IESA DeepTech Hackathon

Team: E-LITE

SR. NO	ROLE	NAME	ACADEMIC YEAR
1	Team Leader	Vishal Singh Kushwah	1st year
2	Member 1	Shaurya Singh	1st year
3	Member 2	Md Saad Khan	1st year
4	Member 3	Dhruv Choudhary	2nd year

COLLEGE NAME

Indian Institute of technology - Madras

TEAM LEADER CONTACT NUMBER

+91 9773465661

TEAM LEADER EMAIL ADDRESS

24f3100291@es.study.iitm.ac.in

IESA DeepTech Hackathon

Problem Statement Addressed

Semiconductor Wafer Defect Classification (Edge-AI):

Semiconductor fabrication involves hundreds of tightly controlled steps where microscopic defects can significantly reduce yield or cause functional failure. Modern fabs generate terabytes of inspection images per day, making centralized or manual analysis impractical.

Key challenges with current inspection approaches:

- *High latency due to centralized processing*
- *Network bandwidth bottlenecks*
- *High infrastructure cost (GPU/cloud)*
- *Poor scalability for real-time production lines*

Need:

A low-latency, edge-deployable AI system capable of classifying wafer/die defects in real time, aligned with Industry 4.0 manufacturing.

Semiconductor wafer inspection requires fast and accurate defect detection under strict resource constraints. Existing solutions are costly and unsuitable for real-time edge deployment. This project addresses the challenge by developing a lightweight CNN-based Edge-AI model capable of classifying multiple wafer defect types with high accuracy and low memory footprint, optimized for deployment using ONNX on NXP edge platforms.

IESA DeepTech Hackathon

DATASET INFORMATION – Dataset Plan & Class Design

The dataset consists of 1,188 grayscale wafer images across 8 classes (6 defect types + Clean + Other), balanced using weighted loss and augmentation, split 50/50 for training and validation, and optimized for edge deployment. (Google Drive): Click here to download dataset

📂 Training Set Summary

	Class Name	Number of Images
0	Bridge	50
1	Clean	48
2	Cmp-Scratch	50
3	Crack	80
4	Mal-Farmed-Vias	100
5	Other	73
6	Oxide	130
7	Pattern-Collapse	55

Total Training Set Images: 586

📂 Validation Set Summary

	Class Name	Number of Images
0	Bridge	50
1	Clean	48
2	Cmp-Scratch	50
3	Crack	80
4	Mal-Farmed-Vias	100
5	Other	73
6	Oxide	130
7	Pattern-Collapse	55

Total Validation Set Images: 586

DATASET STRUCTURE:

```
Train/
├── Bridge/
├── Clean/
├── CMP-Scratch/
├── Crack/
├── Mal-Farmed-Vias/
├── Other/
├── Oxide/
└── Pattern-Collapse/

Validation/
└── Same structure as Train/
```

IESA DeepTech Hackathon

Proposed Solution - Baseline Model, Architecture

Lightweight CNN-based solution for semiconductor wafer defect classification, optimized for edge deployment on NXP i.MX RT series devices.

MODEL ARCHITECTURE

Model Name: phase1Model.ipynb

A custom edge-optimized convolutional neural network built for real-time semiconductor wafer inspection.

Key Features:

- *Batch normalization applied after each convolution layer*
- *Hardware-efficient ReLU activation functions*
- *Max pooling layers for progressive spatial reduction*
- *Adaptive average pooling for flexible feature aggregation*
- *Designed to support INT8 quantization for efficient deployment on edge hardware*

MODEL STATISTICS

- *Total Parameters: 1,475,912*
- *Trainable Parameters: 1,475,912*
- *FP32 Model Size: 5.63 MB*
- *INT8 Model Size: ~1.41 MB*
- *ONNX Compatible: Yes*

TRAINING CONFIGURATION

- *Epochs: 30*
- *Batch Size: 16*
- *Optimizer: AdamW*
- *Learning Rate: 0.1176*
- *Scheduler: ReduceLROnPlateau*
- *Loss: Class-weighted CrossEntropyLoss*
- *Training Device: CPU*
- *Framework: PyTorch 2.1*
- *Training Time: ~30-60 minutes*

EDGE DEPLOYMENT

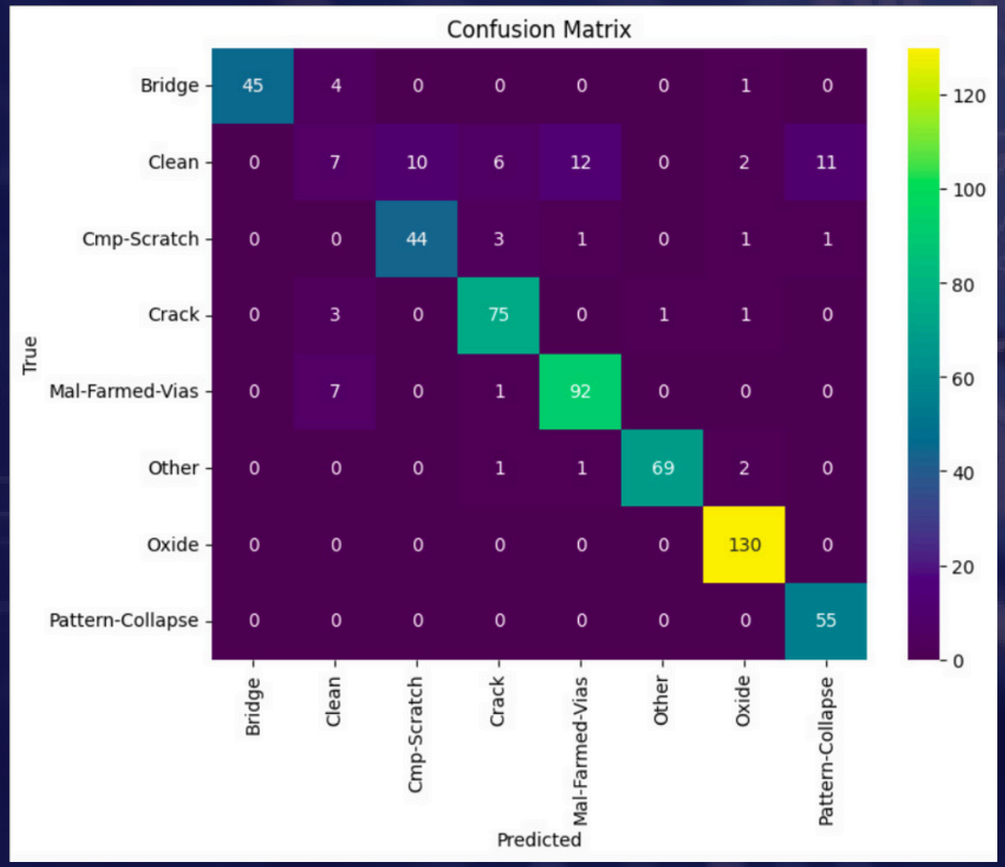
- *Target Hardware: NXP i.MX RT Series*
- *Framework: NXP eIQ Toolkit*
- *Model Format: ONNX*
- *Quantization: INT8-ready*

IESA DeepTech Hackathon

Impact and Results

CLASSIFICATION REPORT AND CONFUSION MATRIX:

Class	Precision	Recall	F1-Score	Support
Bridge	100.00%	92.00%	0.9583	50
Clean	47.62%	41.67%	0.4444	48
CMP Scratch	90.00%	90.00%	0.9000	50
Crack	95.00%	95.00%	0.9500	80
Mal-Formed Vias	93.62%	88.00%	0.9072	100
Other	97.26%	97.26%	0.9726	73
Oxide	94.85%	99.23%	0.9699	130
Pattern Collapse	84.62%	100.00%	0.9167	55

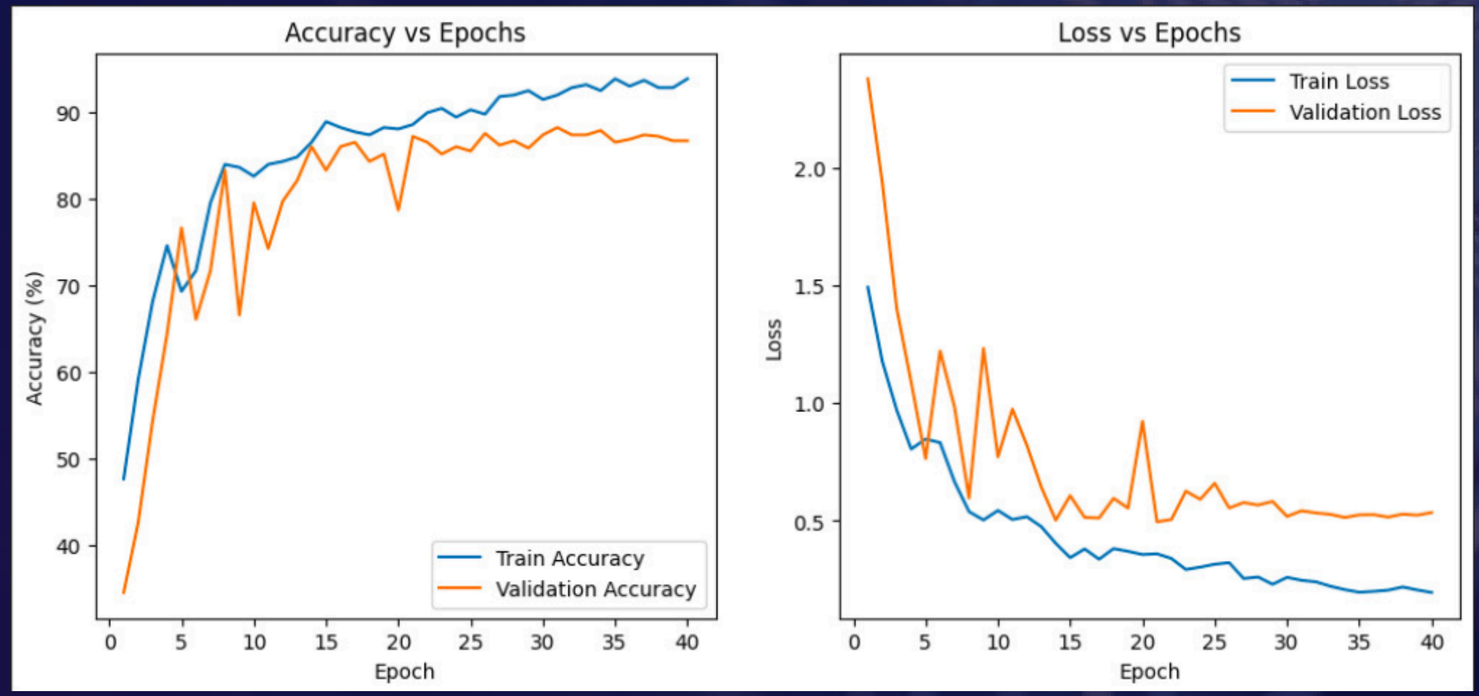


IESA DeepTech Hackathon

Results & Visualizations

Project Summary:

- Test Accuracy: 92.83%
- Validation Accuracy: 90.27%
- Classes: 8 defect types
- Dataset Size: ~1200 images
- Input: Grayscale (64x64, resized during training)
- ONNX Model Size: 5.63 MB
- INT8 Model Size (Estimated): ~1.41 MB



IESA DeepTech Hackathon

GitHub & Dataset Link

Dataset drive

Dataset Drive Link

GitHub Repository

GitHub Source Code Link

ONNX model link:

ONNX Modal Link

IESA DeepTech Hackathon

Research and References

Research Background & Methodology

- *CNN-based visual inspection: Convolutional Neural Networks are widely proven for automated defect detection in semiconductor and industrial inspection due to strong spatial and texture feature learning.*
- *Grayscale SEM image processing: SEM images are inherently structural; using grayscale inputs reduces computation while preserving critical defect information.*
- *Lightweight edge-focused architecture: The model follows embedded CNN design principles (SEMNet-style) to balance accuracy and memory for edge deployment.*
- *Class imbalance handling: Weighted cross-entropy loss is used to improve minority defect and "Clean" class recognition, a standard practice in industrial ML systems.*
- *Edge optimization via quantization: INT8 post-training quantization applies established model compression theory, achieving major size reduction with minimal accuracy loss for NXP eIQ deployment.*

References & Citations

NXP eIQ Toolkit

Sample images and documentation

Semiconductor Wafer Particle Defect Classification Based on Deep Learning and Multimodal Feature Fusion