ViT-CD: A Vision Transformer Approach to Vehicle Collision Detection

Fahad Salim Dalwai
School of Computer Science and Engineering
Vellore Institute of Technology, Vellore
Vellore, Tamil Nadu, India
fahadsalim.dalwai2020@vitstudent.ac.in

Goutham Senthil Kumar
School of Computer Science and Engineering
Vellore Institute of Technology, Vellore
Vellore, Tamil Nadu, India
gouthamsenthil.kumar2020@vitstudent.ac.in

Zain Sharief Ashraf Sharief
School of Computer Science and Engineering
Vellore Institute of Technology, Vellore
Vellore, Tamil Nadu, India
zain.sharief2020@vitstudent.ac.in

Swathi Jamjala Narayanan
School of Computer Science and Engineering
Vellore Institute of Technology, Vellore
Vellore, Tamil Nadu, India
jnswathi@vit.ac.in

***Abstract*—Traditional methods for accident detection often struggle to handle the complexities of dynamic scenes and varying lighting conditions, leading to sub-optimal performance in accident detection. Subsequently, Vision Transformers' self-attention mechanisms can be utilized to capture spatial relationships and contextual information within video frames, introducing an innovative solution for real-time vehicle collision detection in CCTV footage. This paper discusses the methodology, preprocessing techniques, feature extraction architecture and the required modifications for it, the classification procedure of the accidents, and also explores optimization techniques for model training and performance analysis and evaluation of ViT-CD in comparison with convolutional neural networks (CNNs). The findings demonstrate the efficiency and accuracy of the proposed approach in relatively improving accident detection accuracy conclude with high-level insights into future research directions highlight the potential impact of ViT-based systems on enhancing road safety and underscore the need for their continued exploration and development.**

***Index Terms*—Self-Attention Mechanism, Computer Vision, Image Classification, Feature Extraction, Crash Detection, CCTV Footage**

I. Introduction

One of the most common and everyday occurrences of modern-day society is vehicle traffic. Traffic Signals, roundabouts, and express carriageways are all extremely prone zones for accidents to happen, which cause a large number of casualties and loss of property every day all over the globe. According to the World Health Organization (WHO), traffic accidents take over 1.2 million lives annually, significantly contributing to the top ten global causes of death [6]. In fact, in 2018, India accounted for a staggering 6% of international road accidents, despite having just 1% of the world's vehicles. India alone also witnessed around 73% of the South Asia region's road accident fatalities [6].

With such high dangers, efforts to improve road monitoring systems have always been a significant concern, with various systems being deployed to mitigate these issues such as speed cameras, policemen, Automatic Breaking Systems, etc. But even then, a large number of accidents still occur. Thus, the focus has increased now from accident prevention to accident detection. Detection is one of the most essential ways to reduce fatalities by allowing the first responders to arrive save victims of accidents and reduce the casualties.

Projections indicate that the initial figures will increase by about 65% over the next 20 years unless there is a new commitment to detection [5]. As a result, it has become clearly evident that a new methodology to improve systems is required as soon as possible.

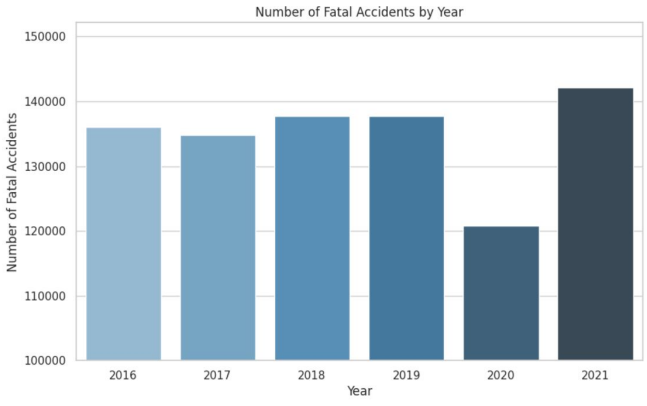

Fig. 1. Number of Fatal Accidents Over 6 year period by MRTH, India [5]

Vision transformers, first proposed in 2021, are a remarkable innovation in the field of machine learning and computer vision, and have opened up a wide array of possibilities for public services across various domains [8]. Their application ranges span from traffic management, where they can revolutionize accident detection and congestion monitoring, to urban planning, where they facilitate efficient surveillance and infras-

tructure maintenance. In law enforcement, vision transformers can augment facial recognition systems, enhancing public safety through improved suspect identification. Additionally, in healthcare, these models enable the interpretation of medical images, aiding in early disease detection and diagnosis [9]. In public safety and disaster management, they can analyze satellite imagery and aerial footage to assess the extent of natural disasters and coordinate rescue operations swiftly.

Hence, In order to improve the existing machine learning approach systems, this paper introduces the benefits of vision transformers in revolutionizing accident detection, paving the way for safer roadways and enhanced traffic management.

II. Related Work

In 2022, A. Ajith, A. O. Philip, S. Sreedhar, and M. U. Sreeja introduced real-time accident detection using deep learning and CCTV footage. Their work showed that a hybrid CNN-GRU model yields superior results for sequence-based accident classification, addressing the critical need for prompt response systems in reducing fatalities. This proved critical for research to determine whether the need for a vision transformer system was required. As GRU can effectively store and access long-term dependencies, they make excellent candidates for CCTV footage, which has temporal-based information, however, it is observed that there was still a possibility for improvement [1].

In the same year, another paper presented an automated method for detecting accidents in traffic videos. Their process involves extracting frames from video shots, identifying keyframes through histogram differences, extracting features using VGGNET, and classifying them as accident or non-accident. The effectiveness of their approach had been validated against existing methods, and it demonstrated its superior performance. [2].

Going back to 2021, V. S. Sindhu focused on real-time Vehicle Detection using Computer Vision and YOLOv4 in their paper. A novel dataset evaluated the model under diverse conditions. Initial dataset preprocessing was followed by vehicle image collection. The project culminated with assessing model performance using metrics like Precision, Mean Average Precision (mAP), and Average Intersection Over Union (IoU), enabling the recognition of diverse vehicle types. The paper presented potential applications for real-time vehicular accident detection framework development. [3].

In the same year, another paper suggested making a system that employs CCTV cameras for post-crash detection, facilitating prompt emergency response. Their system's applicability extended to security in educational and residential campuses. Moreover, accident-prone intersections are monitored, enabling timely notifications to users or authorities, making it a truly remarkable system. [4].

In 2023, another paper explored the practicality of deploying Vision Transformers (ViTs) on mobile devices, given their resource-intensive nature. While many in the field focus on creating larger ViTs for improved performance, this study takes a different approach by investigating how small a ViT can be while still maintaining a balance between accuracy and inference latency, making it suitable for mobile deployment. The authors examine several ViT designs tailored for mobile applications, noting that these adaptations often involve modifying the transformer's architecture or combining Convolutional Neural Networks (CNN) with transformers. They also discuss recent efforts to create sparse ViT networks and alternative attention mechanisms [12].

III. Existing Systems

Existing crash detection systems use various technologies, however, each has its drawbacks and as a result, the case of detecting crashes becomes a matter of efficiency versus accuracy. Some of the different existing systems deployed around the world are:

A. Sensor Based Systems

Sensor-based systems are technological setups that utilize a variety of sensors, such as cameras, radar, and LiDAR, to gather real-time data about their surroundings [10]. They analyze this data to identify and respond to specific events or conditions, often in the context of safety or automation, such as in self-driving cars. However, a disadvantage of sensor-based systems is their susceptibility to false positives, where inaccurate sensor readings can lead to unnecessary or incorrect actions, potentially impacting the system's reliability.

B. V2X communication

V2X (Vehicle-to-Everything) systems are communication setups that enable vehicles to exchange information with each other and their surrounding infrastructure. These systems facilitate data sharing about road conditions, traffic patterns, and more, enhancing overall road safety and traffic efficiency [14]. However, a disadvantage of V2X systems is their reliance on network connectivity, which can be affected by signal disruptions or cyber-attacks, potentially compromising the accuracy and reliability of the exchanged information.

C. Convolution Neural Networks

Convolutional Neural Networks (CNNs) are specialized deep learning architectures designed for processing and analyzing visual data, such as images and videos, and are also the most direct comparison to the proposed vision transformer-based approach. These networks use convolutional layers to automatically learn and extract features from the input data, enabling them to recognize patterns, objects, and structures within the images [13]. Despite their effectiveness in image-related tasks, the major drawback of CNNs is their computational complexity, which can require significant computational resources for training and inference, limiting their deployment in resource-constrained environments.

IV. Methodology

As can be seen, traditional methods struggle with dynamic scenes and lighting variations. This means the aim of this paper is to create an efficient, lightweight method that employs self-attention mechanisms to capture spatial relationships, addressing these challenges effectively.

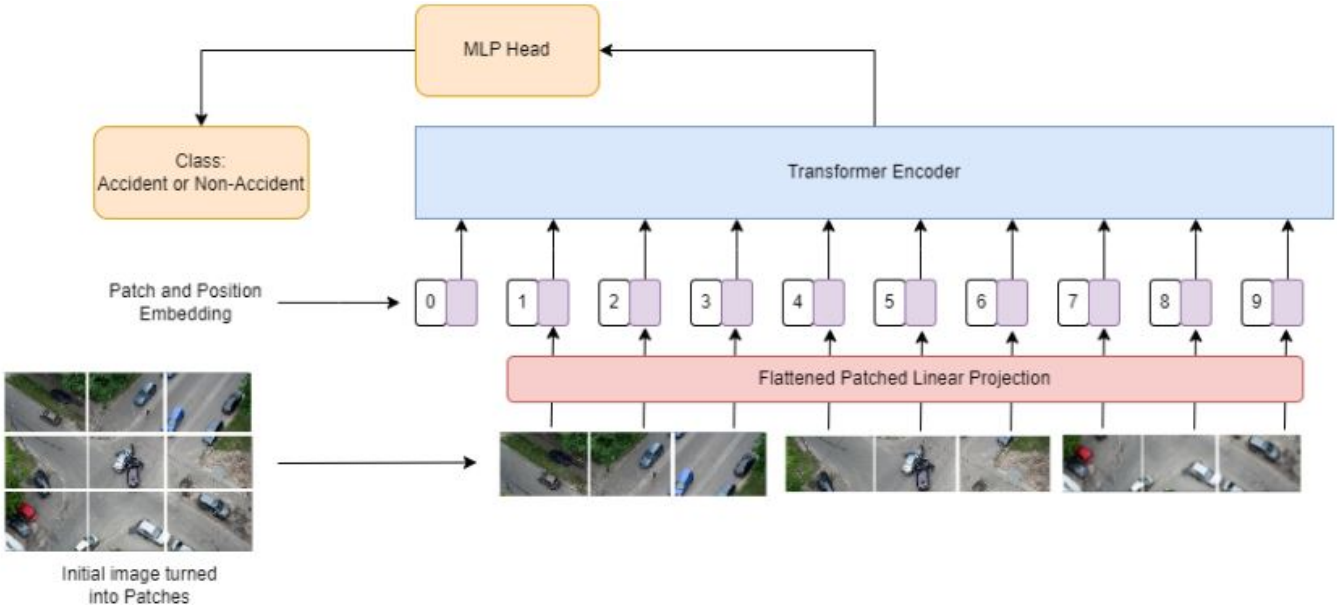

Fig. 2. Architecture Diagram of Proposed Vision Transformer

A. Architecture

The architecture pattern of the Vision Transformer (ViT) works on a novel approach to computer vision tasks [8]. It revolves around the concept of transforming images into sequences of patches, which are then processed using a transformer architecture, originally designed for natural language processing. The proposed model functions in the following way:

1) Initial image is subjected to augmentation to enhance model resilience and diversify the image dataset.
2) These augmented images are divided into segments known as patches or chunks. Each patch represents a section of the original image.
3) The creation of these image patches can be accomplished using Convolutional Layers with suitable kernel sizes and strides. Alternatively, TensorFlow's Image module can also be utilized for direct patch extraction.
4) These image patches undergo processing by a Dense Layer, which learns not only the image data but also captures the positional embedding of each patch.
5) The introduction of positional embedding facilitates the network's comprehension of the arrangement of all patches as a unified image.
6) Subsequently, the transformed data is fed into a Transformer Network for a specified number of iterations. This network encompasses a normalisation layer followed by a Multi-Head Attention layer, facilitating the understanding of inter-patch relationships.
7) The resultant outputs are combined with the original inputs using residual learning, enabling smoother gradient flow during backpropagation.
8) The refined outputs undergo further normalisation, followed by processing through a Multi-Layer Perceptron (MLP). The MLP not only learns class distinctions but also identifies attention-worthy regions at lower levels [15].

Lastly, the outputs from the final Transformer Block are channeled into an MLP, which yields the ultimate class probabilities for the image classification task.

B. Preprocessing

Preprocessing involves the efficient extraction of crucial video frames. Primarily based on scene alterations, the process commences by segmenting a video into smaller shots using the Python library openCV. Subsequently, frames from each distinct shot are extracted. Addressing these challenges can be facilitated through the implementation of keyframes, which encapsulate the video's content in a reduced number of frames. Essential modules for this task encompass openCV, cv2, and numpy.

C. Feature Extraction Architecture

The proposed model's specific architecture involves a sequence of structured layers for efficient image analysis. Starting with input data, normalisation is applied, adapting to training data. Augmented data undergoes further process patch patch extraction. These patches are encoded through a PatchEncoder, enhancing their representations. A Transformer Network then employs multiple blocks, integrating multi-head attention and neural networks for comprehensive feature extraction. After normalisation and dropout, an MLP layer adapts features, leading to the final classification. The output layer generates predictions. This architecture optimises image analysis through data augmentation, patch encoding, and transformer-based context understanding, culminating in enhanced classification outcomes.

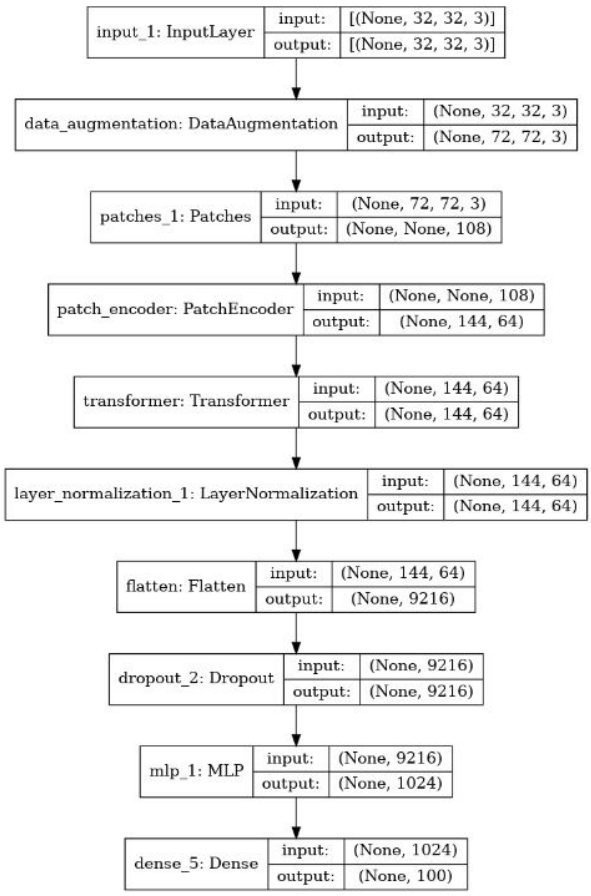

Fig. 3. Proposed TensorFlow Model Summary With Layers

D. Classification

From the dense layer output, a 100-dimension vector is observed, with which the model feeds into a specified loss function known as Sparse Categorical Crossentropy ('SCCe'):

- Sparse Categorical Crossentropy:
  SCCe is a loss function commonly used in machine learning for multi-class classification tasks. Specifically designed for scenarios where classes are mutually exclusive, such as image classification, it calculates the difference between predicted probabilities and true labels. Unlike standard cross-entropy, 'SCCe' optimizes memory efficiency by considering only the index of the true class instead of encoding it as a one-hot vector. This streamlines computations and reduces memory requirements, making it suitable for large datasets with a high number of classes. The function's output measures the dissimilarity between predictions and ground truth, guiding the model's learning process.

Finally, in assessing the model's performance, the evaluation metric employed is Accuracy ('Acc'). It quantifies the proportion of correctly predicted instances among the total number of instances in the dataset.

E. Optimization Technique

In place of the conventional stochastic gradient descent method, AdamW is an optimization technique that is used to iteratively update network weights depending on training data [7].

AdamW, short for Adaptive Moment Estimation with Weight Decay is an optimization algorithm frequently employed in training neural networks. It builds upon the popular Adam optimizer by incorporating weight decay regularization, which counteracts overfitting by penalizing large weight values. AdamW adapts the learning rates of individual model parameters by estimating the first and second moments of the gradients, enhancing convergence and generalization. The model was trained on 100 epochs to get an adequate accuracy measure.

V. Experimental Study and Result Analysis

A. Experimental Environment

The implementation of this proposed model is carried out in a Google Colab Notebook running Intel(R) Xeon(R) CPU @ 2.00GHz and an NVIDIA Tesla P100 GPU with 16GB RAM.

B. Experimental Dataset and Setup

In the case of the accident detection system, the dataset is a collection of frames captured from YouTube videos involving accidents from all around the world [11]. This diverse dataset allows the proposed model to generalize effectively across different accident types and environmental conditions. The Train Test Validation ratio is capped at roughly 80:10:10.

TABLE I
Dataset Information

Dataset Type	Number of Labelled Images
Train	791
Test	100
Validation	98

C. Result Analysis

The model training process requires a duration of 450 seconds to complete 100 epochs on the virtual computing infrastructure. To gain insights into the model's performance characteristics, a random subset of outcomes was extracted from the training process. This subset was then utilized to generate a visual representation, enabling a qualitative assessment of the model's predictive capabilities.

```
Epoch 100/100
8/8 [==============================] - 5s 200ms/step - loss: 0.3070 - Accuracy: 0.8609 - T
op-5-Accuracy: 1.0000 - val_loss: 0.2092 - val_Accuracy: 0.9388 - val_Top-5-Accuracy: 1.00
00
```

Fig. 4. Output of Model Training

Fig. 5. Sample Results of Model on Validation Images

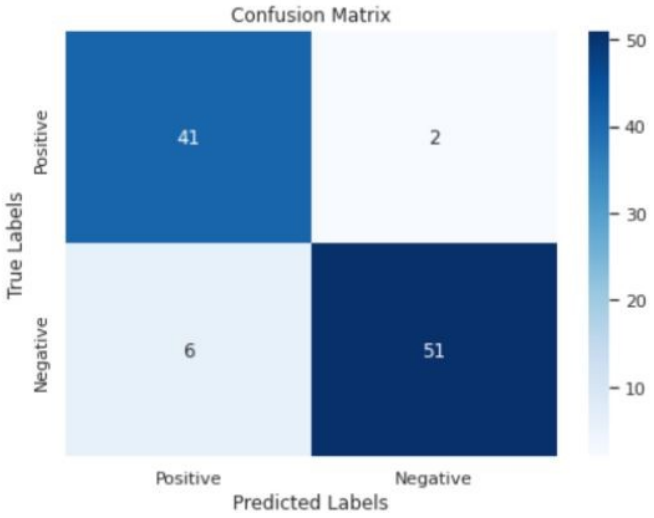

Fig. 6. Confusion Matrix Results

D. Evaluation Performance

Various metrics were used to calculate how the ViT performance is for the purpose of crash detection. These can be seen in the following cases:

1) Accuracy
It is a metric that measures the percentage of correct predictions made by a model, relative to the total predictions it made.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

Where TP (True Positives) are correctly predicted positive instances, TN (True Negatives) are correctly predicted negative instances, FP (False Positives) are falsely predicted as positive, and FN (False Negatives) are falsely predicted as negative by the model.
The proposed model gives an accuracy of 86.09% for the validation dataset.

2) Precision
Precision serves as a numerical representation of the model's ability to correctly identify positive instances among the instances it predicts as positive. It plays a critical role in assessing the model's performance in terms of false positive predictions. A higher precision value indicates that the model is making fewer false positive predictions, resulting in more accurate positive predictions.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

The model gives a precision of 87.23% for the dataset.

3) Recall
It serves as a numerical representation of a model's overall errors, effectively assessing its performance. When there are significant defects, the loss value increases, indicating poor model performance. Conversely, better model performance is reflected by lower loss values.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

The proposed model gives a precision of 95.34% for the dataset.

4) Model Loss
It serves as a numerical representation of the model's overall errors, effectively assessing its performance. When there are significant defects, the loss value increases, indicating poor model performance. Conversely, better model performance is reflected by lower loss values.

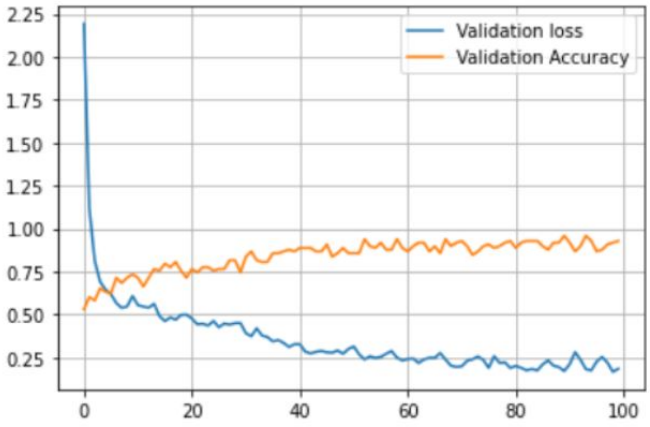

Fig. 7. Model Loss and accuracy values over 100 epochs

VI. Performance Comparison with CNN

It can be observed that as compared to a CNN-based model built on top of MobileNetV4, the accuracy for the Training Set with the same number of epochs (100) for ViT is relatively higher, mainly due to the inability of the CNN architecture to recognize enough features as compared to the vision transformers patching approach. There is a possibility

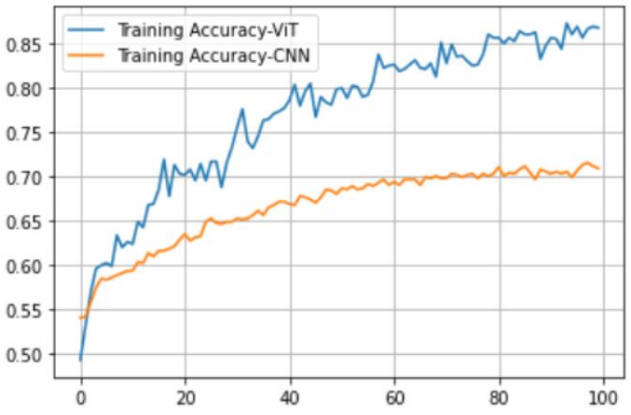

Fig. 8. Accuracy of Proposed ViT to CNN Architecture

for better performance from CNN-based systems if more custom layers are added. But this in turn would result in more computation, increasing processing requirements compared to ViT.

By comparing the obtained results with those of the results obtained from other papers, a comparison table is created, which clearly shows the advantage of ViT over the other architecture patterns.

TABLE II
MODEL ACCURACY COMPARISON

Architecture Approach	Accuracy Percentage
Vision Transformer	86.85%
CNN (MobileNetV4+3 CNN layers)	70.92%
VGG19 Detection Approach	81.00%
Spatio Temporal Approach	77.50%

VII. CONCLUSION AND FUTURE WORKS

In conclusion, this study introduced a pioneering approach to accident detection utilizing vision transformers within CCTV surveillance systems. The architecture's effectiveness in feature extraction and competitive accuracy was evident through its adeptness in real-time accident identification. By amalgamating pre-processing techniques, patch-based encoding, and the power of self-attention mechanisms in transformers, the model exhibited strong potential for practical accident prediction. Looking forward, there are plenty of possibilities for advancement. The model's performance could be further augmented through expanded dataset utilization, transfer learning, and advanced data augmentation methods. Experimentation with diverse transformer architectures and the integration of temporal information between frames might enhance motion-related insights. Collaborations with traffic management systems to incorporate accident alerts and response mechanisms would improve the model's real-world impact. Overall, this research lays a foundation for progress in accident detection and vision transformers, paving the way for safer roadways through state-of-the-art computer vision advancements.

REFERENCES

[1] A. Ajith, A. O. Philip, S. Sreedhar and M. U. Sreeja, "Road Accident Detection from CCTV Footages using Deep Learning," 2022 3rd International Conference on Smart Electronics and Communication (ICOSEC), Trichy, India, 2022, pp. 1354-1361

[2] R. Babu and B. Rajitha, "Accident Detection through CCTV Surveillance," 2022 IEEE Students Conference on Engineering and Systems (SCES), Prayagraj, India, 2022, pp. 01-06

[3] V. S. Sindhu, "Vehicle Identification from Traffic Video Surveillance Using YOLOv4," 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2021, pp. 1768-1775

[4] N. Yadav, U. Thakur, A. Poonia and R. Chandel, "Post-Crash Detection and Traffic Analysis," 2021 8th International Conference on Signal Processing and Integrated Networks (SPIN), Noida, India, 2021, pp. 1092-1097, doi: 10.1109/SPIN52188.2021.9552366.

[5] Ministry of Road Transport & Highways, Government of India, pp.6

[6] World Health Organization Global Report on Road Safety, 20 June 2022

[7] L. Jiang, Z. Zhou, T. Zhang, S. Zhao, and Y. Gao, "On the Convergence of Adam and Beyond," IEEE Transactions on Neural Networks and Learning Systems, vol. 32, no. 5, pp. 2113-2125, May 2021, doi: 10.1109/TNNLS.2020.3047554

[8] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," arXiv:2010.11929 [cs.CV], 2021.

[9] A. Parvaiz, M. A. Khalid, R. Zafar, H. Ameer, M. Ali, M. M. Fraz, "Vision Transformers in medical computer vision—A contemplative retrospection," Engineering Applications of Artificial Intelligence, vol. 122, 2023, p. 106126, ISSN 0952-1976, https://doi.org/10.1016/j.engappai.2023.106126.

[10] P. Josephinshermila, S. Sharon Priya, K. Malarvizhi, R. Hegde, S. Gokul Pran, and B. Veerasamy, "Accident detection using Automotive Smart Black-Box based Monitoring system," Measurement: Sensors, vol. 27, 2023, p. 100721, ISSN 2665-9174, https://doi.org/10.1016/j.measen.2023.100721.

[11] C. Kumar, "Accident Detection From CCTV Footage," Kaggle, 2020. [Online]. Available: https://www.kaggle.com/dsv/1379553. [Accessed: 09/04/2023].

[12] N. Alam, S. Kolawole, S. Sethi, N. Bansali, and K. Nguyen, "Vision Transformers for Mobile Applications: A Short Survey," arXiv preprint arXiv:2305.19365, 2023.

[13] S. Ghosh, S. J. Sunny and R. Roney, "Accident Detection Using Convolutional Neural Networks," 2019 International Conference on Data Science and Communication (IconDSC), Bangalore, India, 2019, pp. 1-6, doi: 10.1109/IconDSC.2019.8816881.

[14] Ribeiro B, Nicolau MJ, Santos A. Using Machine Learning on V2X Communications Data for VRU Collision Prediction. Sensors (Basel). 2023 Jan 22;23(3):1260. doi: 10.3390/s23031260. PMID: 36772299; PMCID: PMC9920954.

[15] M.-C. Popescu, V. Balas, L. Perescu-Popescu, and N. Mastorakis, "Multilayer perceptron and neural networks," WSEAS Transactions on Circuits and Systems, vol. 8, July 2009.