International Journal of Innovative Research in Computer Science and Technology (IJIRCST)
ISSN (Online): 2347-5552, Volume-13, Issue-3, May 2025
DOI: https:/doi.org/10.55524/ijircst.2025.13.3.25
Article ID IRP-1648, Pages 179-187
www.ijircst.org

A Real-Time Road Anomaly Detection via Masked Autoencoder-Enhanced Vision Transformers

Ahmad Khalid Hussain[1], and Fati Oiza Ochepa[2]

[1] B.Sc. Scholar, Department of Computer Science, Federal University Lokoja, Nigeria
[2] Lecturer, Department of Computer Science, Federal University Lokoja, Nigeria

Correspondence should be addressed to Ahmad Khalid Hussain; ahmadkhalidhussain408@gmail.com

Received 28 April 2025; Revised 13 May 2025; Accepted 28 May 2025

ABSTRACT- This study presents a deep learning solution for detecting road anomalies via a hybrid architecture consisting of a Masked Autoencoder (MAE) and a Vision Transformer (ViT) model. It presented a framework for dual road classification, namely an intact road (good) and defected road (bad) where defected roads are characterized by anomalies such as potholes or cracks. The target road anomaly classification model was trained and tested using publicly available datasets of road condition images. The model demonstrated good feature extraction as well as good generalization with a training accuracy of 99.79% and a test accuracy of 98.29%. Furthermore, we integrated the road anomaly detection model into a web-application providing real-time road anomaly detection, exemplifying the possible benefits of applying computer vision and machine learning algorithms to improve road maintenance in Nigeria.

KEYWORDS- Anomaly detection, Deep learning, Image processing, Masked autoencoders, Transport safety, Vision Transformers

I. INTRODUCTION

Poor condition of roads is one of the significant issues in transportation engineering and the possible negative impacts on safety, comfort, and overall experience of road users. Roads in a bad state degrade the vehicle, increase the risk of accidents, and increase the maintenance costs of road agencies [1]. To improve roadway safety and better manage infrastructure requires a validated, reliable non-intrusive, and less costly alternative to the traditional methods of manual inspections. As the use of computer vision, coupled with machine learning techniques has emerged recently as a unique way of inspecting road anomalies, automatic systems that identify road faults can improve operations and have greater effects as they reduce repair costs and make roads safer for traffic as discussed in [2] and [3]. The systems identify road faults well, also in areas that are inaccessible or risky. It is faster and improved compared to manual checking by humans. Creating a reliable system to identify road faults can contribute considerably to traffic safety. It reduces expenses from damage of vehicles and enhances travel. Conversely, traditional methods of identifying faults by hand are time-consuming and involve a lot of work, and they are prone to errors. This is why there has been interest in automatic systems that employ vision and machine learning to identify and categorize faults.

Despite the significant advances that have taken place in research on road anomaly detection, issues still crop up, especially with ensuring detection system accuracy and reliability [2]. Real-time detection is also essential in delivering timely alerts for both drivers and road authorities [4]. While automated systems have significant potential for improved road safety and reduced costs of maintenance, further research is necessary for improving their resilience, especially in real-world setups [5].

Traditional machine learning models, like CNNs, have demonstrated promise in the detection of road anomalies. Nonetheless, they are often unable to detect complex patterns and contextual information in images, thus diminishing their performance in certain situations. Masked Autoencoders (MAE) is a further developed mechanism for feature learning that involves training the model on predicting the missing parts of an image, forcing the model to learn richer and more significant features. This feature supports the system in detecting minute road imperfections, even for those that are not directly observable in images. This is especially beneficial in the detection of small, localized imperfections, like potholes or fissures, that are hard to detect with traditional approaches [16]

In numerous applications of computer vision, Vision Transformers (ViTs) have been shown to perform exceptionally, particularly with big and complicated data. While CNNs, on the contrary, extract local features, the ViTs are able to extract universal contextual relations in an image since they process the image entirely in parallel. It is for this reason that the ViTs are appropriately suited for detecting large-scale road anomalies that require spatial relations, for example, the progress of cracks or the interaction of different types of damage [17]. The combination of MAE for feature learning and the use of ViTs for anomaly detection and image classification is a stronger system that rectifies the deficiencies of previous methods. Below are the primary contributions of the research:

- Road Anomaly Detection Framework Development using Deep Learning Algorithm: This system applies Masked Autoencoders (MAEs) for feature learning and uses Vision Transformers (ViTs) for anomaly detection, providing a mechanism for detecting faults in Nigerian roads. This is an improvement over the traditional methods in that it increases the effectiveness and precision in detecting road anomalies.
- Real-time web-based road monitoring system: The proposed deep learning frameworks are incorporated in a real-time system in order to identify road irregularities in real time. This system provides timely intervention by road maintenance authorities and facilitates effective management of road safety.
- Overall performance assessment: The system is tested on different machine learning criteria in order to consider its reliability, stability, and feasibility in real-world large-scale road repair activities.

II. REVIEW OF RELATED WORKS

Machine learning and computer vision have also been investigated in a number of studies for the detection of irregularities on roads. For instance, Ramesh et al.[6] applied deep learning models like MobileNet, Inception-v3, and YOLOv5 for real-time detection of potholes with impressive accuracy. Road surface flaws were found to be a significant risk for vehicle integrity and traffic safety in another study by R. Bibi et al. [5]. These flaws are made worse by climate change, subpar building materials, and growing traffic volumes. In order to enable autonomous identification of traffic irregularities, the authors presented a system that combines deep learning (DNN) algorithms with in-vehicle sensors. The technology enables autonomous vehicles to identify and categorize road imperfections, including potholes, cracks, and bumps, and relay this information to neighboring vehicles by utilizing Edge AI and vehicular ad hoc networks (VANETs). Using publicly available information, experimental results showed that models such as ResNet-18 and VGG-11 were very accurate in identifying different types of road surface conditions.

Kim et al. [7] divided automated pothole detection systems into vision-based, vibration-based, and 3D reconstruction-based techniques, highlighting how deep learning can improve the detection's precision. Likewise, Tahir et al [1] proposed FactorNet, a slim deep learning architecture, that was found to outperform Detectron2 in terms of detection efficiency. Furthermore, research has examined sensor-based and hybrid detection methods, for example, Ben-Shoushan & Brook [8] combined thermal and RGB imagery with YOLOv5 CNN, enhancing pothole detection in poor illumination conditions. Ramesh et al. [9] designed a cloud-based system with a combination of YOLOv5 and LSTM for tracking road conditions with smartphone sensors. Hassan et al. [10] created a predictive maintenance system by applying R-CNN in analyzing large-scale images of roads in Norway.

Nonetheless, in tackling harsh conditions, Bucˇko et al. [11] proved YOLOv3's performance in the detection of potholes in low-light and harsh weather conditions. Salaudeen & Celebi [12] employed the image enhancement with the use of an ESRGAN-based algorithm in upgrading pothole detection in complicated conditions. Jakubec et al. [13] use of CNN-based models in relation to pothole spotting in various real-world conditions is investigated in this manuscript. Contemporary Region-Based CNN (R-CNN) as well as You Only Look Once (YOLO) variants are some of the models tested. Under conditions of rain, evening, night, and intense sunshine, the YOLO models had faster learning rates and enhanced detection precision. On the other hand, R-CNN models performed better in conditions with poor visibility, mostly during night. They contribute to the result of the emerging structure in the literature, as well as in aiding in the choice of all deep models suitable for road surface anomaly detection in varied environments.

In Nigeria, for instance, Ezeibe et al. [14] in their research in underscored the importance of road transport in the social and economic environment, while potholes were determined to be the leading cause of accidents. Their research underscored the importance of upgraded detection systems for tackling the recognition of speed limit signs and abandoned road humps. Potluru et al. [15] proposed a new framework, an algorithm that applied a hybrid technique for road anomaly detection employing the Swin Transformer and YOLOv8. The Swin Transformer applies a shifting window-based self-attention mechanism with an accurate achievement rate of 97.64% while YOLOv8 was surpassed by the Deep Convolutional Neural Network (DCNN) with a 98.75% achievement rate. Their research identifies that while both models are performing, the combination of DCNN-YOLOv8 proves stronger in terms of robustness and precision, notably in real-world applications. This research reaffirms the potential for high-end neural models in upgrading autonomous road infrastructure monitoring and presents a promising avenue for future intelligent transportation systems.

III. METHODOLOGY

The research uses a sequential, iterative pipeline of data gathering, preparation, modeling, testing, and deployment as shown in Figure 1. Road images were obtained, inspected, and split into training and testing sets. A hybrid structure of Masked Autoencoder (MAE) for feature learning and a Vision Transformer (ViT) for categorization was employed. The model was trained and tested with its performance measured in terms of accuracy, precision, recall, and F1-score. Deployment in web-based real-world applications is in the pipeline to facilitate real-time road anomaly detection.

Figure 1: General Pipeline of the Proposed System

A. System Design

The proposed Road Anomaly Recognition System uses an integrated deep learning network, as depicted in Figure 2. It integrates Masked Autoencoder (MAE) and Vision Transformer (ViT). The MAE is used for unsupervised feature learning by predicting masked areas in input road images and permitting the model to learn influential semantic and structural features that are necessary in identifying road faults (He et al., 2022) in [16]. The ViT receives this information and is very efficient in detecting global spatial relationships and dependencies that runs through the image (Dosovitskiy et al., 2021) in [17]. This integration enhances the system's ability in detecting hard-to-detect abnormalities with high precision and generalization, with examples being potholes and other road faults. The transformer-based methodology is suitable for real-time applications in intelligent transportation systems since it supports scalable and efficient processing.

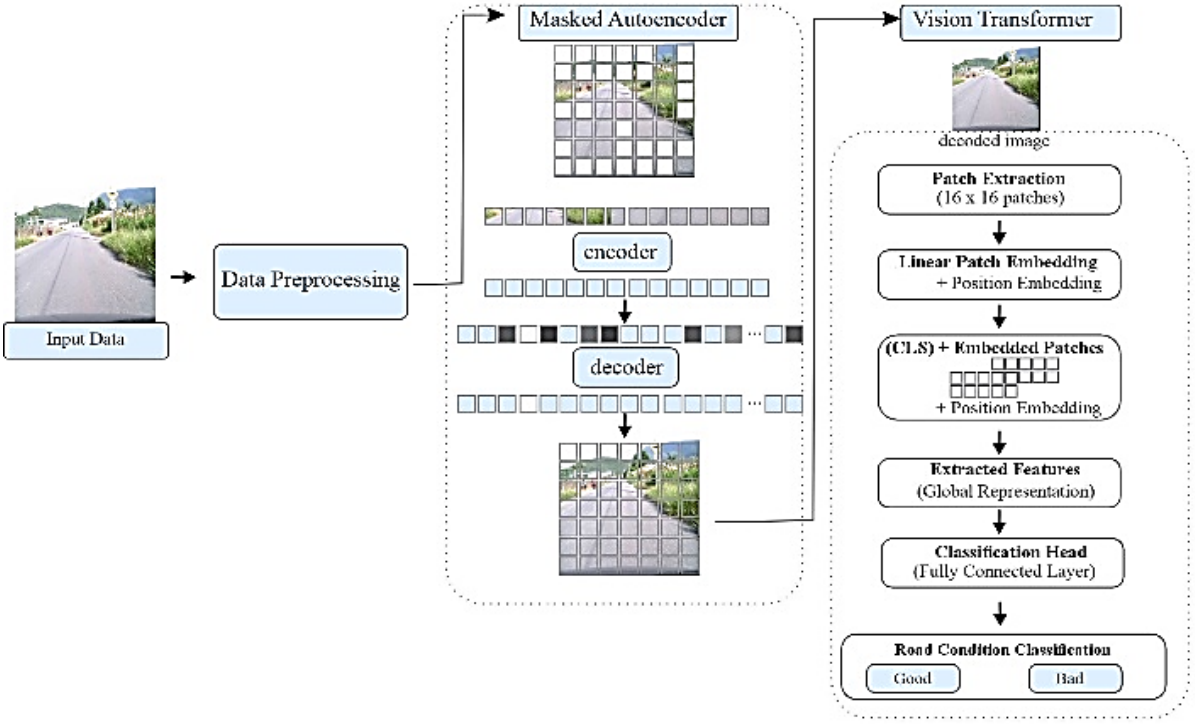

Figure 2: Mae-Vit Hybrid Road Anomaly Detection Framework

B. Masked Autoencoders

For this research, a DeepConvAutoencoder was employed for unsupervised representation learning and reconstruction of images with applications in anomaly detection for roads. The model consists of two main components, an encoder that compresses the input to a small latent representation and a decoder that reconstructs the input from the resultant representation. Assuming a kernel size of 3×3, stride = 2, and padding = 1, the encoder's convolution with ReLU activation in its four layers gradually reduces the spatial dimensions of the image but enhances its feature depth from three channels (RGB) alone to 128 channels. This arrangement enables the network to learn spatially localized as well as hierarchal representations of the feature. The decoder upsamples the feature maps and reconstructs the input image by imitating this topology with ConvTranspose2d layers. For normalized image data, its last layer of output restricts pixel values to the range of [0,1] with a Sigmoid activation.

Mean squared error (MSE) loss function is employed for model training in order to minimize pixel-to-pixel variance between the input and output images. For the purpose of optimization, the Adam Optimizer is employed, and the learning rate is 0.003. The architecture of the overall model is illustrated in Table 1, and a visualization of the behavior of the model during training is presented in Figure 3, highlighting a clear reduction in the reconstruction error with every 10 epochs of training, demonstrating successful convergence.

Table 1: Architecture of the Deep Convolutional Autoencoder

Component	Layer Type	Output Channels	Kernel Dimension	Stride	Padding	Activation
Encoder	Conv2d	16	3×3	2	1	ReLU
	Conv2d	32	3×3	2	1	ReLU
	Conv2d	64	3×3	2	1	ReLU
	Conv2d	128	3×3	2	1	ReLU
Decoder	ConvTranspose2d	64	3×3	2	1	ReLU
	ConvTranspose2d	32	3×3	2	1	ReLU
	ConvTranspose2d	16	3×3	2	1	ReLU
	ConvTranspose2d	3	3×3	2	1	Sigmoid

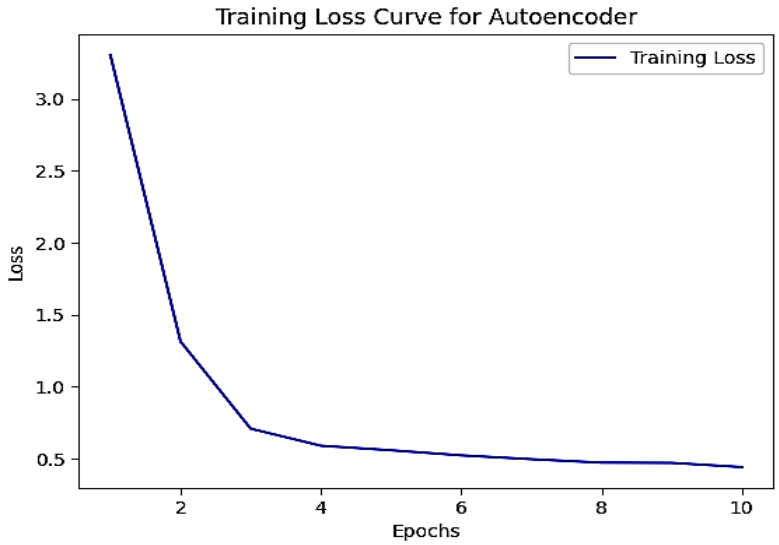

Figure 3: Training loss curve of the deep convolutional autoencoder over 10 epochs.
The steady drop in the loss indicates that the model is learning well and improving over time.

C. Vision Transformer

The model employed in the proposed system incorporated the Vision Transformer (ViT) - google/vit-base-patch16-224-in21k pretraining architecture - for predicting road surface conditions in terms of two classes, namely, good (intact) and bad (defective). This pre-trained transformer model is trained on the ImageNet-21k dataset and further fine-tuned on the road anomaly dataset in order to learn high-level and discriminative visual features. Every input image $x \in R^{(H\times W\times C)}$ is divided into a set of fixed-size, non-overlapping patches of size $16\,x\,16$. For a resized image of size 224 x 224, this means we have $N = \frac{224\times 224}{16\times 16} = 196$ patches. The individual patches are projected linearly into a feature space of size D, with a learnable classification token being added to the sequence. Positional embeddings are added in order to preserve the spatial structure as shown in Equation 1,

$$z_0 = \left[x_{\text{cls}}; x_p^1 E; x_p^2 E; \ldots; x_p^N E\right] + E_{\text{pos}} \qquad (1)$$

where x_p^1 is the i^{th} image patch, $E \in R^{(P^2\cdot C)\times D}$ is the linear projection matrix, and E_{pos} are learnable positional embeddings. A series of Transformer encoder layers is used to process this embedded sequence. This embedded sequence of images is processed in a sequence of Transformer encoder layers. Each of the layers consists of a Multi-Layer Perceptron (MLP) and Multi-Head Self-Attention (MHSA) both of which are preceded by Layer Normalization and followed by residual connections. The forward pass in each of the encoder layers is described in Equation 2,

$$z_l' = \text{MSA}\big(\text{LN}(z_{l-1})\big) + z_{l-1}\ ,$$
$$z_l = \text{MLP}\big(\text{LN}(z_l')\big) + z_l' \qquad (2)$$

where l is the index of the layer. Following the last layer, the output that corresponds with class token z_L^0 is taken, then passed through a linear classification head followed by a sigmoid activation function for binary classification as presented in Equation 3.

$$\hat{y} = \sigma(w^T z_L^0 + b) = \frac{1}{1 + e^{-(w^T z_L^0 + b)}} \qquad (3)$$

Here, the weight and bias of the classification layer are represented by w and b. Equation 4 presents the Binary Cross-Entropy loss function is used in order to train the model:

$$L = -\big[y \log P\,(\text{anomaly}|x) + (1 - y)\log\big(1 - P(\text{anomaly}|x)\big)\big] \qquad (4)$$

where y is the true label (0 for normal roads, 1 for anomalies) and $P(\text{anomaly}|x)$ as the predicted probability of anomaly by the model. Equation 5 presents approximations of the first and second moments of the gradients, upon which the Adam optimizer decides the learning rate for each parameter.

$$w = w - \eta \cdot \frac{m_t}{\sqrt{v_t} + \epsilon} \qquad (5)$$

Where η is the learning rate, while first and second moments estimates are m_t and v_t respectively, and is a small quantity for numerical stability. All images are resized to 224 x 224 x 3 with bilinear interpolation to be compatible with the expected input of ViT's. For intermediate output of varying channel depth from masked autoencoder, 1×1 convolutional projection layer is employed for conversion into 3 channels.

D. Evaluation Metrics

The system's capacity for detecting road irregularities is quantified by applying conventional measures of classification that include accuracy, precision, recall, and F1-score as shown in Figures 6, 7, 8, and 9 respectively.

$$\text{Accuracy} = \frac{TN + TP}{TP + TN + FP + FN} \qquad (6)$$

$$\text{Precision} = \frac{TP}{FP + TP} \qquad (7)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (8)$$

$$F1 = 2\,x\,\frac{\text{Recall x Precision}}{\text{Precision} + \text{Recall}} \quad (9)$$

Where TN is the number of properly identified normal roads (true negatives), FP is the number of wrongly identified normal roads as anomaly (false positives), FN is the number of anomaly roads wrongly identified as normal (false negatives), and TP is the number of anomaly roads properly identified as anomaly (true positives).

IV. EXPERIMENTAL DESIGN

A. Dataset Description

The dataset consists of 1,755 road images gathered from a public repository, labeled as normal and anomalous roads. Suitable for preliminary training, its small size and limited variability can, however, limit the model's capacity for generalizing on diverse road conditions due to variability in resolution, environmental conditions, and texture of surfaces. To minimize these challenges, data augmentation methods were employed, such as geometric transforms (flipping, rotation, cropping) for improved variability in spatial terms, and photometric transforms (brightness, contrast, saturation) for illumination adaptation improvement. Figure 4 shows a subset of the acquired dataset. The dataset was divided into 80% for training and 20% for testing for model development and verification, as indicated in Figure 5. Notwithstanding the foregoing, adding road images from different geographically widespread regions of the world would further enhance the generalization and real-world performance of the model.

Figure 4: Dataset sample

Figure 5: Dataset splitting

Therefore, there is an organized preprocessing pipeline that is employed in order to process the images for deep models. The unwanted regions of the image are cropped, with the region of interest being the road surface, as in Equation 9.

$$X_{cropped} = X[\text{crop\_coordinates}] \quad (9)$$

Labels are encoded as binary values (0 for normal, 1 for anomalous)

$$y_{labelled} = \begin{cases} 0, & no\ anomaly \\ 1, & anomaly\ detected \end{cases} \quad (12)$$

B. Simulation

The platform for the training environment was Python in Google Collaboratory, on a system (2.6GHz, 12GB RAM,

500GB storage). Installed on this was Anaconda and Visual Studio Code (VS Code), on Microsoft Windows 10, with a web browser. During the setup of the project, the system installed the latest version of important machine learning libraries on it, which included OpenCV for image processing, PyTorch for deep learning, Transformers for processing pre-trained models, and NumPy for numerical computations. Data preprocessing and visualization as well as model evaluation dependencies were also installed.

V. PERFORMANCE EVALUATION

A. Model Performance

The proposed hybrid model based on a Masked Autoencoder and a Vision Transformer, was tested for its robustness in categorizing road conditions into two classes: good roads (intact) and bad roads with potholes or cracks. As seen in Figure 4, the process of training improved both the training and the test accuracy. The model reached a training accuracy of 99.79%, while the test accuracy was 98.29%, reflecting both efficient feature extraction and learning from images of the road surface. Additionally, the test loss kept declining with training, and there was little or no gap in the difference between the train and the test metrics, indicating excellent generalization capacity. Early stopping was enforced at epoch 16, maintaining model stability and avoiding overtraining past the optimal point of convergence.

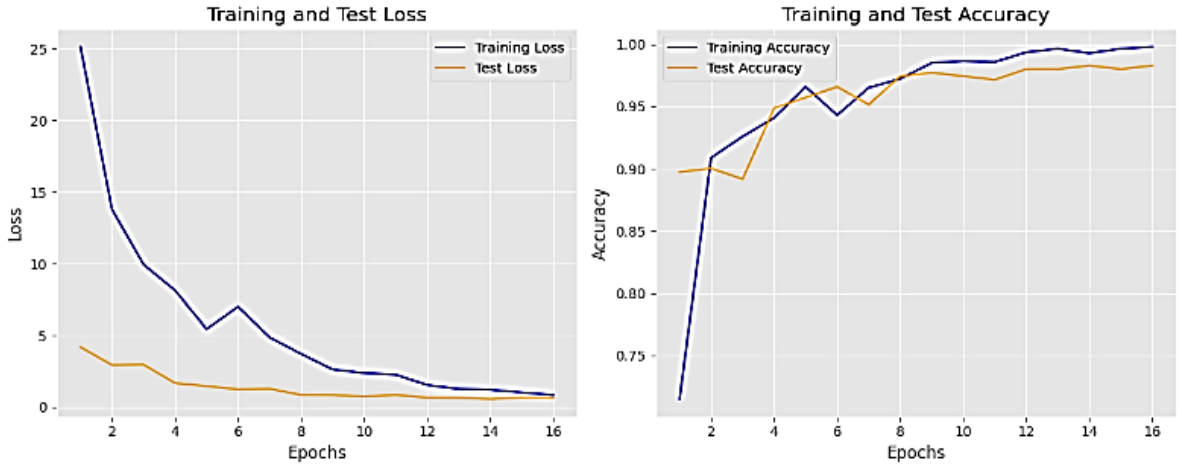

Figure 6: Training and evaluation curves of the proposed hybrid Mae-Vit model

Table 2 shows the performance analysis of the model by classifying roads as "Good" or "Bad,". The model recorded an impressive performance with an overall validation accuracy of 98%, meaning it correctly predicts road conditions in 98% of cases. Precision was 97% for good roads and 100% for bad roads. Recall for bad roads was 96% and 100% for good roads, F1-score for bad roads was 98% and 99% for good roads. The high-performance metrics recorded reflects robust overall performance of the model.

Table 2: Model Performance Analysis

Metric	Good Roads	Bad Roads
Precision	97%	100%
Recall	100%	96%
F1-score	99%	98%
Accuracy		98%

B. Error Analysis

In order to analyze the performance of the classification even better, Figure 7 presents the confusion matrix where the model has classified 144 bad roads and 201 good roads accurately, 6 bad roads are wrongly classified as good (false positives) and this might lead to undetected road faults, however, there are no good roads classified as bad roads (false negatives).

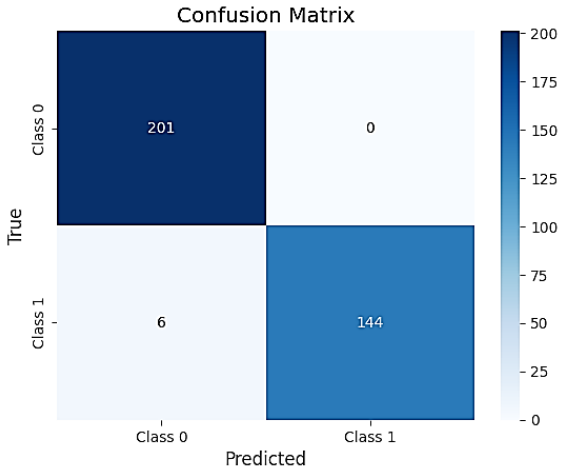

Figure 7: Confusion matrix for road condition classification

C. Deployment

In the final phase of the research, the trained model is deployed as a web-based application that is able to identify road anomalies in real-time, as illustrated in Figure 8. The users are able to upload images of the road surfaces, and the system classifies them as "good" or "bad" (normal or abnormal). The tool for achieving this is created with HTML, CSS, and JavaScript for the front-end, and Flask for the back-end for handling image uploads, processing,

and serving model predictions. The time for a response by the server, from the point the user uploads an image until the time the result of the classification is displayed, typically ranging from a few hundred milliseconds to a few seconds, depending on the performance of the server, size of the image, and processing by the model. Once the model has been loaded upon initialization, the response for successive requests is faster, as the model does not have to be loaded again. The tool was also tested with real-road images, which had different illumination conditions, quality, and road surfaces. The tests revealed that the model is able to perform well on new images, maintaining high classification accuracy. The user interface was smooth, with images being uploaded and displayed along with the outcome promptly, and the tool is a valuable asset for road maintenance teams. The tool's capacity for classifying road images in real-time, together with its friendly interface and stable performance on different images, enables the road maintenance teams to respond promptly to road abnormalities.

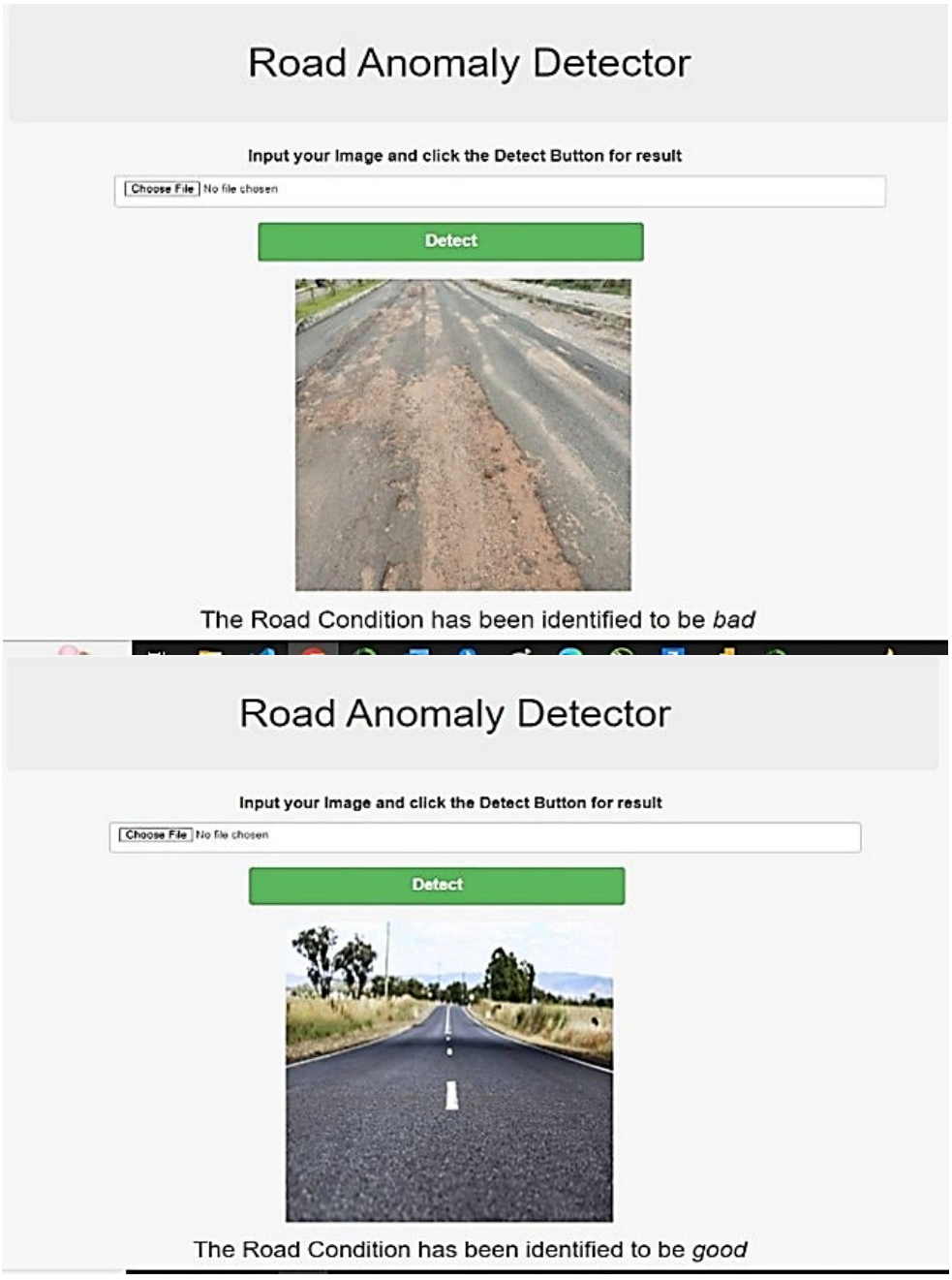

Figure 8: Web Application illustrating the road anomaly detection system user interface, presenting real-time road condition analysis with identified anomalies

VI. DISCUSSION

The model, built on masked autoencoders and vision transformers, has demonstrated solid performance in detecting road anomalies. Detailed in Table 2, the model had a precision of 97% on good roads and 100% on bad roads, showcasing capability in accurately separating road conditions. Recall on bad roads was 96%, supporting the effectiveness of the model in detecting faulty roads with minimal false negatives. F1-score on bad roads was 98%, indicating a well-balanced performance on precision and recall, especially for detection of faults. The model

performed equally well in the case of classifying good roads, with precision of 97%. This balanced performance on both classes supports that the model is suitable for real-world applications in which properly classifying good and bad roads is critical.

A deeper examination of the confusion matrix further shows that 144 bad roads were classified accurately along with 201 good roads. It also incorrectly classified 6 bad roads as good (false negatives), potentially resulting in undetectable road flaws. Of interest is that there were no false positives, i.e., there were no good roads classified as bad. The misclassifications indicate avenues of improvement, specifically in the ability of the model to separate intact roads from faulty roads in different conditions. While the model shows high performance, there is potential for improvement, specifically in the optimization of inference speed in order to accommodate real-time applications in mobile platforms. Implementing the model on mobile devices can dramatically improve its usability in field activities. Also, the use of multiple models or hybrid methods may further enhance the model's precision and resilience.

Now deployed as a web-based tool, the system allows for real-time road anomaly detection, with a user-friendly HTML front-end and a Python-based backend that allows for efficient decision-making. All of these serve users with a simple-to-use interface for tracking the conditions on the road and decision-making through the outputs of the model. Future efforts can be on the integration of multi-model frameworks for increased precision and for expanding the scope of the model.

VII. CONCLUSION

In this work, a road anomaly detection model with masked autoencoders and vision transformers was proposed, with high detection accuracy for road defects. The model showed excellent precision and recall values, specifically for bad roads, achieving an accuracy of 98% and an F1-score of 98%. The confusion matrix indicated a high number of properly classified roads, with suggestions for improvement, including less false negatives in detecting faulty roads. The easy integration of the model as a web-based tool allows for real-time detection of road anomalies, with an easy user interface for tracking road conditions. This system provides useful insights on how road conditions can be improved in terms of road safety and road maintenance. In the future, we will be optimizing the inference speed of the model for mobile applications, investigate hybrid methods for further improvement in terms of accuracy, as well as combining multiple models for enhanced robustness. Generally, the model is promising for use in real-world applications in the monitoring of road conditions, with further potential refinement for improved accuracy and effectiveness in actual applications.

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest.

REFERENCES

[1] T. Tahir, H. Hassam, J. Choi, and E.-S. Jung, "Lightweight deep learning model for road pothole detection," *Applied Intelligence*, vol. 56, pp. 1234–1245, 2022.

[2] K. Muhammad, J. Ahmad, Z. Lv, P. Bellavista, P. Yang, and S. W. Baik, "Automated Road defect and anomaly detection for traffic safety," *Sensors*, vol. 23, no. 12, p. 5656, 2023. Available from: https://doi.org/10.3390/s23125656

[3] M. M. Q. U. H. T. Anwar, S. M., "Deep learning for road anomaly detection: A survey," *arXiv preprint*, 2022.

[4] A. Martinez-Ríos, M. R. Bustamante-Bello, and L. A. Arce-Saénz, "A review of road surface anomaly detection and classification systems based on vibration-based techniques," *Applied Sciences*, vol. 12, no. 19, p. 9413, 2022. Available from: https://doi.org/10.3390/app12199413

[5] R. Bibi, Y. Saeed, A. Zeb, T. M. Ghazal, T. Rahman, R. A. Said, S. Abbas, M. Ahmad, and M. A. Khan, "Edge AI-Based Automated Detection and Classification of Road Anomalies in VANET Using Deep Learning," *Computational Intelligence and Neuroscience*, vol. 2021, Article ID 6262194, 2021. Available from: https://doi.org/10.1155/2021/6262194

[6] R. Basher, A. R. Ayon, A. Gharamy, A. A. Zayed, and M. S. Y. Ibna Zaman, "Real-time pothole detection using deep learning models: Mobilenet, Inception-V3, and YOLOv5," B.Sc. thesis, Dept. of Computer Science and Engineering, BRAC University, Dhaka, Bangladesh, 2022. Available from: https://tinyurl.com/hutawa5r

[7] Y.-M. Kim, Y.-G. Kim, S.-Y. Son, S.-Y. Lim, B.-Y. Choi, and D.-H. Choi, "Review of recent automated pothole-detection methods," *Applied Sciences*, vol. 12, no. 11, p. 5320, 2022. Available from: https://doi.org/10.3390/app12115320

[8] R. Ben-Shoushan and A. Brook, "Fused Thermal and RGB Imagery for Robust Detection and Classification of Dynamic Objects in Mixed Datasets via Pre-Trained High-Level CNN," *Remote Sensing*, vol. 15, no. 3, p. 723, 2023. Available from: https://doi.org/10.3390/rs15030723

[9] Ramesh, D. Nikam, V. N. Balachandran, L. Guo, R. Wang, L. Hu, G. Comert, and Y. Jia, "Cloud-Based Collaborative Road-Damage Monitoring with Deep Learning and Smartphones," *Sustainability*, vol. 14, no. 14, p. 8682, 2022. Available from: https://doi.org/10.3390/su14148682

[10] M. U. Hassan, O.-M. H. Steinnes, E. G. Gustafsson, S. Løken, and I. A. Hameed, "Predictive Maintenance of Norwegian Road Network Using Deep Learning Models," *Sensors*, vol. 23, no. 6, p. 2935, 2023. Available from: https://doi.org/10.3390/s23062935

[11] Bučko, E. Lieskovská, K. Zábovská, and M. Zábovský, "Computer vision-based pothole detection under challenging conditions," *Sensors*, vol. 22, no. 22, p. 8878, 2022. Available from: https://doi.org/10.3390/s22228878

[12] H. Salaudeen and E. Çelebi, "Pothole detection using image enhancement GAN and object detection network," *Electronics*, vol. 11, no. 12, p. 1882, 2022. Available from: https://doi.org/10.3390/electronics11121882

[13] M. Jakubec, E. Lieskovská, B. Bučko, and K. Zábovská, "Comparison of CNN-Based Models for Pothole Detection in Real-World Adverse Conditions: Overview and Evaluation," *Applied Sciences*, vol. 13, no. 9, p. 5810, 2023. Available from: https://doi.org/10.3390/app13095810

[14] Ezeibe, C. Ilo, C. Oguonu, A. Ali, I. Abada, E. Ezeibe, C. Oguonu, F. Abada, E. Izueke, and H. Agbo, "The impact of traffic sign deficit on road traffic accidents in Nigeria," *International Journal of Injury Control and Safety Promotion*, vol. 26, no. 1, pp. 1–9, 2018. Available from: https://doi.org/10.1080/17457300.2018.1456470

[15] S. S. Potluru, R. Mohammad, and R. Mande, "Road anomaly detection utilizing Swin transformer and deep convolutional neural networks with yolov8," in *Innovations in Computational Intelligence and Computer Vision*, Springer, 2024, pp. 375–393. Available from: https://link.springer.com/chapter/10.1007/978-981-97-6992-6_28

[16] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. Available from: https://tinyurl.com/52bnm4z4

[17] Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, ... & N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *International Conference on Learning Representations (ICLR)*, 2021. Available from: https://arxiv.org/pdf/2010.11929/1000