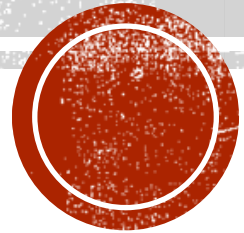
ROBUST MODELS FOR DETECTING BRIDGE DEFECTS UNDER NATURAL DISTRIBUTION SHIFTS

Team 4: Eric Rodriguez, Nicholas Miller, Kanitta Srichan, Daniel Anthony



MOTIVATION

- Traditional vs. Machine Learning Approach

Aspect	Traditional Methods	Our ML Approach
Efficiency	Labor-intensive, time-consuming	Automated, faster image analysis
Accuracy	Prone to human error	Improved defect detection accuracy
Scalability	Limited by manual effort	Easily scalable with additional data
Defect Detection	Subjective, based on inspector's experience	Consistent, data-driven analysis
Challenges	High cost, inconsistent outcomes	Addressed dataset quality and generalization
Real-World Deployment	Difficult to standardize	Robust to distribution shifts with diverse data

DEFINING DISTRIBUTION SHIFT & ROBUSTNESS

Components of Natural Distribution Shift present our data:

- Background color and Shading
- Focus
- Size/shape of crack

Standard Data (D)

7002-48.jpg

7002-49.jpg

7002-63.jpg

7002-64.jpg

Shifted Data (P)

095-86.jpg

095-87.jpg

095-94.jpg

095-95.jpg

- Relative Robustness = Measure of Robustness to Natural Distribution Shift
  - Difference between Accuracy of Model with Intervention and Accuracy of Model without Intervention
  - Interventions used: Size, Diversity, Arch

Robustness to Adversarial Attacks

- Gradient Sign Attack aka Fast Gradient Sign Method (FGSM)
- Implemented with Foolbox
  - Adds perturbations to the image by evaluating how each pixel contributes to the loss
  - Uses the gradients of the loss with respect to the input image to create a new image that maximizes the loss

EXPERIMENT

- (3) Training Sets
  - A = Large & Diverse (D & P)
  - B = Small & Diverse (D & P)
  - C = Small & Standard (D only)

- (2) Test Sets
  - D = Standard
  - P = Shifted

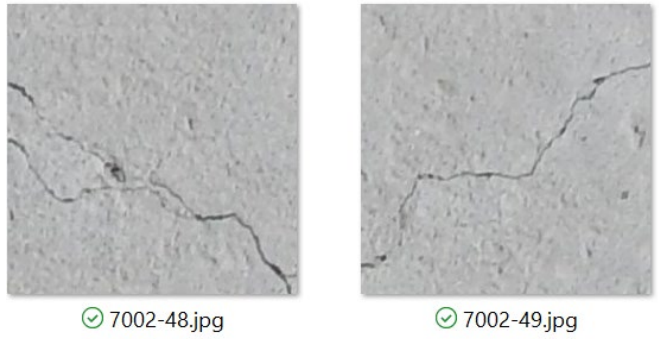
- (3) Interventions
  - Size: Compare A to B
  - Diversity: Compare B to C
  - Architecture: Compare A' to A

- (2) Architectures: Baseline (A) & Experimental (A')
  - For each architecture, a model is trained on each training set and evaluated on the two test sets

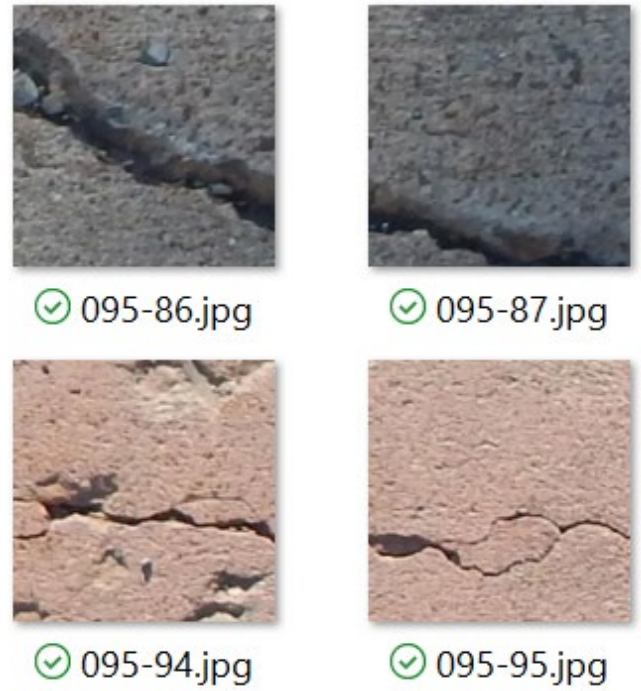
Table 3: Training Dataset statistics after preprocessing.

Training Set	Number of D Images	Number of P Images
A	5,141	9,733
B	514	973
C	5,141	0

D: Standard Images



P: Shifted Images



MODEL ARCHITECTURE & METHODS

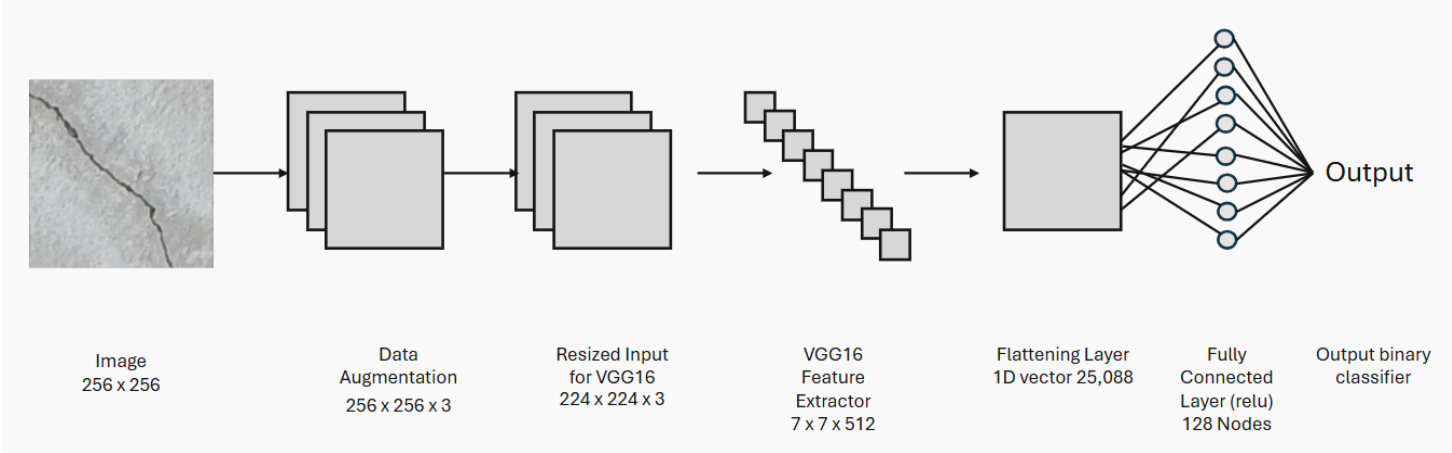


Fig.1 Base model

- Regularization Techniques:
  - Data Augmentation
  - Early Stopping
  - Dropout

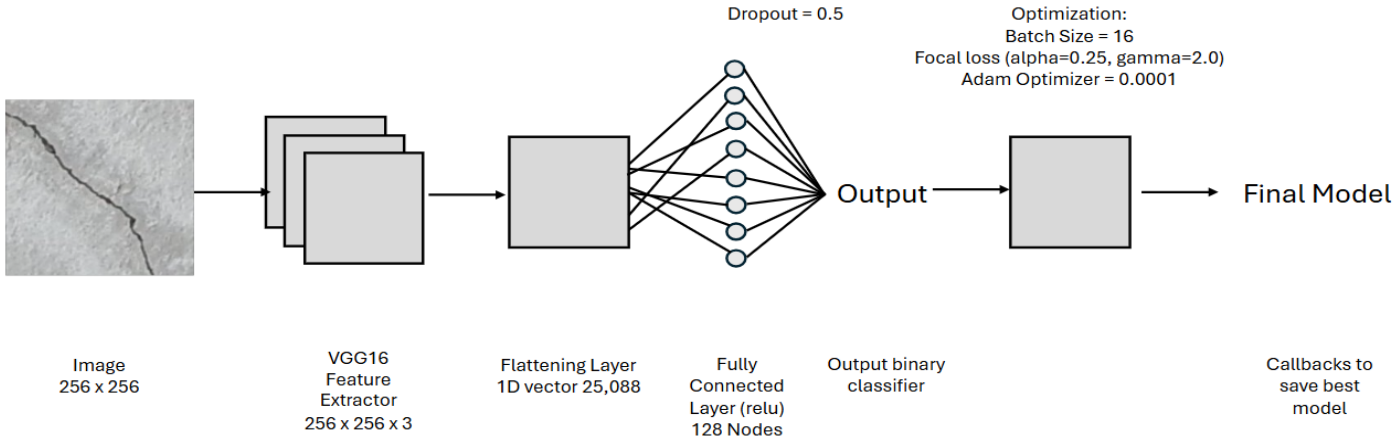
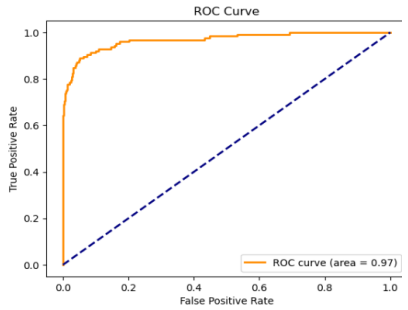


Fig.2 Experimental model

- Model Improvements:
  - Additional Fine-tuning of final 10 layers
  - Threshold optimization

RESULTS/CONCLUSION

- Training on Larger Datasets *improves* model performance (on experimental arch only) but *decreases* robustness to adversarial attacks
- Training on Diverse Datasets improves robustness to distribution shifts but decreases traditional performance metrics (accuracy)
- Additional fine-tuning improves model performance
- Transfer learning is a viable approach for binary image classification with structural health monitoring (SHM) tasks



Best model could achieve up to 97% accuracy

Table 4: Baseline Architecture Performance: Accuracy and Robustness Metrics

Model	Train-Test Combination	Accuracy (%)	Adversarial Robustness
A	Best_Model_Transfer_A_to_D	96.0	0.31
A	Best_Model_Transfer_A_to_P	90.0	0.52
B	Best_Model_Transfer_B_to_D	93.0	0.51
B	Best_Model_Transfer_B_to_P	90.0	0.49
C	Best_Model_Transfer_C_to_D	97.0	0.16
C	Best_Model_Transfer_C_to_P	87.0	0.24

Table 5: Experimental Architecture Performance: Accuracy and Robustness Metrics

Model	Train-Test Combination	Accuracy (%)	Adversarial Robustness
A'	Best_model_transfer_A_to_D	97.0	0.39
A'	Best_model_transfer_A_to_P	95.0	0.40
B'	Best_model_transfer_B_to_D	93.0	0.81
B'	Best_model_transfer_B_to_P	93.0	0.57
C'	Best_model_transfer_C_to_D	97.0	0.34
C'	Best_model_transfer_C_to_P	91.0	0.29

Table 6: Relative Robustness Chart

Intervention	Models Compared	Relative Robustness
Size	A - B on P	0.0
Size	A'-B' on P	2.0
Diversity	B - C on P	3.0
Architecture	A' - A on P	5.0

FUTURE WORK

- Cost Analysis: Is the increase from 87% to 91% on the shifted dataset via fine-tuning more cost-effective than procuring a training dataset with diversity or size and achieving accuracies of 93-95%
- Include an evaluation of effective robustness (Taori et al., 2020)
- Include a comparison of adversarial attacks
- Include additional techniques for improving the baseline model
- Additional classes of structural defects such as spalling and corrosion
- Prepare for distribution shifts caused by weather or lighting