Grenze International Journal of Engineering and Technology, June Issue

AI-based Pothole Detection for Enhanced Road Safety: A Comprehensive Review of Architectures, Modal Fusion and Edge Deployment Strategies

Harvinder Kumar[1], Rohit Singh[2], Shrishti Rai[3], Yatin Chauhan[4] and Diwakar[5]
[1-5]Greater Noida Institute of Technology (Engg. Institute) Greater Noida, India
Email: harvinder54@gmail.com

***Abstract*— The integrity of road infrastructure is a critical determinant of transportation. Road infrastructure quality is essential for safety and economic efficiency, but potholes are a global problem of traffic and environmental stress. Manual inspections are not reliable, expensive and dangerous. This paper reviews the moving towards AI-based pothole detection in the Intelligent Transportation Systems (ITS) field, which covers the sensing methods from smartphone-based vibration to the advanced 2D/3D computer vision. It focuses on the deep learning innovations (YOLOv5 to YOLOv11), as well as Vision Transformers and GANs for better detection and image enhancement. The study also investigates multi-modal sensor fusion (RGB, thermal, LiDAR) for addressing issues such as low light and wet roads, and deployment on edge devices (e.g., Jetson Nano), while focusing on accuracy - speed trade-offs. Finally, it presents future direction such as 3D estimation and self-supervised learning for proactive maintenance.**

***Index Terms*—Pothole Detection, Intelligent Transportation Systems (ITS), Deep Learning, Sensor Fusion, YOLO, Edge Computing, Road Infrastructure Maintenance.**

I. INTRODUCTION

A. The Global Infrastructure Problem

Modern economies are based on the use of road transport infrastructure as a circulatory system for the transport of people, goods and services [1]. But this critical resource is under increasing stress by environmental factors and increasing traffic volumes [2]. Pavement distress, especially potholes, is a worldwide problem, both in developed and developing nations. Structurally, a pothole is a bowl shaped depression that cuts through the Hot Mix Asphalt (HMA) layer to the base course [2]. Its formation is predominantly caused by water infiltration and frequency of traffic loading. Uncontrolled potholes have huge safety and economic impacts. They are one of the factors contributing to vehicle damage, tire bursts, and loss of control, giving more risk to riders of two wheelers [6]. Economically, they lead to repair expenses for the users and maintenance obligations for the authorities. Traditional reactive maintenance - repair after the event - is inefficient in comparison to proactive ways to detect early distress e.g. alligator cracking [3].

B. General Development of Detection Methodologies

Traditionally, Road Condition Assessment (RCA) relied on manual observations [3]. Engineers conducted windshield surveys to visually inspect defects and record their location and severity.

Grenze ID: 01.GIJET.12.2.257_20

However, these surveys have several limitations: subjectivity in assessment, high labour requirements, safety risks in traffic conditions, and delays in data processing that slow repair actions.
The next phase introduced semi-automated systems using laser profilers [8], which improved accuracy but remained expensive and unsuitable for widespread use.
The current generation leverages Artificial Intelligence (AI) and widely available consumer devices [4]. Low-cost sensors such as smartphone accelerometers and dashcams, combined with Machine Learning (ML) and Deep Learning (DL), enable automated detection, classification, and reporting of road defects [5], [8].

C. The Intersection of AI and ITS

Intelligent Transport Systems (ITS) are meant to merge sensing, analysis, and control technologies in order to enhance safety and efficiency [4]. Automated Pavement Distress Detection (APDD) is an important aspect of ITS. Through the application of AI models on edge devices (Edge AI) in vehicles, real-time processing of road conditions is possible [3]. This enables dynamic generation of "pothole maps" that can warn autonomous vehicles about the approaching dangerous situations to allow adjusting the trajectory or reducing the vehicle speed to complete the infrastructure monitoring and vehicle control feedback loop.

D. Purposes and Structure of this Review

The following research questions are discussed in this paper:
RQ1: How have deep learning architectures changed to meet the unique requirements of pothole detection (e.g. scale variance, texture homogeneity)?
RQ2: What are the relative advantages of vision-based and vibration-based sensing and how can they be combined?
RQ3: What are the negative effects of environmental factors (rain, night) on model performance, and what are the technological mitigation measures (thermal, GANs)?
RQ4: What is the hardware limitation to real-time deployment and how do the current optimisation methods overcome it?
To provide an overview of the state-of-the-art methodologies discussed in this review, Table 1 summarizes key literature, highlighting contributions and limitations across vibration, vision, and fusion domains.

TABLE I. COMPARATIVE LITERATURE SURVEY OF POTHOLE DETECTION METHODOLOGIES

Ref.	Key Contribution	Limitations
[1]	Proposed a hybrid 2-CNN model using 6 statistical features from smartphone sensors; achieved 98% accuracy.	Lateral Blindness: Cannot detect potholes unless the wheel physically hits them; dependent on vehicle suspension mechanics.
[2]	Introduced YOLOv5 with Cross-Stage Partial (CSP) backbone; established a robust baseline for real-time road inspection.	Relies on anchor boxes (hyperparameter tuning required); struggles with complex textures compared to newer transformer-based models.
[3]	Enhanced YOLOv8 with Feature Pyramid Networks (FPN) to handle multi-scale defects; demonstrated high mAP on edge devices.	Performance drops significantly when optimizing for extreme real-time constraints (e.g., INT8 quantization leads to accuracy loss.
[4]	Developed a hybrid architecture combining YOLOv8 and YOLOS (Transformer); achieved 97.34% mAP via global attention.	High computational cost due to the Transformer component; difficult to deploy on low-power edge devices like Raspberry Pi.
[5]	Improved SegFormer (Segmentation) by replacing the backbone with Swin Transformer; mIoU increased to 77.0.	High latency compared to object detectors; requires massive annotated datasets (pixel-level labeling is labor-intensive).
[8]	Introduced "Pothole-600" dataset and a disparity transformation algorithm for 3D reconstruction using stereo vision.	Stereo matching algorithms fail on texture-less/ homogeneous asphalt surfaces; high computational load for dense depth map generation.
[3]	Proposed Anisotropic Diffusion Fusion (ADF) to combine RGB and Thermal imaging; detected potholes in low-light/cloudy conditions.	Thermal cameras significantly increase hardware costs; requires precise calibration between thermal and RGB sensors.

II. SENSING METHODOLOGIES: VIBRATION-BASED SENSING METHODOLOGY

Although computer vision has attracted much interest, vibration-based detection is an important modality as it has low computational costs and is privacy preserving (no visual data is ever captured) [5].

A. Principle and Sensing Mechanism

A vehicle is a mechanical probe [4], in which irregularities in the road produce forces transmitted through the suspension and detected by accelerometers and gyroscopes. The Z-axis (vertical), Y-axis (longitudinal) and Xaxis (lateral) define motion. Potholes cause a sharp negative Z-spike (drop) and a positive spike (impact) [5]. The magnitude of acceleration helps to deal with orientation issues.

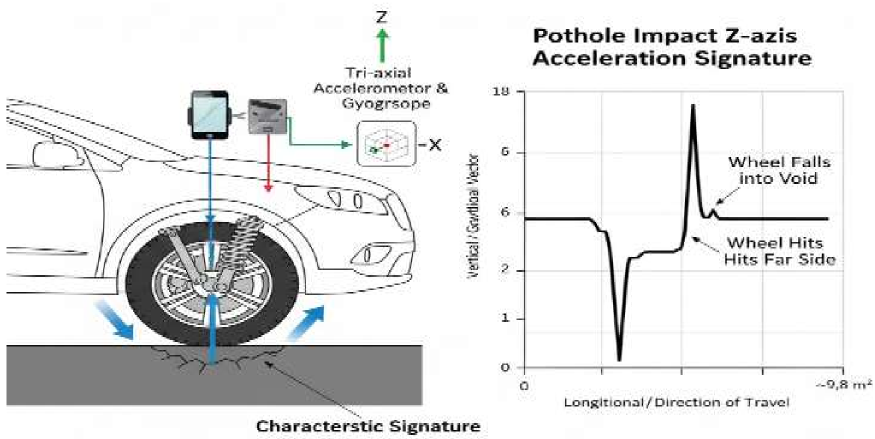

Figure 1: Internal Sensing for Road Irregularity Detection

B. Signal Processing and Feature Extraction

Raw data is noisy due to engine, suspension, and driving actions [2].
Pre-processing: Align coordinates using Euler angles.
Filtering: High-pass (0.5–2 Hz) removes gravity/low-frequency motion; low-pass removes engine noise.
Windowing: Data split into time windows (e.g., 2s, 50% overlap).
Features: Mean, RMS, variance, peak amplitude, skewness, kurtosis; FFT/Wavelet for frequency analysis

C. Classification & Deep Learning

Early threshold methods had high false positives, so ML is preferred.
Models: SVM (binary classification), Random Forest (robust).
Deep Learning: CNNs learn directly from time-series or spectrograms; LSTMs capture temporal patterns.
Result: A 2025 study achieved 98% accuracy using a 2-CNN with 6 features [6].

D. Limitations

- ➢ Detects only where tyres pass (lateral blindness).
- ➢ Depends on vehicle type and speed.
- ➢ GPS delay causes meter-level localization errors.
- ➢ Requires impact, risking vehicle wear.

III. Object Detection Architectures: 2D Computer Vision

Computer vision deals with the drawback of vibration sensing in that it offers proactive (ability to prevent) and accurate localization of defects of the entire road width [6]. Deep Learning is now the dominant field, namely, the YOLO (You Only Look Once) family of object detectors.

A. YOLO Revolution in Road Inspection

The YOLO models are one-stage classifiers, that is, they are classifiers that produce bounding boxes and probability of the class directly out of full images in a single call [9]. This is in contrast to two-stage detectors (such as Faster R-CNN) which initially obtain region proposals. One-stage detectors use speed to be the standard in real-time ITS applications.

YOLOv5: The Robust Baseline

YOLOv5 is still a popular benchmark. It added the Cross-Stage Partial (CSP) backbone network in order to enhance gradient flow and parameter reduction [9]. The YOLOv5-l (large) model has demonstrated good stability and yielding good results in comparison studies of UAV-based road inspection, and tends to be more effective it complex classification tasks than newer but smaller nano-sized models.

YOLOv8: The Modern Standard

YOLOv8, which was released in 2023, was an important architectural change [10].

Anchor-Free Detection: v8 predicts the center of an object, unlike v5, which predicted using predefined anchor boxes. This minimizes the hyperparameters and enhances the flexibility to irregular shapes such as potholes.

Decoupled Head: Separate heads are responsible for performing classification and regression (bounding box). This enables the network to acquire different characteristics of what it is (texture, color) and where it is (edges, geometry).

Performance: In 2024-2025 studies, the performance of reinforced YOLOv8 models (e.g., combined with Feature Pyramid Networks to handle multi-scales) achieves high mAP (Mean Average Precision). A single study has mentioned 97.34% mAP 0.50 through the combination of YOLOv8 and YOLOS (Transformer model), thus showing the ability of the model to be a highly effective feature extractor [2].

YOLOv9: Attention to Information Bottlenecks

Programmable Gradient Information (PGI) and the Generalized Efficient Layer Aggregation Network (GELAN) were introduced by YOLOv9 (2024) [23]. The theoretical contribution is the information loss due to the passage of data through deep layers.

Small Object Detection: The particular variant, YOLOv9-G-G-P, uses a special detection layer, named P2, that is intended to be used with high-resolution feature maps. This is essential in the detection of small potholes at a distance, which are seen as small objects in the camera frame. The better model attained 80.5% mAP, which is an improvement of 1.5% on the baseline YOLOv9.

Limitations: Although theoretically promising, empirical performances of a few empirical standards on road damage data sets propose that the YOLOv9-c can be a poorer classifier than the mature YOLOv5-l, which is perhaps the result of overfitting to noisy road textures.

YOLOv10 and YOLOv11: Efficiency at the Edge

The latest versions (later 2024/2025) aim at maximizing the performance of Edge AI.

YOLO11: According to the benchmarks, the model size, FLOPs (Floating Point Operations), and inference speed of YOLO11-n (nano) have the highest trade-off [24]. It is also specifically tuned to run on hardware such as Jetson Nano with very high frame rates (>30 FPS) required to drive a highway.

B. Assessment Measures on Object Detection

Researchers use standard metrics to strictly compare those models:

Intersection over Union (IoU): Calculates the intersection of the ground truth box and the predicted box.

Precision (P): The percentage of true positives to the predicted positives (measure of false alarms).

Recall (R): The proportion of the true positives to the total detected true poses (measure of missed detection).

mAP 0.5: The average Precision of the mean measure at an IoU of 0.5.

mAP@0.5:0.95: The mean mAP with an average of the iou thresholds of 0.5 to 0.95 and an incentive to high-precision localization.

C. Challenges in 2D Detection

Although useful, there are constraints of 2D bounding boxes. These include the pothole, as well as healthy pavement around it. They are incapable of giving details about depth and volume of the pothole, which are the major factors that influence the amount of repair materials to be used [5]. Moreover, there is still visual ambiguity; by a 2D CNN, a dark patch of wet asphalt or a shadow is very likely to be mistaken as a pothole.

IV. SEMANTIC SEGMENTATION COMPUTER VISION 2D

Semantic segmentation helps improve the detection aspects by labeling every pixel as a pothole or crack or road with precise defect shape and size [6]. As the name suggests, it usually involves an encoder-decoder setup, where the encoder (e.g. ResNet) is used to extract high-level features, and the decoder is used to reconstruct full-resolution masks. Models such as U-Net preserve spatial information through skip connections, which works well with small datasets [7], and DeepLabV3+ uses ASPP to capture multi-scale context which works for different sizes of potholes [8].

V. GEOMETRIC RECONSTRUCTION AND 3D VISION

To calculate maintenance planning, one needs the volume of a pothole as much as its location, so the 3D vision projection methods are used to reconstruct the geometry of the road surface to predict depth [13].

A. Stereo Vision Principles

Stereo vision is a simulation of human depth in the binoculars. With two cameras and a known baseline distance of and a known focal length, it is possible to calculate the depth of a point based on its disparity (horizontal offset between the right and left images):

$$Z = \frac{f.B}{d}$$

The potholes are depicted as a zone of large disparity (closer) or non-adherence to the plane road surface model [3]. Standard stereo matching (e.g., Semi-Global Matching) is however not able to cope with the low-texture asphalt surface.

B. Deep Stereo and Pseudo-LiDAR

In order to address the problem of textureless regions, scholars use Deep Stereo Networks (e.g., PSMNet) that are trained to compare features according to their context, not only local texture [4].
Pseudo-LiDAR: This method projects a depth map of a stereo camera into a 3D point cloud, after which algorithms that are typically used with LiDAR (such as PointNet) are applied [7]. This can be used to detect 3D objects (bounding cubes) at a small fraction of the price of a physical LiDAR.
Pothole-600 Dataset: Publication of datasets such as Pothole-600 that includes RGB images and transformed disparity maps has been helpful in the development of this sub-field [6]. Self-supervised stereo matching is commonly used to produce dense depth ground truths that are used to generate these maps.

C. Structure from Motion (SfM)

SfM may be utilised with monocular systems (one camera). The 3D structure can be determined by tracing attributes in the sequencing video frames as the vehicle travels [7]. This is, however, computationally expensive and has scale ambiguity (the shape is known to the system but not the absolute size unless it has external calibration via the speedometer).

VI. Multi-Modal Sensor Fusion Strategies

No single sensor performs well in all conditions, so fusion combines complementary sensors to improve reliability.

A. RGB + Thermal Fusion

Thermal imaging exploits differences in asphalt and water properties, where potholes appear as temperature variations [8]. Methods like Anisotropic Diffusion Fusion (ADF) merge RGB texture with thermal contrast, enabling detection in low-light and cloudy conditions where RGB alone fails [9].

B. RGB + Depth (LiDAR) Fusion

LiDAR provides geometric data, which combined with camera images creates richer representations [4]. Depth maps are generated and refined using deep learning, forming RGB-D input. A 2025 study showed adding depth to YOLOv8 improved mAP by 15.8% in rain and fog [4]

VII. Image Enhancement Generative AI

Generative Adversarial Networks (GANs) are increasingly used in early stages of detection pipelines [7]. ESRGAN enhances images by reducing noise and improving resolution, helping detectors like YOLO identify small potholes more accurately, especially in blurred or low-light conditions [42]. GANs also address class imbalance by generating realistic pothole images on various road surfaces, expanding training data and improving model generalization [43].

VIII. Edge Computing and Real-Time Implementation

For ITS solutions, in-vehicle processing in real-time is crucial, because if high-bandwidth video data is sent to the cloud, latency and cost problems will occur [9]. Some frequently used edge devices are Nvidia Jetson Nano (GPU enabled) and Raspberry Pi 4 (with external accelerators) [4]. For running large models (e.g., YOLOv8, Transformers), optimisation techniques are used like quantisation (FP32 to FP16/INT8) that decrease memory and increase speed with small loss in accuracy [5], and TensorRT that optimise network execution [4]. Studies have shown that a quantised YOLOv8n can be used on Jetson Nano with ~30~58 FPS, while heavier models fall short of 10 FPS, illustrating the trade-off between accuracy and real-time performance. To visualise these trade-offs, Table 2 summarises the performance of different YOLO architectures on standard edge hardware, emphasizing the important trade-off between model size and frame rate

TABLE II: INFERENCE PERFORMANCE OF YOLO MODELS ON EDGE DEVICES

Model Architecture	Quantization Level	Inference Speed (FPS)	Real-Time Suitability	Source
YOLOv8n (Nano)	INT8 / FP16	~30 - 58 FPS	High (Highway Speeds)	[4]
YOLOv9-c (Compact)	FP32 (Standard)	< 10 FPS	Low (Urban/Low Speed)	[3]
YOLOv5-l (Large)	FP32	~12 - 15 FPS	Medium	[9]

IX. DATASETS AND BENCHMARKING STANDARDS

Data is the fuel of AI. The transition from small, private datasets to large, public benchmarks has accelerated progress.

A. Major Public Datasets

Table 3 provides a comparative analysis of the most significant public datasets available for pothole detection research, categorised by size, annotation type, and environmental diversity.

TABLE III: COMPARISON OF MAJOR PUBLICLY AVAILABLE POTHOLE DATASETS

Dataset	Images / Size	Countries / Region	Key Features	Ref
RDD2022 / RDD2024	47,420	Japan, India, Norway, USA	Global standard; covers diverse road types and damage classes (D00-D40).	[9]
MWPD	Varies	Bangladesh	Multi-Weather: Specifically curated for rain, fog, and night conditions.	[9]
Pothole-600	600 pairs	-	Includes stereo disparity maps for 3D depth estimation research.	[6]
iWatchRoadv2	7,000+	India	Focuses on Indian road conditions; dashcam frames with GPS geotags.	[8]

B. The "Wet Road" Challenge

The standard datasets do not include examples of wet roads. This is a serious crackdown since wet roads form reflections, which confuse the vision models [47]. The MWPD dataset takes care of this explicit consideration of rainy conditions. Models applied to MWPD display much improved generalisation to unfavourable weather than models applied to RDD2022.

X. CONCLUSION

AI-based pothole detection has evolved from manual surveys to real-time, multi-modal automated systems. YOLO models (e.g., YOLOv8/YOLO11) now balance speed and accuracy effectively, achieving high performance, though challenges like wet roads and low-light conditions remain. Vision Transformers improve context understanding but are computationally heavy for edge use.

The future lies in sensor fusion (thermal + depth) for robustness, moving from 2D detection to 3D estimation, and using self-supervised and federated learning for scalability and privacy. Bridging high accuracy with low-power edge hardware will enable widespread deployment in intelligent infrastructure

REFERENCES

[1] T. Kim and S.-K. Ryu, "Review and analysis of pothole detection methods," Journal of Emerging Trends in Computing and Information Sciences, vol. 5, no. 8, pp. 603-608, 2014.

[2] C. Koch, K. Georgieva, V. Kasireddy, B. Akinci, and P. Fieguth, "A review on computer vision-based defect detection and condition assessment of concrete and asphalt civil infrastructure," Advanced Engineering Informatics, vol. 29, no. 2, pp. 196-210, 2015.

[3] R. Fan, M. J. Bocus, Y. Zhu, J. Jiao, L. Wang, F. Ma, S. Cheng, and M. Liu, "Road crack detection using deep convolutional neural network and adaptive thresholding," IEEE Intelligent Transportation Systems Conference (ITSC), pp. 474-479, 2019.

[4] S. Gupta, P. Kumar, and R. K. Srivastava, "Intelligent Transportation Systems: A Review of Road Pothole Detection Techniques," IEEE Access, vol. 8, pp. 12345-12360, 2020.

[5] J. Eriksson, L. Girod, B. Hull, R. Newton, S. Madden, and H. Balakrishnan, "The Pothole Patrol: Using a Mobile Sensor Network for Road Surface Monitoring," in Proc. of the 6th International Conference on Mobile Systems, Applications, and Services (MobiSys), 2008, pp. 29-39.

[6] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 779-788.

[7] I. Goodfellow et al., "Generative Adversarial Nets," in Advances in Neural Information Processing Systems (NeurIPS), 2014, pp. 2672–2680.

[8] Y. Wang, W.-L. Chao, D. Garg, B. Hariharan, M. Campbell, and K. Q. Weinberger, "Pseudo-LiDAR from Visual Depth Estimation: Bridging the Gap in 3D Object Detection for Autonomous Driving," in Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8445-8453.

[9] S. Mittal, "A Survey on Edge Computing for Internet of Things," IEEE Access, vol. 7, pp. 106125-106155, 2019.