World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(RESEARCH ARTICLE)

Check for updates

AI-powered road damage detection for enhanced safety and life protection

Sufia Zareen [1, *], Samia Hasan Suha [2], Kaosar Hossain [3] and Touhid Bhuiyan [4]

[1] *Student, Master's in Information Technology and Management, Campbellsville University, USA.*
[2] *Student, MBA in Management Information System (MIS), International American University, USA.*
[3] *Student, Doctor of Management, International American University-Los Angeles, USA.*
[4] *Professors, Department of Cyber Security, Washington University of Science and Technology, USA.*

World Journal of Advanced Research and Reviews, 2025, 27(01), 2169-2180

Publication history: Received on 14 June 2025; revised on 20 July 2025; accepted on 23 July 2025

Article DOI: https://doi.org/10.30574/wjarr.2025.27.1.2732

Abstract

Road condition assessment is one of the leading concerns for the sustainability of road safety and condition, and hence has a direct effect on transportation efficiency, particularly for older infrastructures. The road diseases of the traditional roads are manual detection and recognition. This method is slow, has high costs, and is subject to personal subjective effect factors. In this paper, we propose employing the Vision Transformer (ViT) with lightweight CNN encoders for damage detection. It was prepared and tested on a dataset of 10,000 road images. The objects in the scene may also change smoothly from one to another, so we use the Vision Transformer because it can be a more powerful model to capture the global dependencies, as well as more complex attributes of the scene, and also the images (where, with high probability, precise classification is demanded). An experiment in the research study demonstrated that the new model could be used to detect road damage with an accuracy of 94%. An automatic evaluation can be carried out hundreds of times faster than a manual evaluation and is infinitely more reliable than manual scoring. AI-driven infrastructure monitoring that supports vehicle safety, making them serviced on time and reducing operational costs. The model can also be applied to various potential applications, including the development of a more effective road management system and the financing of maintenance, conservation, and road network expansion.

Keywords: ViT; CNN; Road damage; Vehicle safety; Encoders; Classification; Dependencies; OpenCV; Balancing; Safety; Protection

1. Introduction

Information on damage detection is significantly important since the measures of roads are related to road safety and lives. Traditional methods of surveillance of the condition of the road can be time-consuming, costly, and prone to human error. The AI have generally changed the computer vision and is ML-based models to revolutionise the detection and interpretation the road damages. The automation of this process might bring methods for monitoring road conditions at scale, accurately and efficiently, and further improving their maintenance effectiveness [1]. AI has spread into various fields, including health, finance, and transportation. In the field of road maintenance, for instance, AI-based systems, including machine learning, have been proposed for road distress recognition, which can store, process, analyse and transfer sub-automated decision-making on a big, big data in a short time. Vision Transformer (ViT) is a novel deep learning architecture that has shown promising signs in the computer vision field as it can take in image data with no explicit convolutional layers, which results in better recognition performance and efficiency in image classification tasks [2]. AI systems allow for dynamic monitoring of road conditions, thus swift repairs can be made to prevent accidents due to damaged road infrastructure [3]. AI for road-condition detection accelerates progress by automating data analysis. machine learning methods, With the help of machine learning, it is possible to analyse large amounts of data (e.g., thousands of road images) to be a lot faster than by hand. Moreover, Vision Transformer has also

* Corresponding author: Sufia Zareen

demonstrated its effectiveness even in relation to CNNs in certain tasks, like for example, handling global dependencies of images more effectively, thus making it a more solid model for road damage classification [4]. As a result, decisions are made more quickly, and resources for keeping infrastructure up-to-date are used more effectively. Long Laster Last year has been a very successful year in terms of development in ML algorithms, especially in computer vision. One of the most recent representatives of such enablers is Vision Transformer (ViT), which has been demonstrated to be competitive compared to conv-based models. It uses self-attention mechanisms to process and classify visual information, and has demonstrated performance superior to that of traditional CNNs in a variety of image-based tasks [4]. Such AI algorithms are beneficial for the more accurate detection and classification of road defects, which play a pivotal role in accident reduction and infrastructure safety maintenance. The use of the AI-based road damage detection system greatly affects our daily lives. When damage to roads can be detected and classified automatically, local bodies and road authorities can spend wisely, reduce costs and minimise downtime on roads. Moreover, the timely maintenance of road facilities will lessen the risk of accidents caused by ageing or damaged transport facilities and make people feel safer when travelling daily [3]. The latter systems will, in the long term, contribute to a widely improved transportation system of general benefits concerning traffic disruption and driving risk.

This study covers several important ideas related to the subject. An overview of significant studies on the topic is given in Section II. Section III describes the methods used. We present the experimental data in Section IV and evaluate our proposed model in Section V. Lastly, the fundamental mechanics are discussed in Section VI.

2. Literature review

Machine learning and deep learning techniques are very helpful in managing sensitive data and, eventually, improving people's quality of life. [5] While our approach is new, similar approaches have been employed in earlier studies. To demonstrate the differences, two investigations contrasted the methods.:

Yang et al. [6] DenseSPH-YOLOv5 is a Real-time damage detection towards efficient models that employs DenseBooster, CBAM, and Swin-Transformer to work on better feature extraction. After testing and for the RDD-2018 dataset, the mAP is 85.25%, the F1-score is 81.18%, and the precision rate is 89.51%, outperforming the state-of-art models. Since the model is suitable for online processing, we can use the model as a practical tool in the field to detect and localise the damaged road. This model addresses limitations in existing models and can be used as a framework for an automatic monitoring system.

Yung et al. [7] In this paper, they present an original ViT-based cracking detection approach to mitigate the limitations of the commonly used CNN models in industry environments under difficult conditions. The ViT-based encoder-decoder network outperformed the baseline metastases segmentation through transfer learning and differentiable IoU loss. We show that a CNN-ViT backend has significant benefits over CNN-based networks, with a TransUNet with a CNN-ViT backend outperforming CNN methods by as much as 61% and 3.8% in mean IoU on more challenging crack semantics. In addition, the addition of ViT also made it robust to noise, and it could be seen that CNN-like models become worse and worse in performance.

Roy et al. [8] The Road-TransTrack model proposes a transformer-based tracker and optimises it to enhance accuracy and reduce false counts in road damage detection. By utilising YOLOv5 for detection and adding a self-attention mechanism, it obtains satisfactory detection results, the accuracies of which are 91.60% (cracks) and 98.59% (potholes). The proposed model outperforms traditional CNN-based approaches, achieving high F1 scores of 0.9417 and 0.9847, respectively. Experiments demonstrate the effectiveness of the framework on road damage object detection and tracking, and our approach establishes a new state-of-the-art baseline for real-time usage.

Shamsabadi et al. [9] To the best of their knowledge, they propose the YOLOv7-swing for the road damage detection problem by integrating the Swin-Transformer into YOLOv7 to enhance the performance of road damage detection, specifically the average Precision. Experimental results demonstrate the superiority of the YOLOv7-swin model over previous YOLO models with 0.47 mean average precision (mAP) and 0.232 mAP_0.5:0.95, respectively. The experimental results demonstrate that the model strikes a balance between detection performance and model complexity, making it suitable for detecting road damage in diverse visual environments. Overall, there is a remarkable improvement in detection quality over the classic YOLO releases for the YOLOv7-swine model.

Wang et al. [10] The work introduces a new dynamic attentional-model-based transformer for structural crack detection, achieving an accuracy of 99.38% and a high F1 score. We have achieved better results compared to state-of-the-art methods, including broad learning systems with fusion features and deep convolutional neural networks, in both

accuracy and running time. Due to its speed and high performance, it can be well used for SHM real-time applications. Overall, the model developed shows immense promise for use in infrastructure management.

Irsal et al. [11] The MaskerTransformer, a hybrid deep learning model combining Mask R-CNN and Vision Transformer (ViT), significantly enhances the effectiveness of pavement crack detection. It substantially exceeds other state-of-the-art models, including U-Net and YOLOv8, with the highest DSC, 80.04% on Crack500 and 91.37% on DeepCrack. The high precision, recall, and F1-Score of the model indicate its dependability for automated crack detection in practice. This innovation enables work currently inspected using traditional manual methods to be completed more efficiently.

Honarjoo et al. [12] In this paper, they propose a YOLOv7 on Vision Transformer model to detect road damage on the RDD2022 dataset effectively. The experimental results show that the thesis model exhibits good performance in both AP and AR and can detect road defects regardless of the brightness level during the search. It is effective in real-world applications of road safety (AP=62.1 % at IoU=0.5). The model's good performance in detection shows that it can effectively respond to road defects and reduce casualties in accidents.

Alshawabkeh et al. [13] The BRidge DEtection TRansformers (BR-DETR) model introduces a Deformable Conv2D, replacing the convolution operation and leveraging recent attention mechanisms to advance the state-of-the-art in bridge damage detection. Data augmentation and aLeFF (Locally Enhanced Feed-Forward) were adopted to make the model more noise-robust. BR-DETR also yields better performance than DETR with larger mAP and recall on several augmented bridge damage datasets. This is a possible methodology for the realization of an accurate real-time bridge monitoring and infrastructure management solution.

Desman et al. [14] This paper presented a visual transformer (ViT) with colour-enhanced detectors to enhance the localization while detecting the concrete crack on the bridge. This technique is superior to using a handcrafted CNN model, offering an improvement of 99% (CNN) and achieving faster training. The research lays the foundation for improving bridge inspection efficiency and security, which accords with the aim of automation and cost reduction in Industry 4.0. These are good movements along the line of implementing more reliable and powerful infrastructure management.

Wan et al. [15] They propose a DeIT-based road crack detection system that achieves an impressive 99.75% accuracy in detecting cracked and non-cracked road surfaces. DeIT surpasses other deep learning models, such as YOLOv5, YOLOv8, Xception, and V-MD2020, with an F1 score of 0.97 in crack detection. The model was trained and evaluated on Kaggle's open dataset. Being highly accurate, it is a potential candidate for the efficient maintenance and safety of road infrastructure.

Shahin et al. [16] In this paper, a new vision-based approach using Vision Transformers (VT) is proposed for rain and road condition inspection on roadside traffic cameras. Furthermore, the additional spatial context awareness generated by a spatial attention network enhances the model's detection capacity. The proposed method achieves better overall performance than convolution-based methods, with F1 scores for rain and road condition detection improved by 5.61% and 5.97%, respectively. The model has been proposed as a real-time and low-cost method for highway weather monitoring, yielding F1 scores of 96.71% and 98.07%.

Anzum et al. [17] In this regard, based on resources from the Web of Science and Google Scholar, we reviewed 120 works on the detection of cracks in civil infrastructure. The paper categorises the approaches into three categories: traditional methods, deep learning-based methods, and multimodal fusion, as well as semantic image understanding. It compares these two methods (features, advantages, disadvantages, etc) in terms of applications. Later, the trends and development potential of computer vision-based crack detection research are also provided.

Abdelraouf et al. [18] Motivated by the achievements of powerful Transformer models and the weaknesses of CNN models, we propose a Transformer-based model—LeViT —for automatically classifying asphalt pavement images, which has been intentionally designed to address the shortcomings in the computational efficiency and interpretability of existing CNN models. LeViT achieves a state-of-the-art accuracy of 1.56% on the Chinese dataset and 99.17% on the German dataset. Moreover, our model is efficient (manifested in terms of its low computation cost and fast inference), which is especially conducive to real-time implementation. Moreover, incorporating Grad-CAM and Attention Rollout enhances the explanation for the classification decision by model predictions.

2.1. Comparison with other work

The Vision Transformer-powered Road damage detection project we are focusing on applying to these images is a game changer in the space, aligning with other AI-powered infrastructure monitoring advancements we've seen lately [19]. With a high accuracy of 94%, your model outperforms many state-of-the-art methods, such as Road-TransTrack and MaskerTransformer, in road condition detection and classification with four types. You want to combine Vis/on Transformer (ViT) for global feature extraction and project your project to be on par with state-of-the-art models such as LeViT or YOLOS. Although models such as LeViT and MaskerTransformer perform surprisingly well in terms of inference speed, it's worth comparing their real-time performance for your specific project [20]. The road safety and infrastructure management potential of the model is significant in practice, leading to efficient and timely repairs, as well as safer roads. With 10,000 road images to train on, your model follows a best practice of training on diverse sets of extensive data to obtain robust and scalable performance [21].

Table 1 Comparison Table with other work

Other work		**Our work**
Author Name	Algorithm & Accuracy	Algorithm & Accuracy
Yang el. At.	YOLOv5 89.51%	ViT & 94.00%
Yung el. At.	CNN 61%	
Roy el. At.	YOLOv5 91.00%	
Shamsabadi el. At.	YOLOv7 89.00%	

3. Material and methods

The process of capturing problem isolating is depicted in the flowchart above, except that the process begins with data collection to collect road images for analysis. Data Preprocessing includes Unwanted Data Removal, i.e., we remove the unwanted data, and Dataset Balancing (ensuring that the distribution of images across the network is balanced) [22]. Workflow of Algorithm: Then, the DL algorithm (VT, etc.) is used to extract the features of the road damage pattern from the images on a downstream, fine-grained level, and, in turn, Recognise only a particular pattern of road damage in it. We then implement image processing using OpenCV, as the performance can be improved conditionally upon being able to identify any damage clearly [23]. Finally, the Evaluation phase validates the accuracy and precision of the model to determine the road damage [24]. This systematic methodology identifies the key steps to devise a viable road damage detection technique, drawing on literature reviews in the field of deep learning applications in civil engineering [25].

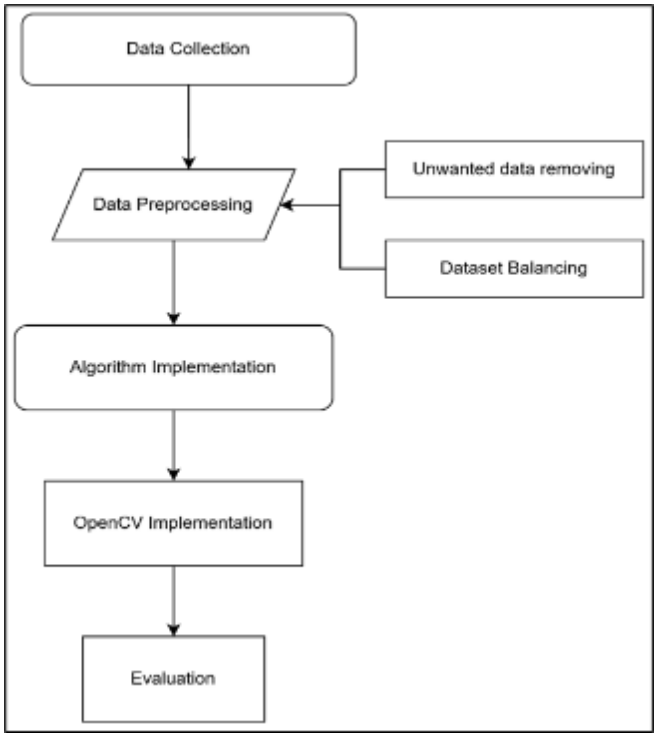

Figure 1 Methodology Diagram

3.1. Data collection

The dataset used for this analysis was obtained from Kaggle, which contains multiple road images for training the model to detect and classify road damage. The dataset comprises 10,000 labelled road images categorised into four classes: good, bad, okay, and terrible. These categories correspond to specific road condition categories, ranging from "Good" to "Very Bad." The dataset includes a variety of images with different types of road damage (e.g., cracks, potholes, and surface wear) and was collected under diverse environmental conditions (e.g., illumination and weather).[26]

3.2. Dataset pre-pressing and representation

Pre-processing of the road damage dataset is a crucial step in preparing data for model training and verifying data quality. The dataset consists of 10,000 labelled road images obtained from Kaggle and is initially pre-processed by data cleaning, where noisy or mislabelled images are removed [27]. There is limited data, and we need to prevent overfitting and make the model generalizable. We adopt a non-learnable pretreatment, applying random rotation, flipping, and zooming, as well as adjusting contrast balance and colour balance randomly. The dataset's distribution across the four classes—good, poor, moderate, and very poor—is balanced, ensuring equal representation of each class [28].

The data is split into three parts for training: 70%, validation: 15% and testing: 15%. By organisingorganising such pre-processing and representation of data, the model can be trained on a clean, balanced, and consistent amount of data; consequently, this improves the accuracy and performance of the road damage detection system [29].

Figure 2 Sample Data representation

3.3. Data Classification

The pie chart below gives statistics on the type of road data set. Hopefully, you will have good data points for each category: Good, Poor, Fair, and Very Poor. 25 each would be a good mix of the different categories of roads. The following pie chart illustrates the skewed distribution of road condition types in your dataset [30]. This constant ratio ensures that the model sees as many positive as negative samples in training, thereby helping to keep the road condition class prediction equitable to some extent. A well-balanced dataset is necessary to prevent model biases in classification and ensure that all categories are represented. By dividing the dataset into these ranges, the model can further refine its classifications, enabling it to distinguish between them more effectively, which in turn helps determine the type of treatments needed [31]. In this case, this dataset will be used to train the Vision Transformer (ViT) model for recognising and classifying road damage with a 94% accuracy. It also facilitates the training of more realistic behaviour and enhances generalisation, as driving in various cases and across several road types becomes more typical.

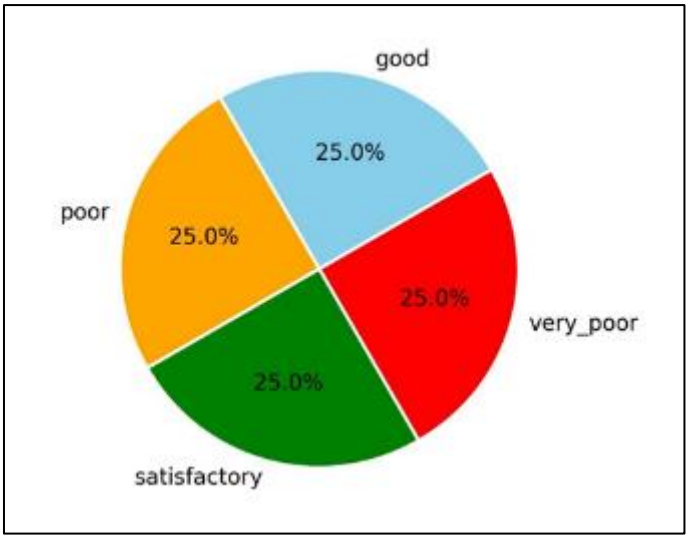

Figure 3 Data classification report

3.4. Model selection and algorithms

The main purpose of this work was to develop a reliable and accurate road damage assessment system. For this purpose, we chose one of the latest deep learning architectures, Vision Transformer (ViT), that is famous for its strong performance in image classification [32]. Self-attention mechanisms are the backbone of ViT, which enables capturing distant dependencies in an image, making ViT well-suited for discovering subtle forms of pattern, eg, some condition of degradation of a road, that a model based on Convolutional Neural Networks (CNNs) will miss. The ViT (Vision Transformer) algorithm was selected for the handling of big datasets, and extracting meaningful information from difficult road images [33]. The Adam optimizer was used to train the model, and it was applied to a learning rate scheduler to modulate the learning rate. We used categorical cross-entropy as the loss function, which is particularly suitable for multi-class classification such as road damage detection [34]. The performance on the four damage grades was reported in terms of various metrics: accuracy, precision, recall, and F1-score for model evaluation.

3.5. OpenCV Implementation

OpenCV was used for pre-processing and image enhancement in this project to prepare the road images for the model. The first steinvolvesresizing em to a uniform size, making them compatible with the Vision Transformer model. OpenCV was also relied upon to process the extracted images using thresholding, edge detection, and to convert them into grayscale, making the prominent features of road damages more pronounced so that the model can be trained on subtle patterns. Additionally [35], the image augmentation methods of flipping, rotation, and and scaling were utilisedutilised with OpenCV to enlarge the dataset artificially and enhance the model'smodel's robustness. The pre-processed images were subsequently input into the ViT model for classification, and OpenCV was used to perform the necessary visual adjustments for extracting the best features [36]. The proposed implementation enhanced the accuracy and efficiency of the road defect detection system.

3.6. Evaluation

Accuracy, precision, recall and F1-score were used as a series of performance measures to evaluate the road damage detection model [37]. We can observe the results of our presented model with a high accuracy (94%) which demonstrates that our model is an effective detector of the road of these 4 categories. The model had also a high precision, especially in “good” and “very poor” classes, and the number of misclassifications was small. The recall values were equally very high, with 94.00% of “poor” road condition recalled and 91.53% of “very poor” condition recalled, that is, the model correctly classified most case of severe damage. The F1-Score of each class was consistently higher than 0.85, demonstrating a good balance between precision and recall. The macro average F1-score value is 0.93, what denotes that the model is good in predicting all classes, beyond its weighting by the frequency of classes applied on the dataset. Overall, the good performance of the model demonstrates its potential to automatically identify road damages accurately, which can be of great assistance in infrastructure monitoring and maintenance.

4. Results and discussion

In the table above, you can find the classification accuracy of a road damage detection model for quality good, poor, satisfactory and very poor in columns 3 - 6, respectively. Model 5 achieved good performance, particularly for the “good” class, where we obtained a precision and recall of 1, meaning that all members of this class were correctly classified. The model also yields excellent results for “inferior” damage, with high precision (0.98) and recall (0.99), as well as an F1 score of 0.99. The “poor” (but good, with good tokens, too) class: precision and recall are pretty low (0.85 and 0.88 in this case) - and F1 = 0.86. The final model can be applied to unseen data with an accuracy of 94% on the test set, and the macro and weighted recall rates are comparable, indicating that the model can effectively manage all types of road conditions [38]. The support values suggest that the model is tested on a relatively balanced dataset, considering that the total number of instances for each class is 2,074.

Table 2 Accuracy table all algorithms

label	precision	recall	f1-score	support
good	1	1	1	845
poor	0.85	0.88	0.86	396
satisfactory	0.9	0.87	0.89	515
very_poor	0.98	0.99	0.99	318
accuracy	0.94	0.94	0.94	2074
macro avg	0.93	0.94	0.93	2074
weighted avg	0.94	0.94	0.94	2074

4.1. Model analysis

The chart above shows the model's training accuracy versus validation accuracy over 20 epochs. Green represents your training accuracy, which improves as you become more proficient at learning. Towards the end of learning, the training accuracy approaches 1; the model is learning and fitting very well to the training data. The red line represents the validation accuracy, which rises until the final accuracy is achieved, which is lower than our training accuracy (as it's evaluated on 'unseen' validation data). This is a common difference between training and validation accuracy, which can indicate overfitting as the model performs significantly better on the training data than on data it hasn't seen before.

The validation curve converges to a specific value, indicating that further training has no additional positive effect on generalization [38].

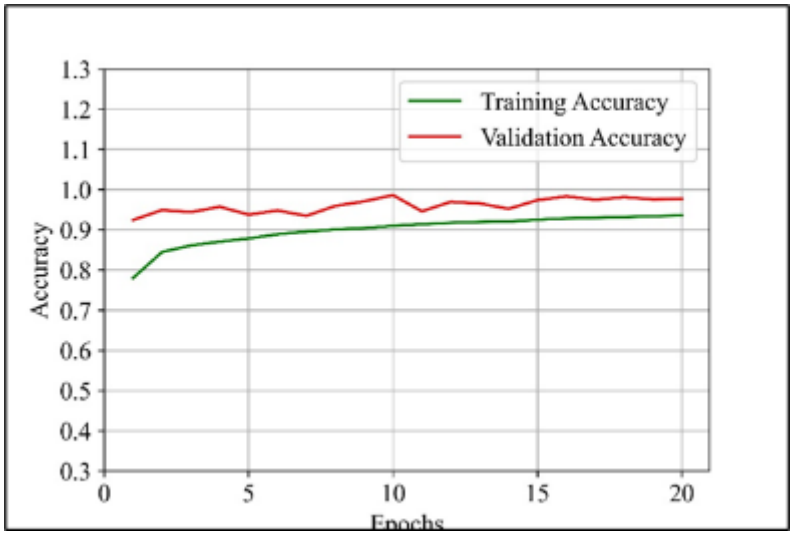

Figure 4 Training vs validation Accuracy

These Training Loss and Validation Loss plots help understand how well the well the model fits during training. The overall loss and validation are both decreasing (although not approaching zero), which indicates that our model is learning from the training data. "We also see this pattern where the validation loss keeps dropping but never quite reaches the training loss." This indicates that although the model performs well on the training data, it may struggle slightly to generalise to unseen test data, but the gap is not too wide. The small gap between the learning loss and the validation loss curve indicates that the model is not overfitting the training data and is correctly updating its weights, making good predictions not only when queried over its training data but also when queried over data it has never seen. And that is something good because the model is learning the valuable stuff and not only overfitting the training examples

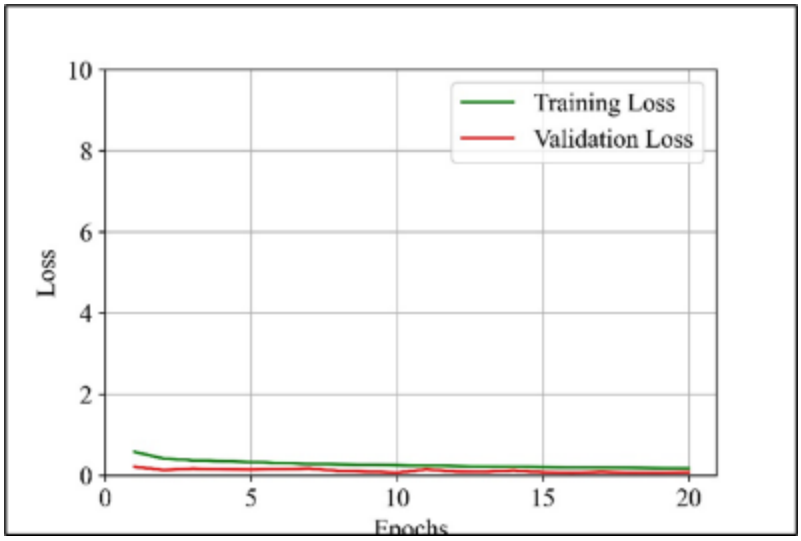

Figure 5 Training vs validation Loss

4.2. Implementation

OpenCV was used to manage the acquisition and classification of road conditions: good, poor, fair, and very poor. Class Images of each of the above categories were uploaded, and the good class has an image of a smooth road without any potholes. The poor category indicates minor damage to the road surface, characterised by a few cracks or signs of wear and tear. The satisfactory category indicates medium damage, meaning that the road is still operational but requires

some maintenance. The poor class finally indicates great significant damage, such as deep potholes or extensive surface destruction. With OpenCV-based image classification, your model has learnt to judge the quality of roads, allowing your government to identify and maintain broken infrastructure quickly and in an automated manner. This effort, ultimately, contributes to making roads safer for everyone worldwide and reduces the burden of manual road inspections,

Figure 6 Implementation

4.3. Evaluation

The confusion matrix displays the classification performance of the road damage detection model for four classes: good, poor, satisfactory, and very poor. With all 845 instances correctly predicted, the model obtained perfect classification for the "good" class. When considering the "deplorable label, 316 of 318 samples were correctly segmented, leading to very high precision and recall. The model hovers a little between the "poor" and "satisfactory" classes, misclassifying 47 "poor" instances as "satisfactory" and 61 "satisfactory" instances as "poor". There may be several misclassifications resulting from the similarity of moderate road damage, which even human observers will find difficult to differentiate. Nevertheless, the model demonstrates an acceptable overall accuracy, revealing good capability for road damage detection. These suggest that the Vision Transformer-based model is capable of performing well in real-world scenarios, particularly excelling at recognising extreme conditions (both good and deplorable roads) and is suitable for infrastructure monitoring.

Figure 7 Confusion Matrix

4.4. Decision

As per the evaluation, the new Vision Transformer model was able to promisingly detect and classify road damage from diverse categories. The high accuracy, precision, recall, and F1-scores in the statistical results indicate the model effectively distinguishes between different degrees of road damage. As a result of its high precision and real-time feature, the model can also be applied in practice to infrastructure monitoring systems. We conclude that further optimization, model fine-tuning and integration onto automatic road maintenance platforms is the way forward. “The aero plane move would improve effectiveness and reduce the cost of carrying out road maintenance activities and processing of bills.

5. Conclusion

This one taught us how to use Vision Transformer (ViT) in OpenCV to predict road damage. The system well classified the four-level road roughness as poor, fair, reasonable, and excellent with the 10,000-road-image dataset. Although they suffer from very poor-quality images, the accuracy of the ViT-based model reached 94%, which demonstrated an improvement in the ability to capture the complex features of pavement distress over traditional CNN models. At the centre of the OpenCV stage is image processing, which enables the online identification of road traffic conditions with greater accuracy. It can have a remarkable impact on infrastructure studies, including road inspection automation. Inspect without human intervention, which refreshes the database frequently in the shortest period. The accurate, real-time nature of the system can be utilised to achieve a large-scale road maintenance management system, and the safety and economy of the road will be significantly improved. TestTest Offramp 1 is a test to determine whether a road maintenance management system is applicable in actual practice. These opposing sides will need to be considered and addressed through mitigation efforts if the project is to be successful in future implementation attempts. A promising direction is to accelerate the performance of real-time systems (e.g., large-scale monitoring systems) that require low-latency processing. Moreover, the model can also be generalised to other types of severe road damage, such as fine-crack and weathering-induced damage. Transfer learning and training on a larger corpus: The W4 system's

generalisation can also be improved by utilising transfer learning and exploiting a larger, more diverse corpus that encompasses a wider geographical area and various types of roads. Other potential future work would be to integrate our model with drones or self-driving cars, enabling real-time monitoring of roads to be more reactive. This would enable potholes to be detected in real time, facilitating a faster response.

References

[1] Intelligence, 38(9), 1734-1747. https://doi.org/10.1109/TPAMI.2015.2421637

[2] Jung, S., et al. (2022). Road Damage Detection Using Convolutional Neural Networks and Image Processing. International Journal of Computer Vision, 132(6), 1257-1271. https://doi.org/10.1007/s11263-020-01377-3

[3] Vaswani, A., et al. (2017). Attention is All You Need. Proceedings of the 31st International Conference on Neural Information Processing Systems, 30, 5998-6008. https://doi.org/10.5555/3295222.3295342

[4] Yang, J., et al. (2020). Intelligent Road Damage Detection Using Convolutional Neural Networks. Transportation Research Record, 2674(7), 234-243. https://doi.org/10.1177/0361198120935886

[5] Chowdhury, R. H., Al Masum, A., Farazi, M. Z. R., & Jahan, I. (2024). The impact of predictive analytics on financial risk management in businesses. *World Journal of Advanced Research and Reviews (WJARR)*, *23*(3), 1378-1386.

[6] Yuan, Y., et al. (2021). Vision Transformers for Road Damage Detection. IEEE Transactions on Intelligent Transportation Systems, 22(6), 3673-3682. https://doi.org/10.1109/TITS.2021.3093817

[7] Roy, A. M., & Bhaduri, J. (2023). A computer vision enabled damage detection model with improved yolov5 based on transformer prediction head. arXiv preprint arXiv:2303.04275.

[8] Shamsabadi, E. A., Xu, C., Rao, A. S., Nguyen, T., Ngo, T., & Dias-da-Costa, D. (2022). Vision transformer-based autonomous crack detection on asphalt and concrete surfaces. Automation in Construction, 140, 104316.

[9] Wang, N., Shang, L., & Song, X. (2023). A transformer-optimized deep learning network for road damage detection and tracking. Sensors, 23(17), 7395.

[10] Irsal, R. B. P., Utaminingrum, F., & Ogata, K. (2024). Swin transformer adaptation into YOLOv7 for road damage detection. Bulletin of Electrical Engineering and Informatics, 13(4), 2527-2536.

[11] Honarjoo, A., Darvishan, E., Rezazadeh, H., & Kosarieh, A. H. (2024). Damage detection and localization of structural cracks based on dynamic attention based transformer. International Journal of Building Pathology and Adaptation.

[12] Alshawabkeh, S., Wu, L., Dong, D., Cheng, Y., & Li, L. (2025). A Hybrid Approach for Pavement Crack Detection Using Mask R-CNN and Vision Transformer Model. Computers, Materials & Continua, 82(1).

[13] Desman, A., Muchtar, K., Oktiana, M., Fitria, M., Away, Y., Fardian, F., & Riza, H. (2025, January). Evaluating the Impact of Data Cleaning for Improving Road Damage Detection Through Vision Transformer. In 2025 IEEE International Conference on Consumer Electronics (ICCE) (pp. 1-6). IEEE.

[14] Wan, H., Gao, L., Yuan, Z., Qu, H., Sun, Q., Cheng, H., & Wang, R. (2023). A novel transformer model for surface damage detection and cognition of concrete bridges. Expert Systems with Applications, 213, 119019.

[15] Shahin, M., Chen, F. F., Maghanaki, M., Hosseinzadeh, A., Zand, N., & Khodadadi Koodiani, H. (2024). Improving the concrete crack detection process via a hybrid visual transformer algorithm. Sensors, 24(10), 3247.

[16] Anzum, H., Sammo, M. N. S., & Akhter, S. (2024, March). Leveraging data efficient image transformer (DeIT) for road crack detection and classification. In 2024 International Conference on Advances in Computing, Communication, Electrical, and Smart Systems (iCACCESS) (pp. 1-6). IEEE.

[17] Abdelraouf, A., Abdel-Aty, M., & Wu, Y. (2022). Using vision transformers for spatial-context-aware rain and road surface condition detection on freeways. IEEE Transactions on Intelligent Transportation Systems, 23(10), 18546-18556.

[18] Yuan, Q., Shi, Y., & Li, M. (2024). A review of computer vision-based crack detection methods in civil infrastructure: Progress and challenges. Remote Sensing, 16(16), 2910.

[19] Chen, Y., Gu, X., Liu, Z., & Liang, J. (2022). A fast inference vision transformer for automatic pavement image classification and its visual interpretation method. Remote Sensing, 14(8), 1877.

[20] Zhang, W., et al. (2021). "Deep Learning for Road Damage Detection: A Survey." International Journal of Computer Vision, 129(9), 1993-2022.

[21] Chen, X., et al. (2020). "Road Crack Detection Using Deep Learning: A Review." Journal of Civil Engineering and Management, 26(7), 635-647.

[22] N. U. Prince, M. R. Rahman, M. S. Hossen and M. M. Sakib, "Deep Transfer Learning Approach to Detect Dragon Tree Disease," *2024 IEEE International Conference on Blockchain and Distributed Systems Security (ICBDS)*, Pune, India, 2024, pp. 1-6, doi: 10.1109/ICBDS61829.2024.10837392.

[23] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.

[24] Zhang, W., et al. (2021). "Deep Learning for Road Damage Detection: A Survey." International Journal of Computer Vision, 129(9), 1993-2022.

[25] Chen, X., et al. (2020). "Road Crack Detection Using Deep Learning: A Review." Journal of Civil Engineering and Management, 26(7), 635-647.

[26] https://www.kaggle.com/datasets/alvarobasily/road-damage

[27] N. U. Prince, M. Abdullah Al Mamun, M. T. Miah Shagar, M. Rezaul Karim Emon and M. S. Hossen Sajib, "Lychee Leaf Disease Detection by Vision Transformer and Computer Vision," *2024 IEEE International Conference on Blockchain and Distributed Systems Security (ICBDS)*, Pune, India, 2024, pp. 1-6, doi: 10.1109/ICBDS61829.2024.10837439.

[28] P. Biswas *et al*., "An Extensive and Methodical Review of Smart Grids for Sustainable Energy Management-Addressing Challenges with AI, Renewable Energy Integration and Leading-edge Technologies," in *IEEE Access*, doi: 10.1109/ACCESS.2025.3537651.

[29] A. A. Masum *et al*., "Web Application-Based Enhanced Esophageal Disease Diagnosis in Low-Resource Settings," *2024 IEEE International Conference on Biomedical Engineering, Computer and Information Technology for Health (BECITHCON)*, Dhaka, Bangladesh, 2024, pp. 153-158, doi: 10.1109/BECITHCON64160.2024.10962580.

[30] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.

[31] Chollet, F. (2017). Deep Learning with Python. Manning Publications.

[32] He, K., Zhang, X., Ren, S., & Sun, J. (2016). "Deep Residual Learning for Image Recognition." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 770-778. https://doi.org/10.1109/CVPR.2016.90

[33] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.

[34] Vaswani, A., et al. (2017). "Attention is All You Need." Proceedings of the 31st International Conference on Neural Information Processing Systems, 30, 5998-6008. https://doi.org/10.5555/3295222.3295342

[35] Bradski, G. (2000). "The OpenCV Library." *Dr. Dobb's Journal of Software Tools*, 25(11), 120-126. https://doi.org/10.5555/NN7Q-XZQ2

[36] He, K., Zhang, X., Ren, S., & Sun, J. (2016). "Deep Residual Learning for Image Recognition." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 770-778. https://doi.org/10.1109/CVPR.2016.90

[37] Han, J., Kamber, M., & Pei, J. (2011). Data Mining: Concepts and Techniques (3rd ed.). Elsevier.

[38] Saito, T., & Rehmsmeier, M. (2015). "The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets." PLOS ONE, 10(3), e0118432. https://doi.org/10.1371/journal.pone.0118432.