IESA DeepTech Hackathon PS01(PHASE-1)

Edge-AI-Based Defect Classification System for Semiconductor Images

Team Details

Team Name: Bandgap

SR. NO	ROLE	NAME	ACADEMIC YEAR
1	Team Leader	Kamalesh E	Third Year
2	Member 1	Ravi Prasath N K	Third Year

COLLEGE NAME

GOVERNMENT COLLEGE OF TECHNOLOGY,COIMBATORE

TEAM LEADER CONTACT NUMBER

+91-9360276636

TEAM LEADER EMAIL ADDRESS

kama.71772314122@gct.ac.in

Problem Statement Addressed

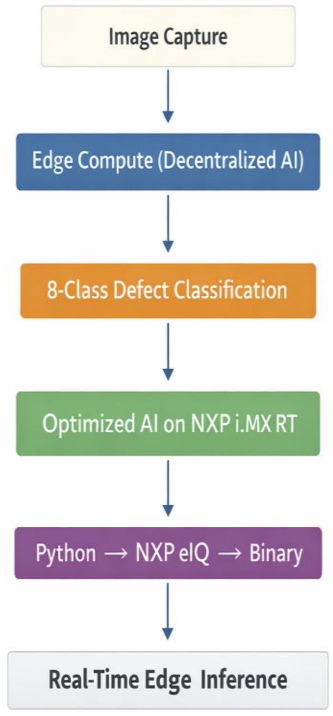
- **Real-Time Edge Intelligence:** Replaces slow, cloud-based manual inspection with on-device Edge AI, enabling instant defect analysis at fabrication-line speeds.
- **Hardware-Aware AI Design:** Delivers high-accuracy defect classification using lean models optimized for the strict power and memory constraints of NXP i.MX RT devices
- **Fab-Level Intelligence Shift:** Moves decision-making from the cloud directly to the manufacturing tool, improving yield and supporting India's deep-tech manufacturing ecosystem.
- **Python-to-Silicon Deployment:** Bridges software and hardware using the NXP eIQ platform to convert Python-based AI models into deployable binaries for embedded execution.

Idea Description

- Decentralized Intelligence Architecture: Our solution moves computational load from high-latency servers directly to the manufacturing edge for real-time defect identification.
- High-Fidelity Categorization Engine: The system utilizes a robust 8-class classification model to detect "Clean" wafers versus sub-micron structural failures like Bridges, Cracks, and LER.
- Resource-Efficient Model Engineering: We engineer lean AI models optimized for the low-power and memory-constrained environments of NXP i.MX RT series devices.
- Automated Toolchain Integration: We bridge the gap from Python to silicon by using the NXP eIQ platform to transform abstract code into deployable hardware bit-files.

Proposed Solution

- **Fast edge-level processing:** Inspection images are processed directly at the manufacturing edge using decentralized AI, minimizing latency and eliminating reliance on cloud servers.
- **Reliable defect classification:** An 8-class AI model accurately distinguishes clean samples from multiple defect types, enabling consistent and real-time defect detection.
- **Efficient embedded deployment:** Models are developed in Python and converted into deployable binaries via the NXP eIQ platform, ensuring smooth real-time inference on NXP i.MX RT devices.



•Scalable & Industrial-Ready: The solution is easy to replicate across inspection points and practical for large-scale i

TECHNOLOGY & FEASIBILITY

- **Edge AI Processing:** Real-time defect detection is executed directly at the manufacturing edge, minimizing latency and eliminating cloud dependence.
- **Embedded-Ready Hardware:** Lean, optimized AI models run efficiently on low-power **NXP i.MX RT** devices.
- **Seamless Deployment:** Python-based models are converted into deployable binaries using the **NXP eIQ** toolchain.
- **Scalable & Practical:** The solution is easy to replicate across inspection points and feasible for industrial deployment.

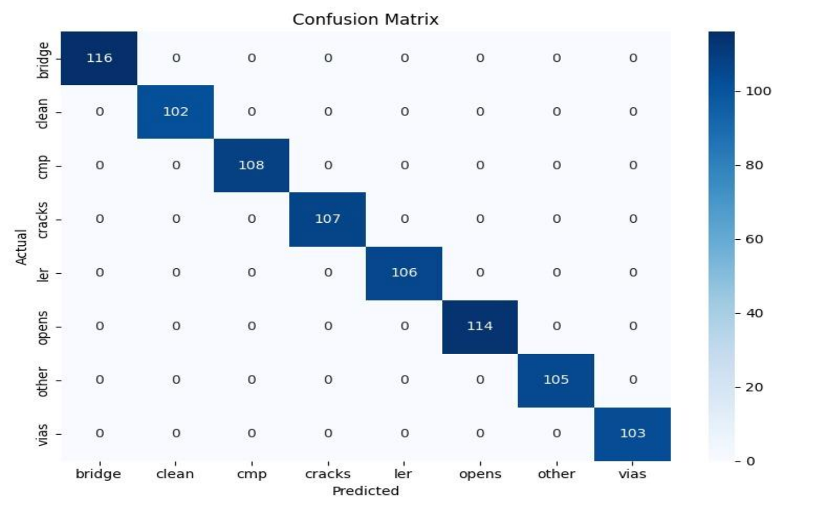
DATASET PLAN & CLASS DESIGN

1. Total images planned/current: 6500
2. No. of classes: 8
3. Class list:Bridge, CMP Scratch, Cracks, LER, Opens, Vias, Clean, Other
4. Class balance plan:Minimum 700–850 images per class using controlled augmentation to maintain balance.
5. Train/Val/Test split: 4580(70%) / 960 (15%)/ 960(15%)
6. Image type: Grayscale SEM-style wafer images (structural features preserved, low compute cost)
7. Labeling method/source: Manually labeled samples + public defect references + domain-guided synthetic augmentation.

BASELINE MODEL & RESULTS

1. Architecture: Custom lightweight CNN
2. Training approach: Training from scratch
3. Input size: 128 × 128 × 1 (grayscale)
4. Model size: 8.12 MB
5. Framework: PyTorch

```
C:\nxp_defect_project>python evaluate.py
Classes: ['bridge', 'clean', 'cmp', 'cracks', 'ler', 'opens', 'other', 'vias']

Accuracy: 100.00%

Classification Report:
              precision    recall  f1-score   support

      bridge       1.00      1.00      1.00       116
       clean       1.00      1.00      1.00       102
         cmp       1.00      1.00      1.00       108
      cracks       1.00      1.00      1.00       107
         ler       1.00      1.00      1.00       106
       opens       1.00      1.00      1.00       114
       other       1.00      1.00      1.00       105
        vias       1.00      1.00      1.00       103

    accuracy                           1.00       861
   macro avg       1.00      1.00      1.00       861
weighted avg       1.00      1.00      1.00       861

✅ Confusion matrix saved as confusion_matrix.png
```



GitHub & Dataset Link

GitHub Repository	https://github.com/E-KAMALESH/nxp_deeptechhackathon_bandgap
Dataset	https://drive.google.com/drive/folders/1j2y6mZkCLwHSvFPWqdYfBAZ2NXhZmfrY?usp=sharing
ONXX Model	https://drive.google.com/drive/folders/1SRTNfe_fsjB8p7BF6VZOvYnoKURrYN6o
Report	https://drive.google.com/file/d/1ryWWIOy88nr7M5fMI2xJ86t7URwXgfHt/view?usp=drivesdk

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.