DOI: 10.21917/ijdsml.2026.0190

COMPUTER VISION-BASED MULTICLASS ROAD DISTURBANCE CLASSIFICATION USING DEEP LEARNING

A.V. Satyanarayana[1], Anjnay Mahajan[2], Prem Prakash Vuppuluri[3] and C. Patvardhan[4]
[1,2]Department of Mechanical Engineering, Dayalbagh Educational Institute, India
[3,4]Department of Electrical Engineering, Dayalbagh Educational Institute, India

Abstract

Modern preview control-based suspension and advanced driver assistance systems increasingly rely on accurate road disturbance classification to anticipate hazards and optimize vehicle responses. This paper studies the application of deep learning models for classifying road anomalies. To this end, three state-of-the-art pre-trained deep learning architectures: MobileNetV2, ViT, and InceptionV3 are applied for the road anomaly classification task. The effectiveness of the three models is studied on a novel dataset developed as part of this study. The dataset comprises seven distinct categories of road disturbances typically encountered on urban and suburban roads in the Indian subcontinent. ViT achieves testing accuracy of 97.3% and Precision, Recall, and F1 scores of 0.97, demonstrating superior classification capabilities vis-à-vis the other models. MobileNetV2 achieves an accuracy score of 97.33%, but with relatively higher misclassification rates. InceptionV3 also exhibits robust performance, balancing accuracy and generalizability. This study demonstrates the superiority of ViT over InceptionV2 and MobileNetV3 for this problem, thereby highlighting the potential of vision transformer-based learning for road anomaly classification. The novel road disturbance dataset developed in this work would contribute to further related research in intelligent transportation systems. Furthermore, all code and datasets developed as part of this study are made openly available to support further research in this and related domains.

Keywords:

Deep Learning, MobileNetV2, ViT, InceptionV3, Preview Control, Road Anomaly

1. INTRODUCTION

Road disturbance detection and classification are essential for two primary motivations: prioritizing infrastructure maintenance works and rendering comfort and safety in intelligent transportation systems, autonomous vehicles and preview-based controllers. Structural damage to roads has been posing a significant safety hazard to vehicles since the early years of mass road transportation. Prioritising the resources and identification of precise locations have been challenging tasks which were traditionally performed manually.

Gradually, there had been a shift towards vibration-based devices such as accelerometers to detect the road anomalies. The complexity involved in the quick and reliable detection of road disturbances makes this a challenging task. In recent years, there has been an increase in the application of advanced machine learning techniques for the detection of road defects for maintenance purposes. Most current studies in this domain have focused on two categories: potholes and cracks. Initially, efforts were focused on identifying cracks in road infrastructure. Subsequently, more advanced methods have been developed, such as Gabor filter banks with image classifiers [1] and supervised deep convolutional neural networks (CNNs) for road crack detection [2]. These models are also valuable for road maintenance systems for detecting other anomalies affecting road quality and safety, and are essential for vehicle navigation, passenger safety, and comfort[3]. Yebes et al. [4] applied deep learning models, such as InceptionV2, on a dataset built using images of real-world roads and achieved a mean average accuracy of over 75% for pothole detection. Maeda et al. [5]studied object detection using DL on a large-scale dataset and classified road damage into eight categories. Of the eight classes, five were for different types of road cracks, and two were for blurred white lines. However, they were combined into a single class because of the difficulty in distinguishing the four types of road disturbances (bumps, potholes, rutting, and separation).

Some of the high-end automobiles are now being produced with advanced technologies such as active vehicle suspensions that respond to road disturbances by continuously adjusting the suspension characteristics thereby yielding superior ride comfort and handling. Preview controllers enhance active suspensions by utilising look-ahead technologies, including laser sensors and stereo cameras, to reconstruct the road profile and make real-time suspension adjustments. For preview-based controllers, it is necessary to detect and classify road conditions such as garbage / heap of leaves/ debris, concrete tiles, waterlogging, etc. instead of restricting to the road defects alone (that are detected for road maintenance prioritization). These systems need real-time information on additional anomalies such as garbage/heap of leaves, cracks, build of certain roads, such as concrete tiled roads, which render different drive characteristics in road handling and ride comfort. Inoue et al. [6] estimate the road profile for preview suspension control by applying supervised machine learning. They analyze the reduction in root-mean-square errors in estimating the road profile using point cloud data obtained from LIDAR sensors deployed at two locations: in the front and on the vehicle rooftop. However, when road anomalies such as water-filled potholes and obstacles such as speed breakers are encountered, the performance of the preview controllers deteriorate, owing to impaired prediction and a reduced timespan for response. To overcome this limitation, an integrated system is visualized in which the preview controller is augmented by road anomaly detection and classification in real-time. A conceptual framework describing such a system is presented in Fig.1.

The next leap in the automotive industry is towards autonomous and self-driven vehicles, that operate on the connected vehicle technologies, enabling vehicles to share detected road data over the cloud. The scope of road anomalies has thus been expanded to include multiple anomalies to ensure optimal performance of autonomous and active suspension-based preview controllers. Automating the detection of road disturbances plays a vital role in ensuring a safe driving experience and serves as an enabler for intelligent transportation systems [7]. Road disturbance detection methods can be classified

as computer vision-based, vibration-based, and 3D reconstruction -based [8]. Vision-based methods are the most cost-effective, utilizing image classification, object recognition, or semantic segmentation algorithms to identify disturbances, such as potholes, in road images. Vibration-based methods use accelerometers to detect potholes and speed-breakers. 3D construction-based methods leverage cloud data to detect and localize road surface defects; however, reconstruction-based methods are more complex and expensive. Deep learning is an effective tool for various computer vision-based applications, including object detection, scene recognition and image classification. Some recent studies have focused on road profiling [9]and estimation, [10] which are essential road data for navigation and traffic management. The results highlight the scope for further research on the application of DL tools for object detection.

A major challenge in DL models is their requirement for extensive datasets to improve their efficacy. For instance, a significant challenge was reported in the preparation of a well-annotated dataset for training CNNs [11]. Some studies have focused on pothole detection in Indian roads using contemporary DL models such as InceptionV2, and obtained 97% accuracy [12]. However, the development of datasets that include various other road disturbances typically encountered on suburban and rural roads, especially in countries such as India, which has a vast and varied road infrastructure, has not been sufficiently investigated.

Previous research in this area also does not focus on detecting and classifying a wide range of road disturbances. The literature review shows that most models have been built and trained using datasets from developed countries, where road surfaces that are relatively uniform and environmental noise is limited. As a result, these models might not be sufficiently robust when deployed in regions with unstructured roads, different types of pavement materials, varying lighting conditions, and a higher occurrence of unexpected obstacles. Overcoming these limitations would require the development of datasets that reflect real-world variability, and training models capable of handling such complexity.

The main contributions of this work are as follows:

- The novelty of the study's lies in broadening the classification of road disturbances to identify seven types of road anomalies commonly found on urban and suburban roads in developing countries like India.
- A detailed comparison and effectiveness evaluation of three state-of-the-art DL models in the road disturbance classification in the Indian context is presented.
- The data set used in the present study comprises of images taken manually on various roads in Agra, India, in combination with similar images scraped online. To the authors' knowledge, no previous work has focused on developing such a dataset, representative of typical road disturbances encountered in urban and rural roads in the Indian context.
- The results of this study show the effectiveness of pre-trained CNN models and Vision Transformer for the road anomaly classification problem.

The subsequent sections of this paper are structured as follows. Section 2 describes the methodology of this study involving DL models. Section 3 presents the results and discusses the findings of implementing the latest DL architectures for this problem. Section 4 summarizes the conclusions and the scope for future work.

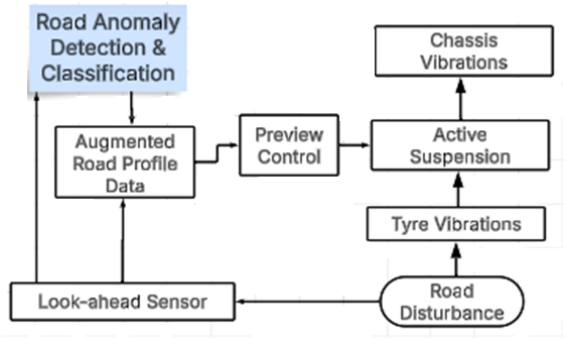

Fig.1. Augmented Preview Control System

2. METHODOLOGY

This section outlines the methodology employed to classify road surface anomalies through a systematic workflow, as illustrated in Fig.2. The methods for this study are systematically divided into three key components: models, dataset preparation, and model training, discussed in three sections – A , B and C. Section A involves shortlisting suitable DL models that are investigated to address the road disturbance classification task. It outlines the unique characteristics of each chosen model. The second section focuses on the data preparation, i.e., acquisition, preprocessing, and augmentation of road disturbance images to evaluate the selected models. Finally, Section C provides details of the training configurations and metrics for assessing their performance. In order to address the risk of overfitting (on account of the limited size of the road disturbance dataset), all the pre-trained models are employed as fixed-feature extractors. For each model, the original ImageNet classification head is removed, all backend parameters are frozen, and a task-specific classifier is appended. This step also includes hyperparameter tuning and performance analysis based on accuracy and computational efficiency. During fine-tuning, the newly added classifier head is the sole trainable component. Input images are resized to conform to the native resolution required by each architecture (224×224 for MobileNetV2 and ViT; 299×299 for InceptionV3).

2.1 MODELS

Three pre-trained models are employed for fixed-feature extraction in this work. In each model, the original ImageNet classification head is removed, all backbone parameters are frozen, and a task-specific classifier is appended. During fine-tuning, the newly added classifier head is the sole trainable component. Input images are resized to conform to the native resolution required by each architecture (224×224 for MobileNetV2 and ViT; 299×299 for InceptionV3).

2.1.1 MobileNetV2:

MobileNetV2 is a lightweight convolutional architecture designed for computational efficiency on mobile and embedded platforms. It improves on MobileNetV1 [2] by introducing inverted residual blocks and linear bottlenecks, which enhance

representational capacity while significantly reducing computational cost.

Depthwise Separable Convolutions: MobileNetV2 replaces standard convolutions with depthwise separable convolutions. These consist of a depthwise convolution, which applies one spatial filter per input channel, followed by a pointwise (1×1) convolution.

For an input tensor

$$X \in \mathrm{R}^{H \times W \times M} \tag{1}$$

with M input channels and N output channels, the cost of a standard convolution with kernel size K is $\mathrm{Cost}_{\mathrm{std}}=K^2MN$.

Depthwise separable convolution significantly reduces the convolution cost further, to:

$$\mathrm{Cost}_{\mathrm{dw\text{-}sep}} = K^2M + MN \tag{2}$$

For instance, when K=3, this results in cost reduction by a factor of over eight. This factorization is the key contributing factor to the computational efficiency of MobileNetV2.

Inverted Residuals and Linear Bottlenecks: Unlike ResNet [3], which forms residual connections between wide layers, MobileNetV2 introduces inverted residuals, where shortcuts connect thin bottleneck layers. Each block expands the channel dimension via a 1×1 convolution, applies a depthwise 3×3 convolution, and then projects the representation back to a low-dimensional linear bottleneck. Residual connections are used when input and output dimensions match (stride = 1). The linear bottleneck avoids nonlinear distortion in compact feature spaces, preserving essential information.

- **Activation and Normalization**: All convolutions except the final projection are followed by Batch Normalization and ReLU6 activation:

$$\mathrm{ReLU6}(x)=\min(\max(0,x),6). \tag{3}$$

This "clipped" activation in ReLU6 stabilizes training at low precision and helps improve model quantization. Transfer Learning: All convolutional layers are frozen, and a task-specific classifier was appended consisting of a GAP layer, a 1024-unit ReLU dense layer, and a final six-class softmax output. Only this classification head was trained for road disturbance classification.

2.1.2 InceptionV3:

InceptionV3 [4] extends the original GoogLeNet/Inception architecture [5] through improved factorized convolutions, multi-branch modules, and enhanced regularization mechanisms. It is designed to extract multi-scale spatial features efficiently while maintaining high representational capacity.

- **Inception Modules and Factorization**: Each Inception module comprises parallel branches employing 1×1, 3×3, and factorized convolutions (e.g., 1×3 followed by 3×1), alongside pooling operations. Given inputs X, the module output is:

$$Y = \mathrm{Concat}(Y_1, Y_2, Y_3, Y_4) \tag{4}$$

where each branch $Y_i=f_i(X)$ captures features at a specific spatial scale.

Factorizing large kernels reduces computational complexity. For instance, replacing a 5×5 kernel with two 3×3 kernels reduces computations significantly, by a factor of $(2\cdot32)/5 \approx 0.72$. This represents savings of around 28%, while preserving receptive field size.

- **Auxiliary Classifiers**: During training, InceptionV3 incorporates auxiliary classifiers connected to intermediate layers to improve gradient flow. Each auxiliary branch contributes an auxiliary loss, L_{aux} given by:

$$L_{\mathrm{aux}} = -\sum_i y_i \log \hat{y}_i^{(\mathrm{aux})} \tag{5}$$

where the $\hat{y}_i^{(\mathrm{aux})}$ are the respective auxiliary head predictors, i being an index into the target classes) and $y_i \in \{0,1\}$. When weighted by $\alpha \in [0.3,0.4]$, the overall training loss is given as:

$$L_{\mathrm{total}} = L_{\mathrm{main}} + \alpha L_{\mathrm{aux}} \tag{6}$$

where L_{main} is the main classifier loss. Further, since these branches are part of the original ImageNet head, they are discarded when include top=False is used for transfer learning.

- **Transfer Learning Setup**: The pretrained InceptionV3 model (299×299 input) was used with all convolutional layers frozen. A GAP layer, flattening layer, 1,024-unit ReLU dense layer, and six-class softmax output are added. Only the appended classifier was trained, while the Inception backbone functioned as a frozen feature extractor.

2.1.3 Vision Transformer (ViT):

The Vision Transformer (ViT) [7] adapts the Transformer encoder architecture [8] to vision tasks by representing images as sequences of patch embeddings and modeling long-range dependencies through self-attention, instead of spatial convolutions.

- **Patch Embedding**: An input image $X \in \mathrm{R}^{H \times W \times C}$ is partitioned into $N=HW/P^2$ non-overlapping patches of size $P \times P$, flattened and projected into a D-dimensional embedding space $z_0^i = x_p^i E$, where the x_p^i's represent the flattened patches. A learnable class token z_0^{cls} is prepended to the sequence, and positional embeddings Epos are added to retain spatial information:

$$Z_0 = [z_0^{\mathrm{cls}}; z_0^1; \ldots; z_0^N] + E_{\mathrm{pos}} \tag{7}$$

- **Transformer Encoder**: Each of the L encoder layers applies Multi-Head Self-Attention (MHSA) and a Feed-Forward Network (FFN), each preceded by Layer Normalization and followed by residual connections:

$$Z_l' = \mathrm{MHSA}(\mathrm{LN}(Z_{l-1})) + Z_{l-1}, \quad Z_l = \mathrm{FFN}(\mathrm{LN}(Z_l')) + Z_l' \tag{8}$$

For each attention head, the input sequence Z is first linearly transformed into three different representations: a query vector Q_h, a key vector K_h, and a value vector V_h by multiplying Z with three separate learned projection matrices W_h^Q, W_h^K and W_h^V. This is expressed as follows:

$$Q_h = ZW_h^Q \tag{9}$$

$$K_h = ZW_h^K \tag{10}$$

$$V_h = ZW_h^V \tag{11}$$

The attention mechanism for head h is computed by taking the dot product between each query and all keys, scaling the result by the factor $\sqrt{d_k}$ (normalized dimensionality of key vectors) to

maintain numerical stability, and applying the softmax function to obtain a set of normalized attention weights.

$$\text{Attention}_h = \text{softmax}\left(\frac{Q_h K_h^{\cdot}}{\sqrt{d_k}}\right) V_h \quad (12)$$

- **Transfer Learning**: A pretrained ViT-Base/16 model (224×224 input) was used with its ImageNet classification head replaced by a newly initialized linear layer producing six class logits. All transformer encoder parameters are frozen, and only the classification head was fine-tuned for the target task.

2.2 DATASET PREPARATION

A dataset was compiled comprising seven classes of road objects relevant to both urban and semi-urban environments in India. Data were collected in two stages: manual and web scraping. Most images were manually captured during field surveys using a 50 MP (f/1.8, wide) camera with the Samsung JN1 primary sensor designed for high-resolution and precise image capture. The camera was complemented by a 2 MP (f/2.4) macro camera for close-up photography and a 2 MP (f/2.4) depth sensor to enhance portrait effects. Photographs of road conditions, such as potholes, solid waste, sand, dirt, and speed breakers, were taken at angles varying from 10° to 75 °and at distances ranging from 1 m to 20 m.

Additionally, the contextual information of the data was enhanced for future use by geotagging each image to document its location. Web scraping was employed to enhance specific classes, such as manhole covers and water-filled potholes, where it was challenging to manually collect sufficient numbers of images, while ensuring the compilation of only open-source images. Additional images are scraped to ensure that the dataset was balanced. A wide range of conditions and perspectives are captured in the images to generate a realistic dataset.

Post the data collection phase, each image was labelled as one of seven predefined classes: potholes, water-filled potholes, speed breakers, sand and dirt, manhole covers, cement tiles, and solid waste. Multiple reviewers participated in this labelling process to ensure high accuracy, employing cross-verification to eliminate errors. Sample images are shown in Fig.3. The final distribution of the images for the respective classes is listed in the Table.1. Subsequently, data preprocessing was conducted to enhance the quality and usability of the dataset. This involved cleaning the dataset by removing duplicate or irrelevant images and discarding low-quality images. All photos were standardized to the JPEG format.

Image augmentation, as described below, was applied to enhance the diversity of the dataset.

- Flipping: The images were horizontally flipped.
- Brightness adjustment: The brightness was adjusted by ±25.
- Saturation Adjustment: Saturation was adjusted by ±25.

Normalization was performed to speed up the convergence of the CNN by scaling the pixel values in the range [0, 1]. The processed dataset was then split into training and validation datasets in a 70:30 ratio. This Road Surface Anomaly dataset is now available for public use at https://www.kaggle.com/datasets /road-surface-anomaly-dataset.

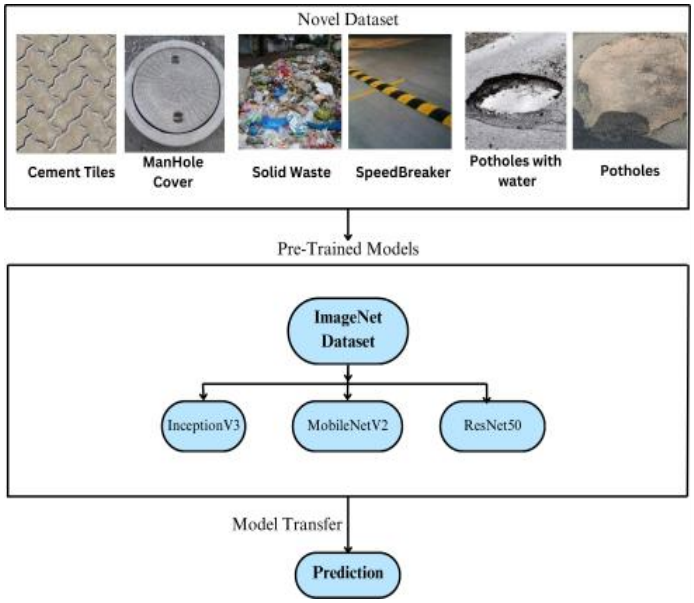

Fig.2. General workflow of the project

2.3 MODEL TRAINING AND TESTING

The ViT, InceptionV3, and MobileNetV2 architectures, implemented in Python using Tensorflow and Keras libraries, were initialized with pre-trained weights from the ImageNet Dataset. Transfer learning was implemented by preserving the initial layers of these models to retain their feature extraction capabilities, The final layers were then replaced and fine-tuned for to perform the classification of road objects. The hyperparameters of the models used are illustrated in Table.2.

The hyperparameter configurations were meticulously selected to optimize the efficacy of the model and the efficiency of the computational process. Keras Tuner was employed to optimize the hyperparameters, and involved examining changes in epochs, learning rate, dense layer units, optimizers, and activation functions.

An early stopping approach was employed to determine the most effective hyperparameter combinations. The training duration for each trial was 15 epochs, and the tuner executed a maximum of 30 trials. For both MobileNetV2 and InceptionV3, the final dense layer was configured with 1024 units. The models were trained using the Adam optimizer with a learning rate of 0.001 and ReLU activation.

Fig.3. A sample of images from the dataset

Table.1. Classwise distribution of images in Road Surface Anomaly Dataset

Class	Total number of images
Cement Tiles	1050
Manhole Cover	1050
Potholes	1050
Sand and Dirt	1050
Solid Waste	1050
Speed Breaker	1050
Water-filled Potholes	1050

Table.2. Hyperparameters of DL models

Model	Input Shape	Number of Layers	Filter Size
MobileNetV2	Varies	50	3 x 3
InceptionV3	(299,299,3)	48	Varies
ViT	(224,224,3)	12 Trans-former layers	16 x 16 patch embeddings

For a comprehensive evaluation of the DL models, standard metrics such as Precision, Recall, and F1 score [20],[14] are also used alongside the accuracy scores. These are described briefly as follows:

- Precision measures the accuracy of the model's positive predictions. A high precision indicates that the model has a low false positive rate. It is mathematically expressed as

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \tag{13}$$

- Recall assesses the model's ability to identify all the relevant positive instances in the dataset. A high recall means the model captures most of the actual positives. It is mathematically expressed as

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \tag{14}$$

- F1 Score, which is a harmonic mean of precision and recall, indicates a robust balance between the two metrics. This is a critical metric when false positives and negatives have significant consequences. It is mathematically expressed as

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{15}$$

3. RESULTS AND DISCUSSION

This section provides a detailed analysis of the performance of the various deep learning models evaluated in this study. All models were implemented in Python 3 using the Keras framework. The experimental work was conducted as described in Section 2, and the performance of the models was assessed using the average training accuracy, validation accuracy, Precision, Recall, and F1 scores. The results are presented in Table.3. After training the models, confusion matrices were plotted to evaluate their prediction and generalization abilities. The confusion matrices for the InceptionV3, MobileNetV2, and ViT models are shown in Figs. 4, 5 and 6, respectively. MobileNetV2 achieved slightly higher testing accuracy, but its confusion matrix reveals significant misclassifications compared to InceptionV3 and MobileNetV2. This is further demonstrated by the Precision, Recall, and F1 scores of ViT, which are 0.97, in contrast to 0.93 for InceptionV3 and 0.94 for MobilenetV2, clearly illustrating ViT's superior performance.

Table.3. Road Anomaly Classification Results

Model	Training Accuracy	Testing Accuracy	Precision	Recall	F1-score
InceptionV3	97.57	93.33	0.94	0.93	0.93
MobileNetV2	99.72	97.33	0.93	0.93	0.93
ViT	94.26	97.3	0.97	0.97	0.97

A class wise examination of the efficacy of the three models exhibits substantial variance, as revealed by the Precision, Recall, and F1 metrics shown in Fig.7-Fig.9, and is discussed as follows. The results show that model performance varied across different road anomaly classes:

- All models accurately identified cement tiles, achieving perfect precision (1.0) and high recall (ranging from 0.92 to 0.96), with F1 scores reaching 0.96, as indicated by the nearly converging model representation lines.
- For manhole covers, InceptionV3 and MobileNetV2 reported lower accuracy and F1 scores (accuracy = 0.96, F1 = 0.96) as compared to ViT, which clearly outperformed the other two models, achieving near-perfect results (precision = 1.0, recall = 0.96, F1 = 0.98). This is clearly illustrated by the blue line representing ViT, which is noticeably closer to the 1.0 node compared to the green line of MobileNetV2 and the red line of InceptionV3.
- ViT's transformer framework enhanced context modeling and feature boundary differentiation, thereby reducing the number of false positives typically observed in CNN-based models. This is seen in the perfect scores (Precision = 1.0, Recall = 1.0, F1 = 1.0) achieved by ViT in the identification of the solid waste category. Similar performance was obtained by MobileNetV2, underscoring the strong class separation capabilities of these two models. In contrast, InceptionV3 experienced a decrease in precision (0.93) and F1 score (0.96) due to the misclassification of other disturbances, such as solid waste.
- For the case of dry potholes, InceptionV3 achieved a perfect performance with precision, recall, and F1 scores all at 1.0. ViT also performed commendably, with precision, recall, and F1 scores each at 0.96. In contrast, MobileNetV2 had the weakest performance in this category, with precision at 0.92, recall at 0.88, and an F1 score of 0.90, indicating a high rate of misses and misclassifications.
- Water-filled potholes and speed breakers posed the most significant challenges and accounted for the majority of misclassifications. Several examples of these are illustrated in Fig.10, with further details provided in Table.4. For speed breakers, ViT excelled, achieving perfect recall (1.0) and a balanced precision (0.93), resulting in an F1 score of 0.96, as depicted in blue in Fig.6-Fig.8. In contrast, MobileNetV2

experienced a notable performance decline, with a precision of 0.74, recall of 0.92, and an F1 score of 0.82.

In identifying water-filled potholes, ViT achieved an F1 score of 0.98, with perfect recall (1.0) and commendable precision (0.96). Conversely, InceptionV3 and MobileNetV2 missed 20% of these anomalies, both exhibiting lower recall (0.80) and F1 scores (0.87). ViT's ability to detect subtle cues often overlooked by traditional CNNs, such as surface depressions and water reflections, is attributed to its advanced transformer-based context aggregation. This has clear implications for car preview controllers: achieving a balance in precision is necessary to avoid unnecessary interventions, while high recall is crucial for safety to ensure no anomalies are missed. Further, ViT poses fewer risks, as demonstrated by its higher recall and overall F1 score, especially when handling unclear road anomalies.The difference in performance is further analyzed statistically. As the F1 score uses a balanced approach invoking the values of precision and recall, further analysis is done based on this metric. Non-parametric tests are performed on the class-wise F1 scores of ViT, InceptionV3, and MobileNetV2 to compare their classification performance on road anomaly detection. The Friedman test [15]is used to evaluate overall differences due to the small sample size (six classes) and non-normal distribution of scores. Wilcoxon signed-rank tests are then used for pairwise comparisons [16].

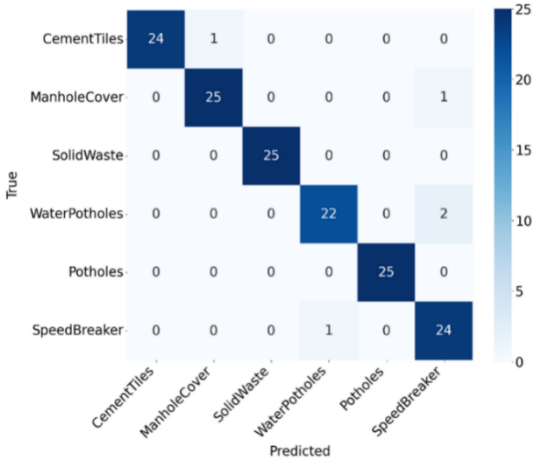

Fig.4 Confusion Matrix of InceptionV3

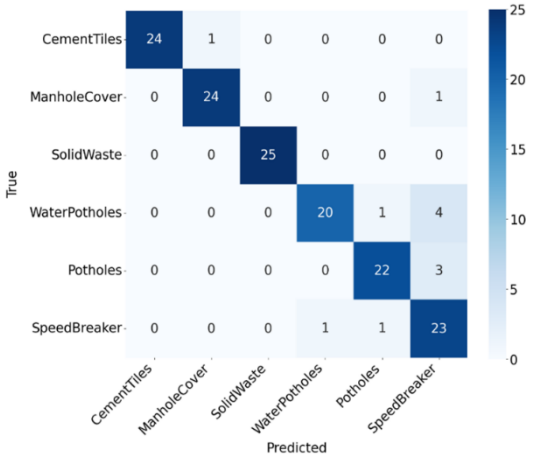

Fig.5 Confusion Matrix of MobileNetV2

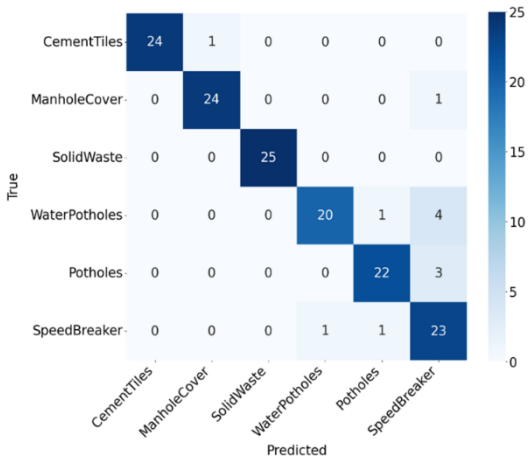

Fig.6 Confusion Matrix of ViT

The average F1 scores are as follows.

- ViT mean F1 ≈ 0.973
- InceptionV3 mean F1 ≈ 0.946
- MobileNetV2 mean F1 ≈ 0.921

So on the whole, ViT > InceptionV3 > MobileNetV2.

ViT had better F1 scores than InceptionV3 in four of six classes. In only one class (Potholes), InceptionV3 did better The Friedman test statistic (χ^2) was approximately 6.5 with a p-value < 0.05, indicating that there is a statistically significant difference in classification performance across at least two of the models. This suggests that at least one model consistently outperforms the others across the six classes, necessitating further pairwise analysis.

Pairwise Wilcoxon signed-rank tests were used to find specific model differences, which gave the following insights.

- **ViT vs. InceptionV3**: ViT had better F1 scores than InceptionV3 in four of six classes. In only one class (Potholes), InceptionV3 did better than ViT. However, the signed-rank test showed a statistically significant difference ($p < 0.05$) in favour of ViT. ViT's design gives a better balance of precision and recall in most anomaly classes than InceptionV3, especially in critical cases like potholes filled with water and speed breakers.
- **ViT vs. MobileNetV2**: ViT did better than MobileNetV2 in five out of six classes and tied in one. The test showed a significant difference ($p < 0.05$) in favour of ViT. Thus, ViT is much better than MobileNetV2, which has trouble with classes that need more detailed feature representation (like speed breakers).
- **InceptionV3 vs. MobileNetV2**: The differences between the two are less significant. The Wilcoxon test revealed a marginal difference, not statistically significant at the 0.05 level ($p \sim 0.1$). While InceptionV3 frequently outperforms MobileNetV2, the difference is inconsistent enough to claim supremacy.

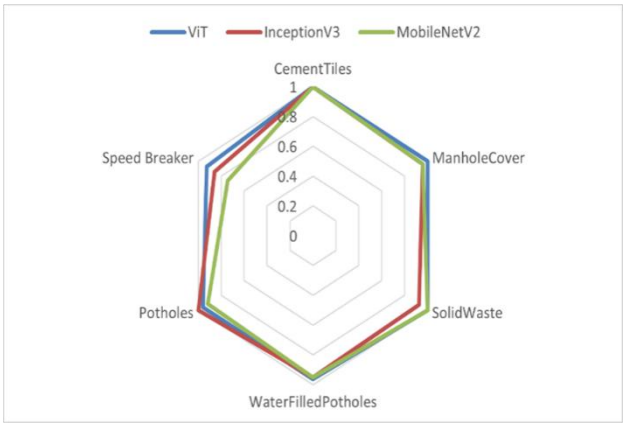

Fig.7. Precision values of classwise predictions

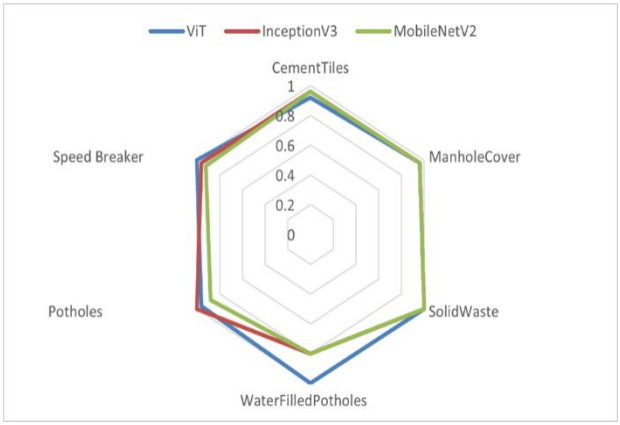

Fig.8. Recall values of class wise predictions

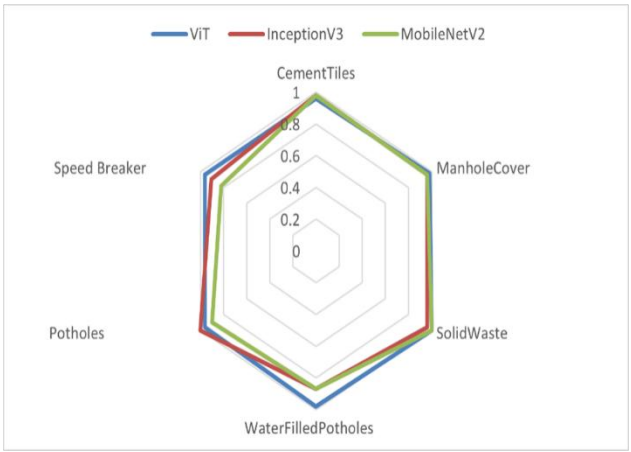

Fig.9. F1 scores of classwise predictions

The Friedman test rejects the null hypothesis of equal model performance, and the Wilcoxon signed-rank tests show that ViT has a statistically significant advantage over both InceptionV3 and MobileNetV2. The Friedman test statistically validates that ViT is the best-performing model in the road anomaly classification problem based on F1 scores across multiple classes, reflecting its superior capability to balance precision and recall.

Vision Transformers exhibits the best overall performance, combining high accuracy and fewer misclassified samples, followed closely by InceptionV3 and MobileNetV2. These results indicate ViT's superior adaptability to the dataset's characteristics. MobileNetV2 uses depth-wise separable convolutions to create a lightweight model, making it computationally efficient. MobileNetV2's architecture contributes to its superior accuracy while maintaining a balance with computational efficiency, making it a suitable model for real-time preview control augmentation. ViT surpasses MobileNetV2 and InceptionV3 in anomaly classification by effectively capturing long-range dependencies, dynamically emphasising anomalies, and leveraging large-scale pretraining for improved generalisation. However, ViT may demand higher computational cost and memory usage due to the quadratic scaling of its complexity involving the self-attention mechanism [15].

(a) Misclassified prediction 1

(b) Misclassified prediction 2

(c) Misclassified prediction 3

Fig.10 Challenging predictions – a few instances

Table.4. Misclassified predictions – a few instances

Ground Truth	ViT	InceptionV2	MobileNetV3
Water-filled Potholes (Fig.10(a))	Water-filled Potholes	Water-filled Potholes	Solid Waste
Water-filled Potholes (Fig.10(b))	Water-filled Potholes	Solid Waste	Water-filled Potholes

Speed breaker (Fig.10(c))	Speed breaker	Speed breaker	Water-filled Potholes

4. CONCLUSIONS

This research investigates the application of deep learning models for multiclass road disturbance classification, addressing the need to detect various road anomalies in intelligent transportation systems. Using a novel dataset focusing on Indian road conditions developed as part of this work, the study compares the performance of three DL models: ViT, MobileNetV2 and InceptionV3 for road disturbance classification. The research findings reveal that ViT achieves exceptionally high testing accuracy with superior precision, recall and F1 scores compared to the other models. MobileNetV2 and InceptionV3 also demonstrate robust performance, effectively balancing accuracy and generalizability. While the study is based on data compiled from Indian road networks, the conclusions drawn are relevant to other developing countries having similar road profiles. Further, the dataset developed as part of this study is made publicly available to support further research in this domain. Future work could investigate alternative architectural combinations to enhance accuracy and computational efficiency. Expanding the scope from image classification to object detection could enable real-time applications in autonomous driving, surveillance, and active vehicle suspension. Additionally, real-world applications, such as preview control augmentation, use high-resolution images and videos, in which case computational time and resources play a vital role. MobileNetV2 may be significant in this context due to its lightweight architecture. Future studies focusing on computational time and resources can bring additional insights into this aspect. This study lays a foundation for advancing road disturbance detection systems and improving vehicle vibration control and suspension systems, promoting safer and more efficient mobility.

REFERENCES

[1] E. Zalama, J. Gomez-Garcia-Bermejo, R. Medina and J. Llamas, "Road Crack Detection using Visual Features Extracted by Gabor Filters", *Computer-Aided Civil and Infrastructure Engineering*, Vol. 29, No. 5, pp. 342-358, 2014.

[2] L. Zhang, F. Yang, Y. Daniel Zhang and Y.J. Zhu, "Road Crack Detection using Deep Convolutional Neural Network", *Proceedings of International Conference on Image Processing*, pp. 3708-3712, 2016.

[3] Z. Tong, J. Gao and H. Zhang, "Innovative Method for Recognizing Subgrade Defects based on a Convolutional Neural Network", *Construction and Building Materials*, Vol. 169, pp. 69-82, 2018.

[4] J.J. Yebes, D. Montero and I. Arriola, "Learning to Automatically Catch Potholes in Worldwide Road Scene Images", *IEEE Intelligent Transportation Systems Magazine*, Vol. 13, No. 3, pp. 192-205, 2021.

[5] H. Maeda, Y. Sekimoto, T. Seto, T. Kashiyama and H. Omata, "Road Damage Detection and Classification using Deep Neural Networks with Smartphone Images", *Computer-Aided Civil and Infrastructure Engineering*, Vol. 33, No. 12, pp. 1127-1141, 2018.

[6] M. Inoue, Y. Kawasaki, T. Suzuki, Y. Washimi, T. Tanimoto and M. Takahashi, "Point Cloud Interpolation by RGB Image to Estimate Road Surface Profile for Preview Suspension Control", *Proceedings of International Symposium on Advanced Vehicle Control*, pp. 672-678. 2024.

[7] Samuel Ochai Audu-War, Sylvanus Okwudili Anigbogu, Kenechukwu Sylvanus Anigbogu, Gloria Nkiru Anigbogu and Doris Chinedu Asogwa, "Pothole Detection using Image Surveillance System: A Review", *World Journal of Advanced Engineering Technology and Sciences*, Vol. 9, No. 2, pp. 214-222, 2023.

[8] L. Xu, X. Zhou, Y. Tao, L. Liu, X. Yu and N. Kumar, "Intelligent Security Performance Prediction for IoT-Enabled Healthcare Networks using an Improved CNN", *IEEE Transactions on Industrial Informatics*, Vol. 18, No. 3, pp. 2063-2074, 2022.

[9] J.H. Jeong, H. Jo and G. Ditzler, "Convolutional Neural Networks for Pavement Roughness Assessment using Calibration-Free Vehicle Dynamics", *Computer-Aided Civil and Infrastructure Engineering*, Vol. 35, No. 11, pp. 1209-1229, 2020.

[10] M.H. Kim, J. Park and S. Choi, "Road Type Identification Ahead of the Tire using D-CNN and Reflected Ultrasonic Signals", *International Journal of Automotive Technology*, Vol. 22, No. 1, pp. 47-54, 2021.

[11] Y. Safyari, M. Mahdianpari and H. Shiri, "A Review of Vision-based Pothole Detection Methods using Computer Vision and Machine Learning", *Sensors*, Vol. 24, No. 17, pp. 1-7, 2024.

[12] C. Saisree and U. Kumaran, "Pothole Detection using Deep Learning Classification Method", *Procedia Computer Science*, pp. 2143-2152, 2022.

[13] D.J. Hand, P. Christen and N. Kirielle, "F*: An Interpretable Transformation of the F-Measure", *Machine Learning*, Vol. 110, No. 3, pp. 451-456, 2021.

[14] O. Rainio, J. Teuho and R. Klen, "Evaluation Metrics and Statistical Tests for Machine Learning", *Scientific Reports*, Vol. 14, No. 1, pp. 1-7, 2024.

[15] J. Demsar, "Statistical Comparisons of Classifiers Over Multiple Data Sets", *Journal of Machine Learning Research*, Vol. 7, pp. 1-30, 2006.

[16] A. Benavoli, G. Corani and F. Mangili, "Should We Really Use Post-Hoc Tests based on Mean-Ranks?", *Journal of Machine Learning Research*, Vol. 17, No. 1, pp. 152-161, 2016.

[17] Alexey Dosovitskiy, Lucas Beyer and Alexander Kolesnikov, "An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale", *Proceedings of International Conference on Computer Vision and Pattern Recognition*, pp. 45-67, 2021.