constrained edge SoCs** with CPU/NPU acceleration.
- **High implementation feasibility:** Built using **mature, open-source frameworks** and validated on local GPU hardware, ensuring scalability and ease of deployment.

METHODOLOGY USED

- **SEM-like data generation:** Generate inspection images from **VLSI GDS layout crops** and apply image processing techniques (grayscale conversion, noise injection, edge and line edge roughness modeling) to approximate SEM characteristics.
- **Defect creation and annotation:** Manually introduce defects (Open, Short, Crack, LER, Protrusion) and annotate all images using **Label Studio**.
- **Detection-based model training:** Train a **YOLO-based multi-class detector** to perform simultaneous defect localization and classification.
- **Model optimization:** Optimize the trained model for **low-latency edge inference** & maintaining detection accuracy.
- **Edge deployment and inference:** Export the model to **ONNX** and perform real-time defect analysis directly at the edge, transmitting only defect metadata.

GitHub & Video Link

GitHub Repository

https://github.com/Baalekshan/TRIT

Prototype / Simulation Video

{Paste your Video Link here showing simulation or working prototype}

Research and References

Research Background & Methodology

- Semiconductor wafer and die inspection requires precise detection of microscopic defects to ensure yield and reliability.
- Centralized inspection systems introduce high latency, bandwidth overhead, and limited scalability.
- SEM-like synthetic defect datasets are generated from VLSI layout data using image processing techniques.
- A lightweight YOLOv8-Nano model is trained for simultaneous defect detection and classification.
- The optimized model is deployed at the edge for real-time inspection with reduced data transfer.

References & Citations

1. **Redmon et al., "You Only Look Once: Unified, Real-Time Object Detection"** – Introduces the YOLO framework for real-time object detection.
   🔗 https://arxiv.org/abs/1506.02640
2. **Li et al., "Automatic Wafer Defect Detection Using Deep Learning"** – Discusses deep learning–based defect detection in semiconductor manufacturing.
   🔗 https://www.sciencedirect.com/science/article/pii/S240584402501761X
3. **Xu et al., "Edge AI for Industrial Inspection: Challenges and Opportunities"** – Explores the role of Edge AI in real-time industrial inspection systems.
   🔗 https://www.sciencedirect.com/science/article/pii/S016636152100164X