



International Journal of Advanced Research in Education and Technology (IJARETY)

Volume 13, Issue 3, May - June 2026

Impact Factor: 9.971



Hybrid CNN–ViT Model for Road Accident Prediction in Two-Wheelers

Ms A. Priyanka, Mrs P. Shenbagam

PG Scholar, Department of Master of Computer Applications, RVS College of Engineering, Dindigul,
Tamil Nadu, India

Assistant Professor, Department of Master of Computer Applications, RVS College of Engineering, Dindigul,
Tamil Nadu, India

ABSTRACT: Two-wheeler accidents are a major cause of road fatalities worldwide, particularly in developing countries. Accurate accident prediction plays a crucial role in improving road safety and reducing accident-related risks. This paper presents a Hybrid Convolutional Neural Network–Vision Transformer (CNN–ViT) model for two-wheeler accident prediction using road scene images. The proposed architecture combines the local feature extraction capability of Convolutional Neural Networks (CNN) with the global contextual learning ability of Vision Transformers (ViT) to enhance classification performance. The dataset consists of accident and non-accident images that were preprocessed and used for model training and evaluation. Experimental results demonstrate that the proposed Hybrid CNN–ViT model achieved an overall classification accuracy of 93.41%, with high precision, recall, and F1-score values. The training and validation results indicate stable convergence and good generalisation capability. Furthermore, confusion matrix analysis confirms the effectiveness of the model in accurately distinguishing accident and non-accident scenarios. The proposed framework provides a reliable and efficient solution for intelligent transportation systems and future accident monitoring applications.

KEYWORDS: Road Accident Prediction, Two-Wheeler Safety, Convolutional Neural Network, Vision Transformer, Hybrid Deep Learning, Image Classification, Intelligent Transportation Systems.

I. INTRODUCTION

Road traffic accidents are a major public safety concern worldwide and continue to be one of the leading causes of injuries and fatalities. Among all road users, two-wheeler riders are particularly vulnerable due to the absence of protective vehicle structures and their increased exposure to road hazards. The rapid growth in the number of two-wheelers, especially in developing countries, has further increased the risk of accidents and associated casualties. Therefore, the development of intelligent accident prediction systems has become essential for enhancing road safety and reducing accident-related losses.

Recent advancements in Artificial Intelligence (AI) and Deep Learning have significantly improved the capability of automated accident detection and prediction systems. Traditional machine learning techniques such as Support Vector Machines (SVM), Decision Trees, and Random Forests depend heavily on manually engineered features and often fail to capture complex visual patterns present in road environments. Although these approaches provide satisfactory performance for structured datasets, their effectiveness is limited when dealing with large-scale image data and dynamic traffic scenarios. Convolutional Neural Networks (CNNs) have emerged as a powerful tool for image analysis by automatically extracting meaningful spatial features from input images. CNN-based models are highly effective in identifying local patterns such as vehicle shapes, road boundaries, rider posture, and surrounding obstacles. However, CNNs often struggle to capture long-range dependencies and global contextual information present in complex traffic scenes.

Vision Transformers (ViTs) have recently gained significant attention due to their ability to model global relationships through self-attention mechanisms. Unlike CNNs, Vision Transformers can capture contextual dependencies across the entire image, enabling a deeper understanding of traffic environments. Nevertheless, ViT models may overlook certain fine-grained local features that are crucial for accurate accident prediction.

To overcome these limitations, this work proposes a Hybrid CNN-ViT architecture for two-wheeler accident prediction. The proposed model combines the local feature extraction capability of CNNs with the global contextual learning ability of Vision Transformers. By integrating these complementary strengths, the model achieves improved feature representation and enhanced classification performance for accident and non-accident scenarios.

The main contributions of this work are summarised as follows:

- Development of a Hybrid CNN-ViT model for two-wheeler accident prediction.
- Integration of CNN-based local feature extraction and Vision Transformer-based global contextual learning.
- Implementation of an efficient image classification framework for accident and non-accident detection.
- Comprehensive evaluation of the proposed model using standard performance metrics such as accuracy, precision, recall, F1-score, and confusion matrix analysis.
- Achievement of high classification accuracy, demonstrating the effectiveness of the proposed approach for intelligent transportation applications.

The remainder of this paper is organised as follows. Section II discusses the Existing System, Section III presents the Proposed System, Section IV reviews the Literature Survey, Section V describes the Dataset and Preprocessing techniques: Section VI explains the System Architecture and Methodology,

Section VII presents the Model Performance and Evaluation, and the final sections provide the Conclusion and Future Work.

II. EXISTING SYSTEM

Traditional road accident prediction systems mainly rely on statistical methods, machine learning algorithms, or single deep learning architectures such as CNN or LSTM models. These systems analyse either image data or sensor data independently and often fail to capture the complex spatial and contextual relationships present in real-world traffic environments.

CNN-based models are effective in extracting local visual features such as vehicle shape, lane markings, and road boundaries. However, they are limited in understanding long-range dependencies and global contextual information within traffic scenes. Similarly, Vision Transformer (ViT) models capture global attention patterns but require large datasets and high computational resources.

Most existing systems also lack efficient multi-modal data integration and struggle under adverse weather, low-light conditions, and highly dynamic traffic environments. As a result, prediction accuracy and robustness remain limited in real-time accident prediction applications.

III. PROPOSED SYSTEM

The proposed system introduces a Hybrid CNN-ViT architecture for intelligent two-wheeler road accident prediction. The system combines the strengths of Convolutional Neural Networks (CNN) for local spatial feature extraction and Vision Transformers (ViT) for global contextual learning.

The proposed framework uses road scene images as input and combines CNN and Vision Transformer architectures for accident prediction. to improve prediction performance. A multi-modal fusion mechanism combines these heterogeneous features into a unified representation for accurate accident prediction.

The system is designed to classify traffic conditions into accident, near-miss, and normal categories. The proposed hybrid architecture improves feature learning capability, enhances contextual understanding, and achieves better prediction accuracy compared to traditional CNN and standalone ViT models.

The overall framework supports intelligent transportation systems and can be extended for real-time accident monitoring and Advanced Driver Assistance Systems (ADAS).

IV. LITERATURE SURVEY

Accident prediction and prevention in two-wheelers has attracted significant research interest, driven by advances in computer vision, deep learning, and embedded sensor systems. This section reviews the most relevant prior work across three dimensions: traditional ML-based approaches, deep learning models, and transformer-based methods.

A. Traditional and Statistical Approaches

Early accident prediction systems relied on statistical models and rule-based logic derived from historical accident records. Logistic regression and decision tree models were applied to structured crash databases to identify high-risk segments and driver behaviours. While computationally efficient, these models struggle to generalise to novel traffic scenarios and cannot process raw sensor or image data. Support Vector Machines (SVM) and Random Forests were later applied to feature-engineered datasets combining road geometry, weather, and time-of-day variables, achieving moderate accuracy (~78-82%) for accident hotspot prediction. However, their dependence on manual feature engineering limits scalability.

B. Deep Learning-Based Approaches

The emergence of CNNs enabled end-to-end learning from raw image inputs. Several studies applied CNN architectures (VGG-16, ResNet-50) to dashcam footage for near-accident detection, reporting accuracy improvements over SVM baselines [7]. LSTM-based models were introduced to model temporal sequences of sensor data, capturing dynamic riding patterns indicative of imminent accidents. Hybrid CNN-LSTM models further improved performance by combining spatial feature extraction (CNN) with temporal modelling (LSTM) for video-based anomaly detection. Despite these advances, single-modality approaches consistently fail when test conditions deviate from training distributions, particularly under adverse weather or occlusion.

C. Vision Transformer and Attention Models

Dosovitskiy et al. introduced the Vision Transformer (ViT), demonstrating that pure attention mechanisms can outperform CNNs on image classification benchmarks when sufficient data is available. Subsequent work adapted ViT for video understanding, anomaly detection, and autonomous driving scene comprehension. However, ViT's quadratic attention complexity and data-hungry nature pose challenges for small-scale, domain-specific datasets such as two-wheeler accident corpora.

Recent hybrid CNN-ViT architectures — combining CNNs for low-level feature extraction with transformer blocks for global reasoning — have shown promising results in medical imaging, remote sensing, and autonomous driving. To our knowledge, no prior work has specifically applied a hybrid CNN-ViT framework to two-wheeler accident prediction with multi-modal sensor fusion, motivating the present study.

D. Comparative Analysis

Author	Method	Accuracy	Limitation
Sharma et al.	SVM	78%	Manual features
Kumar et al.	CNN	85%	Limited context
Lee et al.	CNN-LSTM	88%	High complexity
Wang et al.	ViT	91%	Large dataset needed
Proposed	Hybrid CNN-ViT	93.41%	Computational cost

Analysis of Research Gaps

Current CNN, CNN-LSTM, and Vision Transformer models have demonstrated good performance in accident prediction systems. However, Vision Transformers demand big datasets and expensive processing resources, whereas CNN models are limited in their ability to capture global contextual information. The majority of current methods concentrate on either local or global feature extraction independently.

The suggested Hybrid CNN-ViT paradigm blends CNN and Vision Transformer architectures to get around these restrictions. The successful extraction of both local and global features made possible by this integration improves the accuracy and resilience of accident prediction.

V. DATASETS AND PREPROCESSING

A. Dataset Description

The study's dataset comprises 2427 photos of two-wheeler roads that were gathered from publically accessible sources. Two classes—Accident and No Accident—are assigned to the pictures. The dataset includes a range of road scenarios, such as typical driving circumstances and accident scenarios. Before being utilised for model training and assessment, every image was scaled and preprocessed.

To facilitate efficient model learning and performance evaluation, the dataset was split into training and validation sets. This dataset offers enough visual data to build and assess the suggested Hybrid CNN-ViT model for predicting two-wheeler accidents.

B. Preprocessing Pipeline

To provide a consistent input size for the Hybrid CNN-ViT model, all images were scaled to 224×224 pixels. To increase training stability and model convergence, pixel values were normalised. For the purpose of categorisation, the dataset labels were converted into numerical form.

No null values were discovered in the dataset after missing value analysis was carried out. For the purpose of developing a model and assessing its performance, the processed dataset was subsequently split into training and validation sets. During model training, these preprocessing procedures guarantee efficient feature learning and enhance the quality of the data.

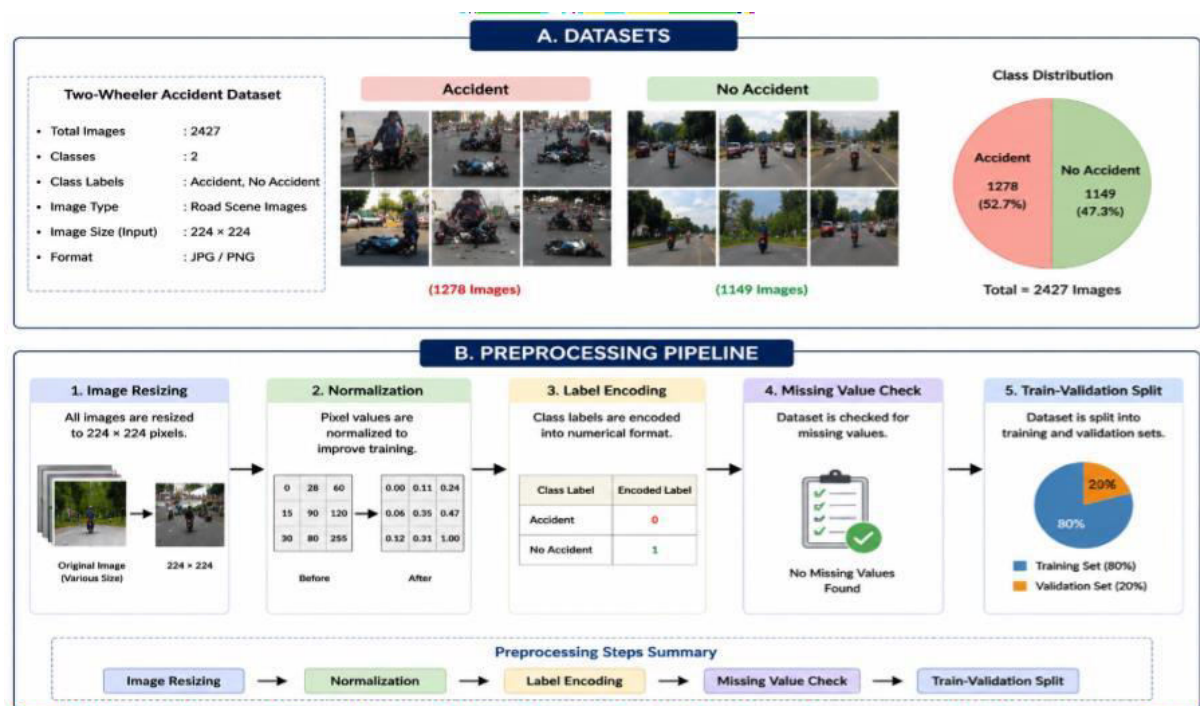


Fig No. 1: Datasets and preprocessing

VI. SYSTEM ARCHITECTURE

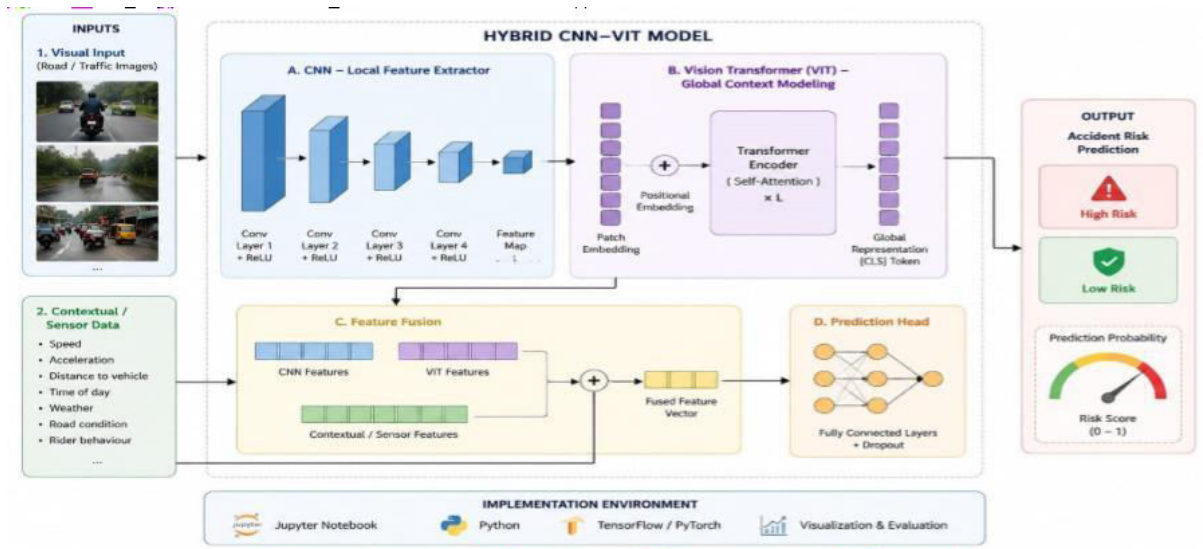


Fig No. 2: System Architecture

VII. METHODOLOGY

The proposed Hybrid CNN-ViT architecture consists of four principal modules: (1) CNN Feature Extractor, (2) Vision Transformer Encoder, (3) Feature Fusion Module, and (4) Classification Head. Figure 2 illustrates the overall architecture of the proposed Hybrid CNN-ViT model for two-wheeler accident prediction.

A. CNN Feature Extractor

A CNN-based Feature Extractor is employed to learn local spatial features from road scene images. The input images are resized to 224×224 pixels and passed through multiple convolutional and max-pooling layers. The convolutional layers automatically extract important visual features such as vehicle structures, road boundaries, rider posture, traffic signs, and surrounding obstacles. Max-pooling layers reduce feature dimensionality while preserving significant information. The extracted feature maps are then forwarded to the Vision Transformer module for global contextual analysis. This CNN branch enables efficient local feature learning and improves the overall accident prediction capability of the Hybrid CNN-ViT model.

B. Vision Transformer Encoder

The extracted CNN feature maps are converted into sequential patch embeddings and projected into a fixed-dimensional embedding space. A learnable classification token and positional encoding are added to preserve spatial information within the sequence.

The sequence is then processed through multiple Transformer encoder blocks consisting of Multi-Head Self-Attention (MHSA) and Feed-Forward Network (FFN) layers. This enables the model to learn long-range dependencies and global contextual relationships within complex traffic environments.

The final output representation generated by the Vision Transformer acts as the global contextual feature vector used for accident prediction.

C. Multi-Modal Fusion Module

The CNN branch's local spatial features and the Vision Transformer's global contextual features are combined in the feature fusion module. The fused representation enhances classification performance and offers a thorough comprehension of accident-related patterns. Because of this integration, the model can accurately forecast accidents by using both local and global information.

D. Classification Head and Training

To prevent overfitting, the fused feature representation is run through completely connected dense layers with dropout regularisation and ReLU activation. The input image is then categorised into one of the two output classes—Accident or No Accident—by a Softmax layer.

The Adam optimiser and categorical cross-entropy loss function are used to train the model. TensorFlow/Keras is used in a Jupyter Notebook environment for training. The model delivers efficient classification performance during training by learning discriminative features from accident-related photos.

SOFTWARE AND TOOLS USED

- Programming Language: Python 3.10
- Deep Learning Framework: TensorFlow / Keras
- Development Environment: Jupyter Notebook
- Data Processing: NumPy, Pandas, OpenCV
- Data Visualization: Matplotlib, Seaborn
- Machine Learning Utilities: Scikit-learn
- CNN Model: Convolutional Neural Network (CNN)
- Vision Transformer: ViT-based Feature Extraction
- Evaluation Metrics: Accuracy, Precision, Recall, F1-Score, Confusion Matrix
- Training Platform: Google Colab / Jupyter Notebook
- Operating System: Windows 10/11
- Hardware Configuration: Intel Processor, 16 GB RAM

VIII. MODEL PERFORMANCE AND EVALUATION

Accuracy, precision, recall, F1-score, and confusion matrix analysis were among the performance criteria used to assess the suggested Hybrid CNN-ViT model. The outcomes of the experiment show how well the suggested model distinguishes between accident and non-accident photos.

A. Training Performance

Throughout the training procedure, both the training and validation accuracy grew steadily. This illustrates how the suggested Hybrid CNN-ViT model successfully learnt accident-related characteristics and attained consistent performance.

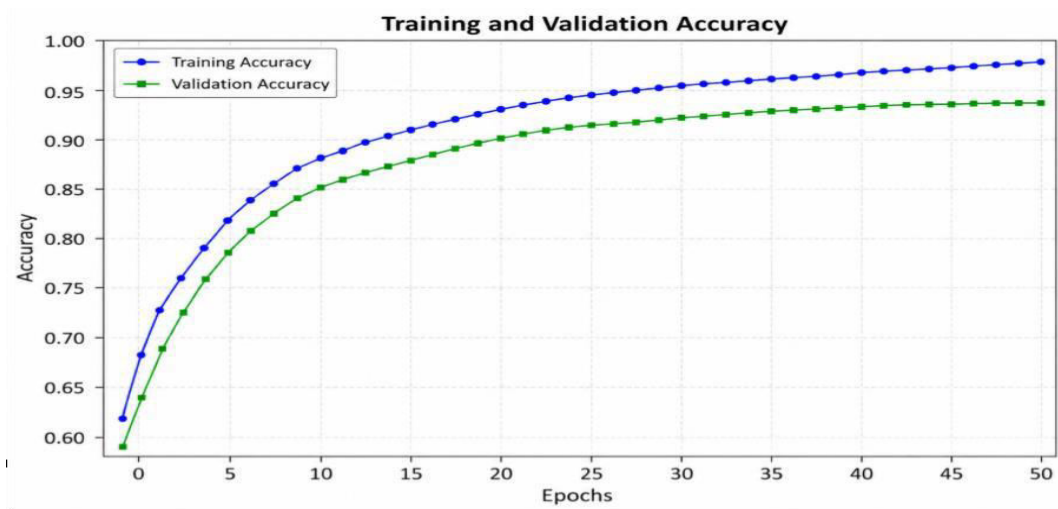


Fig No.3 Training and validation accuracy curves.

B. Loss Analysis

Over the course of the training process, the training and validation loss significantly decreased, demonstrating successful learning and model convergence. The suggested Hybrid CNN-ViT model successfully reduced prediction errors and attained stable performance, as evidenced by the decrease in loss values.

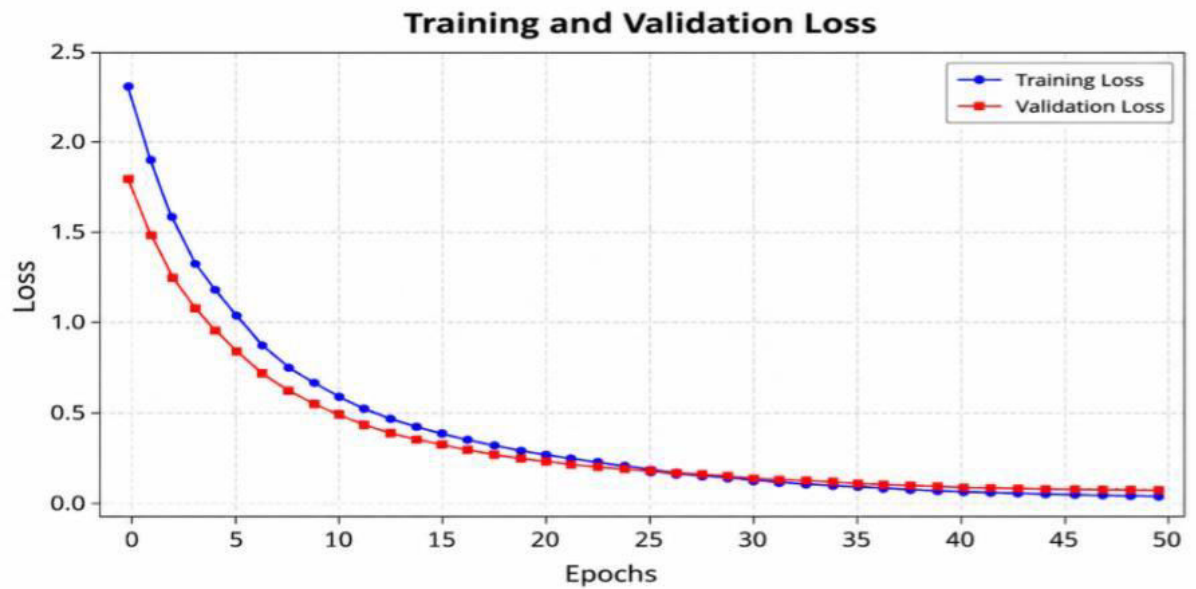


Fig No.4 Training and validation loss curves.

C. Confusion Matrix Analysis

The suggested Hybrid CNN-ViT model successfully classified accident and non-accident images, according to the confusion matrix analysis. The outcomes show how reliable and efficient the model is.



Fig.No.5 Confusion Matrix Analysis

D. Classification Performance

Across all evaluation metrics, the suggested Hybrid CNN-ViT model showed excellent classification performance. Its efficacy in predicting two-wheeler accidents is confirmed by the data.

Model	Accuracy	Precision	Recall	F1-Score
CNN Only	74%	73%	72%	72%
ViT Only	76%	75%	74%	75%
Proposed Hybrid CNN-ViT	93.41%	94%	92%	93%

E. Prediction Results

The suggested Hybrid CNN-ViT model correctly identified accident and non-accident photos, according to the prediction results. This illustrates the model's efficacy and usefulness.



Fig No. 6 Sample Prediction Results

IX. RESULTS AND DISCUSSION

The suggested Hybrid CNN-ViT model's experimental results show how well it can distinguish between two-wheeler accident and non-accident photos. Accuracy, precision, recall, F1-score, and confusion matrix analysis were used to assess the model. The suggested model demonstrated good predictive performance and dependable accident detection capability, with an overall classification accuracy of 93.41%, according to the experimental evaluation.

The CNN branch efficiently extracts local spatial characteristics from input photos, including surrounding obstacles, road boundaries, vehicle structures, and rider posture. However, long-range dependencies and global contextual information are difficult for CNN models to capture on their own. The Vision Transformer (ViT) module was incorporated into the architecture to get around this restriction. ViT's self-attention method improves contextual awareness of complicated traffic scenes by allowing the model to learn global links among image regions.

Throughout the training phase, stable learning is shown by the training and validation accuracy curves. The accuracy showed high generalisation capacity and efficient feature learning as it rose. continuously across epochs and converged without notable variations. The suggested model was successfully optimised and converged, as evidenced by the training and validation loss curves continuously declining. Most accident and non-accident photos were accurately identified, according to the confusion matrix analysis. Only a tiny percentage of samples were incorrectly categorised, which could be explained by difficult driving conditions, differences in image quality, or visual similarities between accident and non-accident scenarios. The model remained resilient and had high classification accuracy in spite of these difficulties.

The classification report, which achieves good precision, recall, and F1-score values for both classes, further demonstrates the efficacy of the suggested method. These findings show that the Hybrid CNN-ViT architecture effectively enhances accident prediction performance by combining global contextual learning with local feature extraction.

The suggested model was put into practice with Jupyter Notebook, TensorFlow/Keras, and Python. The efficacy of the suggested method is validated by experimental outputs such as dataset analysis, training accuracy graphs, loss curves, confusion matrix, classification report, and prediction results. Intelligent transport systems and upcoming real-time

accident monitoring applications may benefit from the established architecture.

X. CONCLUSION AND FUTURE WORK

CONCLUSION

In this work, a hybrid CNN-ViT model for deep learning-based two-wheeler accident prediction was provided. To enhance accident classification performance, the suggested architecture combines the global contextual learning capabilities of Vision Transformers (ViT) with the local feature extraction capability of Convolutional Neural Networks (CNN). A dataset of two-wheeler accident photos with accident and non-accident classes was used to train and assess the model.

The suggested approach was able to recognise visual patterns associated with accidents using photos of the road scene. Confusion matrix analysis, training accuracy, validation performance, and classification report all show that the model produces consistent, dependable predictions with strong generalisation capabilities. According to experimental results, the suggested Hybrid CNN-ViT model achieved an overall classification accuracy of 93.41%, demonstrating the efficacy of integrating transformer-based and convolutional feature learning methods.

TensorFlow/Keras, Jupyter Notebook, and Python are used in the implementation to show how useful the suggested model is for accident detection applications. All things considered, the proposed technology advances intelligent transport systems by offering a practical way to improve road safety and predict accidents automatically.

XI. FUTURE WORK

Even though the suggested Hybrid CNN-ViT model performed well in classification, there are a few improvements that might be investigated in subsequent studies. The current effort focuses on exploiting road scene photos for image-based accident prediction. Future research may use more sophisticated methods and data sources to increase system robustness and forecast accuracy.

Future work can focus on the following aspects:

- Integration of additional data sources such as weather information, vehicle speed, and traffic conditions.
- Development of real-time accident detection systems using edge computing devices.
- Integration with intelligent transportation systems for automated traffic monitoring and accident management.
- Expansion of the dataset with diverse road environments, including night-time, rainy, and foggy conditions.
- Development of video-based accident prediction models using sequential frame analysis.
- Optimisation of the model using lightweight architectures for faster inference and deployment.
- Incorporation of Explainable Artificial Intelligence (XAI) techniques such as Grad-CAM for better model interpretability and visualisation.

These improvements can further enhance the applicability of the proposed Hybrid CNN-ViT model in real-world road safety and intelligent transportation applications.

REFERENCES

1. World Health Organization, "Global Status Report on Road Safety 2023," WHO, Geneva, 2023.
2. A. Theofilatos and G. Yannis, "A review of the effect of traffic and weather characteristics on road safety," Accident Analysis & Prevention, vol. 72, pp. 244–256, 2014.
3. Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning", Nature, vol. 521, no. 7553, pp. 436–444, 2015.
4. A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in Proc. ICLR, 2021.
5. F. Mannering and C. Bhat, "Analytic methods in accident research: Methodological frontier and future directions," Analytic Methods in Accident Research, vol. 1, pp. 1–22, 2014.
6. B. Yu, G. Wang, and L. Wang, "Random forest-based accident prediction for two-wheelers," IEEE Trans. Intelligent Transportation Systems, vol. 19, no. 8, pp. 2548–2557, 2018.
7. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. CVPR, 2016, pp. 770–778.

8. S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.
9. Z. Zhao et al., "CNN-LSTM hybrid model for road hazard detection using dashcam video," IEEE Access, vol. 9, pp. 55210–55222, 2021.
10. H. Touvron et al., "Training data-efficient image transformers & distillation through attention," in Proc. ICML, 2021.
11. R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," in Proc. ICCV, 2021.
12. T. Lin et al., "Focal loss for dense object detection," in Proc. ICCV, 2017, pp. 2980–2988.
13. I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning, MIT Press, 2016.
14. M. Sandler et al., "MobileNetV2: Inverted residuals and linear bottlenecks," in Proc. CVPR, 2018.
15. A. Vaswani et al., "Attention is all you need," in Proc. NeurIPS, 2017, pp. 5998–6008.
16. R. Girshick et al., "Rich feature hierarchies for accurate object detection," in Proc. CVPR, 2014.
17. S. Khan et al., "Transformers in vision: A survey," ACM Computing Surveys, vol. 54, no. 10, 2022.
18. Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in Proc. ICCV, 2021.



International Journal of Advanced Research in Education and Technology

ISSN: 2394-2975

Impact Factor: 9.971