Elinvo (Electronics, Informatics, and Vocational Education)
Vol 10, No. 2, 2025, pp. 147-168
ISSN 2580-6424 (printed) | ISSN 2477-2399 (online)

A Performance Comparison of YOLOv5 and YOLOv8 for Road Damage Object Detection on a Mixed GRDDC–PUPR Dataset

Yuniar Indrihapsari [*], Danang Wijaya [2], Satya Adhiyaksa Ardy [3], Ikhwan Inzaghi Siswanto [4], Dhista Dwi Nur Ardiansyah [5], Widya Ardianto [6]

[1,4,5,6]Universitas Negeri Yogyakarta, Yogyakarta, Indonesia
[2]National Central University, Taoyuan, Taiwan
[3]National Taiwan University of Science and Technology, Taipei, Taiwan
[1] yuniar@uny.ac.id *; [2] danangwijaya750@gmail.com; [3] satyaadhiyaksa@gmail.com; [4] ikhwaninzaghi999@gmail.com; [5] dhistadna@gmail.com; [6] widyaardianto.9090@gmail.com
* corresponding author

Article Info

Article history:
Received August 02, 2025
Revised December 17, 2025
Accepted December 30, 2025
Available Online dd, yyy

Keywords:
YOLOv5; YOLOv8; Deep learning; GRDDC 2020 Dataset; Object detection

Abstract

Accurate road damage detection is vital for ensuring road safety and supporting timely infrastructure maintenance. However, the question remains open as to which YOLO variant offers the best trade-off between accuracy and efficiency for road damage detection under Indonesian conditions when models are trained on a mixed international–local dataset. This study evaluates and compares the performance of four YOLO models: YOLOv5-S, YOLOv5-M, YOLOv8-S, and YOLOv8-M, for detecting road damage types, including Alligator Cracks, Longitudinal Cracks, Transverse Cracks, Potholes, and Lateral Cracks. The models are trained on a combined dataset from GRDDC 2020 and the Ministry of Public Works and Housing (PUPR) of the Republic of Indonesia, addressing challenges such as class imbalance and diverse road and lighting conditions. Results show that YOLOv8-M achieves the highest mAP@0.5 (0.435), with strong precision and recall for prominent damage types, making it the most reliable option for high-accuracy applications. The YOLOv5-M is generally well-balanced in terms of precision and recall, while the YOLOv5-S focuses more on the concept of recall, thus being appropriate in situations where more cases of the damaged type need to be detected. It is also noted that all models still have problems with less significant kinds of road damage, especially Lateral Cracks, which has a likelihood of being identified under the Background category. Through comparison, it was determined that YOLOv8-M has the highest accuracy among the models using the mixed GRDDC-PUPR scheme, aside from still having improvements in the minority categories.

INTRODUCTION

Indonesia has registered major growth in the mileage coverage of roads over the past two decades, making the need for monitoring the conditions necessary for timely maintenance more pressing. Statistics Indonesia (BPS) data show a total mileage coverage increased by 18.73%, from 542,160 km in 2019 to 550,735 km in 2023 [1]. Even with the growth, the task of maintaining the relevant infrastructure has become difficult, given that the condition of the roads can result in reduced transportation safety and efficiency.

Road transport is the main mode used in goods distribution in Indonesia, with road conditions playing an important role in economic performance. Bad road conditions accelerate the chances of traffic accidents, logistics processing, and economic loss [2]. In 2022, BPS documented traffic accidents

amounting to 139,258 cases with estimated economic loss around IDR 280 billion [3]. Moreover, poor road conditions and inadequate road construction quality performance had been associated with the deterioration of road longevity due to decreased performance quality [4], [5]. This highlights the significance associated with accurate and appropriate scaled performance in road condition assessment.

Traditional manual road surveying practices, remaining common in Indonesia until now, also have limitations in terms of scalability, impartialness, and objectiveness. Moreover, the findings of the survey may also differ from surveyor to surveyor and are dependent on the available time and labor. On the other hand, the use of computer vision-based AI for the automation process of road damage detection can help address the maintenance issue through faster and more impartial evaluation for model selection for the automation process to be implemented in the Indonesian setting. Therefore, a common framework for the comparison of the YOLO models is crucial for the identification of the most suitable models for a balance between detection precision and real-time applicability in the Indonesian context.

In real-time object detection models, the You Only Look Once (YOLO) series of models have gained popularity across the board owing to their overall efficiency and processing speed in equal measure [6]. Competitive efficiency of YOLO series-based models in road damage recognition tasks across multiple nations and sets of research datasets had been identified in past studies [7]. However, their efficiency can be affected in domain-specific scenarios in terms of variations in light conditions, road texture, and complex road crack patterns, among others, typical of tropical road scenarios. YOLO series models, specifically YOLOv5, had shown efficient performance in multiple road damage recognition tasks but could potentially struggle in very harsh lighting conditions with complex road patterns such as alligator-cracked roads [8], [9]. Recent models like YOLOv8 showed architectural improvements for enhanced efficiency in road damage recognition tasks [10], [11].

Although research on road damage detection using YOLO-based approaches has emerged, there is a lack of empirical evidence comparing YOLOv5 and YOLOv8 within a common training and testing scheme, particularly for Indonesian roads. Therefore, this paper aims to provide a systematic comparison of YOLOv5 and YOLOv8 on a mixed GRDDC–PUPR dataset under Indonesian tropical road conditions. Specifically, we compare YOLOv5-S, YOLOv5-M, YOLOv8-S, and YOLOv8-M trained and evaluated on a combined dataset that integrates international data from GRDDC 2020 [12] and local road-damage images provided by the Indonesian Ministry of Public Works and Public Housing (PUPR). The main objective is to identify which of these models offers the most favorable balance between detection accuracy and inference speed for road-damage detection in Indonesia. The contributions presented in this research are threefold. First, this research is interested in exploring the impact of training and testing YOLOv5 and YOLOv8 models on a hybrid GRDDC-PUPR dataset which is a more realistic representation for Indonesia. Second, this research is interested in providing benchmark results for comparison on standard evaluation metrics for detection such as mAP at 0.5, precision, recall, and F1 score. Third, this research offers insights into some challenges for future development for creating Indonesia-specific AI-assisted road maintenance systems using machine learning models.

METHODS

Flowchart

This paper uses a structured approach in evaluating and comparing four different YOLO models, namely YOLOv5-S, YOLOv5-M, YOLOv8-S, and YOLOv8-M [13] on the task of road damage detection in Indonesia. From Figure 1, it is observed that the analysis began with an Exploratory Data Analysis phase involving GRDDC 2020 and Indonesian PUPR. This phase helped in analyzing the distribution of road damage issues in the GRDDC 2020 and Indonesian PUPR dataset.

ISSN 2580-6424 (printed) | ISSN 2477-2399 (online)

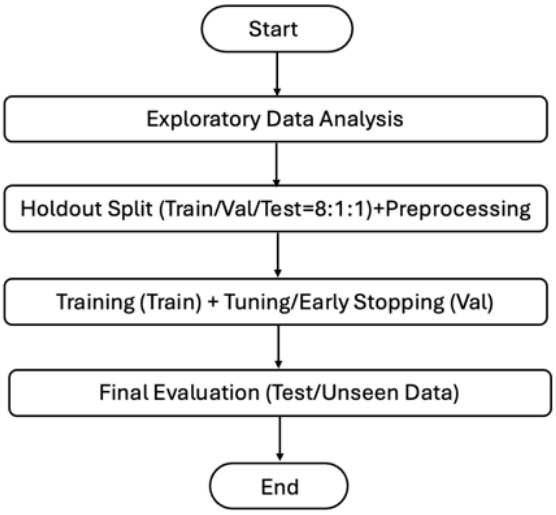


Figure 1. Research Flowchart

The dataset was then split into a train, validation, and test set using a Holdout method with an 8:1:1 split [14]. This was followed by data augmentation, but only on the train set. The hyperparameters were then tuned using the validation set.

All the YOLO models were then trained on the train set, and the selection process was performed on the validation set using Precision, Recall, and mAP [15]. Finally, the best models were evaluated on the test set, thus the unseen set constituted 10% of the entire task. The results are then plotted using Precision–Recall curves and F1 Confidence Analysis at different confidence thresholds. In this manner, the analysis is repeatable and allows for a fair comparison of the models.

Experiment Design

For evaluation, the current research will choose YOLOv5 and YOLOv8, given the efficiency demonstrated by the models in real-time object detection, according to Sami et al. [16]. YOLOv5 is widely used as a reliable baseline model that offers a trade-off between speed and accuracy, while YOLOv8 is more advanced with architecture improvements to optimize computational cost for practical uses [11].

In Figure 2 below, the basic components universally found in both models are described. The YOLOv5 benefits from features extraction defined by the application of CSPDarknet with BottleneckCSP and SPP, a PANet for multi-resolution features fusion, and the YOLOv5 detection head [17].

ISSN 2580-6424 (printed) | ISSN 2477-2399 (online)

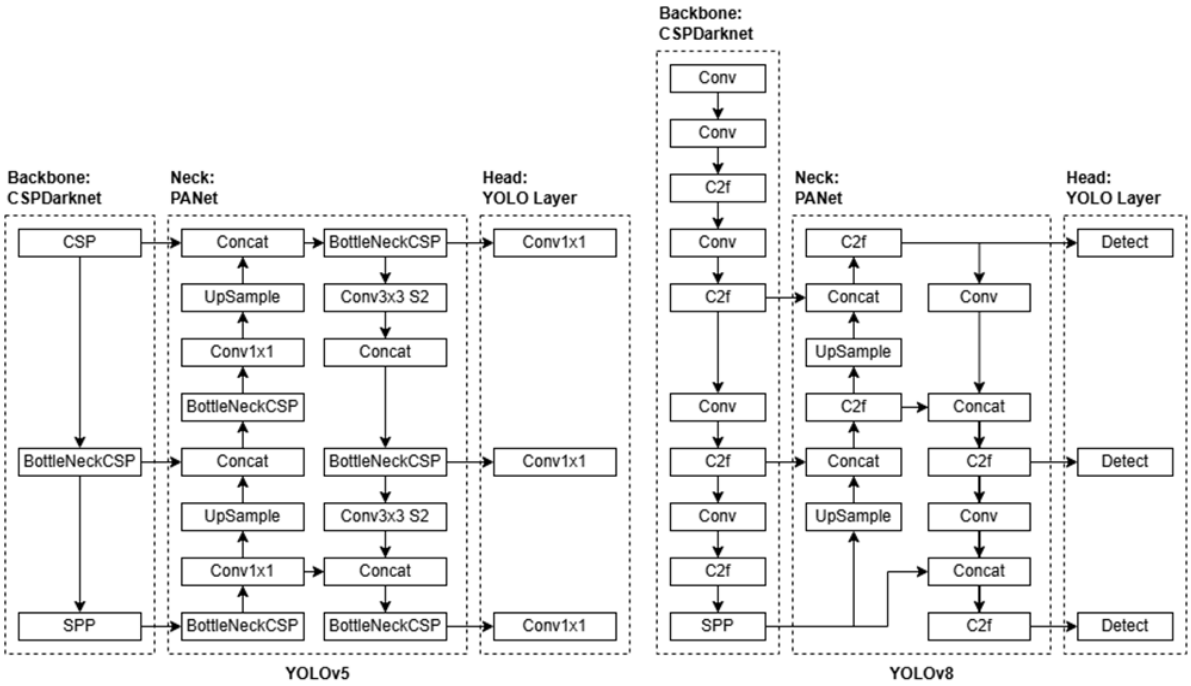


Figure 2. YOLOv5 and YOLOv8 architectures (backbone–neck–head), reproduced/adapted from Ultralytics resources [29], [18].

In YOLOv8, the features are potentially improved with the adoption of C2f features in the backbone and neck, coupled with the application of a detection head [18].

For each model, an equal training protocol and data augmentation process was used on the training split (random cropping, rotation, and color jitter). Handling the problem of imbalance between the classes was considered in the training process through sampling and/or weighting of the classes. Model evaluation was performed on the validation split, while the final results are expressed on the test split (unseen) for each model. The protocols considered are helpful in comparing the YOLOv5 and YOLOv8 models in the Indonesian road setting.

Evaluation Metrics

In the bid to ensure the reproducibility of the evaluation for every version of the YOLO, the performance of the model is tested through the metrics of IoU, precision, recall, F1 score, and the average precision. This is because the above-mentioned parameters are commonly used in the object detection tasks.

Intersection over Union (IoU). IoU measures the overlap between a predicted bounding box B_p and the corresponding ground-truth box B_{gt}, and is defined as:

$$IoU = \frac{|B_p \cap B_{gt}|}{|B_p \cup B_{gt}|} \quad (1)$$

A prediction is considered a true positive when its IoU with the matched ground-truth box exceeds a predefined threshold and the predicted class is correct; otherwise, it contributes to false positives or false negatives depending on matching outcomes.

Precision, Recall, and F1-score. Based on the counts of true positives (TP), false positives (FP), and false negatives (FN), precision and recall are computed as:

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

The F1-score provides a single measure that balances precision and recall, which is particularly informative under class imbalance:

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision \cdot Recall} \quad (3)$$

Average Precision (AP) and mean Average Precision (mAP). For each class, the Precision–Recall (PR) curve is obtained by sweeping the confidence threshold over the detector outputs and computing precision and recall at each operating point. The Average Precision (AP) for a class is defined as the area under its PR curve:

$$AP = \int_0^1 Precision(Recall)\, dRecall. \quad (4)$$

The mean Average Precision (mAP) is then computed by averaging AP values across all evaluated classes:

$$mAP = \frac{1}{K} \sum_{k=1}^{K} AP_k \quad (5)$$

where K denotes the number of classes.

Following common object-detection practice, we report mAP@0.5, which uses an IoU threshold of 0.5, and mAP@0.5:0.95, which averages AP over multiple IoU thresholds from 0.50 to 0.95 with a step of 0.05 (COCO-style evaluation). In addition, we report PR curves to visualize model behavior under varying confidence thresholds, and we analyze F1–confidence curves to identify the confidence level that yields the best balance between precision and recall for each model.

The definitions of IoU, precision, recall, F1-score, AP, and mAP used in this study follow standard formulations commonly adopted in deep learning–based object detection benchmarks and prior work [30].

Dataset

On account of the analogous circumstances in Indonesia and Indian cases relating to roads, this research combines the GRDDC 2020 dataset with the Indonesian PUPR dataset. The GRDDC 2020 dataset is a set of 25, 336 labeled pictures from both Japan and India. On the contrary, the PUPR dataset consists of anonymized pictures of damaged roads in different Indonesian provinces [19], [20]. These PUPR pictures have been collected with permission from the relevant institution, and the timeframe collected was similar in both cases. There are several factors that affect the pattern of damage, including overuse, drainage issues, and rainfall.

After filtering, harmonizing, and implementing quality control, the end result is the fusion of a dataset of 15,581 images gathered from both the GRDDC and the PUPR data sources. Images with inconsistent labels, heavily corrupted images, and images lacking labels were removed from the final dataset. The dataset of 15,581 images comprises images from the GRDDC-2020 source: 11,123 images, and the PUPR source: 4,458 images.

To ensure label consistency across all datasets, a label harmonization process was applied. Damage categories from the GRDDC and PUPR datasets were mapped into a unified taxonomy consisting of eight road damage classes. Semantically equivalent labels were merged, while classes that had no corresponding instances after harmonization were excluded from further analysis. Of the eight

retained classes, three classes contain no instances in the final dataset. In practice, the model was defined with eight labels, but training was effectively performed on five labels with non-zero instances.

Exploratory Data Analysis and Preprocessing

The combined dataset includes 15,581 images and is split 8:1:1 into training (12,453), validation (1,571), and testing (1,557) sets [21]. The split was performed at the image level using stratified random sampling to preserve the class distribution across splits, with a fixed random seed for reproducibility. Although the unified taxonomy defines eight damage categories, only five classes contain non-zero instances in the final combined dataset: Alligator Crack, Longitudinal Crack, Transverse Crack, Pothole, and Lateral Crack. Figure 3 summarizes the distribution of annotated instances for these five active classes in the combined GRDDC–PUPR training set.

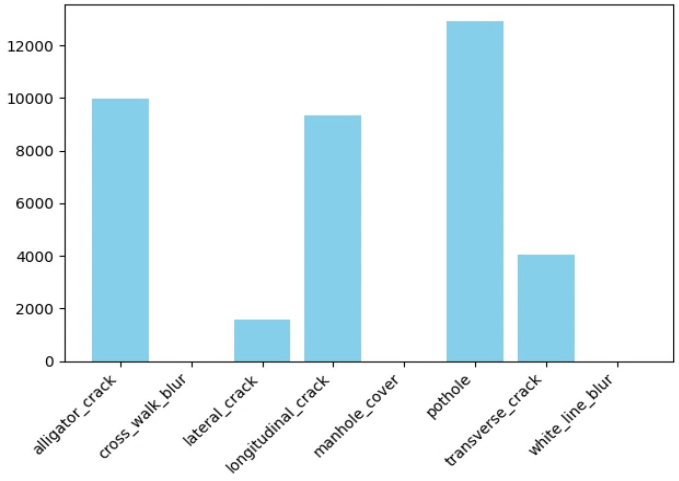


Figure 3. Class distribution of annotated road-damage instances in the combined GRDDC–PUPR training set.

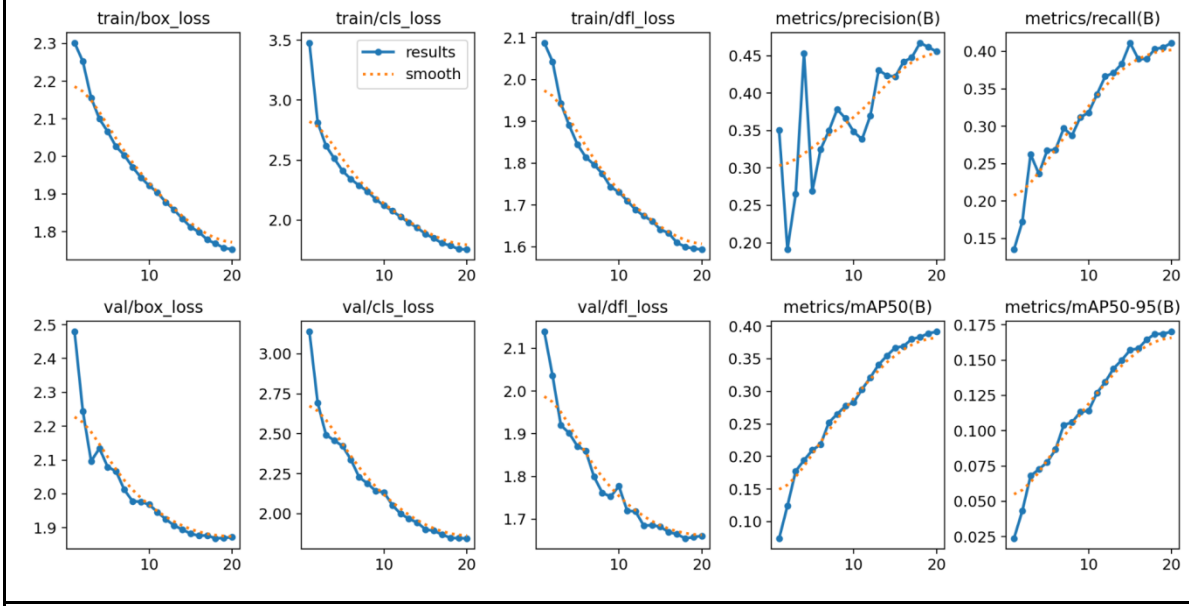
As shown in Figure 3, the biggest group of cracks in the dataset is that of the pothole type, with more than 12,000 examples, followed by longitudinal and alligator cracks with around 9,000 examples for each type. Lateral and transverse cracks are under-represented. The distribution of the cracks is not equally proportioned. This is consistent with previous studies employing GRDDC, in which the normalization exhibited a significantly greater proportion of the population for the pothole type compared with the other [22], [23]. The above-mentioned uneven distribution may affect the training and test of the classifiers. To prevent the majority-class bias, several techniques for addressing the imbalance were employed during training. These include class-aware oversampling of the minority classes and class-weighted loss [24], [25].

RESULT AND DISCUSSION

Training and Validation Result

YOLOv5-S

YOLOv5-S was trained for 10 epochs on the RTX 4060 Ti card with 16 GB of VRAM at a batch size of 8. During both the train and validation phases, there was an evident reduction in all major loss components, including box loss, loss on the classification, and Distribution Focal Loss. This pointed towards the convergence of the entire learning process during the train phase (Figure 4). There were no major oscillations in both the train and validation loss curves.



(a) Yolov5-S training result

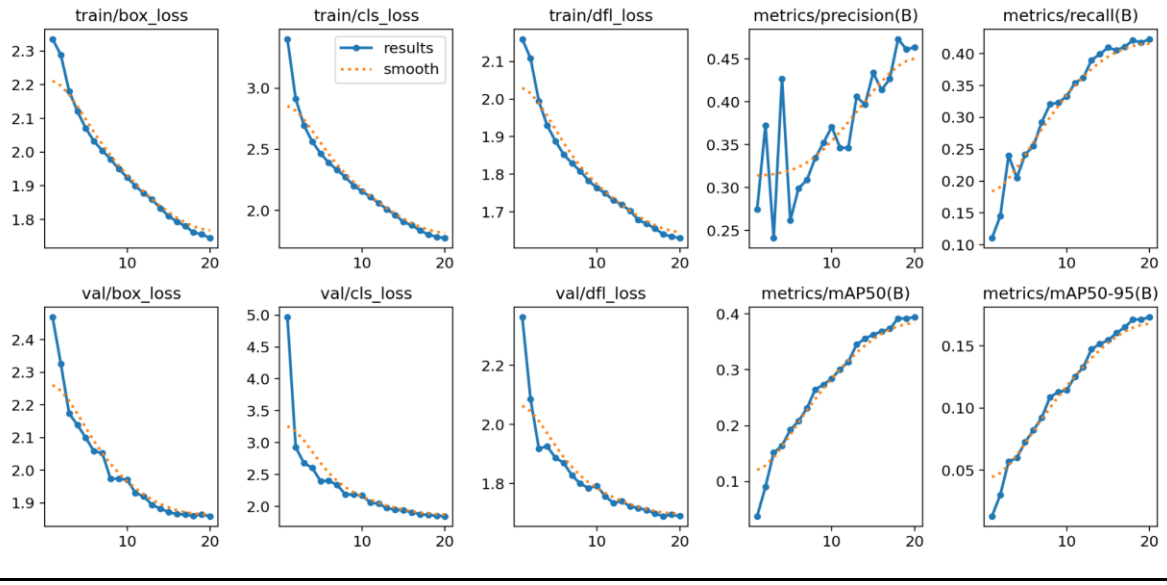
Figure 4. Training and validation curves on the hybrid GRDDC–PUPR dataset for YOLOv5-S

On the quantitative side, YOLOv5-S displayed the strongest improvement regarding the metric of recall, with this metric increasing from an initial training value of 0.135 to 0.426 after the first 10 epochs, while precision reached 0.466 and mAP@0.5 0.404 (Table 1). The overall mAP scores of all variants also elevated from somewhat low values initially to relatively higher, although modest, values. These trends support the hypothesis that YOLOv5-S prioritizes recall compared with the larger YOLOv5-M model.

Suitability/Interpretation: YOLOv5-S is suitable for applications requiring a maximal coverage of detections (e.g., initial screening or safety inspection applications, where a missed detection has an expensive consequence). The training process indicates a reliable enhancement of recall with a small number of training iterations.

YOLOv5-M

In YOLOv5-M, the same set of experiment parameters were used (10 epochs, RTX 4060 Ti, batch size of 8), and there were also declines in both train and validation loss values, reflecting healthy learning behavior (Figure 5). There was a smooth convergence of loss terms, and progress was also seen in validation metrics.



(b) Yolov5-M training result

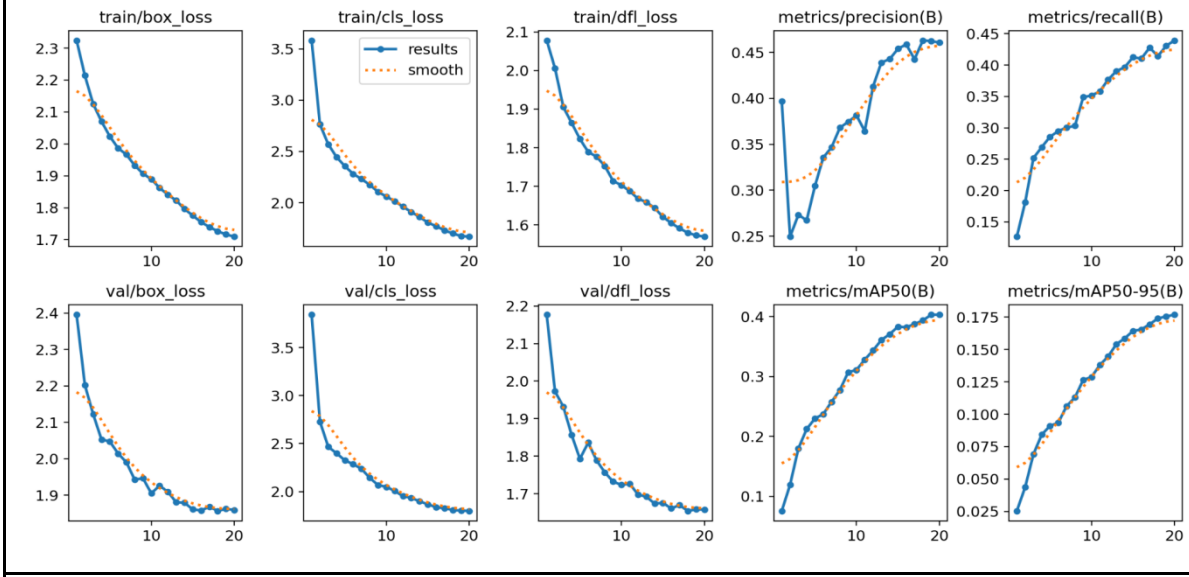
Figure 5. Training and validation curves on the hybrid GRDDC–PUPR dataset for YOLOv5-M

Quantitatively, YOLOv5-M increased recall from 0.110 to 0.421 over the 10 epochs, and achieved a validation precision of 0.488 with mAP@0.5 0.415. These improvements in recall, precision, and mAP (Table 1) indicate that YOLOv5-M attains a favorable trade-off between sensitivity and false positives compared with the smaller YOLOv5-S model.

Interpretation / suitability: YOLOv5-M provides a balanced option when both detection coverage and acceptable false positive rates are required. It is a good candidate for deployments that require moderate inference speed with improved detection robustness over the smallest variant.

YOLOv8-S

During training of the YOLOv8-S, which was done under the same hardware and hyperparameters as in other experiments, there was a steady drop in the loss components, as well as improvements in the validation metrics (Figure 6). The training curves show that there were no sudden divergence issues in the optimization of the model.



(c) Yolov8-S training result

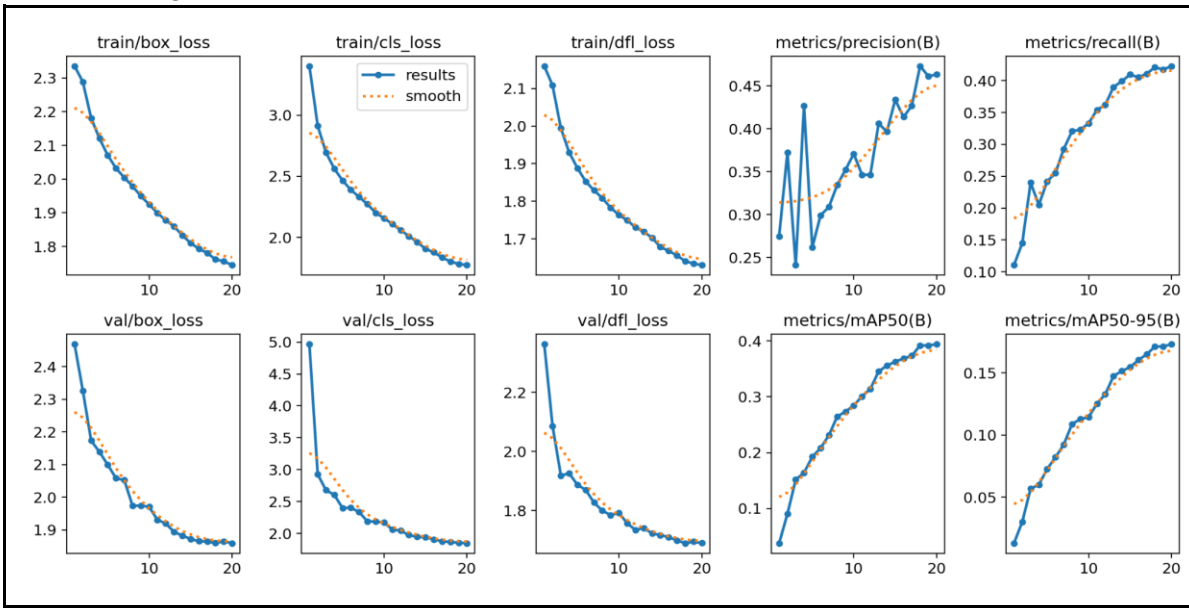
Figure 6. Training and validation curves on the hybrid GRDDC–PUPR dataset for YOLOv8-S

After 10 epochs, YOLOv8-S reached a validation precision of 0.494, recall of 0.437, and mAP@0.5 of 0.430 (Table 1), with all loss components decreasing steadily (Figure 6). The combination of higher precision and substantial mAP gains indicates that YOLOv8-S reduces false positives compared with the YOLOv5 variants, while simultaneously improving overall detection accuracy.

Interpretation/suitability: This variant will be more suited to situations where precision is of key importance and false positives incur significant costs. Based on its behavior, it seems to be acting conservatively and accurately within its budget of epochs.

YOLOv8-M

YOLOv8-M also demonstrated stable convergence of training and validation losses across the 10 epochs on the RTX 4060 Ti (Figure 7-model panel for YOLOv8-M). Loss reductions were smooth, and validation metrics trended upward with epoch number, indicating effective optimization under the selected settings.



(d) Yolov8-M training result

Figure 7. Training and validation curves on the hybrid GRDDC–PUPR dataset for YOLOv8-M

YOLOv8-M achieved the highest validation precision among all models (0.536), with a recall of 0.438, mAP@0.5 of 0.435, and mAP@0.5:0.95 of 0.191 (Table 1). The corresponding training and validation curves in Figure 7 show smooth loss reduction and monotonic improvements in the metric. The combination of high precision with modest mAP gains suggests YOLOv8-M produces fewer false positives while offering incremental improvements in overall detection accuracy.

Interpretation / suitability: YOLOv8-M is most suitable for applications that prioritize precision and low false positive rates (for example, automated decision systems where false alarms are costly). Its behavior suggests conservative but reliable detection performance within the given epoch budget.

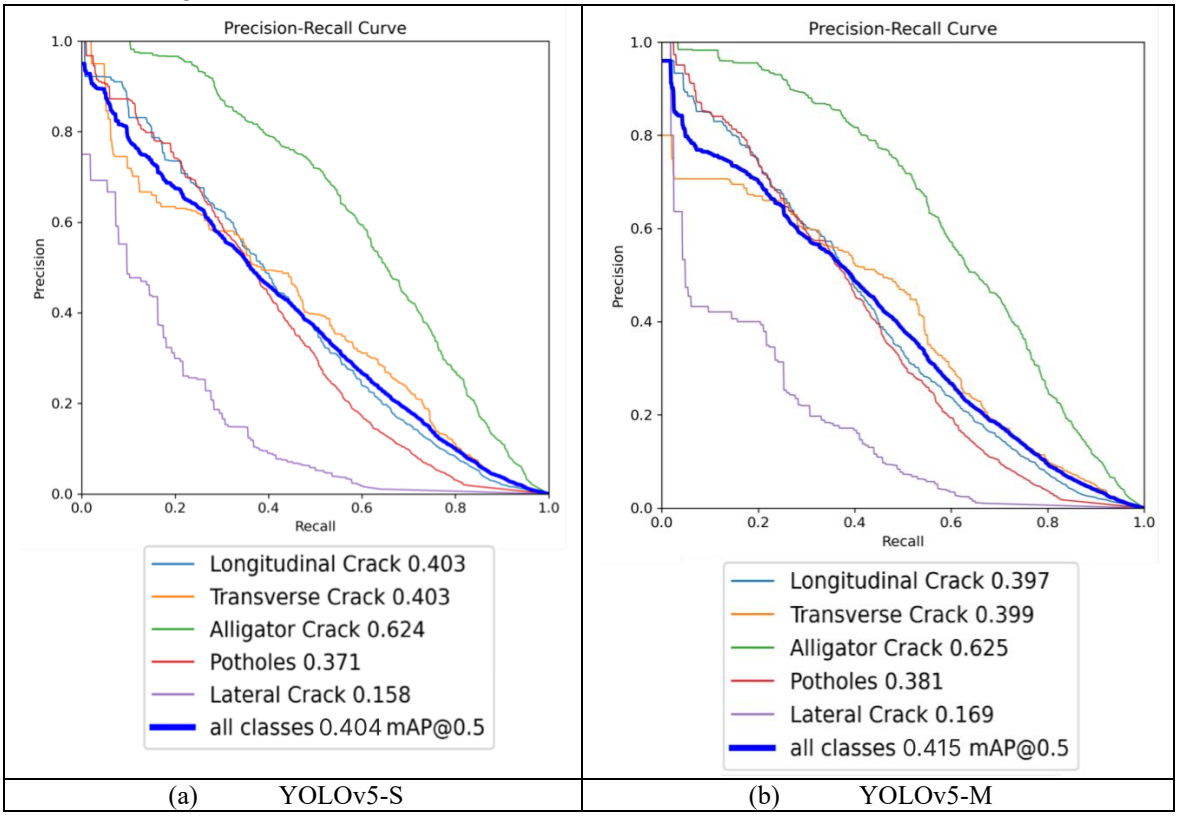
Table 1. Comparison of Training and Validation Metrics for YOLOv5 and YOLOv8 Models

Metric	YOLOv5-S	YOLOv5-M	YOLOv8-S	YOLOv8-M
Recall	0.426	0.421	0.437	**0.438**
Precision	0.466	0.488	0.494	**0.536**
mAP@50	0.404	0.415	0.430	**0.435**
mAP@50-95	0.176	0.182	0.186	**0.191**

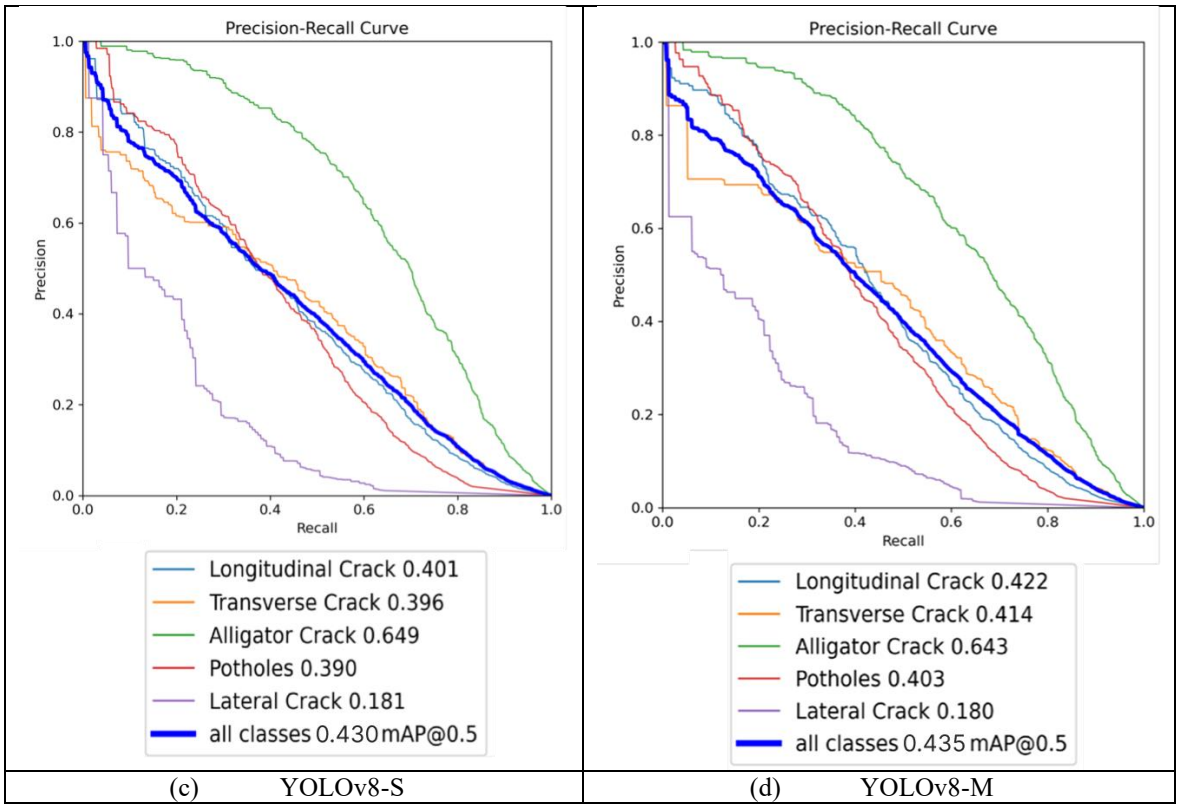
Precision-Recall

For better assessment of classification performance with varying levels of imbalance between the classes, we proceeded to calculate the Precision-Recall (PR) curves for all models. Although metrics such as accuracy and/or ROC curves are commonly used for comparing model performance, they are less insightful compared to PR curves, which are especially useful with less frequent classes [25], [26].

Figure 8 below illustrates the PR curves of all four models: YOLOv5-S, YOLOv5-M, YOLOv8-S, and YOLOv8-M, while Table 2 below provides their mean average precision at IoU ≥ 0.5 (mAP@0.5) for all five damage classes.



(a) YOLOv5-S (b) YOLOv5-M



(c) YOLOv8-S (d) YOLOv8-M

Figure 8. Precision-recall curves for: (a) YOLOv5-S, (b) YOLOv5-M, (c) YOLOv8-S, (d) YOLOv8-M

Table 2. Per-class Average Precision (AP@0.5) on the Test Set for Each YOLO Variant

Class	YOLOv5-S	YOLOv5-M	YOLOv8-S	YOLOv8-M
Alligator Cracks	0.403	0.377	**0.649**	0.643
Transverse Cracks	0.403	0.377	0.396	**0.414**
Longitudinal Cracks	0.403	0.377	0.402	**0.422**
Potholes	0.377	0.373	0.390	**0.403**
Lateral Cracks	0.158	0.169	**0.181**	0.180

Across models, YOLOv8-M achieved the highest overall mAP@0.5 on the test set (0.435), followed by YOLOv8-S (0.430), YOLOv5-M (0.415), and YOLOv5-S (0.404) (Table 4). At the class level (Table 2), YOLOv8 variants performed exceptionally well on Alligator Cracks, with AP@0.5 scores of 0.649 (YOLOv8-S) and 0.643 (YOLOv8-M). In contrast, all models struggled with Lateral Cracks, which yielded the lowest AP@0.5 values (0.158–0.181). In contrast, all models struggled with Lateral Cracks, which yielded the lowest AP@0.5 values (0.158–0.181). Visually, the PR curves for YOLOv8 variants appear smoother across confidence thresholds, suggesting a more stable precision–recall trade-off.

F1-Confidence Curve

Apart from analyzing the precision-recall curves, we also analyzed the F1 vs. confidence curve to find the confidence threshold that yields the best trade-off between precision and recall. Table 3 shows the peak F1 scores attained by each of the YOLO models along with the corresponding confidence threshold values on the test data.

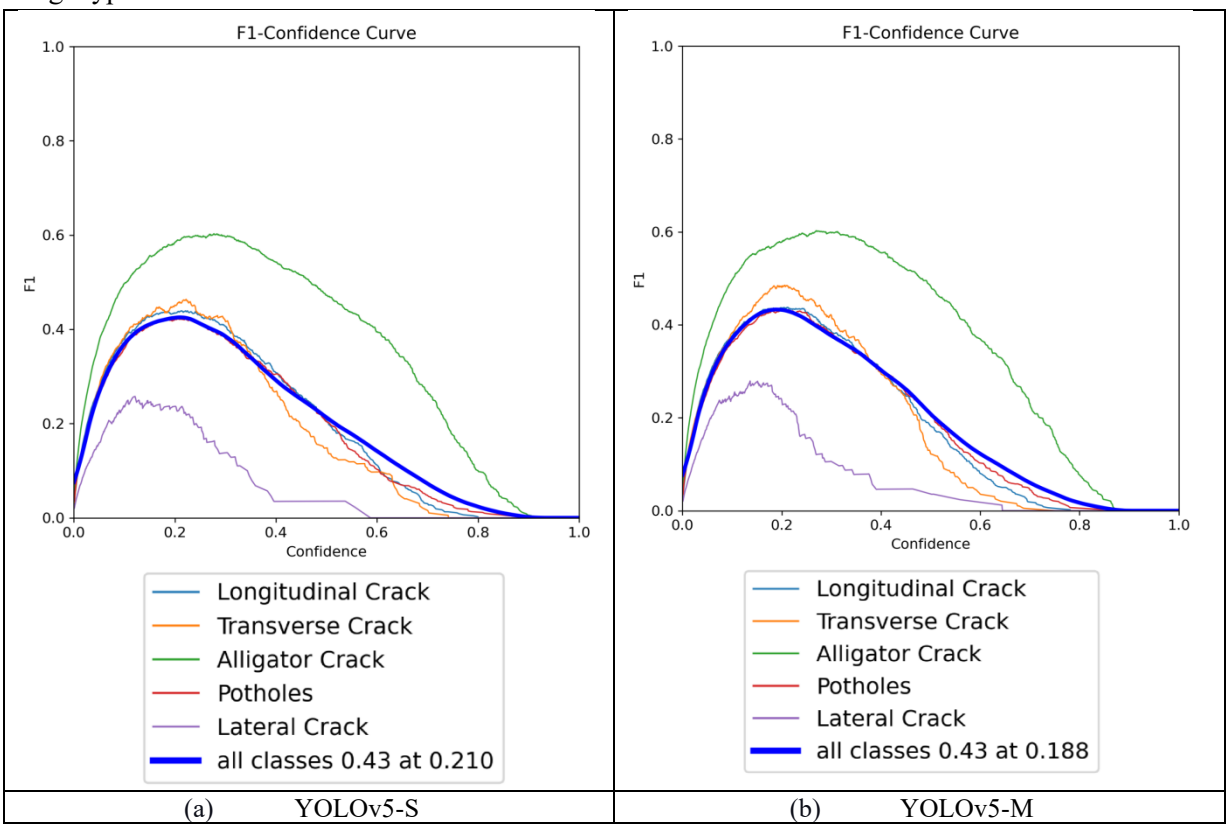
Table 3. Peak F1-score and Corresponding Confidence Threshold on the Test Set

Model	Peak F1-Score	Confidence threshold at peak F1	Best-performing class at peak
YOLOv5-S	0.43	0.210	Longitudinal Crack

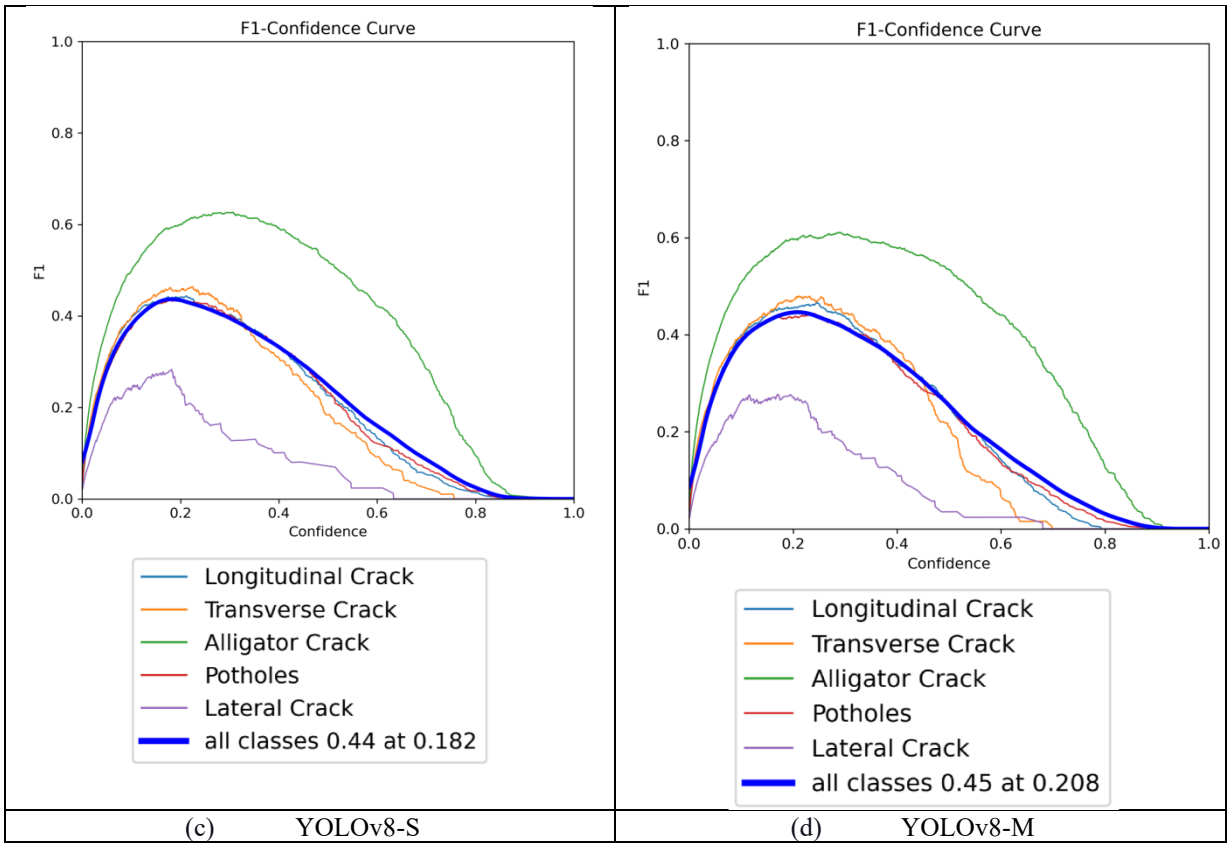
ISSN 2580-6424 (printed) | ISSN 2477-2399 (online)

YOLOv5-M	0.43	0.188	Transverse Crack
YOLOv8-S	0.44	0.182	Alligator Crack
YOLOv8-M	0.45	0.208	Alligator Crack

The table reveals that YOLOv8-M yields the highest peak F1 score of 0.45 with a confidence threshold of 0.208, while YOLOv8-S yields a peak F1 score of 0.44 with a confidence threshold of 0.182. For the YOLOv5 models, the peak F1 scores attained are 0.43, which occur at confidence thresholds of 0.210 for YOLOv5-S and 0.188 for YOLOv5-M, respectively. Figure 9 depicts the F1 vs. confidence curves of all models, which show the F1 scores attained by each of the models for each damage type at different confidence thresholds.



(a) YOLOv5-S (b) YOLOv5-M



(c) YOLOv8-S (d) YOLOv8-M

Figure 9. The F1-Confidence Curve for Each YOLO Variant

Performance Comparison Across YOLO Variants

This analysis evaluates and compares the performance of four different YOLO variants, namely YOLOv5-S, YOLOv5-M, YOLOv8-S, and YOLOv8-M, for road damage detection tasks. The performance analysis is carried out using four main performance measures, including precision, recall, mean Average Precision at IoU 0.5 (mAP@0.5), and mean Average Precision from IoU 0.5 to 0.95 (mAP@0.5:0.95) as shown in Table 4.

Table 4. Performance Comparison of YOLO Variants

Model	Precision	Recall	mAP@0.5	mAP@0.5:0.95
YOLOv5-S	0.466	0.426	0.404	0.176
YOLOv5-M	0.488	0.421	0.415	0.182
YOLOv8-S	0.494	0.437	0.430	0.186
YOLOv8-M	**0.536**	**0.438**	**0.435**	**0.191**

As shown in Table 4, the YOLOv8 variants consistently outperformed the YOLOv5 counterparts across all metrics. Notably, YOLOv8-M achieved the highest detection performance with a precision of 0.536, recall of 0.438, an mAP@0.5 of 0.435, and an mAP@0.5:0.95 of 0.191. This was closely followed by YOLOv8-S, which achieved an mAP@0.5 of 0.430 and mAP@0.5:0.95 of 0.186. In comparison, YOLOv5-M and YOLOv5-S obtained slightly lower results, with mAP@0.5 scores of 0.415 and 0.404, respectively.

Table 5. Per-class Precision on the Test Set

Class	YOLOv5-S	YOLOv5-M	YOLOv8-S	YOLOv8-M
Longitudinal Crack	0.458	0.473	0.447	**0.483**
Transverse Crack	0.435	**0.455**	0.428	0.446

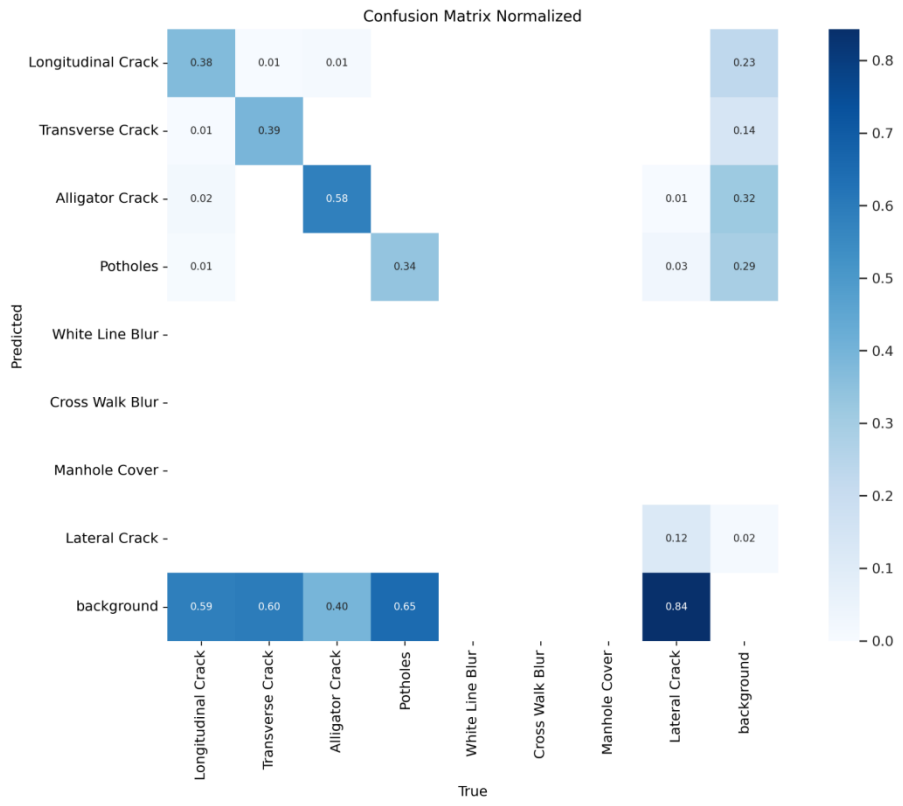
ISSN 2580-6424 (printed) | ISSN 2477-2399 (online)

Class	YOLOv5-S	YOLOv5-M	YOLOv8-S	YOLOv8-M
Alligator Crack	0.554	0.535	0.536	**0.563**
Potholes	0.472	0.463	0.457	**0.488**
Lateral Crack	0.366	0.383	**0.430**	0.410

Table 6. Per-class Recall on the Test Set

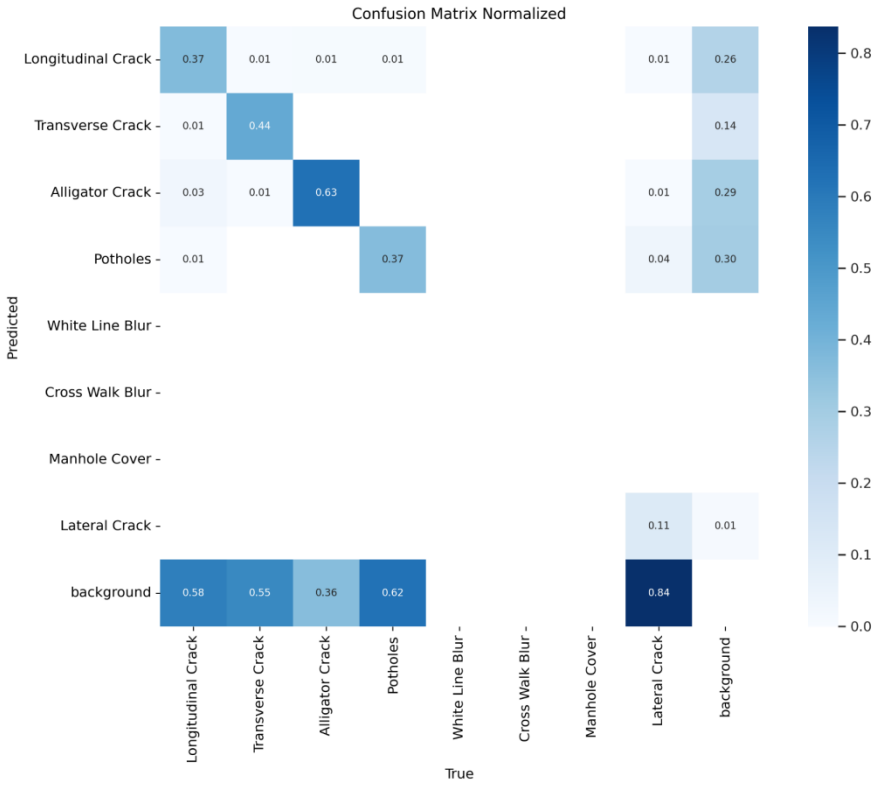
Class	YOLOv5-S	YOLOv5-M	YOLOv8-S	YOLOv8-M
Longitudinal Crack	0.413	0.399	**0.433**	0.432
Transverse Crack	0.469	**0.515**	0.487	0.512
Alligator Crack	0.620	0.625	**0.671**	0.635
Potholes	0.380	0.399	**0.414**	0.396
Lateral Crack	0.175	0.181	0.193	**0.199**

Figure 10 also indicates the results in the form of normalized confusion tables for the discovery of patterns of misclassification for the four versions of YOLO, highlighting the variation in performance for low-contrast damage classes like longitudinal and transverse cracks.

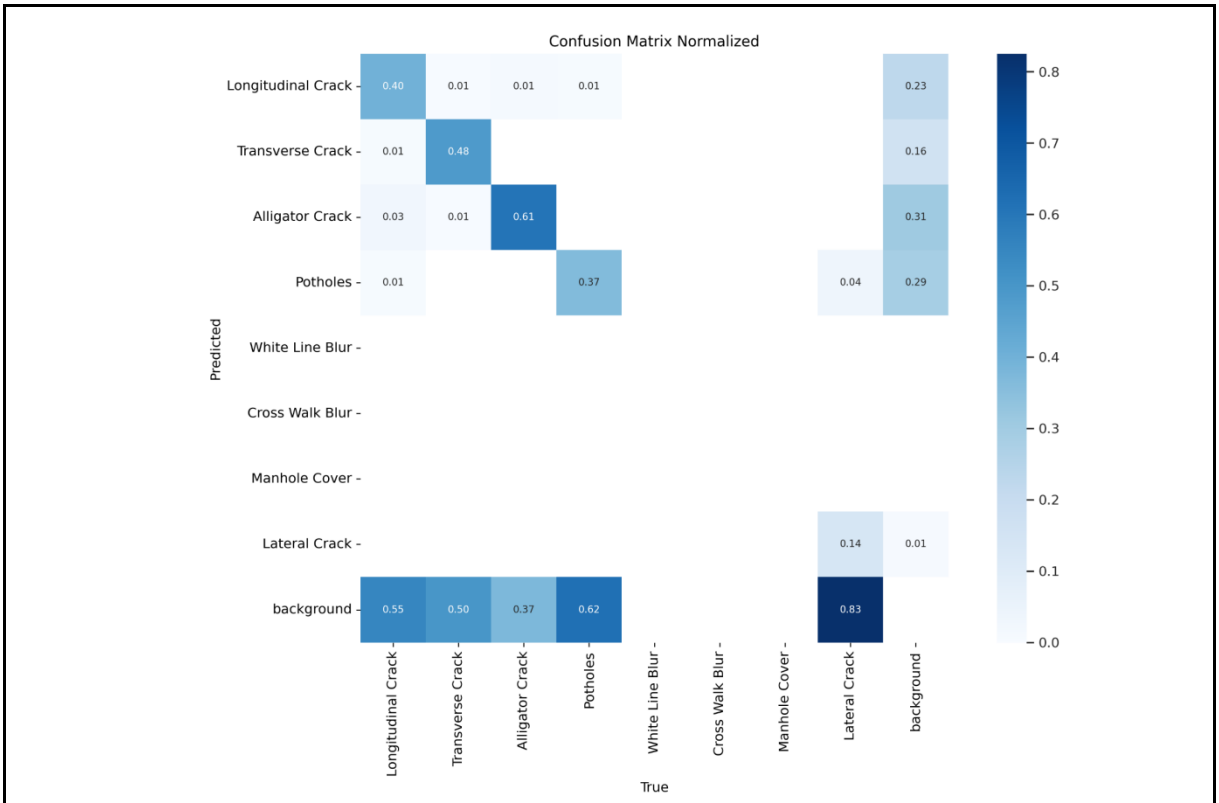


(a) Conf matrix YOLOv5-S

(b) Conf matrix YOLOv5-M



(c) Conf matrix YOLOv8-S



(d) Conf matrix YOLOv8-M

Figure 10 (a) YOLOv5-S, (b) YOLOv5-M, (c) YOLOv8-S, and (d) YOLOv8-M. Particularly for subtle damage types, such as longitudinal and Transverse Cracks, which are frequently misidentified as background, darker colors suggest a higher frequency of misclassification.

To offer qualitative observations of model performance, Figure 11 shows the ground truth annotation alongside the YOLOv5-S and YOLOv8-M predictions.

Figure 11. Visual comparison of the experimental results for road damage detection: (a) Ground truth annotation, (b) YOLOv5-S predictions, and (c) YOLOv8-M predictions. YOLOv8-M shows higher confidence and less false positives, especially in the longitudinal and alligator crack samples.

Discussion

Interpretation of overall trends and deployment-oriented implications

As given in Table 1, YOLOv5-S prioritizes Recall, making it suitable for applications where achieving maximum detection is a priority. YOLOv8-M has been found to offer maximum precision and is suitable for applications where low false positive values are required. YOLOv5-M and YOLOv8-

S offer balanced solutions for trade-offs based on which they should be utilized for achieving maximum speed, accuracy, and precision.

YOLOv8-M has been found to offer maximum precision for almost all types of damage, especially for Alligator Cracks and Potholes. In comparison, YOLOv5-S still holds relevance for applications with consistent detection within specific classes, especially Transverse Cracks. To improve detectability within all classes, there is still a need for optimal optimization, especially for those classes with low performance, including Lateral Cracks.

The YOLOv8 model has shown significant improvement in road damage detection with increased precision and location accuracy and has been found to be competitive in retaining similar levels of Recall. The negligible levels of performance variation among YOLOv8-M and YOLOv8-S imply trade-offs in simplicity and speed.

YOLOv8-M is found to be more accurate than YOLOv5-S but with increased computational complexity. In comparison, YOLOv5-S promotes increased inference speed and is deemed suitable for applications with real-time inputs. These results imply trade-offs among computational complexity, accuracy, and speed for different variants of YOLO models.

F1-confidence interpretation and threshold behavior

This analysis aims to evaluate the F1-Confidence Curve in an attempt to establish the efficacy capabilities of a variety of confidence levels. The F1-score is an appropriate measure of efficacy in cases involving two-class imbalance issues relevant in the application of this research. A trade-off between precision and recall has been recognized in past studies. Precision-recall curves are found to provide more detailed information than an ROC curve in cases involving imbalanced data [25]. The trade-off between the F1-score and threshold control has also been suggested to provide a proper measure of the actual efficacy capability of models [27]. Previous studies have also clarified the effectiveness of applying the F1-score in combination with deep learning approaches on a large scale, reaching peaks of 0.58 and 0.57 [28] in identifying overall pavement damage.

Subfigures 9(d) and 9(a) in Fig. 9 indicate the F1-Confidence curves of the YOLOv8-M and YOLOv5-S models, respectively, and the curves are found to indicate similarities. The effectiveness of each YOLO variant varies across the five damage classes, as illustrated by the F1-confidence curves in Figure 9 and further supported by the confusion matrices in Figure 10. YOLOv8 approaches indicate significant consistency and high F1-scores at every confidence interval, especially with Alligator Cracks, according to the data. These results effectively verify the hypothesis that YOLOv8 models are more relevant and applicable in cases involving extensive applications, considering both high efficacy capability and consistency at confidence levels.

Class-specific challenges and the role of imbalance

Accuracy of detection for all YOLO versions significantly varied with the damage type, as given in Table 5. Alligator Cracks always demonstrated the highest recall and precision due to their distinguishable characteristics and their adequate number of samples in the database. Lateral Cracks, however, demonstrated the lowest detection performance among all models due to the challenges caused by their low-contrast appearance and class dominance. The contrast highlights the necessity for the employment of data augmentation techniques and special training methods for the improvement of the detection rates of minor classes, for instance, the Lateral Cracks.

Error patterns and qualitative interpretation

Figure 10 shows the misclassification tendencies of YOLOv5-S and YOLOv5-M, most notably for fragile damage types like Longitudinal and Transverse Cracks, commonly misclassified as

background. This reflects that the YOLOv5 versions are unable to derive discriminative characteristics for less noticeable damage types, a problem probably exacerbated by class unbalance in the training dataset. Comparatively, YOLOv8-S and YOLOv8-M substantially reduce these problems, the YOLOv8-M version in particular registering the lowest rates of misclassification in separating fine cracks from the background. This highlights the improved architectural abilities of YOLOv8 when identifying optically challenging or low-contrast types of damage.

The results of this analysis are that the architectural innovations of YOLOv8, particularly in the medium (M) form, are responsible for more accurate and repeatable road damage classification in all categories.

As illustrated in Fig. 11(a) ground truth annotations are the golden standard used for model prediction assessment. Fig. 11(b) indicates multiple missed detections in the output of YOLOv5-S, whereas Fig. 11(c) shows the better localization of YOLOv8-M and increased confidence value. These subfigures highlight the comparative performance of the models in varying environmental conditions, wherein YOLOv8-M is always able to produce fewer false positives and yield better prediction confidence.

Comparison with Previous YOLO-based Road Damage Studies

Compared with prior YOLO-based road damage detection studies, our findings are broadly consistent with the performance patterns reported in the literature. Earlier works using GRDDC 2020 or similar benchmark datasets have shown that YOLOv3 and YOLOv5 achieve competitive mAP@0.5 values for road damage categories under controlled, single-domain settings [7], [9], [11]. In our experiments on the hybrid GRDDC–PUPR dataset, the best-performing model (YOLOv8-M) achieved an mAP@0.5 of 0.435 and an mAP@0.5:0.95 of 0.191 on the test split (Table 4). We expect these scores to be lower than the strongest results reported on pure GRDDC benchmarks because our setting combines international and local images, introduces domain shift, and includes heavily imbalanced crack classes typical of Indonesian roads.

Several pavement-distress studies have reported that thin or low-contrast cracks are systematically harder to detect than prominent and visually salient damage, such as alligator cracking or potholes [7], [8], [10]. Our per-class precision and recall (Tables 5 and 6) follow the same trend: alligator cracks and potholes achieve the highest detection performance across all YOLO variants, whereas lateral cracks remain the most challenging class, with noticeably lower precision and recall. This convergence with previous work indicates that the main limitation arises not from the proposed models themselves, but from intrinsic visual characteristics of the damage and the severe class imbalance in the dataset.

Our comparison between YOLOv5 and YOLOv8 aligns with recent object-detection studies, which show that newer YOLO variants generally improve detection accuracy at the cost of higher computational demand [10], [11]. In our hybrid GRDDC–PUPR setting, YOLOv8-M provides the best overall precision and mAP, while YOLOv5-S and YOLOv5-M offer more favorable trade-offs for resource-constrained or real-time deployments. These results confirm, in the Indonesian road-damage context, the architectural trade-offs reported in earlier YOLO research and provide additional evidence based on a mixed international–local dataset.

Limitation and Future Work

We acknowledge several limitations in this study. First, we trained all models for only 10 epochs because of computational constraints. This restricted training schedule may prevent the models from reaching their full performance potential; extending the number of epochs or conducting more extensive hyperparameter tuning could further improve detection accuracy. Second, although the hybrid GRDDC–PUPR dataset reflects a variety of road and environmental conditions, we did not perform external

validation on independent datasets or video streams. As a result, the models' generalization to unseen regions and weather conditions remains uncertain. Third, even with oversampling and class-weighted loss functions, class imbalance continues to affect the detection of underrepresented categories, particularly lateral cracks. In future work, we plan to investigate more advanced imbalance-handling strategies, incorporate additional baseline detectors such as Faster R-CNN or SSD, evaluate newer YOLO variants, and conduct real-time field trials on mobile or edge platforms to assess deployment readiness in operational road-inspection scenarios.

CONCLUSION

This work aims to explore which YOLO variant works best in the automated road damage detection task with the use of the hybrid GRDDC 2020-PUPR dataset. Four models, namely YOLOv5-S, YOLOv5-M, YOLOv8-S, and YOLOv8-M, have been tested on the five active damage classes (Alligator, Transverse, Longitudinal, Pothole, and Lateral Cracks) through an overall training and testing process utilizing an 8:1:1 hold-out split.

From the experimental results, it is clear that these four models can effectively locate major types of damage with acceptable accuracy. On the test sets, the YOLOv8-M model yielded overall superior performance with precision at 0.536, recall at 0.438, mAP@0.5 at 0.435, and mAP@0.5:0.95 at 0.191, which made it the optimal choice regarding accuracy and reliability. On the other hand, YOLOv5-S offered lower accuracy but with comparable recall rate at 0.426 with the lowest computational requirements, which makes it an attractive choice for real-time applications with resource-constrained devices. YOLOv5-M and YOLOv8-S presented balancing perspectives regarding both accuracy and efficiency, which allows multiple options according to deployment constraints.

These results confirm that the main objective of identifying the most accurate model for Indonesian road-damage detection is only partly fulfilled since it leads to an overall comparison among the four models, with YOLOv8-M being the ideal choice under certain conditions, mainly regarding accuracy and reliability. YOLOv5-S is more appropriate under other conditions where execution speed and devices efficiency become major issues. However, these models still lack effectiveness under underrepresented classes with minimal contrast levels, particularly for the "Lateral Cracks" class. Moreover, these models remain sensitive regarding class imbalances and domain variations. Future work should be directed towards longer train durations, novel methods for dealing with class imbalances, additional baselines, as well as evaluations at the real-world scale aiming at narrowing the gap between benchmarking performance and real road examination requirements. These findings indicate that, for Indonesian road networks, robust automated damage screening is feasible using off-the-shelf YOLO models when trained on a hybrid GRDDC–PUPR dataset, although detection of minority crack classes still requires further improvement.

ACKNOWLEDGMENT

This study received no funding from external parties. The authors used OpenAI's ChatGPT to enhance the quality of the written work. After that, the text was carefully checked, revised, and completed by the authors, who are solely responsible for the accuracy and integrity of the paper.

REFERENCES

[1] B. P. S.-S. Indonesia, “Panjang Jalan Menurut Provinsi dan Tingkat Kewenangan Pemerintahan (km),” in *Last modified August 13*. [Online]. Available: https://www.bps.go.id/en/statistics-table/3/U0VOeFZEZFNiVnByUkdGMlNrOTFVVGRHY1ZkVGR6MDkjMw==/panjang-jalan-menurut-provinsi-dan-tingkat-kewenangan-pemerintahan–km—2022.html?year=2023.

[2] P. J. Gertler, M. Gonzalez-Navarro, T. Gračner, and A. D. Rothenberg, “Road maintenance and local economic development: Evidence from Indonesia’s highways,” *J Urban Econ*, vol. 143, p. 103687, Sep. 2024, doi: 10.1016/j.jue.2024.103687.

[3] B. P. S.-S. Indonesia, "Traffic Accident, Killed Person, Seriously Injured, Slight Injured and Expected of Material Losses Value." [Online]. Available: https://www.bps.go.id/en/statistics-table/2/NTEzIzI%3D/traffic-accident–killed-person–seriously-injured–slight-injured-and-expected-of-material-losses-value.html.

[4] D. Willar, A. A. D. P. Dewi, and F. P. Makalew, "Reviewing Quality Control Management of Road Construction Projects," in *Proceedings of the 5th International Conference on Sustainable Civil Engineering Structures and Construction Materials (SCESCM 2020*, vol. 215, S. Belayutham, C. K. I. C. Ibrahim, A. Alisibramulisi, H. Mansor, and M. Billah, Eds., Singapore: Springer, 2022, pp. 1261–1271. doi: 10.1007/978-981-16-7924-7_82.

[5] Y. Sari and M. H. Yudhistira, "Bad light, bad road, or bad luck? The associations of road lighting and road surface quality on road crash severities in Indonesia," *Case Stud Transp Policy*, vol. 9, no. 3, pp. 1407–1417, Sep. 2021, doi: 10.1016/j.cstp.2021.07.014.

[6] P. Jiang, D. Ergu, F. Liu, Y. Cai, and B. Ma, "A Review of Yolo Algorithm Developments," *Procedia Comput Sci*, vol. 199, pp. 1066–1073, 2022, doi: 10.1016/j.procs.2022.01.135.

[7] D. Arya *et al.*, "Deep learning-based road damage detection and classification for multiple countries," *Autom Constr*, vol. 132, p. 103935, Dec. 2021, doi: 10.1016/j.autcon.2021.103935.

[8] Norsuzila Ya'acob, Mohamad Danial Ikmal Zuraimi, Amirul Asraf Abdul Rahman, Azita Laily Yusof, and Darmawaty Mohd Ali, "Real-Time Pavement Crack Detection Based on Artificial Intelligence," *Journal of Advanced Research in Applied Sciences and Engineering Technology*, vol. 38, no. 2, pp. 71–82, Jan. 2024, doi: 10.37934/araset.38.2.7182.

[9] Z. Diao, X. Huang, H. Liu, and Z. Liu, "LE-YOLOv5: A Lightweight and Efficient Road Damage Detection Algorithm Based on Improved YOLOv5," *International Journal of Intelligent Systems*, vol. 2023, no. 1, Jan. 2023, doi: 10.1155/2023/8879622.

[10] M. Khan, M. A. Raza, G. Abbas, S. Othmen, A. Yousef, and T. A. Jumani, "Pothole detection for autonomous vehicles using deep learning: a robust and efficient solution," *Front Built Environ*, vol. 9, p. 1323792, Jan. 2024, doi: 10.3389/fbuil.2023.1323792.

[11] N. Chitraningrum *et al.*, "Comparison Study of Corn Leaf Disease Detection based on Deep Learning YOLO-v5 and YOLO-v8," *Journal of Engineering and Technological Sciences*, vol. 56, no. 1, pp. 61–70, Feb. 2024, doi: 10.5614/j.eng.technol.sci.2024.56.1.5.

[12] D. Arya *et al.*, "Global Road Damage Detection: State-of-the-art Solutions," in *2020 IEEE International Conference on Big Data (Big Data)*, IEEE, Dec. 2020, pp. 5533–5539. doi: 10.1109/BigData50022.2020.9377790.

[13] G. Jocher, "Ultralytics/YOLOv5: V6.0—YOLOv5n 'Nano' Models, Roboflow Integration, TensorFlow Export, OpenCV DNN Support," *Zenodo*.

[14] A. Rácz, D. Bajusz, and K. Héberger, "Effect of Dataset Size and Train/Test Split Ratios in QSAR/QSPR Multiclass Classification," *Molecules*, vol. 26, no. 4, p. 1111, Feb. 2021, doi: 10.3390/molecules26041111.

[15] J. Miao and W. Zhu, "Precision–recall curve (PRC) classification trees," *Evol Intell*, vol. 15, no. 3, pp. 1545–1569, Sep. 2022, doi: 10.1007/s12065-021-00565-2.

[16] A. A. Sami, S. 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, Sep. 2023, doi: 10.3390/a16090452.

[17] Q.-H. Phan, V.-T. Nguyen, C.-H. Lien, T.-P. Duong, M. T.-K. Hou, and N.-B. Le, "Classification of Tomato Fruit Using Yolov5 and Convolutional Neural Network Models," *Plants*, vol. 12, no. 4, p. 790, Feb. 2023, doi: 10.3390/plants12040790.

[18] J. Torres, "YOLOv8 Architecture." [Online]. Available: https://yolov8.org/yolov8-architecture/.

[19] K. Indra Sari, B. Sugiarto Waloejo, and I. Widyawati Agustin, "Comparative Study: Level of Service in Indonesia and India," *International Journal of Science and Research (IJSR)*, vol. 11, no. 7, pp. 1333–1337, Jul. 2022, doi: 10.21275/SR22718055914.

[20] A. Kanoungo, U. Sharma, A. Goyal, S. Kanoungo, and S. Singh, "Assessment of Causes of Pothole Development on Chandigarh Roads," *Journal of The Institution of Engineers (India): Series A*, vol. 102, no. 2, pp. 411–419, Jun. 2021, doi: 10.1007/s40030-021-00520-5.

[21] D. Arya, H. Maeda, S. K. Ghosh, D. Toshniwal, and Y. Sekimoto, "RDD2020: An annotated image dataset for automatic road damage detection using deep learning," *Data Brief*, vol. 36, p. 107133, Jun. 2021, doi: 10.1016/j.dib.2021.107133.

[22] L. P. Ingrassia, P. Spinelli, G. Paoloni, and F. Canestrari, "Top-down cracking in Italian motorway pavements: A case study," *Case Studies in Construction Materials*, vol. 13, p. e00442, Dec. 2020, doi: 10.1016/j.cscm.2020.e00442.

[23] Y. Wang, A. W. Z. Chew, and L. Zhang, "Building damage detection from satellite images after natural disasters on extremely imbalanced datasets," *Autom Constr*, vol. 140, p. 104328, Aug. 2022, doi: 10.1016/j.autcon.2022.104328.

[24] M. R. Rezaei-Dastjerdehei, A. Mijani, and E. Fatemizadeh, "Addressing Imbalance in Multi-Label Classification Using Weighted Cross Entropy Loss Function," in *2020 27th National and 5th International Iranian Conference*

on Biomedical Engineering (ICBME), IEEE, Nov. 2020, pp. 333–338. doi: 10.1109/ICBME51989.2020.9319440.

[25] T. Saito and M. Rehmsmeier, "The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets," *PLoS One*, vol. 10, no. 3, p. e0118432, Mar. 2015, doi: 10.1371/journal.pone.0118432.

[26] J. Davis and M. Goadrich, "The relationship between Precision-Recall and ROC curves," in *Proceedings of the 23rd international conference on Machine learning - ICML '06*, New York, New York, USA: ACM Press, 2006, pp. 233–240. doi: 10.1145/1143844.1143874.

[27] P. Branco, L. Torgo, and R. Ribeiro, "A Survey of Predictive Modelling under Imbalanced Distributions," May 2015, [Online]. Available: http://arxiv.org/abs/1505.01658

[28] V. Mandal, A. R. Mussah, and Y. Adu-Gyamfi, "Deep Learning Frameworks for Pavement Distress Classification: A Comparative Analysis," in *2020 IEEE International Conference on Big Data (Big Data)*, IEEE, Dec. 2020, pp. 5577–5583. doi: 10.1109/BigData50022.2020.9378047.

[29] "YOLOv5 Architecture." [Online]. Available:https://github.com/ultralytics/yolov5/issues/280

[30] R. Padilla, W. L. Passos, T. L. B. Dias, S. L. Netto, and E. A. B. da Silva, "A comparative analysis of object detection metrics with a companion open-source toolkit," *Electronics*, vol. 10, no. 3, Art. no. 279, 2021, doi: 10.3390/electronics10030279.