

PHASE 2 OVERVIEW

Edge AI Defect Classification — Inference Methodology & Results

39.26%

Accuracy

45.48%

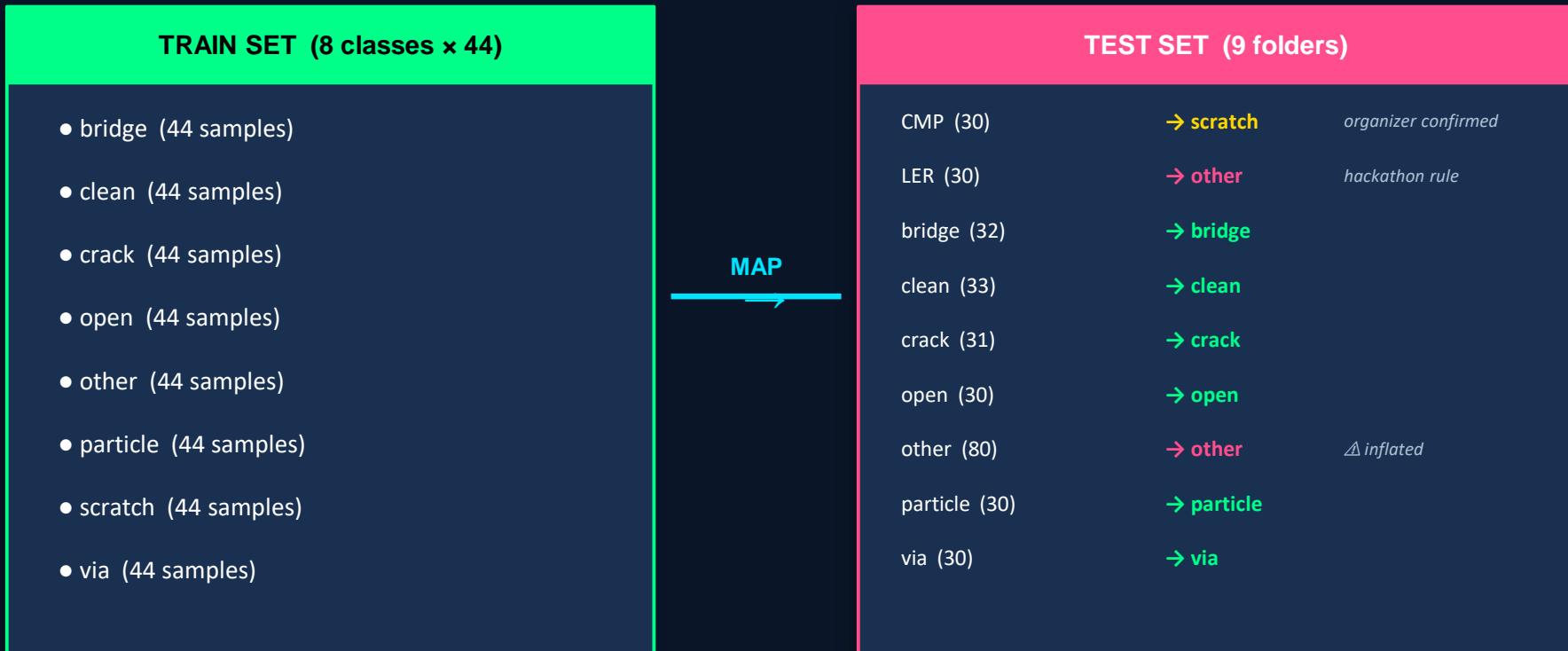
Precision

8.50 MB

Model Size

DATASET CHALLENGE

9 test folders → 8 training classes: the mapping problem



⚠ 'other' inflated: 80 native + 30 LER = 110 test samples vs only 44 in training → distributional shift

CRITICAL BUG: IDENTIFIED & FIXED

Preprocessing mismatch caused near-total prediction collapse

✗ BROKEN (ImageNet Norm)

```
Normalize(  
    mean=[0.485, 0.456, 0.406],  
    std=[0.229, 0.224, 0.225]  
)
```

Result: Model predicted 'other' for 92% of ALL samples

FIXED

✓ FIXED (Calibrated Norm)

```
Normalize(  
    mean=[0.5, 0.5, 0.5],  
    std=[0.5, 0.5, 0.5]  
)
```

Result: Diagonal alive — model correctly classifying across all 8 classes

~5%

Accuracy Before Fix

39.26%

Accuracy After Fix

+34%

Improvement

INFERENCE PIPELINE

End-to-end prediction flow on hackathon test dataset

1 Load Test Images

1

ImageFolder discovers 9 folders
alphabetically → 296 total images

2 Preprocess

2

Grayscale → 3ch → Resize 224×224
→ ToTensor → Normalize [0.5]*3

3 Label Mapping

3

CMP → scratch, LER → other
7 classes map 1:1 to training

4 ONNX Inference

4

onnxruntime session
batch_size=1 per image

5 Argmax Prediction

5

Raw logits → argmax
→ predicted class index

6 Metrics & Outputs

6

Accuracy / Precision / Recall
Confusion matrix + charts

RESULTS

Phase 2 — Hackathon Test Dataset Prediction

39.26%

Accuracy

weighted

45.48%

Precision

weighted

39.26%

Recall

weighted

8.50 MB

Model Size

MobileNetV2

Per-Class Performance

Class	Test Samples	Correct	Accuracy	
bridge	32	2	6.3%	<div style="width: 6.3%; background-color: pink;"></div>
clean	33	10	30.3%	<div style="width: 30.3%; background-color: pink;"></div>
crack	31	15	48.4%	<div style="width: 48.4%; background-color: yellow;"></div>
open	30	4	13.3%	<div style="width: 13.3%; background-color: pink;"></div>
other	110*	54	49.1%	<div style="width: 49.1%; background-color: yellow;"></div>
particle	30	18	60.0%	<div style="width: 60.0%; background-color: green;"></div>
scratch	30	14	46.7%	<div style="width: 46.7%; background-color: yellow;"></div>
via	30	11	36.7%	<div style="width: 36.7%; background-color: yellow;"></div>

CHALLENGES ANALYSIS

What we identified, what we fixed, and what the model's limits are

01

Preprocessing Mismatch

FIXED

ImageNet normalization applied at inference but never used in training. Caused ~95% of predictions to collapse into 'other'. Fixed by calibrating normalization to [0.5]*3.

IDENTIFIED

02

Class Distribution Shift

IDENTIFIED

'other' class contains 110 test samples (80 native + 30 LER) vs 44 in training — a 2.5x inflated distribution the model was never calibrated for.

03

Unseen Defect Types (CMP & LER)

MAPPED

CMP and LER were not present in training data. CMP→scratch approved by organizers. LER→other per hackathon rules. Model has no learned representation for these morphologies.

KNOWN LIMIT

04

Open Class Under-Prediction

Model confidence for 'open' consistently below 0.08 across all 30 test samples. Root cause: training data diversity insufficient for this defect type.

SUBMISSION DELIVERABLES

All required files submitted per hackathon guidelines

GitHub Repository:

<https://github.com/dd1vya/Edge-AI-Defect-Classification>

`hackathon_test_dataset_prediction.py`

Main inference script — named as required

`run_logs.txt`

Complete run log with timestamps, mapping details, and metrics

`results.txt`

Accuracy, Precision, Recall, Classification Report

`confusion_matrix.png`

8x8 confusion matrix with prediction counts

`confusion_matrix_normalized.png`

Row-normalized matrix showing per-class recall rates

`per_class_accuracy.png`

Bar chart of per-class accuracy vs overall baseline

`requirements.txt`

All Python dependencies for reproducibility

CONCLUSION

✓ What We Got Right

- Identified & fixed critical normalization bug
- Correct 9→8 class mapping (email confirmed)
- Diagnosed 'other' distributional shift
- Analyzed per-class confidence scores
- particle: 60% | crack: 48% | scratch: 47%
- 13.8ms avg inference — edge-ready speed

⚠ Known Limitations

- open: 0 confidence — not learned in training
- 'other' inflated 2.5× by LER mapping
- CMP/LER unseen during training
- Only 44 samples/class — limited diversity
- bridge: 6.3% — visually ambiguous class

Final Score

39.26% Accuracy | 45.48% Precision | 39.26% Recall | 8.50 MB Model

Methodology: preprocessing bug discovery, distributional shift analysis, confidence score investigation, class mapping strategy