A Hybrid CNN-Vision Transformer Framework with Behavioural Feature Fusion for Driver Drowsiness Detection

Pallavi T[1],Lavanya G[2],Monika S[3],Gumma Mythri[4],Radhika R[5],Dr Krishna Kumar P R[6]
[1,2,3,4,5,6]*Department of CSE,S.E.A College of Engineering and Technology, Bengaluru,karnataka*

***Abstract*—Driver drowsiness is a major road safety concern that can contribute to traffic accidents. This paper presents HCAF-ViT-M,a hybrid deep learning framework designed to detect driver drowsiness using facial images and behavioural features. The proposed framework integrates a ResNet18-based Convolutional Neural Network (CNN), a Vision Transformer (ViT), and a behavioural feature extraction branch. The CNN captures local visual patterns, while the Vision Transformer learns global facial representations. Behavioural features provide additional information about eye and mouth movements and head position. These complementary representations are fused to classify the driver's state into two categories: Active and Fatigue. The framework is developed using the UTA-RLDD dataset and integrated into a real-time monitoring application. The system aims to support continuous driver monitoring and provide fatigue alerts.**

***Index Terms*—Driver drowsiness detection, Convolutional Neural Network, Vision Transformer, Feature Fusion, Facial Behaviour Analysis, Deep Learning, Road Safety.**

I. INTRODUCTION

Driving requires continuous attention, visual awareness, and quick decision-making. Driver drowsiness can reduce concentration, slow reaction time, and affect the ability to respond to unexpected road situations. It is therefore an important factor in road safety. Developing an effective driver monitoring system can help identify signs of fatigue and provide timely warnings to drivers.

Traditional driver drowsiness detection approaches commonly rely on physical indicators such as eye closure, blinking frequency, yawning, and head movement. Facial feature-based techniques can estimate these indicators using a camera without requiring the driver to wear additional sensors. However, changes in illumination, head orientation, facial appearance, and individual behaviour can make reliable detection challenging.

Deep learning techniques have provided new opportunities for analysing facial images. Convolutional Neural Networks (CNNs) are effective at extracting local visual patterns, while Vision Transformers (ViTs) use self-attention mechanisms to learn relationships between different image regions. Combining these approaches can provide complementary visual representations. In addition, behavioural features such as the Eye Aspect Ratio (EAR), Mouth Aspect Ratio (MAR), and head position can provide useful information about driver alertness.

This paper presents HCAF-ViT-M, a hybrid CNN–Vision Transformer framework with behavioural feature fusion for driver drowsiness detection. The proposed architecture combines three feature branches: a ResNet18-based CNN branch, a Vision Transformer branch, and a behavioural feature branch. Their extracted representations are integrated through a fusion network to classify the driver's condition into Active and Fatigue categories.

The framework is developed and evaluated using the UTA-RLDD image dataset. A real-time monitoring application is also implemented to process facial input through a webcam and generate fatigue-related warnings. The main objective is to investigate the integration of visual and behavioural information in a practical driver monitoring prototype.

The main contributions of this work are as follows:

1. Development of a hybrid architecture combining CNN and Vision Transformer representations.
2. Integration of facial behavioural features with deep visual features through a fusion network.

3. Implementation of a real-time webcam-based driver monitoring application.
4. Evaluation of the framework for binary classification of Active and Fatigue states.

II. LITERATURE REVIEW

Drowsiness detection has been investigated using facial behaviour, temporal analysis, and deep learning techniques. Existing approaches commonly examine indicators such as eye closure, blinking patterns, yawning, and head movement to identify possible signs of fatigue.

Ghoddoosian et al. [1] introduced the UTA Real-Life Drowsiness Dataset (UTA-RLDD), which contains video recordings from 60 subjects and includes alert, low-vigilant, and drowsy conditions. They also presented a temporal model based on blink features for early drowsiness detection. This work provides an important dataset and research foundation for developing driver monitoring systems.

Convolutional Neural Networks have been widely adopted for visual recognition because they learn spatial features from image data. He et al. [2] introduced the Residual Network (ResNet), which uses residual connections to facilitate the training of deeper neural networks. ResNet architectures are useful for extracting visual representations and can serve as feature extraction branches in image classification systems.

More recently, Transformer architectures have been applied to computer vision. Dosovitskiy et al. [3] introduced the Vision Transformer (ViT), which divides an image into patches and uses self-attention to learn relationships between them. Unlike conventional convolutional operations that primarily focus on local patterns, the attention mechanism enables the model to capture relationships across different image regions.

These studies provide complementary foundations for the proposed framework. CNN-based feature extraction, transformer-based image representation, and facial behavioural measurements address different aspects of visual and behavioural analysis. HCAF-ViT-M combines these three sources of information through a feature fusion architecture for binary driver-state classification. The implementation focuses on integrating these representations into a real-time monitoring prototype.

Table I: Comparative Analysis of Existing Driver Drowsiness Detection Approaches

Study / Year	Methodology	Reported performance	Limitations
Ji et al. (2004)	Probabilistic visual monitoring	Real-time detection	Sensitive to lighting.
Ghoddoosian et al. (2019	Temporal deep learning	Baseline model	Limited dataset diversity
Xu et al. (2025)	Ensemble learning	86.52% accuracy	Complex feature extraction
ViT study (2024)	Vision Transformer	99.4% accuracy	Computational requirements
Proposed HCAF-ViT-M (2026)	CNN + ViT + behavioural fusion	93.62% accuracy	Requires real-world validation

B. Research Gaps and Motivation

The existing studies reveal several limitations in driver drowsiness detection. Traditional approaches rely mainly on handcrafted facial features, while CNN-based methods may not fully capture global facial relationships. Vision Transformer approaches provide global representations but may overlook important behavioural information. Additionally, variations in lighting, camera position, and facial characteristics can affect detection reliability.

To address these gaps, the proposed HCAF-ViT-M framework combines ResNet18, Vision Transformer, and facial behavioural features through a unified fusion architecture. This integration aims to utilize complementary visual and behavioural information for improved driver drowsiness classification.

III. PROPOSED METHODOLOGY

A. Overview of the Proposed Framework

The proposed HCAF-ViT-M framework is a hybrid deep learning architecture designed to identify driver fatigue by combining facial image representations and behavioural information. The framework consists of three parallel feature extraction branches: a Convolutional Neural Network (CNN), a Vision Transformer (ViT), and a behavioural feature

extraction network. The outputs of these branches are combined using a feature fusion mechanism to perform binary classification.

The input facial image is resized to 224 × 224 pixels and processed by the visual feature extraction branches. In parallel, facial behavioural measurements are extracted and provided to the behavioural branch. The resulting feature representations are concatenated and passed through a fusion network to predict the driver's state as Active or Fatigue.

B. CNN-Based Feature Extraction

The CNN branch uses a ResNet18 architecture to extract local visual features from facial images. ResNet18 employs residual connections that allow information to pass through the network more effectively during training. This branch learns facial patterns that may be associated with driver fatigue.

The output of the CNN branch is transformed into a 256-dimensional feature representation. These features provide local visual information for the subsequent fusion stage.

C. Vision Transformer-Based Feature Extraction

The second branch uses a Vision Transformer to learn global relationships within the facial image. The input image is divided into smaller patches, which are represented as tokens and processed through transformer layers using self-attention.

The Vision Transformer captures relationships between different image regions, complementing the local visual information extracted by the CNN branch. The resulting representation is projected into a 256-dimensional feature vector.

D. Behavioural Feature Extraction

The behavioural branch processes four facial measurements obtained through facial landmark analysis. These measurements are:

1. Eye Aspect Ratio (EAR): Represents the geometric relationship between the eye landmarks and provides information about eye opening.
2. Mouth Aspect Ratio (MAR): Represents mouth geometry and can help identify mouth-opening patterns associated with yawning.
3. Horizontal Head Position: Represents horizontal facial or head displacement.
4. Vertical Head Position: Represents vertical facial or head displacement.

These measurements are extracted using facial landmark information and standardized before being supplied to a Multi-Layer Perceptron (MLP). The behavioural network generates a 128-dimensional feature representation.

E. Feature Fusion and Classification

The outputs of the three branches are combined through feature concatenation. The CNN branch contributes 256 features, the Vision Transformer contributes 256 features, and the behavioural branch contributes 128 features. Therefore, the combined representation contains 640 features.

The fused representation is processed by the fusion network and classification layer to generate predictions for two classes: Active and Fatigue. This integration allows the model to use complementary visual and behavioural information during classification.

F. Real-Time Monitoring Application

The trained model is integrated into a webcam-based monitoring application. Facial input is captured through the camera, and facial landmarks are used to obtain the required behavioural measurements. The processed image and behavioural features are passed to the trained model to estimate the probability of fatigue.

When the fatigue probability satisfies the configured warning conditions for consecutive frames, the application generates a fatigue warning. The application is developed using Streamlit to provide an interactive interface for real-time monitoring.

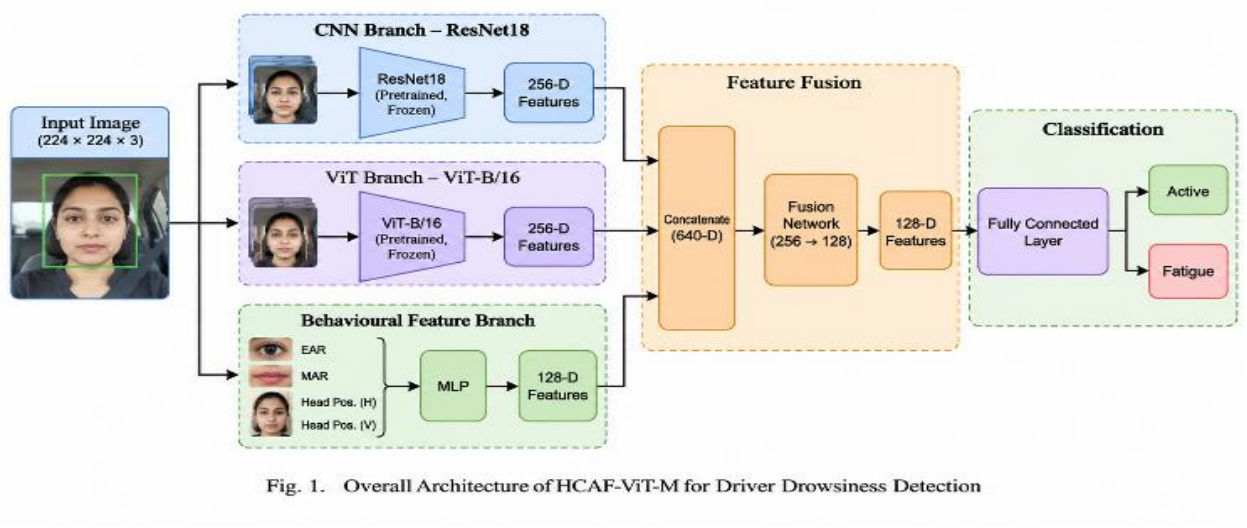


Fig. 1. Overall Architecture of HCAF-ViT-M for Driver Drowsiness Detection

Fig 1: Overall Architecture of HCAF-ViT-M for Driver Drowsiness Detection.

IV. EXPERIMENTAL SETUP

A. Dataset Description

The experiments were conducted using the UTA-RLDD image dataset organized into training, validation, and testing subsets. The dataset contains 11,787 images belonging to two classes: Active and Fatigue. The class labels were encoded as Active (0) and Fatigue (1).

The training subset contains 9,054 images, comprising 5,862 Active and 3,192 Fatigue samples. The validation subset contains 1,824 images, with 912 samples from each class. The testing subset contains 909 images, comprising 456 Active and 453 Fatigue samples. These predefined subsets were used for model development and final evaluation.

B. Behavioural Feature Extraction

Facial behavioural features were extracted from the dataset images using MediaPipe facial landmark detection. Four measurements were selected as inputs to the behavioural branch: Eye Aspect Ratio (EAR), Mouth Aspect Ratio (MAR), horizontal head position, and vertical head position.

These measurements provide complementary information about eye state, mouth movement, and head position. The extracted feature records were merged with the image manifest using image paths. Feature normalization was performed using the mean and standard deviation calculated from the training subset only. The same normalization parameters were then applied to the validation and testing subsets.

C. Proposed Model Configuration

The proposed HCAF-ViT-M framework consists of three feature extraction branches: a ResNet18-based Convolutional Neural Network (CNN), a Vision Transformer (ViT-B/16), and a behavioural feature branch.

The CNN and ViT branches generate 256-dimensional feature representations each. The behavioural branch processes the four input measurements through a multilayer perceptron and produces a 128-dimensional representation. These features are concatenated into a 640-dimensional vector and passed through the fusion network, which maps the combined representation to a 128-dimensional feature vector. A fully connected classifier produces the final two-class prediction.

The ResNet18 and ViT backbone parameters were frozen during training, while the projection, behavioural, fusion, and classification layers were optimized. Input images were resized to 224 × 224 pixels and normalized using ImageNet mean and standard deviation values.

D. Training Configuration

The model was trained using the AdamW optimizer with a learning rate of 0.0001 and a weight decay of 0.0001. A batch size of 16 was used, with a maximum training duration of 15 epochs and an early-stopping patience of five epochs. A weighted cross-entropy loss function was employed to address class imbalance. A ReduceLROnPlateau learning-rate scheduler and automatic mixed precision were also used during training.

The experiments were executed in Google Colab using an NVIDIA Tesla T4 GPU. The best model checkpoint was selected based on validation loss and subsequently used for testing.

E. Evaluation Metrics

The performance of HCAF-ViT-M was evaluated on the held-out testing subset using accuracy, balanced accuracy, precision, recall, F1-score, and the Area Under the Receiver Operating Characteristic Curve (ROC-AUC). The Matthews Correlation Coefficient (MCC) and confusion matrix were also used to provide additional insight into classification performance and class-wise errors.

V. RESULTS AND DISCUSSION

A. Overall Classification Performance

The proposed HCAF-ViT-M framework was evaluated on the held-out test subset containing 909 images. The final checkpoint, selected at epoch 15, achieved an accuracy of 93.62% and a balanced accuracy of 93.61%. The model also obtained a macro F1-score of 93.62% and a ROC-AUC of 0.9811, indicating strong classification performance on the evaluated dataset.

The Matthews Correlation Coefficient (MCC) was 0.8731, reflecting a strong correlation between the predicted and actual class labels. The overall test loss was 0.3100.

Table II. Overall Test Performance of HCAF-ViT-M

Evaluation Metric	Result
Test samples	909
Test loss	0.3100
Accuracy	93.62%
Balanced accuracy	93.61%
Macro F1-score	93.62%
Weighted F1-score	93.62%
ROC-AUC	0.9811
MCC	0.8731

B. Class-Wise Performance Analysis

The model achieved an Active-class precision of 91.98% and a specificity of 95.61%. For the Fatigue class, the precision was 95.40%, while the recall (sensitivity) was 91.61%.

These results indicate that the model identified most Fatigue samples correctly while also maintaining a high rate of correct classification for Active samples. The difference between precision and recall across the two classes suggests that some Fatigue samples were classified as Active, which remains an important consideration for driver drowsiness detection.

C. Confusion Matrix Analysis

The confusion matrix provides further insight into the classification results. Out of 456 actual Active samples, 436 were correctly classified as Active, while 20 were incorrectly classified as Fatigue. Among the 453 actual Fatigue samples, 415 were correctly identified as Fatigue, while 38 were incorrectly classified as Active.

Thus, the model correctly classified 851 out of 909 test samples. The remaining 58 samples were misclassified. In a driver-monitoring application, the 38 Fatigue samples classified as Active are particularly important because such errors may result in missed fatigue warnings.

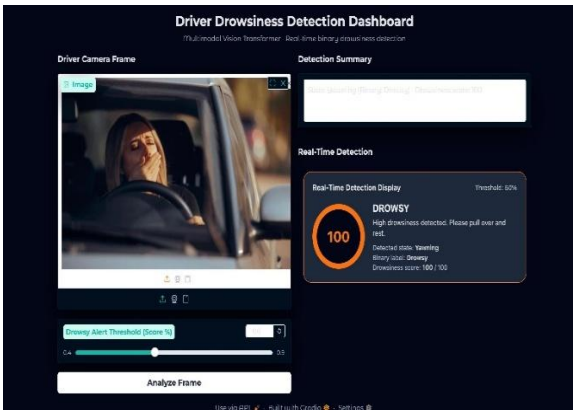


Fig 2: Driver Drowsiness Detection Dashboard

D. Discussion

The experimental results demonstrate the effectiveness of combining visual representations from the CNN and Vision Transformer branches with facial behavioural measurements. The ResNet18 branch extracts local visual patterns, while the ViT-B/16 branch captures broader relationships within the input image. The behavioural branch provides additional information through Eye Aspect Ratio (EAR), Mouth Aspect Ratio (MAR), and horizontal and vertical head-position measurements.

The fusion network combines these complementary representations before classification. The obtained test performance indicates that the integrated framework can distinguish Active and Fatigue states effectively on the evaluated dataset.

However, the results are based on the available dataset and its predefined test split. Further evaluation using additional subjects, varied lighting conditions, different camera positions, and real-world driving scenarios is necessary to assess the model's generalization and practical reliability.

VI. CONCLUSION AND FUTURE WORK

This paper presented HCAF-ViT-M, a hybrid CNN–Vision Transformer framework with facial behavioural feature fusion for driver drowsiness detection. The proposed architecture combines ResNet18-based visual feature extraction, a ViT-B/16 branch, and a behavioural feature branch incorporating Eye Aspect Ratio (EAR), Mouth Aspect Ratio (MAR), and horizontal and vertical head position measurements.

The extracted representations are combined through a feature-fusion network to classify driver states into Active and Fatigue categories. Experimental evaluation on the held-out test subset of 909 images achieved an accuracy of 93.62%, a macro F1-score of 93.62%, and a ROC-AUC of 98.11%. These results indicate the potential of combining deep visual representations with facial behavioural information for drowsiness-related classification.

Despite the promising results, further improvements are required before practical deployment. Future work will focus on evaluating the framework using continuous video sequences, incorporating temporal information such as blink duration and PERCLOS, and testing its performance under different illumination conditions, camera positions, and driver characteristics. Additional evaluation using driver-independent data splits and real-world driving scenarios will also be necessary to assess generalization.

In addition, future development will explore model optimization for real-time inference and deployment on resource-constrained devices. These improvements may help develop a more robust and practical driver-monitoring system for supporting road safety.

REFERENCES

[1] R. Ghoddoosian, M. Galib, and V. Athitsos, "A realistic dataset and baseline temporal model for early drowsiness detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2019, pp. 178–187.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770–778.

[3] Dosovitskiy *et al*., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021.

[4] T. Soukupová and J. Čech, "Real-time eye blink detection using facial landmarks," in *Proc. 21st Comput. Vis. Winter Workshop*, 2016.

[5] C. Lugaresi *et al*., "MediaPipe: A framework for building perception pipelines," arXiv:1906.08172, 2019.

[6] S. Abtahi, M. Omidyeganeh, S. Shirmohammadi, and B. Hariri, "YawDD: A yawning detection dataset," in *Proc. ACM Multimedia Systems Conf. (MMSys)*, 2014, pp. 24–28.

[7] C. Phan, N. H. Q. Nguyen, T. N. Trieu, and T. C. Phan, "An efficient approach for detecting driver drowsiness based on deep learning," *Applied Sciences*, vol. 11, no. 18, Art. no. 8441, 2021.

[8] Xu, W. Huang, J. Liu, and L. Li, "Detecting driver drowsiness using hybrid facial features and ensemble learning," *Information*, vol. 16, no. 4, Art. no. 294, 2025.

[9] R. B. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[10] Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, vol. 25, 2012.

[11] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015.

[12] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 4700–4708.

[13] G. Howard *et al*., "MobileNets: Efficient convolutional neural networks for mobile vision applications," arXiv:1704.04861, 2017.

[14] Vaswani *et al*., "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[15] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015.

[16] Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2019.

[17] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2015, pp. 448–456.

[18] L. N. Bergasa, J. Nuevo, M. A. Sotelo, R. Barea, and M. E. Lopez, "Real-time system for monitoring driver vigilance," *IEEE Trans. Intell. Transp. Syst.*, vol. 7, no. 1, pp. 63–77, Mar. 2006.

[19] Q. Ji, Z. Zhu, and P. Lan, "Real-time nonintrusive monitoring and prediction of driver fatigue," *IEEE Trans. Veh. Technol.*, vol. 53, no. 4, pp. 1052–1068, Jul. 2004.

[20] T. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *J. Big Data*, vol. 6, Art. no. 60, 2019.