QUANTA

Detect . Decide . Deploy

Edge-AI for Real-Time Semiconductor Defect Detection

IESA DeepTech Hackathon 2026

(Phase 1 submission)

Team Name : Quanta

Team members :

Member 1	Sneha BM	2024-28
Member 2	Namit Vajragiri	2024-28
Member 3	Nishitha S	2024-28
Member 4	Siri Sanikop	2024-28

College name : PES University , Ring Road campus, Bangalore

Team contact number : 9108465162

Team email ID : snehaamallikarjun@gmail.com

Problem statement :

- Semiconductor fabs generate **large volumes of wafer/die inspection images** using SEM, AFM, and optical tools
- Microscopic defects such as **bridges, cracks, opens, CMP damage, and LER** can cause yield loss and failures
- Traditional centralized/manual inspection leads to:
  - High latency
  - Expensive infrastructure
  - Network/bandwidth bottlenecks
  - Poor scalability for real-time production
- As technology nodes shrink, **faster and more accurate inspection becomes critical**
- Therefore, **Edge-AI based on-device defect classification** is needed for real-time, low-latency decision-making on the fab floor

KEY CONCEPT & APPROACH

- Edge-AI system for wafer/die defect classification
- **EfficientNet-B0 CNN** trained on grayscale SEM images
- Classifies multiple semiconductor defect types (bridges, cracks, CMP, LER, etc.)
- Designed for low-latency edge inference

SOLUTION OVERVIEW

- Preprocessing -> EfficientNet-B0 pipeline
- Transfer learning for faster convergence
- Real-time, on-device defect prediction
- ONNX export for edge deployment
- Reduces latency and cloud experience
- CNN- based methodology for defect classification

DATASET PLAN & CLASS DESIGN

- Edge- optimised EfficientNet deployment for fab environments
- Custom-curated SEM-style dataset for realistic effect scenarios

Total images : **1087**

Number of Classes : **8** (6 defects + clean + other defects)

Classes :
- Bridges
- Gaps
- CMP(Chemical mechanical polishing)
- Cracks
- LER(Line EDge roughness)
- Foreign Material
- Clean
- Other Defects

Data Split: Train / Val / Test = 734 / 169 / 186

Image Type: Grayscale SEM-style images

Labelling source :Manually labeled dataset curated from academic literature, educational resources, and representative SEM-style defect imagery, supplemented with synthetic defect generation for class balancing

Class Balance: Targeted class balance with ~150 images per class

BASELINE MODEL & RESULTS

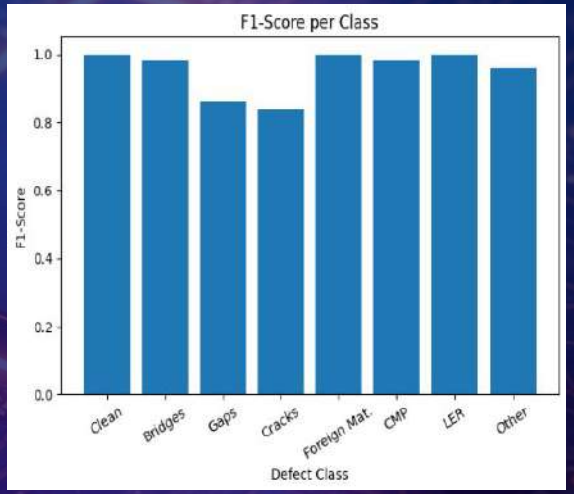
Model: EfficientNet-B0
Training: Transfer Learning
Framework: PyTorch
Input Size: 224 × 224
Model Size: 15.6 MB

Test Performance

Accuracy: **95%**
Macro F1-Score: 0.95
Weighted F1-Score: 0.95

Classification report and Confusion Matrix

```
Classification Report:

                  precision    recall  f1-score   support

           clean       1.00      1.00      1.00        25
         bridges       1.00      0.96      0.98        24
            gaps       0.77      0.96      0.86        25
         cracks        1.00      0.72      0.84        25
foreign_materials      1.00      1.00      1.00        25
             cmp       0.96      1.00      0.98        25
             ler       1.00      1.00      1.00        25
           other       0.92      1.00      0.96        12

        accuracy                           0.95       186
       macro avg       0.96      0.95      0.95       186
    weighted avg       0.96      0.95      0.95       186

Confusion Matrix:

[[25  0  0  0  0  0  0  0]
 [ 0 23  0  0  0  1  0  0]
 [ 0  0 24  0  0  0  0  1]
 [ 0  0  7 18  0  0  0  0]
 [ 0  0  0  0 25  0  0  0]
 [ 0  0  0  0  0 25  0  0]
 [ 0  0  0  0  0  0 25  0]
 [ 0  0  0  0  0  0  0 12]]
```



Edge Suitability / Feasibility

- Lightweight CNN suitable for edge deployment
- Exported to ONNX format
- Supports low-latency inference
- Optimized for grayscale SEM images
- Enables faster inspection and reduced latency in defect detection workflows

Platform used for Training and Inference :

Model training and inference were conducted on a **local Windows-based workstation** using Python 3.10 and the PyTorch framework. **CUDA acceleration** was utilized on an **NVIDIA RTX 4060 GPU**.

No cloud or external infrastructure was used.

Note :

Our system supports unknown defect rejection using **confidence-based open-set handling**. Any sample that does not confidently belong to known defect classes is classified as other.

ARTIFACTS & LINKS

Repository & Code

GitHub Repository: https://github.com/SnooMarzians/sem-defect-classifier

Dataset and Model :
https://drive.google.com/file/d/1NJlA6dN__Vagi_tIThfP5y0t3T5zvsEn/view?usp=sharing

References :

- Tan & Le (2019) — EfficientNet (ICML)
- H. Lee et al. — IEEE wafer defect classification
- Nishi & Doering — Semiconductor Manufacturing Technology Textbook

Thank you