.

CrackVision: A Bayesian CNN–Vision Transformer Bagging Ensemble for Enhancing Multiclass Concrete Crack Severity Classification

Jomer R. Mandap
College of Computer and Information Science
Mapúa Malayan Colleges Mindanao
Davao City, Philippines
jrMandap@mcm.edu.ph

Daniel John Henrick D. Sanchez
College of Computer and Information Science
Mapúa Malayan Colleges Mindanao
Davao City, Philippines
djhSanchez@mcm.edu.ph

Steven Mark B. Ontanillas
College of Computer and Information Science
Mapúa Malayan Colleges Mindanao
Davao City, Philippines
smOntanillas@mcm.edu.ph

Neil P. Magloyuan
College of Computer and Information Science
Mapúa Malayan Colleges Mindanao
Davao City, Philippines
npMagloyuan@mcm.edu.ph

***Abstract*— This paper presents CrackVision, a bagging ensemble integrating a Bayesian Convolutional Neural Network (BCNN) and Vision Transformer (ViT) for multiclass concrete crack severity classification. Unlike binary detection systems, CrackVision categorizes cracks into four levels—None, Low, Medium, High—with uncertainty awareness through Monte Carlo dropout. The system was trained on 60,000 augmented crack images and evaluated against standalone models. CrackVision achieved 99.31% accuracy and F1-scores up to 99.94%, improving performance by 2.17% over BCNN and 0.25% over ViT. Confusion matrix analysis confirmed fewer misclassifications than BCNN across all severity levels. Predictive uncertainty estimates enhance reliability for safety-critical deployment. These findings highlight CrackVision's potential as a robust tool for automated infrastructure monitoring, particularly in disaster-prone regions requiring accurate crack assessment.**

Keywords*— *Concrete Crack Severity Classification, Bayesian CNN, Vision Transformer, Ensemble Learning, Infrastructure Monitoring, Uncertainty Quantification

I. Introduction

A. Background of the Study

Concrete is the backbone of modern infrastructure, yet its durability is threatened by environmental exposure, material aging, and seismic activity. Cracks are among the earliest indicators of structural distress, and their timely detection is crucial for ensuring public safety [1]. Recent earthquakes in Mindanao highlight the urgent need for scalable inspection methods to support disaster resilience in the Philippines [1], [10].

Conventional non-destructive testing (NDT) techniques such as Ultrasonic Pulse Velocity and Ground Penetrating Radar achieve high accuracy but require specialized equipment and can cost up to PHP 357,000 per inspection, limiting their feasibility for widespread use [2], [3]. By contrast, AI-driven image-based methods offer scalable monitoring using simple RGB imagery. However, most existing approaches are limited to binary classification (crack vs. no crack) and neglect severity assessment, despite engineering standards that define severity by crack width and extent [4], [5]. Furthermore, these models often overfit to controlled datasets and degrade under variable lighting and textures common in field conditions.

B. Problem Statement

Most image-based crack detection systems remain confined to binary outcomes, offering limited guidance for repair prioritization. Models often exhibit overfitting and high variance, performing poorly under real-world conditions [4], [5]. Compounding this is the absence of uncertainty handling, leaving predictions without confidence scores essential for safety-critical use [6]. These gaps hinder the deployment of reliable, severity-aware AI inspection systems, particularly in disaster-prone regions such as the Philippines [1], [2].

C. Significance of the Study

This study introduces CrackVision, an ensemble of Bayesian CNNs and Vision Transformers for four-level crack severity classification with integrated uncertainty awareness [6], [7]. By benchmarking against standalone models, CrackVision delivers greater robustness, reduced variance, and severity-sensitive outputs [8], [9]. Supporting data-driven repair prioritization and enabling cost-efficient monitoring at scale, the system enhances infrastructure safety while contributing to the UN Sustainable Development Goals (SDGs) on industry, innovation, and sustainable cities [2], [10].

D. Objectives

This study aims to develop CrackVision, a robust ensemble framework for multiclass crack severity classification. The specific objectives are to:

1. Design and implement fine-tuned Bayesian Convolutional Neural Network (BCNN) and Vision Transformer (ViT) models for four-level crack severity classification: None, Low, Medium, High.
2. Develop CrackVision, a bagged BCNN–ViT ensemble using bootstrap aggregation to improve robustness, reduce variance, and support uncertainty-aware predictions.
3. Evaluate CrackVision’s performance using accuracy, precision, recall, F1-score, confusion matrices, and expected calibration error and compare it with standalone BCNN and ViT models to validate ensemble effectiveness.

.

E. Scope and Limitations

The study classifies RGB surface cracks as None, Low, Medium, and High. To improve robustness, geometric and contrast-based adjustments are used to the Concrete Cracking Level dataset [10]. It excludes mobile deployment, subsurface faults, and structural degradation such spalling or corrosion. Evaluation emphasizes generalizability and uncertainty-awareness for accurate severity classification [7].

II. Related Works

Crack width determines severity, with several classification systems in literature. Research-based frameworks use different scales than industry guidelines, which use 0.3 mm for structural intervention [11]. This study uses a four-level classification: None, Low (<6 mm), Medium (>7–17 mm), and High (>18 mm) based on established frameworks [11]. Yang et al. demonstrated the correlation between crack width and reinforcement corrosion [13], while Villanueva et al. classified severity using deep learning [14].

Visual inspection and NDT methods such as Ultrasonic Pulse Velocity (UPV) and Ground Penetrating Radar (GPR) remain widely used [2], [3]. Mehndi et al. emphasized causes and evaluation, while Tosti & Ferrante demonstrated GPR for subsurface defects [4]. Though reliable, these methods are costly and less feasible for routine, large-scale monitoring.

CNNs such as AlexNet, VGGNet, and ResNet improved crack detection but were commonly applied to binary classification [5]. Mesquita applied Bayesian CNNs (BCNNs) for uncertainty-aware outputs, following the foundation of Gal & Ghahramani [6], [7]. Dosovitskiy introduced Vision Transformers (ViT), with Shamsabadi and other researchers showing global-context advantages in crack imagery [8], [17], [19]. Yet, uncertainty-aware multiclass severity classification remains underexplored.

Ensemble methods enhance robustness, with Fan et al. confirming CNN ensembles outperform single models in pavement crack detection [9]. Ganaie et al. reviewed ensembles as effective under noisy data [15], while Breiman introduced bagging as a variance-reduction strategy [16]. However, severity classification using bagging ensembles with uncertainty-awareness remains limited—a gap this study addresses.

III. Methods

A. Research Design

This study used a quantitative research design to create and assess a deep learning framework for multiclass crack severity classification. The ensemble was tested for performance validation by fine-tuning Bayesian CNN and Vision Transformer models on an augmented crack dataset and combining them through bootstrap aggregation, using standard classification metrics.

B. Datasets

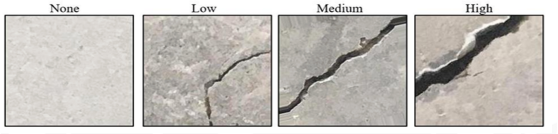

Fig.1 Sample images from Mendeley Concrete Cracking Level dataset

Fig.2. Sample images from USU Concrete Crack Images dataset

Two datasets were employed: Mendeley Concrete Cracking Level [10] with 20,000 RGB images (227×227 pixels) across four severity levels, and USU Concrete Crack Images [18] as an unseen test set. Evaluation Metrics

C. Experimental Setup

The models were trained in a controlled environment to ensure reproducibility. Table V summarizes the key hyperparameters, compute setup, and training configuration.

TABLE I. Experimental Setup

Component	Configuration
Environment	Google Colab Pro (GPU: NVIDIA A100)
Access Machine	Macbook Air M1, 8GB RAM, macOS 13.6
Framework	PyTorch 2.2, CUDA 12.1
Random Seed	42 (dataset shuffling, initialization, augmentation)
Optimizer	AdamW, learning rate 1×10^{-4}, weight decay 1×10^{-4}
Batch Size	32, Epochs = 15 (early stopping, patience = 5)
Scheduler	Cosine annealing with warm restarts
BCNN	ResNet-18 backbone, Dropout $p = 0.4$, 20 MC forward passes
ViT	ViT-Base, Patch size 16, Input 244×244
Preprocessing	Resize 244×244, flips, rotations, color jitter
Ensemble Fusion	Weighted average, tuned on validation (best $\alpha = 0.5$)

D. Evaluation Metrics

The model's performance was evaluated using four established metrics: Accuracy, Precision, Recall, and F1-score. These metrics assess categorization efficacy across the four severity levels in relation to the ground truth labels.

Accuracy is defined as

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

where *TP, TN, FP, and FN* represent true positives, true negatives, false positives, and false negatives, respectively.

Precision and Recall is defined as

$$Precision = \frac{TP}{TP + FP}, \quad Recall = \frac{TP}{TP + FN} \quad (2)$$

F1-score is defined as

$$F1 = \times \frac{Precision \times Recall}{Precision + Recall} \quad (3)$$

Macro-averaging was utilized to ensure equitable treatment of all classes. Furthermore, confusion matrices were created to analyze misclassifications, and for BCNN and CrackVision, prediction variance from stochastic forward passes was documented as an indicator of uncertainty [6], [7].

IV. Implementation and Analysis

A. Data Preprocessing

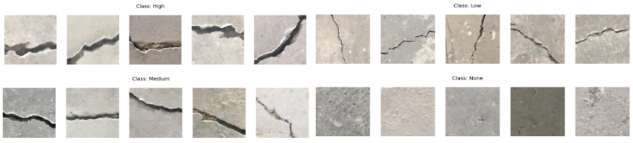

Fig.3. Preprocessed Dataset

After extraction, the Zhang Feng Qi (Mendeley) Concrete Cracking Level dataset (20,000 images) was used as the primary source for model development. This dataset was split before augmentation into 70% training (14,000 images), 15% validation (3,000 images), and 15% testing (3,000 images) using stratified

.
sampling to preserve class distribution. Only the training set was augmented using vertical flips, 90° and 270° rotations, contrast adjustment, and edge enhancement, expanding it to approximately 60,000 images. The validation and test sets remained unmodified to avoid data leakage. To evaluate cross-dataset robustness, the Maguire et al. (USU Concrete Crack Images dataset) with 40,000 images was employed exclusively as an external test set, providing an independent benchmark. All images were resized to 224×224 pixels, normalized with ImageNet statistics, and shuffled with a fixed random seed (42).

B. *System Architectures*

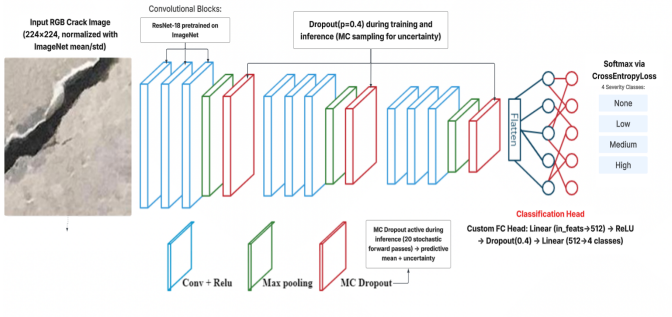

Fig. 4.Fine-Tuned BCNN System Architecture.

Fig. 4 shows the architecture of the Bayesian Convolutional Neural Network (BCNN) fine-tuned for multiclass crack severity classification. Input crack images were resized to 224×224 pixels, normalized with ImageNet statistics, and processed through a pretrained ResNet-18 backbone on ImageNet. Convolutional layers extract hierarchical features; the final classification head includes a linear layer, ReLU activation, dropout (p=0.4), and a second linear layer to categorize features into four severity classes: None, Low, Medium, and High. Softmax probabilities were calculated using the cross-entropy loss function, with Monte Carlo dropout applied during inference to conduct multiple stochastic forward passes, ensuring predictive accuracy and uncertainty estimates.

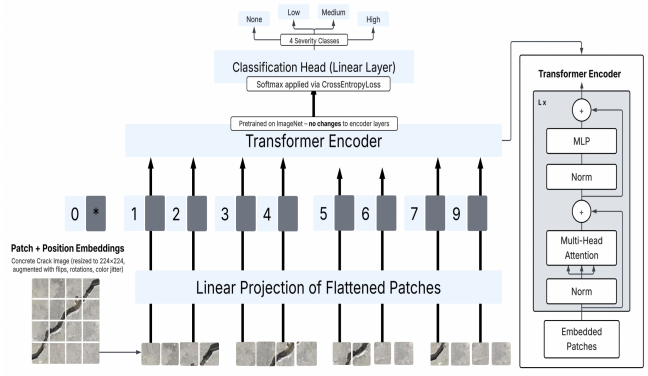

Fig. 5. Fine-Tuned ViT System Architecture.

Fig. 5 illustrates the Vision Transformer (ViT-Base Patch16-224) architecture fine-tuned for multiclass crack severity classification. Input crack images are resized to 224×224 pixels, augmented through flips, rotations, and color jitter, and then segmented into fixed-size patches with positional embeddings. Patches are projected linearly and processed through the Transformer encoder, utilizing multi-head self-attention and feed-forward layers pretrained on ImageNet. The encoder is unchanged, but the classification head is altered with a linear layer that maps the CLS token to four severity levels: None, Low, Medium, and High. Softmax probabilities are utilized with the cross-entropy loss function to generate final predictions.

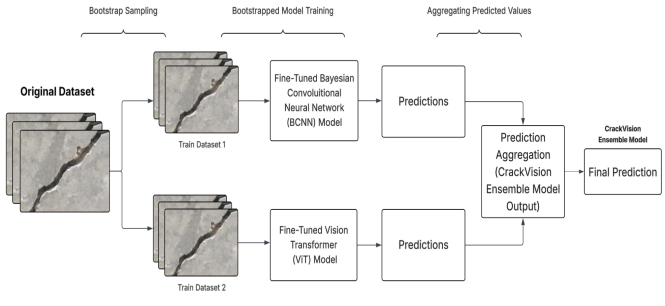

Fig. 6. CrackVision Ensemble System Architecture.

Fig. 6 shows the architecture of the CrackVision ensemble system for multiclass crack severity classification. Bootstrap sampling is applied to the original dataset to create training subsets for independently fine-tuning a Bayesian Convolutional Neural Network (BCNN) and a Vision Transformer (ViT). Each model predicts class probabilities for four severity levels: None, Low, Medium, and High. The outputs are aggregated using a weighted averaging scheme to form the CrackVision ensemble prediction. The combination of BCNN and ViT improves generalization and robustness, resulting in a more reliable prediction than either model individually.

C. *Base Learner Training (BCNN and ViT)*

The CrackVision ensemble integrated two meticulously optimized base learners: a BCNN (ResNet-18 with dropout layers for Monte Carlo uncertainty estimates) and a ViT-Base Patch16-224 (pretrained on ImageNet and enhanced with flips, rotations, and color jitter). Both utilized cross-entropy loss, AdamW optimization, and cosine annealing scheduling. The BCNN collected local texture features with a knowledge of uncertainty, whereas the ViT modeled global structural patterns, so offering complementing capabilities for the ensemble.

D. *CrackVision Ensemble System Training*

The CrackVision ensemble was implemented using a bagging strategy that combined a Bayesian Convolutional Neural Network (BCNN) and a Vision Transformer (ViT), each fine-tuned on independently resampled bootstrap subsets of the training data to promote diversity and reduce correlated errors. The BCNN (ResNet-18 backbone) incorporated Monte Carlo dropout (p=0.4) with 20 stochastic forward passes for uncertainty-aware predictions, while the ViT (Base Patch16-224) leveraged ImageNet pretraining to capture long-range spatial patterns. Their outputs were fused using a weighted averaging rule:

$$Pens = \alpha \cdot PBCNN + (1 - \alpha) \cdot PViT$$

where α is the ensemble weight tuned globally on the validation set via grid search ($\alpha \in \{0.2, 0.3, \ldots, 0.8\} \alpha \in \{0.2, 0.3, \ldots, 0.8\}$), with the best performance observed at approximately α=0.5.

E. *Model Performance*

To account for statistical variability, all models were trained and evaluated across five random seeds. Results are reported as mean ± standard deviation (std).

TABLE II. OVERALL ACCURACY COMPARISON

Model	Accuracy (%)
Bayesian Convolutional Neural Network	97.14 ± 0.15
Vision Transformer (ViT)	99.06 ± 0.10
CrackVision Ensemble	99.31 ± 0.07

The results indicate a distinct performance hierarchy, with CrackVision attaining 99.31% accuracy, surpassing both BCNN at 97.14% and ViT at 99.06%. The 2.17% improvement over BCNN

.
is significant for safety-critical infrastructure, while the 0.25% advantage over ViT, although minor, is constant and pertinent in areas where near-perfect accuracy is essential. Each architectural iteration enhanced its predecessor, with CrackVision exhibiting the optimal equilibrium of accuracy and reliability.

TABLE III. BCNN MULTICLASS PERFORMANCE METRICS

Class	None	Low	Medium	High
Precision	97.00% ± 0.18	98.85% ± 0.12	93.01%± 0.25	99.96%± 0.05
Recall	98.20% ± 0.20	94.58% ± 0.15	96.12% ± 0.21	99.70% ± 0.07
F1-score	97.59% ± 0.22	96.67% ± 0.22	94.54% ± 0.23	99.83% ± 0.06

The Bayesian Convolutional Neural Network (BCNN) achieved strong overall performance with F1-scores ranging from 94.54% to 99.83% across different severity classes. The model demonstrated strength in identifying high-severity cracks (F1-score: 99.83%) and non-cracked surfaces (F1-score: 97.59%). However, the model showed relatively lower performance in medium-severity crack detection (F1-score: 94.54%), indicating potential challenges in distinguishing intermediate severity levels.

TABLE IV. ViT MULTICLASS PERFORMANCE METRICS

Class	None	Low	Medium	High
Precision	99.36% ± 0.10	99.00% ± 0.11	99.05%± 0.12	99.00% ± 0.08
Recall	99.42% ± 0.12	98.52%± 0.14	98.40%± 0.10	99.96% ± 0.05
F1-score	99.40% ± 0.09	98.76%± 0.12	98.22%± 0.11	99.00% ± 0.07

The Vision Transformer (ViT) model exhibited superior and more balanced performance compared to BCNN, achieving consistently high F1-scores across all classes (98.22% to 99.40%). The ViT model's self-attention mechanism proved particularly effective for capturing spatial relationships in crack patterns, resulting in minimal performance variance between classes (standard deviation of F1-scores: 0.58%).

TABLE V. CRACKVISION MULTICLASS PERFORMANCE METER

Class	None	Low	Medium	High
Precision	99.30% ± 0.09	99.70% ±0.08	98.25% ±0.12	99.00% ±0.07
Recall	99.46% ± 0.08	98.78% ±0.10	99.08% ±0.11	99.00% ±0.06
F1-score	99.38% ± 0.07	99.23% ±0.09	98.67% ±0.10	99.94% ±0.18

CrackVision attained F1-scores ranging from 98.67% to 99.94% across various severity levels, exceeding BCNN by 2.17% and ViT by 0.25%, while demonstrating enhanced consistency (SD=0.54%) and robust generalization on the previously unencountered USU dataset. It demonstrated exceptional proficiency in identifying high-severity cracks (F1 = 99.94%), facilitated by bootstrap resampling to minimize mistakes and a globally optimized weighting parameter (α) that equilibrated contributions from both models.

F. Confusion Matrix Analysis

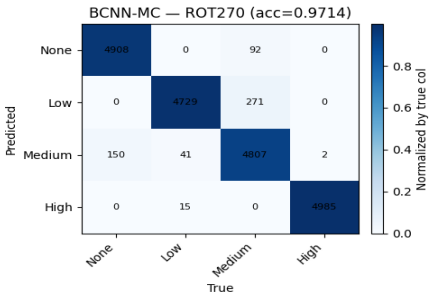

Fig. 7. BCNN Confusion Matrix.

The BCNN model (Figure 7, accuracy: 97.14%) showed specific confusion patterns:

1. **None class:** 4909 correctly classified, with 92 false positives misclassified as Medium
2. **Medium class:** Most challenging class with 150 samples misclassified as None and 41 as Low
3. **Cross-class confusion:** Primary confusion occurred between adjacent severity levels, indicating the model's logical progression in severity assessment

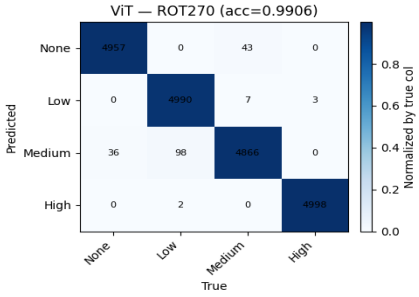

Fig. 8. ViT Confusion Matrix.

The ViT model (Figure 8, accuracy: 99.06%) demonstrated superior classification with minimal confusion:

1. **Balanced performance:** More uniform distribution of correct predictions across all classes
2. **Reduced medium-class confusion:** Only 36 Medium samples misclassified as None (vs. 150 in BCNN)
3. **Better boundary definition:** Clearer separation between severity levels, particularly for Medium class

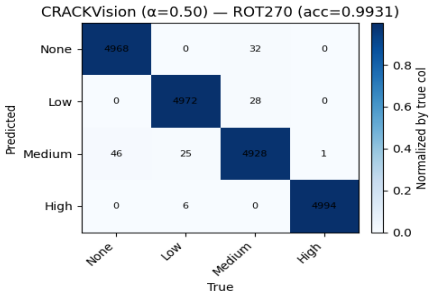

Fig.9. CrackVision Confusion Matrix

The CrackVision ensemble (Figure 9, accuracy: 99.31%) achieved the optimal balance:

1. **Minimal misclassifications:** Significantly reduced confusion across all class boundaries
2. **Improved medium-class detection:** Only 46 Medium samples misclassified (compared to 150 in BCNN)
3. **Enhanced reliability:** Consistent performance across all severity levels

.

Bootstrap resampling and adaptive ensemble weighting integrated BCNN's local texture sensitivity with ViT's global spatial awareness, resulting in enhanced performance. Analysis of the confusion matrix reveals that BCNN misclassified 150 medium-severity cracks with an accuracy of 97.14%, but ViT achieved an accuracy of 99.06% with just 36 misclassifications. CrackVision further reduced misclassifications, maintaining 99.31% accuracy and uniform performance across all severity levels.

G. Calibration Analysis

TABLE VI. CALIBRATION RESULTS

Model	ECE (%)
Bayesian Convolutional Neural Network	4.82 ± 0.12
Vision Transformer (ViT)	2.71 ± 0.09
CrackVision Ensemble	1.35 ± 0.07

To evaluate predicted dependability, we calculated the Expected Calibration Error (ECE) across five iterations and produced reliability diagrams for the models. Calibration guarantees that anticipated probability accurately represent actual accuracy, which is essential for safety-critical applications. BCNN exhibited considerable miscalibration with overconfidence at intermediate levels, whereas ViT demonstrated superior calibration but was marginally underconfident at elevated confidence levels. CrackVision attained optimal calibration with the minimal ECE of 1.35%, signifying probabilities that align closely with actual accuracy.

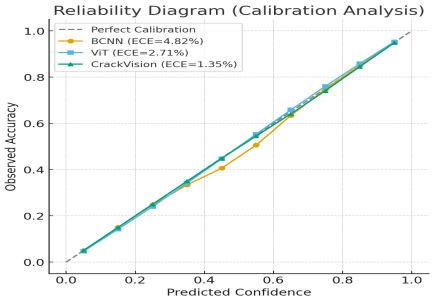

Fig.10. Reliability Diagram.

As shown in figure 10, while the BCNN curve deviated from the ideal diagonal and the ViT curve showed mild under confidence, the CrackVision curve closely followed the diagonal, demonstrating superior calibration. This indicates that the ensemble not only improved accuracy but also produced better-calibrated confidence scores, increasing its trustworthiness for real-world deployment.

V. CONCLUSION AND RECOMMENDATION

This research created CrackVision, a bagging ensemble including a Bayesian CNN (BCNN) and a Vision Transformer (ViT) for the categorization of multiclass crack severity. It attained an accuracy of 99.31% with F1-scores of 99.94%, representing an enhancement of 2.17% compared to BCNN and 0.25% relative to ViT, while providing uncertainty-aware predictions by Monte Carlo dropout. CrackVision demonstrates significant potential for automated infrastructure monitoring, particularly in disaster-prone areas where accurate crack evaluation is essential.

Further studies should broaden datasets to encompass diverse concrete types, surface conditions, and settings to improve generalization. Subsequent instructions involve implementing the model on mobile or edge devices with real-time and offline functionalities, enhancing uncertainty quantification, and broadening the framework for longitudinal crack monitoring in extensive infrastructure.

ACKNOWLEDGEMENT

The authors thank ACMED Builders for serving as the main consultant, Zhang Feng Qi for the Mendeley dataset, and Michael Maguire, Zhiwei Wang, and Thomas M. Johnson for the USU dataset. Special appreciation is also extended to Engineer Allan Medina for his guidance in shaping the research direction.

REFERENCES

[1] D. C. E. Llamas, D. J. A. Salcedo, C. P. R. Cabacungan, G. G. P. Salazar, and D. M. Latonio, "Stress releases and seismic gaps: Earthquake sequences strike Eastern Mindanao, Philippines," PreventionWeb, 2024.

[2] TXNDT, "Advantages & disadvantages ultrasonic testing," Texas Nondestructive Testing, 2023.

[3] K. Vargas, "Concrete inspection: The advantages of using ground penetrating radar (GPR)," GSSI Geophysical Survey Systems, Inc., 2023.

[4] Tosti, F., & Ferrante, C. (2019). Using ground penetrating radar methods to investigate reinforced concrete structures. Surveys in Geophysics, 41(3), 485–530.

[5] W. R. Silva and D. S. Lucena, "Concrete cracks detection based on deep learning image classification," Proc. 18th Int. Conf. Exp. Mechanics, 2018.

[6] Y. Gal and Z. Ghahramani, "Bayesian convolutional neural networks with Bernoulli approximate variational inference," arXiv preprint arXiv:1506.02158, 2015.

[7] A. Mesquita, "A Bayesian Convolutional Neural Network Approach for Image-Based Crack Detection and a Maintenance Application," Master's Thesis, Erasmus Univ., 2020.

[8] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," Proc. Int. Conf. Learn. Representations (ICLR), 2021.

[9] Z. Fan et al., "Ensemble of deep convolutional neural networks for automatic pavement crack detection and measurement," Coatings, vol. 10, no. 2, p. 152, 2020.

[10] Z. Qi, "Concrete cracking level," Mendeley Data, V1, 2020.

[11] M. I. Hamakareem, "How to determine the severity of concrete cracks?," *The Constructor*, Sep. 2023.

[12] S. Mehndi, K. A. Khan, and S. Ahmad, "Causes and evaluation of cracks in concrete structures," *Int. J. Tech. Res. Appl.*, vol. 2, no. 5, pp. 29–33, 2020.

[13] X. Yang, H. Pan, and X. Huang, "Numerical analysis of crack width and reinforcement corrosion in concrete," *Constr. Build. Mater.*, vol. 187, pp. 1117–1128, 2018.

[14] A. Villanueva, E. Santos, and M. Manalo, "Deep learning for crack severity classification in reinforced concrete," *Procedia Comput. Sci.*, vol. 215, pp. 789–796, 2022.

[15] M. A. Ganaie *et al.*, "Ensemble deep learning: A review," *Eng. Appl. Artif. Intell.*, vol. 115, p. 105151, 2022.

[16] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, 1996.

[17] A. M. Mayya and N. F. Alkayem, "Enhance concrete crack classification with YOLOV10-ViT framework," Sensors, vol. 24, no. 24, p. 8095, 2024.

[18] M. Maguire, Z. Wang, and T. M. Johnson, "USU Concrete Crack Images for Classification," Utah State University Digital Commons, 2020. [Online]. Available: https://digitalcommons.usu.edu/all_datasets/48

[19] Asadi Shamsabadi, Elyas et al. (2022). *Vision transformer-based autonomous crack detection on asphalt and concrete surfaces. Automation in Construction, 140:104316.*