International Journal of Advances in Data and Information Systems
Vol. 7, No. 2, August 2026, pp. 907~920
ISSN: 2721-3056, DOI: 10.25008/ijadis.v7i2.1551

Towards Robust Cross-Dataset Pothole Detection Through Multi-Dataset Pretraining

Hanifatus Sa'diyah Widihasaniputri[1], Oddy Virgantara Putra[2], Murhadi[3]
[1,3]Department of Information Technology, Universitas Muhammadiyah Purworejo, Indonesia
[2]Department of Informatics, Universitas Darussalam Gontor, Indonesia

Article Info

Article history:
Received Feb 26, 2026
Revised July 21, 2026
Accepted July 29, 2026

Keywords:
Pothole detection
Cross-dataset evaluation
Domain shift
Generalization gap
Model robustness

ABSTRACT

Pothole detection based on deep learning has achieved high detection accuracy; however, most existing studies evaluate models using the same dataset for both training and testing, providing limited evidence of robustness under unseen data distributions. This study investigated the cross-dataset generalization capability of YOLOv8n using three publicly available pothole datasets with different visual characteristics: the Multi-Weather Pothole Dataset (MWPD), the Jaygala dataset, and the Andrew dataset. The proposed framework evaluated same-dataset and cross-dataset detection performance, quantified robustness through generalization gap analysis, examined the influence of dataset characteristics, and assessed the effectiveness of multi-dataset pretraining. Experimental results showed that the Andrew dataset achieved the highest same-dataset performance (mAP@50 = 0.816) but also exhibited the largest generalization gap (0.246), indicating limited robustness across datasets. In contrast, multi-dataset pretraining reduced the generalization gap for MWPD from 0.093 to 0.048, demonstrating improved cross-dataset robustness, although the improvement was not consistent across all datasets. These findings indicate that same-dataset accuracy alone is insufficient for evaluating model robustness and that cross-dataset evaluation provides a more realistic assessment of deployment performance in heterogeneous road environments.

This is an open access article under the CC BY-SA license.

Corresponding Author:
Hanifatus Sa'diyah Widihasaniputri
Department of Information Technology,
Universitas Muhammadiyah Purworejo,
Jl. K.H. Ahmad Dahlan No. 3, Purworejo, Central Java 54111, Indonesia.
Email: hanifatus@umpwr.ac.id

1. INTRODUCTION

Road surface deterioration remains a significant challenge for transportation infrastructure worldwide. Statistical data from the Indonesian Central Bureau of Statistics (Badan Pusat Statistik—BPS) indicate that 8.96% (48,340 km) of Indonesia's national road network is classified as damaged, while 28.92% (156,044 km) is categorized as severely damaged [1]. Among various types of road defects, potholes are particularly critical because they reduce driving comfort, increase vehicle maintenance costs, and compromise road safety. According to the World Health Organization (WHO), road traffic accidents cause more than 1.19 million deaths annually worldwide, with poor road conditions being one of the contributing factors [2]. Therefore, accurate and timely pothole detection is essential for supporting preventive maintenance and improving road management efficiency [3], [4].

Conventional road inspection primarily relies on manual visual surveys, which are time-consuming, subjective, and labor-intensive. Although alternative approaches such as 3D laser scanning and vibration-based sensing can provide more detailed measurements, they generally require specialized equipment and are less suitable for large-scale real-time deployment [3], [4].

Consequently, computer vision and deep learning have become the dominant approaches for automated pothole detection because they enable efficient and scalable road monitoring [5]–[7].

Recent studies have demonstrated that YOLO-based detectors achieve excellent performance in pothole and road damage detection. Numerous architectural improvements have been proposed to enhance detection accuracy while maintaining computational efficiency. For example, [8], improved YOLOv5 using MobileNetV3 and Coordinate Attention, [9] incorporated Bottleneck Transformer modules, [10], enhanced feature fusion using BiFPN, [11], optimized YOLOv5 through SENet and anchor refinement, [12], proposed YOLOv8-PD with Transformer-based attention, and [13], chieved high detection performance on the MWPD dataset using an enhanced YOLOv9 framework. Collectively, these studies demonstrate that architectural modifications can substantially improve pothole detection accuracy. However, these improvements have almost exclusively been validated using the same dataset for both training and testing, leaving their robustness under unseen data distributions largely unexplored [14].

This limitation is particularly important because road images collected in real-world environments differ considerably in weather conditions, illumination, road surfaces, camera configurations, and geographical locations [15]. Such variations introduce domain shift, where the distribution of deployment data differs from that observed during training, often leading to a significant degradation in detection performance [16], [17]. Consequently, high same-dataset accuracy does not necessarily indicate strong cross-dataset robustness.

To improve robustness under varying conditions, several studies have explored strategies beyond architectural enhancement. [18], employed synthetic weather augmentation to improve detection under adverse environmental conditions, while [19], evaluated a lightweight YOLOv8 framework on an additional dataset to assess its generalization capability. Furthermore,[20], demonstrated that combining multiple datasets can improve robustness by increasing training data diversity and reducing dataset bias. These findings suggest that dataset diversity is an important factor for improving feature generalization across different visual domains.

Nevertheless, several important research gaps remain. First, many recent pothole detection studies primarily focus on improving detection accuracy through architectural modifications and evaluate their methods using a single dataset, providing limited evidence regarding model performance under unseen conditions [21], [22]. Second, model robustness is commonly inferred from high same-dataset detection accuracy rather than being systematically evaluated through cross-dataset experiments, which provide a more realistic assessment of real-world deployment performance [18], [23]. Third, although domain shift has been extensively investigated in general object detection, its influence on pothole detection, particularly the relationship between dataset characteristics and cross-dataset robustness, remains insufficiently explored [16], [17]. Finally, previous studies suggest that increasing training data diversity may improve model generalization; however, the effectiveness of multi-dataset pretraining as a practical strategy for enhancing cross-dataset pothole detection has received limited attention [20], [24].

Therefore, this study investigates the robustness of YOLOv8n under cross-dataset pothole detection scenarios using three publicly available datasets with distinct visual characteristics. The proposed framework systematically evaluates same-dataset and cross-dataset performance, quantifies robustness through generalization gap analysis, examines the influence of dataset characteristics on cross-dataset performance, and evaluates the effectiveness of multi-dataset pretraining. The study provides a comprehensive analysis of cross-dataset generalization and offers practical insights into improving the robustness of deep learning-based pothole detection under heterogeneous real-world conditions.

2. RESEARCH METHOD

2.1. Cross-Dataset Evaluation Framework

The primary objective of this study is to investigate the robustness and generalization capability of pothole detection models across datasets with different visual characteristics. Although deep learning has significantly improved pothole detection performance, most existing studies evaluate models using the same dataset for training and testing, which may overestimate real-world performance and provide limited insight into model generalization [8], [3].

In practical applications, road images are often collected under different environmental conditions, camera settings, and road surfaces. As a result, the data distribution encountered during deployment may differ from the training distribution. This phenomenon, known as domain shift, can substantially degrade the performance of deep learning-based computer vision systems when applied to unseen domains [25], [26].

To address this issue, this study adopts a cross-dataset evaluation framework that systematically measures detection performance across multiple independent pothole datasets [27], [28]. Unlike conventional evaluations, this approach provides a more realistic assessment of model robustness and enables the investigation of how dataset diversity affects generalization performance [19], [29].

As illustrated in Figure 1, the framework consists of two phases. First, baseline YOLOv8n models are trained on individual datasets and evaluated using both same-dataset and cross-dataset scenarios. Second, all datasets are combined for multi-dataset pretraining, followed by dataset-specific fine-tuning and re-evaluation. The resulting performance is analyzed using robustness metrics and dataset characteristic analysis to identify factors influencing cross-dataset generalization.

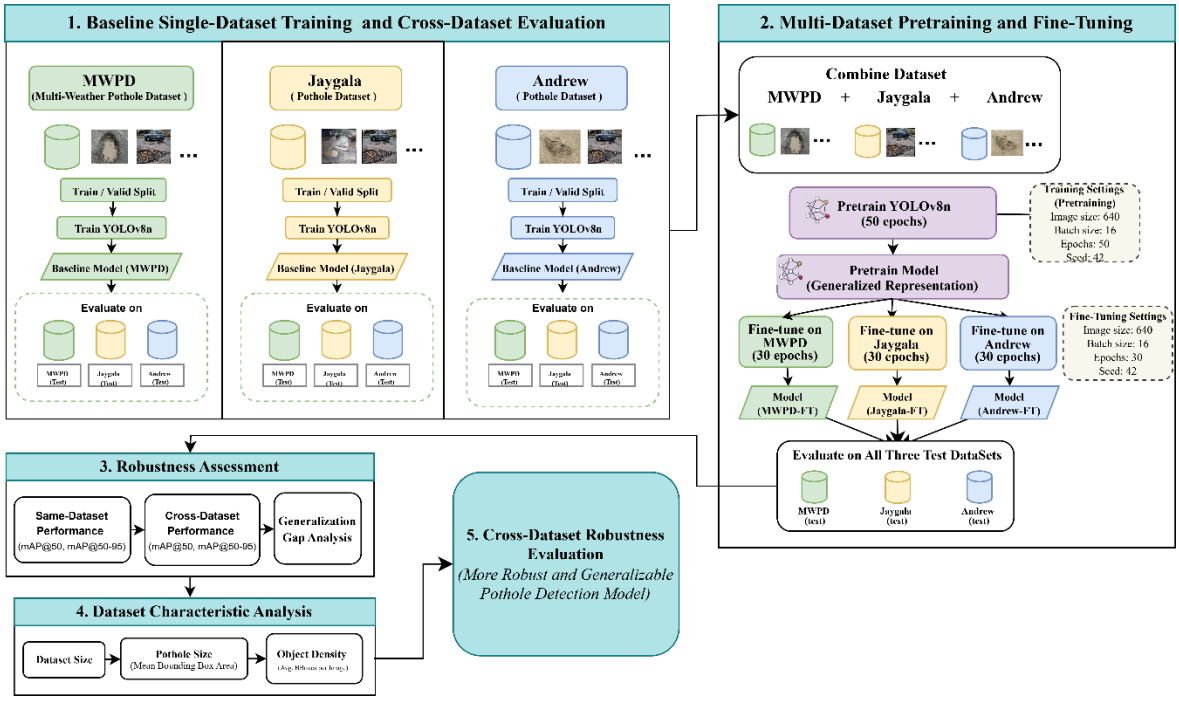


Figure 1. Proposed Cross-Dataset Evaluation Framework

2.2. Pothole Datasets

Three publicly available pothole datasets were used in this study: the Multi-Weather Pothole Dataset (MWPD) [30], the pothole dataset introduced by [31], (hereafter referred to as the Jaygala dataset), and the Andrew pothole dataset [32].These datasets were selected because they exhibit substantial variations in dataset size, pothole scale, object density, environmental conditions, and image acquisition characteristics, thereby providing a suitable benchmark for evaluating cross-dataset generalization under heterogeneous data distributions.

MWPD contains road images collected under diverse weather and illumination conditions. Jaygala contains a relatively high number of pothole instances per image, resulting in dense object distributions. In contrast, Andrew contains fewer images but generally larger pothole instances than the other datasets. To ensure consistency across experiments, all annotations were converted to YOLO format and organized using a unified directory structure. The training and test sets were prepared following the original dataset partitions whenever available, while custom splits were created for datasets without predefined partitions. A summary of the datasets is presented in Table 1.

Table 1. Summary of pothole datasets

Dataset	Training Images	Test Images	Original Annotation Format
MWPD	2,730	97	YOLO
Jaygala	870	187	YOLO
Andrew	465	101	Pascal VOC

2.3. YOLOv8n and Training Strategy

This study employs YOLOv8n as the object detection architecture because it provides a favorable balance between detection accuracy [33], [34] and computational efficiency for real-time and edge-computing applications [35]. YOLO-based detectors perform object localization and classification within a single forward pass, making them widely adopted for real-time object detection [21], [36]–[38]. Lightweight YOLO variants, including YOLOv8n, have demonstrated competitive performance under resource-constrained deployment scenarios [33], [39], [40].

Two training strategies were investigated. The first follows single-dataset training, where separate YOLOv8n models were trained on the MWPD, Jaygala, and Andrew datasets and evaluated using both same-dataset and cross-dataset protocols. Since single-dataset evaluation mainly measures in-distribution performance, cross-dataset testing provides a more realistic assessment of robustness under domain shift [19], [41], [42].

The second strategy employs multi-dataset pretraining, in which all training images were combined to increase visual diversity before model training. Previous studies have shown that training on more diverse datasets promotes generalized feature learning and improves robustness across domains [24], [29], [43].The pretrained model was subsequently fine-tuned on each target dataset to adapt to domain-specific characteristics while retaining generalized knowledge. Fine-tuning is a standard transfer learning strategy for improving performance on limited target data, particularly under domain shift [44], [45].

All experiments were conducted using the Kaggle Notebook environment. The implementation was developed using Python, PyTorch, and the Ultralytics YOLO framework. Identical hardware resources, hyperparameter settings, and a fixed random seed were used across all experiments to ensure fair comparisons.

To assess the robustness of the training process, the same-dataset experiments were repeated using five different random seeds (7, 21, 42, 84, and 123). The reported results correspond to the mean and standard deviation of Precision, Recall, mAP50, and mAP50-95 across these independent runs. Due to the substantially higher computational cost, the cross-dataset and multi-dataset pretraining experiments were conducted using a fixed random seed (42). The experimental environment and training configuration are summarized in Table 2.

Table 2. Experimental Environment and Training Configuration

Parameter	Value
Platform	Kaggle Notebook
Operating System	Ubuntu 22.04.5 LTS
CPU	Intel Xeon CPU @ 2.00 GHz
RAM	31 GB
GPU	2 × NVIDIA Tesla T4 (15 GB each)
Python	3.12.13
PyTorch	2.10.0
Detection Model	YOLOv8n
Pretrained Weight	yolov8n.pt
Input Resolution	640 × 640
Batch Size	16
Baseline Epochs	50
Pretraining Epochs	50
Fine-Tuning Epochs	30
Optimizer	Ultralytics Default
Random Seed	Same-dataset: five seeds (7, 21, 42, 84, 123); others: 42

2.4. Evaluation Protocol

The proposed framework was evaluated using same-dataset and cross-dataset protocols. During the baseline experiment, three independent YOLOv8n models were trained using the MWPD, Jaygala, and Andrew datasets. Each model was first evaluated on its corresponding test set (same-dataset) and subsequently on the remaining two datasets (cross-dataset), resulting in nine evaluation scenarios. For the multi-dataset pretraining experiment, the pretrained model was fine-tuned separately on each target dataset and evaluated using the same testing protocol.

Cross-dataset evaluation simulates practical deployment conditions where the testing data differ from the training distribution. This protocol is widely adopted to assess model robustness under domain shift, where variations in image appearance, illumination, object scale, and acquisition conditions may substantially degrade detection performance [46], [47].

Model performance was evaluated using Precision, Recall, mAP@50, and mAP@50–95, which are standard metrics for object detection evaluation [48], [49]. Precision and Recall measure detection accuracy, whereas mAP@50 and mAP@50–95 evaluate localization performance at an IoU threshold of 0.50 and across multiple IoU thresholds (0.50–0.95), respectively. The evaluation scenarios are summarized in Table 3.

Table 3. Evaluation Scenarios

Training Dataset	Testing Dataset	Evaluation Type
MWPD	MWPD	Same-Dataset
MWPD	Jaygala	Cross-Dataset
MWPD	Andrew	Cross-Dataset
Jaygala	Jaygala	Same-Dataset
Jaygala	MWPD	Cross-Dataset
Jaygala	Andrew	Cross-Dataset
Andrew	Andrew	Same-Dataset
Andrew	MWPD	Cross-Dataset
Andrew	Jaygala	Cross-Dataset

This protocol enables a direct comparison between in-distribution and out-of-distribution performance, providing a systematic assessment of cross-dataset robustness and model generalization.

2.5. Generalization Gap and Dataset Characteristic Analysis

To quantify the impact of domain shift, this study employs the generalization gap, defined as the difference between detection performance under same-dataset and cross-dataset evaluation settings[42], [46]. The generalization gap was computed using Equation (1):

$$Gap = mAP50_{same} - mAP50_{cross} \quad (1)$$

Where $mAP50_{same}$ denotes the detection performance obtained using the same dataset for training and testing, and $mAP50_{cross}$ represents the average mAP@50 obtained from the two cross-dataset evaluation scenarios for the corresponding training dataset. A larger gap indicates stronger dependence on dataset-specific characteristics, whereas a smaller gap suggests better robustness and transferability across different domains [45].

To further investigate factors affecting cross-dataset generalization, three dataset characteristics were analyzed: dataset size, average pothole size measured using normalized bounding-box area, and object density measured as the average number of pothole instances per image. These characteristics have been reported to influence detector robustness and generalization across different visual domains [21], [22].

3. RESULTS AND DISCUSSION

This section presents the results obtained from the baseline same-dataset evaluation, cross-dataset evaluation, generalization gap analysis, dataset characteristic analysis, and multi-dataset pretraining experiments. The objective is to evaluate the robustness of YOLOv8n across datasets with different visual characteristics and to investigate whether multi-dataset pretraining can improve cross-dataset generalization.

3.1. Baseline Same-Dataset Detection Performance

To establish a baseline reference, YOLOv8n was first trained and evaluated on the same dataset. This experiment represents the conventional evaluation protocol commonly used in pothole detection studies. The performance obtained from the same-dataset evaluation is summarized in Table 4.

Table 4. Baseline Same-Dataset Detection Performance

Train	Test	Precision	Recall	mAP50	mAP50-95
MWPD	MWPD	0.778	0.664	0.710	0.361
Jaygala	Jaygala	0.701	0.579	0.656	0.331
Andrew	**Andrew**	**0.807**	**0.750**	**0.816**	**0.515**

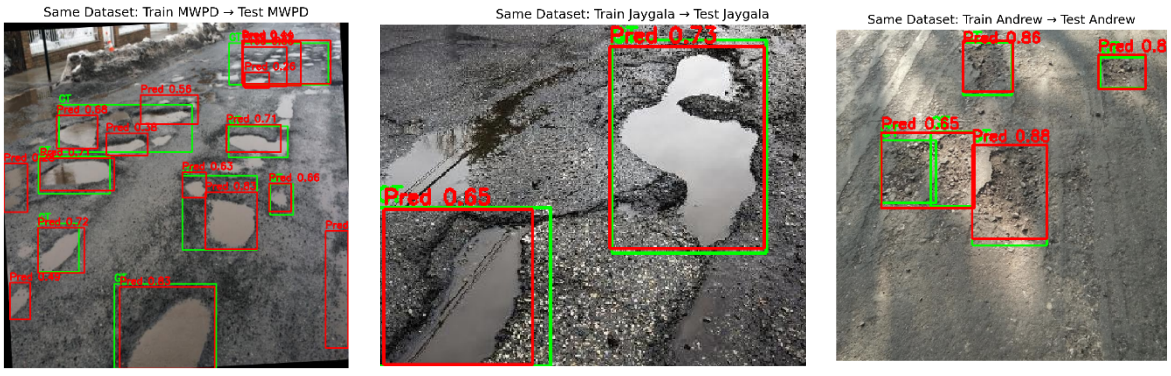
As shown in Table 4, the model achieved satisfactory detection performance on all three datasets under the same-dataset setting. Among them, Andrew produced the highest detection performance (mAP50 = 0.816), followed by MWPD (0.710) and Jaygala (0.656). Although Andrew contains fewer training images than MWPD, its potholes are generally larger, making them easier to localize. In contrast, the greater environmental diversity in MWPD and the denser object distribution in Jaygala increase the detection complexity.

To further evaluate the stability of the training process, each same-dataset experiment was repeated using five different random seeds (7, 21, 42, 84, and 123). The resulting mean and standard deviation of the evaluation metrics are presented in Table 5.

Table 5. Performance Stability Across Five Random Seeds (Mean ± Standard Deviation)

Dataset	Precision	Recall	mAP50	mAP50-95
MWPD	0.748 ± 0.021	0.696 ± 0.008	0.723 ± 0.020	0.366 ± 0.011
Jaygala	0.726 ± 0.028	0.630 ± 0.017	0.700 ± 0.010	0.352 ± 0.005
Andrew	**0.819 ± 0.026**	**0.717 ± 0.051**	**0.800 ± 0.030**	**0.500 ± 0.022**

To evaluate the stability of the training process, each same-dataset experiment was repeated using five different random seeds (7, 21, 42, 84, and 123). As summarized in Table 5, YOLOv8n exhibited relatively small standard deviations across all evaluation metrics, indicating stable and reproducible performance despite different random initializations. The Andrew dataset achieved the highest average detection accuracy (mAP50 = 0.800 ± 0.030 and mAP50-95 = 0.500 ± 0.022), while MWPD showed the most consistent Recall (0.696 ± 0.008), suggesting more stable convergence during training. Overall, these results demonstrate that the reported same-dataset performance is not sensitive to random seed selection. Representative detection results obtained under the same-dataset evaluation are presented in Figure 2.



(a) (b) (c)

Figure 2. Representative Pothole Detection Results Under Same-Dataset Evaluation, (a) Model trained and tested on MWPD, (b) model trained and tested on Jaygala, and (c) model trained and tested on Andrew.

Figure 2 shows that YOLOv8n accurately localizes potholes under the same-dataset setting, with most potholes detected using precise bounding boxes and high confidence scores. These qualitative observations are consistent with the quantitative results in Table 4, confirming that the detector performs well when evaluated within the same data distribution.

3.2. Baseline Cross-Dataset Detection Performance

To evaluate model robustness under distribution shift, each baseline YOLOv8n model was tested on datasets that were not used during training. Unlike same-dataset evaluation, this protocol measures the transferability of learned features across different visual domains. The cross-dataset detection performance is summarized in Table 6.

Table 6. Baseline cross-dataset detection performance

Train	Test	mAP50
MWPD	Jaygala	0.516
MWPD	Andrew	0.718
Jaygala	MWPD	0.584
Jaygala	**Andrew**	**0.905**
Andrew	MWPD	0.536
Andrew	Jaygala	0.604

To facilitate comparison across all dataset combinations, the mAP50 results are visualized as a heatmap in Figure 3.

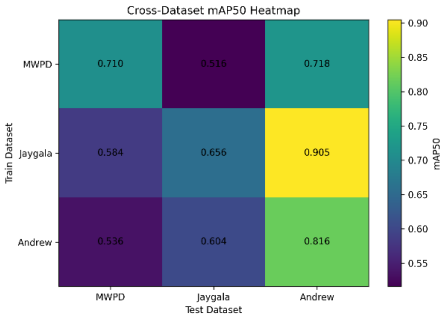
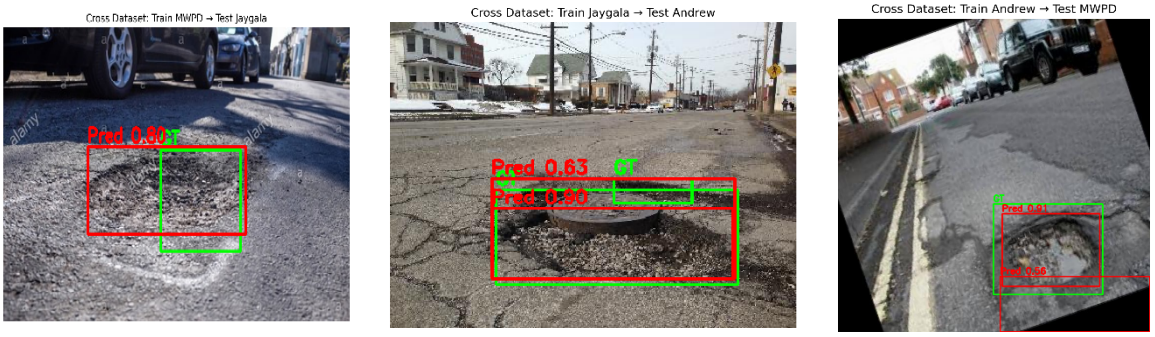


Figure 3. Cross-Dataset Performance Heatmap Based on Map50 Values.

As shown in Table 5 and Figure 3, cross-dataset performance varies considerably across dataset pairs, indicating that the learned representations are sensitive to changes in data distribution. The highest performance was achieved when the model trained on Jaygala was evaluated on Andrew (mAP50 = 0.905), whereas the model trained on Andrew obtained the lowest performance on MWPD (mAP50 = 0.536). These results demonstrate that strong same-dataset performance does not necessarily translate into robust cross-dataset generalization. To provide qualitative evidence of these quantitative results, representative cross-dataset detection examples are shown in Figure 4.



(a) (b) (c)

Figure 4. Representative Pothole Detection Results Under Cross-Dataset Evaluation, (a) Model trained on MWPD and tested on Jaygala, (b) model trained on Jaygala and tested on Andrew, and (c) model trained on Andrew and tested on MWPD.

Figure 4 provides qualitative examples of cross-dataset detection results. Compared with the same-dataset evaluation, more missed detections and less accurate localization are observed, particularly when the testing images differ substantially from the training domain. These observations support the quantitative results presented in Table 5, confirming the impact of domain shift on detection performance.

3.3. Robustness Assessment Through Generalization Gap Analysis

To quantify the impact of domain shift, the generalization gap defined in Equation (1) was computed for each training dataset. The resulting gap values are summarized in Table 7.

Table 7. Baseline Generalization Gap

Training Dataset	Generalization Gap
MWPD	0.093
Jaygala	**-0.088**
Andrew	0.246

Figure 5 provides a visual comparison of the generalization gaps across the three training datasets.

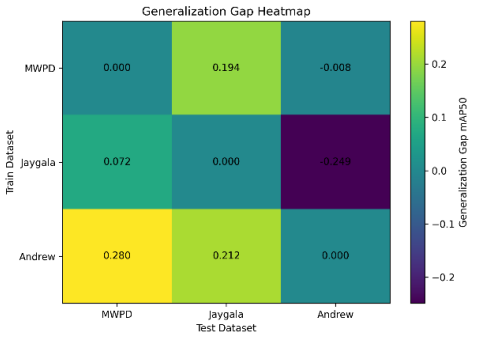


Figure 5. Generalization gap obtained from baseline cross-dataset evaluation.

As shown in Table 6 and Figure 5, Andrew exhibits the largest generalization gap (0.246), indicating substantial performance degradation when evaluated on unseen datasets. In contrast, MWPD achieves the smallest positive gap (0.093), suggesting better robustness and feature transferability. Interestingly, Jaygala produces a negative generalization gap (−0.088), indicating that its average cross-dataset performance exceeds its same-dataset performance. Overall, these results demonstrate that high same-dataset detection accuracy does not necessarily imply strong cross-dataset generalization. Furthermore, the generalization gap provides a simple yet informative measure for assessing model robustness beyond conventional same-dataset evaluation.

3.4. Dataset Characteristic Analysis

To better understand the differences in cross-dataset performance, the characteristics of each dataset were analyzed in terms of pothole size and object density. These characteristics were selected because they may influence object localization and feature learning, thereby affecting model robustness across different datasets. The average pothole size measured using normalized bounding box area is presented in Table 8.

Table 8. Average Pothole Size in the Test Set

Dataset	Mean Bounding Box Area
Andrew	**0.0789**
Jaygala	0.0705
MWPD	0.0621

As shown in Table 8, Andrew contains the largest potholes on average, followed by Jaygala and MWPD. Larger potholes generally provide richer visual information and are easier to localize, which likely contributes to the superior same-dataset detection performance achieved by Andrew. The average object density of the training datasets is presented in Table 9.

Table 9. Average Pothole Density in Training Images

Dataset	Bounding Boxes per Image
MWPD	2.37
Jaygala	**3.26**
Andrew	2.64

Table 8 shows that Jaygala has the highest object density, with an average of 3.26 potholes per image. Detecting multiple potholes within a single image may encourage the model to learn more transferable feature representations, which could explain its relatively strong cross-dataset performance despite achieving lower same-dataset accuracy than Andrew.

Overall, the results suggest that dataset characteristics, particularly pothole size and object density, influence both detection performance and cross-dataset generalization. Larger potholes tend to improve detection accuracy within the same dataset, whereas higher object density may contribute to learning more transferable representations across different datasets.

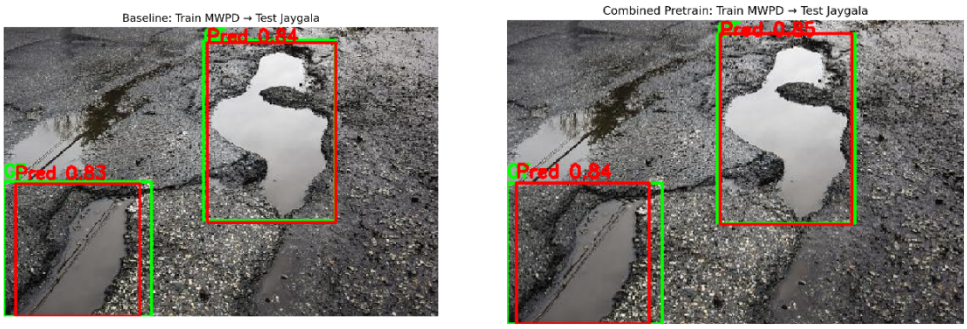
3.5. Effect of Multi-Dataset Pretraining and Fine-Tuning

To investigate whether increasing training data diversity improves cross-dataset robustness, multi-dataset pretraining was performed by combining the training images from all three datasets before fine-tuning on each target dataset. The resulting generalization gaps are summarized in Table 10.

Table 10. Comparison of Generalization Gaps Before and After Multi-Dataset Pretraining

Dataset	Baseline Gap	Combined Gap
MWPD	0.093	**0.048**
Jaygala	-0.088	**-0.122**
Andrew	**0.246**	0.261

As shown in Table 9, multi-dataset pretraining reduced the generalization gap for MWPD from 0.093 to 0.048, indicating improved robustness after exposure to more diverse training data. Jaygala maintained a negative generalization gap, indicating that its cross-dataset performance remained comparable to or slightly better than its same-dataset performance after pretraining. In contrast, Andrew showed a marginal increase in the generalization gap from 0.246 to 0.261. These results suggest that the effectiveness of multi-dataset pretraining depends on the characteristics of the target dataset and does not uniformly improve generalization across all domains.To provide qualitative evidence of the quantitative results, representative detection results before and after multi-dataset pretraining are presented in Figure 6.



(a) (b)

Figure 6. Comparison of pothole detection results before and after multi-dataset pretraining for the MWPD→Jaygala evaluation scenario. (a) Baseline model and (b) multi-dataset pretrained model.

Figure 6 qualitatively illustrates the effect of multi-dataset pretraining on cross-dataset detection performance. Compared with the baseline model, the pretrained model produces more consistent bounding-box localization and slightly higher confidence scores while maintaining

accurate pothole detection. These qualitative observations are consistent with the reduction in the generalization gap reported in Table 9, particularly for the MWPD dataset.

3.6. Discussion in Relation to Previous Studies

To better position the proposed framework within the existing literature, **Table 11** compares representative studies on cross-dataset evaluation, domain generalization, dataset diversity, and detector robustness with the proposed framework.

Table 11. Comparison between representative previous studies and the proposed framework

Study	Main Focus	Key Finding	Relation to This Study
[23]	Cross-dataset evaluation	Same-dataset performance degrades on unseen datasets.	Confirms the domain shift observed in this study.
[17]	Domain adaptation	Domain adaptation reduces distribution mismatch.	Uses multi-dataset pretraining instead of architectural adaptation.
[16]	Cross-domain YOLO	Feature alignment improves cross-domain detection.	Improves robustness through data diversity while preserving YOLOv8n.
[29]	Dataset diversity	Diverse training data improve model generalization.	Consistent with the reduced generalization gap after pretraining.
[24]	Domain generalization	Multi-domain training learns more domain-invariant features.	Supports the proposed multi-dataset pretraining strategy.
[22]	Dataset characteristics	Visual variations influence detection performance.	Extends prior work by quantitatively analyzing pothole size, object density, and robustness.

The cross-dataset experiments confirm the impact of domain shift, as detector performance consistently decreased when evaluated on unseen datasets. Although Andrew achieved the highest same-dataset performance, it also produced the largest generalization gap. This finding agrees with previous studies [23], [17], which reported that high within-dataset accuracy does not necessarily indicate strong cross-dataset robustness. Unlike previous work, this study explicitly quantifies robustness using the generalization gap.

The dataset characteristic analysis further shows that pothole size and object density are associated with cross-dataset performance. Andrew contains larger potholes, whereas Jaygala has the highest object density and maintained a negative generalization gap after pretraining. These findings extend previous studies [22], by quantitatively relating dataset characteristics to detector robustness.

The multi-dataset pretraining results support previous findings that increasing dataset diversity can improve model robustness [29], [24]. The reduced generalization gap observed for MWPD indicates improved cross-dataset performance, whereas the slight increase for Andrew suggests that dataset diversity alone is insufficient to completely overcome domain shift, consistent with previous studies on domain adaptation [17], [16].

Overall, the proposed framework demonstrates that combining cross-dataset evaluation, generalization gap analysis, and multi-dataset pretraining provides a more comprehensive assessment of pothole detector robustness than conventional same-dataset evaluation alone.

4. CONCLUSION

This study investigated the robustness of YOLOv8n for cross-dataset pothole detection using three publicly available datasets with different visual characteristics. The results demonstrate that high same-dataset performance does not necessarily indicate strong cross-dataset generalization. Although the Andrew dataset achieved the highest same-dataset performance (mAP@50 = 0.816), it also produced the largest generalization gap (0.246), confirming the impact of domain shift. The dataset characteristic analysis further showed that pothole size and object density influence model transferability across datasets, while multi-dataset pretraining reduced the generalization gap for MWPD from 0.093 to 0.048 but did not consistently improve robustness across all datasets.

Overall, the findings indicate that cross-dataset evaluation provides a more realistic assessment of pothole detector robustness than conventional same-dataset evaluation. The proposed framework offers a practical approach for evaluating model generalization under heterogeneous road environments and may support the development of more reliable intelligent road inspection

systems. Future research will investigate self-supervised learning and domain adaptation techniques to further improve cross-dataset robustness.

ACKNOWLEDGEMENTS

The authors would like to thank Universitas Muhammadiyah Purworejo for providing research facilities and computational resources that supported this study.

CONFLICT OF INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

DATA AVAILABILITY

The datasets used in this study are publicly available from their official repositories. The Multi-Weather Pothole Dataset (MWPD) is available from Mendeley Data, the Jaygala Pothole Dataset is available on GitHub at https://github.com/jaygala24/pothole-detection, and the Andrew Pothole Dataset is available on Kaggle at https://www.kaggle.com/datasets/andrewmvd/pothole-detection. All datasets were preprocessed into a unified YOLO annotation format to ensure consistency across all experiments.

AUTHOR CONTRIBUTIONS

Hanifatus Sa'diyah Widihasaniputri: Conceptualization; Data Curation; Formal Analysis; Investigation; Methodology; Software; Validation; Visualization; Writing – Original Draft. **Oddy Virgantara Putra:** Conceptualization; Formal Analysis; Methodology; Supervision; Validation; Writing – Review & Editing. **Murhadi:** Conceptualization; Methodology; Supervision; Writing – Review & Editing.

REFERENCES

[1] B. P. Statistik, "Statistik Transportasi Darat 2024," Jakarta, 2024. [Online]. Available: https://www.bps.go.id/id/publication/2025/12/01/ed7ff73d58fc0719ee3df145/statistik-transportasi-darat-2024.html.

[2] A. Salunke, "Road Accident Detection Using AI and ML.," *Interantional J. Sci. Res. Eng. Manag.*, vol. 09, no. 05, pp. 1–9, 2025, doi: 10.55041/ijsrem48619.

[3] A. A. Sami, S. M. N. Sakib, K. Deb, and I. H. Sarker, "Improved YOLOv5-Based Real-Time Road Pavement Damage Detection in Road Infrastructure Management," *Algorithms*, vol. 16, no. 9, p. 452, 2023, doi: 10.3390/a16090452.

[4] K. S. Sri, M. Prasanna, K. V VENKATARAO, K. V Subrahmanyam, and P. Ramesh, "Deep Learning-Based Classification of Road Conditions Using a Customized VGG-Style CNN for Pothole Detection," *Int. J. Eng. Sci. Invent.*, vol. 14, no. 4, pp. 73–80, 2025, doi: 10.35629/6734-14047380.

[5] K. R. Ahmed, "Smart pothole detection using deep learning based on dilated convolution," *Sensors*, vol. 21, no. 24, 2021, doi: 10.3390/s21248406.

[6] K. W. Yuanhang Tang, Kexin Li, "Research on intelligent detection of pavement damage based on CNN," *Math. Model. Eng.*, vol. 8, no. 4, pp. 98–107, 2022, doi: 10.21595/mme.2022.22918.

[7] H. Samma, S. A. Suandi, N. A. Ismail, S. Sulaiman, and L. L. Ping, "Evolving Pre-Trained CNN Using Two-Layers Optimizer for Road Damage Detection from Drone Images," *IEEE Access*, vol. 9, pp. 158215–158226, 2021, doi: 10.1109/ACCESS.2021.3131231.

[8] G. Guo and Z. Zhang, "Road damage detection algorithm for improved YOLOv5," *Sci. Rep.*, pp. 1–12, 2022, doi: 10.1038/s41598-022-19674-8.

[9] G. Yu and X. Zhou, "An Improved YOLOv5 Crack Detection Method Combined with a Bottleneck Transformer," *Mathematics*, vol. 11, no. 10, pp. 1–12, 2023.

[10] J. Wang, K. Wang, and K. Li, "Research on road damage recognition and classification based on improved VGG-19," *Math. Model. Eng.*, vol. 9, no. 4, pp. 115–129, 2023, doi: 10.21595/mme.2023.23455.

[11] S. Zhou *et al.*, "Enhancing autonomous pavement crack detection : Optimizing YOLOv5s algorithm with advanced deep learning techniques," *Measurement*, vol. 240, no. May 2024, p. 115603, 2025, doi: 10.1016/j.measurement.2024.115603.

[12] J. Zeng and H. Zhong, "YOLOv8 ‐ PD : an improved road damage detection algorithm based on YOLOv8n model," *Sci. Rep.*, pp. 1–14, 2024, doi: 10.1038/s41598-024-62933-z.

[13] S. Parvin, F. Munsy, T. Rahat, A. Nahar, and K. Nur, "Results in Engineering Robust multi-weather pothole detection : An enhanced YOLOv9 trained on the MWPD dataset," *Results Eng.*, vol. 28, no. August, p. 107817, 2025, doi: 10.1016/j.rineng.2025.107817.

[14] Y. Liu, S. Zhou, X. Liu, C. Hao, B. Fan, and J. Tian, "Unbiased Faster R-CNN for Single-source Domain Generalized Object Detection," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 28838–28847, doi: 10.1109/cvpr52733.2024.02724.

[15] C. Diaz-Ruiz *et al.*, "Ithaca365: Dataset and Driving Perception under Repeated and Challenging Weather Conditions," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 21351–21360, doi: 10.1109/cvpr52688.2022.02069.

[16] J. Wei, Q. W. Id, and Z. Zhao, "YOLO-G : Improved YOLO for cross-domain object detection," *PLoS One*, vol. 19, no. 9, pp. 1–16, 2023, doi: 10.1371/journal.pone.0291241.

[17] G. Li, Z. Ji, X. Qu, R. Zhou, and D. Cao, "Cross-Domain Object Detection for Autonomous Driving : A Stepwise Domain Adaptative YOLO Approach," *IEEE Trans. Intell. Veh.*, vol. 7, no. 3, pp. 603–615, 2022.

[18] M. Jakubec, E. Lieskovska, B. Bucko, and K. Zabovska, "Pothole detection in adverse weather : leveraging synthetic images and attention ‐ based object detection methods," *Multimed. Tools Appl.*, vol. 83, no. 39, pp. 86955–86982, 2024, doi: 10.1007/s11042-024-19723-6.

[19] S. Lamchoudi and A. Mourhir, "Pothole Detection for Road Safety on Moroccan Roads," *2025 IEEE 8th Congr. Inf. Sci. Technol.*, pp. 336–341, 2025, doi: 10.1109/cist65886.2025.11224074.

[20] R. Buettner, S. Member, S. Bertram, and L. Fischer-brandies, "The Impact of Combining Datasets on the Robustness of Deep Learning Architectures : A Cross-Dataset Analysis," *IEEE Access*, vol. 13, no. August, pp. 151993–152009, 2025, doi: 10.1109/ACCESS.2025.3604689.

[21] A. Vijayakumar and S. Vairavasundaram, "YOLO-based Object Detection Models: A Review and its Applications," *Multimed. Tools Appl.*, vol. 83, pp. 83535–83574, 2024, doi: 10.1007/s11042-024-18872-y.

[22] Y. Zanevych, V. Yovbak, O. Basystiuk, N. Shakhovska, S. Fedushko, and S. Argyroudis, "Evaluation of Pothole Detection Performance Using Deep Learning Models Under Low-Light Conditions," *Sustainability*, 2024, doi: 10.3390/su162410964.

[23] M. Cantone, C. Marrocco, and A. Bria, "Machine Learning in Network Intrusion Detection : A Cross-Dataset Generalization Study," *IEEE Access*, vol. 12, pp. 144489–144508, 2024.

[24] Y. Chen *et al.*, "Achieving domain generalization for underwater object detection by domain mixup and contrastive learning," *Neurocomputing*, vol. 528, pp. 20–34, 2021, doi: 10.1016/j.neucom.2023.01.053.

[25] S. Abdelwahed, B. Sharobim, B. Wasfey, and L. Said, "Advancements in real-time road damage detection: a comprehensive survey of methodologies and datasets," *J. Real-Time Image Process.*, vol. 22, 2025, doi: 10.1007/s11554-025-01683-1.

[26] Y. Zhao, Z. Zhong, N. Zhao, N. Sebe, and G. Lee, "Style-Hallucinated Dual Consistency Learning: A Unified Framework for Visual Domain Generalization," *Int. J. Comput. Vis.*, vol. 132, pp. 837–853, 2022, doi: 10.1007/s11263-023-01911-w.

[27] Z. Demirel, S. T. Nasraldeen, Ö. Pehlivan, S. Shoman, M. Albdairi, and A. Almusawi, "Comparative Evaluation of YOLO and Gemini AI Models for Road Damage Detection and Mapping," *Futur. Transp.*, 2025, doi: 10.3390/futuretransp5030091.

[28] O. Khare, S. Gandhi, A. Rahalkar, and S. Mane, "YOLOv8-Based Visual Detection of Road Hazards: Potholes, Sewer Covers, and Manholes," in *2023 IEEE Pune Section International Conference (PuneCon)*, 2023, pp. 1–6, doi: 10.1109/punecon58714.2023.10449999.

[29] D. Arya *et al.*, "RDD2022: A multi-national image dataset for automatic road damage detection," *Geosci. Data J.*, vol. 11, pp. 846–862, 2022, doi: 10.1002/gdj3.260.
[30] S. Parvin, M. Ahmed, and M. Rahman, "Multi-Weather Pothole Detection Dataset (MWPD)," *Mendeley Data*, 2024. https://data.mendeley.com/datasets/s5hx9n2jc3/2 (accessed Apr. 24, 2026).
[31] P. A. Chitale, K. Y. Kekre, H. R. Shenai, R. Karani, and J. P. Gala, "Pothole Detection and Dimension Estimation System using Deep Learning (YOLO) and Image Processing," in *2020 35th International Conference on Image and Vision Computing New Zealand (IVCNZ)*, 2023, pp. 1–6, doi: 10.1109/IVCNZ51579.2020.9290547.
[32] Andrew MVD, "Pothole Detection Dataset," *Kaggle*, 2023. https://www.kaggle.com/datasets/andrewmvd/pothole-detection (accessed Apr. 28, 2026).
[33] B. Shakya, O. El-Gayar, J. Kaabi, and K. M. Ahmed, "Small-Object Detection at the Edge: A Pareto-Efficient Benchmark of Lightweight YOLO Models on UAV and Overhead Datasets," *IEEE Access*, vol. 14, pp. 528–548, 2026, doi: 10.1109/access.2025.3648968.
[34] H. M. Aljami *et al.*, "Benchmarking YOLOv8 Variants for Object Detection Efficiency on Jetson Orin NX for Edge Computing Applications," *Computers*, vol. 15, pp. 1–20, 2026.
[35] L. Rey *et al.*, "A Performance Analysis of You Only Look Once Models for Deployment on Constrained Computational Edge Devices in Drone Applications," *Electronics*, vol. 14, pp. 1–24, 2025.
[36] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 779–788, doi: 10.1109/cvpr.2016.91.
[37] J. Terven, D.-M. Córdova-Esparza, and J.-A. Romero-González, "A Comprehensive Review of YOLO Architectures in Computer Vision: From YOLOv1 to YOLOv8 and YOLO-NAS," *Mach. Learn. Knowl. Extr.*, vol. 5, pp. 1680–1716, 2023, doi: 10.3390/make5040083.
[38] A. Wang *et al.*, "YOLOv10: Real-Time End-to-End Object Detection," *ArXiv*, vol. abs/2405.14458, 2024, doi: 10.48550/arxiv.2405.14458.
[39] I. Ahmed, M. Ahmad, M. Siddiqi, A. Chehri, and G. Jeon, "Toward AI-Powered Edge Intelligence for Object Detection in Self-Driving Cars: Enhancing IoV Efficiency and Safety," *IEEE Internet Things J.*, vol. 12, pp. 16990–16997, 2025, doi: 10.1109/jiot.2025.3534737.
[40] M. Bakirci, "Enhancing vehicle detection in intelligent transportation systems via autonomous UAV platform and YOLOv8 integration," *Appl. Soft Comput.*, vol. 164, p. 112015, 2024, doi: 10.1016/j.asoc.2024.112015.
[41] M. Cao, Q.-V. Tran, N.-M. Nguyen, and K. Chang, "Survey on performance of deep learning models for detecting road damages using multiple dashcam image resources," *Adv. Eng. Informatics*, vol. 46, p. 101182, 2020, doi: 10.1016/j.aei.2020.101182.
[42] T. Diwan, G. Anirudh, and J. Tembhurne, "Object detection using YOLO: challenges, architectural successors, datasets and applications," *Multimed. Tools Appl.*, vol. 82, pp. 9243–9275, 2022, doi: 10.1007/s11042-022-13644-y.
[43] M. Paiano, S. Martina, C. Giannelli, and F. Caruso, "Transfer learning with generative models for object detection on limited datasets," *Mach. Learn. Sci. Technol.*, vol. 5, 2024, doi: 10.1088/2632-2153/ad65b5.
[44] X. Xu, H. Zhang, Yan, K. Liu, H. Bao, and X. Qian, "TranSDet: Toward Effective Transfer Learning for Small-Object Detection," *Remote. Sens.*, vol. 15, p. 3525, 2023, doi: 10.3390/rs15143525.
[45] P. Goel and A. Ganatra, "Unsupervised Domain Adaptation for Image Classification and Object Detection Using Guided Transfer Learning Approach and JS Divergence," *Sensors (Basel).*, vol. 23, 2023, doi: 10.3390/s23094436.
[46] R. Chakraborty, H. Mitra, P. Shivakumara, and U. Pal, "Towards Robust Cross-Dataset Object Detection Generalization under Domain Specificity," *ArXiv*, vol. abs/2601.09497, 2026, doi: 10.48550/arxiv.2601.09497.
[47] Z. Xue, F. Yang, S. Rajaraman, G. Zamzmi, and S. Antani, "Cross Dataset Analysis of

Domain Shift in CXR Lung Region Detection," *Diagnostics*, vol. 13, 2023, doi: 10.3390/diagnostics13061068.

[48] Y. Gao, W. Luo, S. Zhang, N. S. Ahmad, X. Wang, and P. Goh, "Benchmarking YOLOv8 to YOLOv13 for robust hand gesture recognition in human–robot interaction," *Sci. Rep.*, vol. 15, 2025, doi: 10.1038/s41598-025-23925-9.

[49] M. El-Geneedy, H. E.-D. Moustafa, H. Khater, S. Abd-Elsamee, and S. Gamel, "Advanced real-time detection of acute ischemic stroke using YOLOv12, YOLOv11, and YOLO-NAS: a comparative study for multi-class classification," *Sci. Rep.*, vol. 15, 2025, doi: 10.1038/s41598-025-18997-6.