IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 10, April 2026
Impact Factor: 8.2

An Explainable Real-Time Driver Drowsiness Detection System Using Vision Transformers and Behavioral Fatigue Indicators

Pisolla Rahul[1], Mohammed Afridi[2], Guguloth Venkatesh[3], Dr S Dheeraj[4]
[1 2 3] Department of Computer Science and Engineering
[4] Associate Professor, Department of Computer Science and Engineering
Sreenidhi Institute of Science and Technology, Hyderabad, India.
Correspondence: 23315a0507@cse.sreenidhi.edu.in

Abstract: *Driver drowsiness is a significant cause of road accidents, particularly during prolonged and monotonous driving. Early and reliable detection of fatigue is therefore essential for improving road safety. This paper presents an explainable real-time driver drowsiness detection system that integrates Vision Transformers (ViT) with behavioral fatigue indicators derived from facial analysis. The proposed approach is non-intrusive and operates using a standard real-time camera feed, making it suitable for practical in-vehicle deployment.*

Facial landmarks are extracted using MediaPipe FaceMesh from live video streams to compute interpretable indicators such as Eye Aspect Ratio (EAR), Mouth Aspect Ratio (MAR), blink frequency, and PERCLOS, which capture fatigue-related behavioral patterns. In parallel, facial image regions captured by the camera are analyzed using a Vision Transformer to perform deep learning–based alertness classification. A hybrid decision mechanism fuses heuristic indicators with transformer outputs and applies temporal smoothing to reduce false alarms. Real-time audio and visual alerts are triggered only under sustained drowsiness conditions.

Experimental results on multiple public datasets and real-time camera-based evaluations demonstrate high detection accuracy with low latency. The proposed system bridges the gap between black-box deep learning models and practical, explainable driver safety solutions for intelligent transportation systems..

Keywords: *Driver Drowsiness Detection, Vision Transformer, Behavioral Fatigue Indicators, PERCLOS, Real-Time Systems, Explainable AI, Intelligent Transportation Systems, ADAS*

I. INTRODUCTION

Road traffic accidents remain a major global safety concern, with driver drowsiness identified as one of the leading contributing factors, especially during long-distance and night-time driving. Fatigue reduces a driver's reaction time, situational awareness, and decision-making ability, significantly increasing the risk of severe accidents. As modern vehicles increasingly incorporate Advanced Driver Assistance Systems (ADAS), there is a growing need for reliable, real-time driver monitoring solutions that can detect drowsiness at an early stage.

Traditional driver drowsiness detection methods include vehicle-based measures (such as steering wheel movement and lane deviation) and physiological signals (such as EEG and ECG). While these approaches can provide useful information, they often require specialized sensors, intrusive hardware, or vehicle-specific calibration, limiting their practicality in real-world deployment. In contrast, vision-based methods using cameras offer a non-intrusive, cost-effective, and widely deployable alternative.

Recent advances in computer vision and deep learning have significantly improved the performance of vision-based drowsiness detection systems. Convolutional Neural Networks (CNNs) have been widely used to classify driver

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 10, April 2026 Impact Factor: 8.2

alertness from facial images; however, these models often function as black boxes, providing limited interpretability. This lack of explainability poses challenges for safety-critical applications such as intelligent transportation systems, where understanding the reasoning behind system decisions is essential for trust, validation, and regulatory acceptance. Vision Transformers (ViT) have emerged as a powerful alternative to CNNs by modeling long-range spatial dependencies through self-attention mechanisms. ViT-based models have demonstrated strong performance in image classification tasks, including facial analysis, while offering greater flexibility in feature representation. However, transformer-based approaches alone may still lack transparency when deployed as standalone classifiers for driver monitoring.

To address these limitations, this paper proposes an explainable real-time driver drowsiness detection system that combines Vision Transformers with interpretable behavioral fatigue indicators derived from facial analysis. Using a real-time camera feed, facial landmarks are extracted to compute indicators such as Eye Aspect Ratio (EAR), Mouth Aspect Ratio (MAR), blink frequency, and PERCLOS, which are well-established measures of driver fatigue. In parallel, facial image regions are processed by a Vision Transformer to perform deep learning–based alertness classification.

A hybrid decision mechanism integrates the outputs of heuristic fatigue indicators and the Vision Transformer, incorporating temporal smoothing to ensure robustness and reduce false alarms. Audio and visual alerts are generated only when sustained drowsiness is detected, improving usability and minimizing unnecessary distractions. Experimental evaluation on multiple public datasets and real-time camera-based testing demonstrates that the proposed system achieves high accuracy with low latency while maintaining explainability.

The key contributions of this work are as follows:

Development of a non-intrusive, camera-based real-time driver drowsiness detection system.Integration of Vision Transformers with interpretable behavioral fatigue indicators for explainable decision-making.

A hybrid fusion and temporal smoothing strategy to enhance reliability and reduce false positives.

Validation on public datasets and real-time scenarios, demonstrating practical feasibility for ADAS and intelligent transportation systems.2. Literature Review

Previous methods include yawning detection, eye tracking, and CNN-based models. However, they lack explainability or real-time reliability.

II. PROPOSED SYSTEM

The proposed system presents a hybrid, explainable framework for real-time driver drowsiness detection by combining facial landmark-based fatigue indicators with deep learning–based classification using Vision Transformers (ViT). Unlike traditional approaches that rely solely on either handcrafted features or black-box models, this system integrates both to improve reliability and interpretability.

The system operates on real-time video input captured through a standard camera. Facial landmarks are extracted using MediaPipe FaceMesh, enabling the computation of fatigue-related metrics such as Eye Aspect Ratio (EAR), Mouth Aspect Ratio (MAR), blink frequency, and PERCLOS. These indicators provide human-understandable insights into driver behavior.

Simultaneously, facial image regions are processed using a Vision Transformer, which captures global spatial dependencies through self-attention mechanisms. A hybrid decision fusion strategy combines the outputs of both approaches and applies temporal smoothing to ensure stable predictions. Alerts are generated only when sustained drowsiness is detected, minimizing false alarms.

2.1 System Overview

The system operates on a continuous video stream captured from a camera positioned in front of the driver. Each frame of the video is processed in real time to monitor facial expressions and detect early signs of fatigue.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 10, April 2026 Impact Factor: 8.2

The architecture is designed to be modular, allowing independent functioning of each component. The major stages include:

- Video acquisition
- Face detection and landmark extraction
- Feature computation
- Deep learning-based classification
- Decision fusion
- Alert generation

This modular design improves scalability and allows easy integration with real-world systems such as Advanced Driver Assistance Systems (ADAS).

2.2 Real-Time Video Acquisition

The first stage involves capturing live video using a standard RGB camera. The system processes the incoming stream frame-by-frame to ensure continuous monitoring without interruptions.

Special care is taken to maintain real-time performance by optimizing frame processing speed. The use of commonly available cameras ensures that the system remains cost-effective and easily deployable in practical scenarios.

2.3 Face Detection and Landmark Extraction

Accurate detection of the driver's face is a critical prerequisite for reliable fatigue analysis. In this system, facial detection and landmark extraction are performed using MediaPipe FaceMesh, which is capable of identifying a dense set of facial key points.

A total of 468 landmarks are detected, covering essential facial regions such as:

- Eyes
- Mouth
- Nose
- Jawline

These landmarks provide precise geometric information, enabling the system to track subtle facial movements associated with drowsiness. The robustness of FaceMesh allows the system to function effectively under moderate variations in lighting and head orientation.

2.4 Vision Transformer-Based Classification

In addition to handcrafted features, the system employs a Vision Transformer (ViT) for deep learning–based classification of driver alertness.

Unlike traditional convolutional neural networks, the Vision Transformer processes images by splitting them into smaller patches and applying self-attention mechanisms. This allows the model to capture global relationships across different regions of the face.

The advantages of using ViT include:

- Ability to model long-range dependencies
- Improved feature representation
- Better generalization performance

The model outputs a probability score indicating whether the driver is in an alert or drowsy state.

2.5 Hybrid Decision Fusion Strategy

A key innovation of the proposed system lies in its hybrid decision-making approach. Instead of relying on a single source of information, the system combines:

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 10, April 2026
Impact Factor: 8.2

- Rule-based fatigue indicators (EAR, MAR, PERCLOS)
- Vision Transformer predictions

The fusion strategy ensures that both interpretable features and deep learning outputs contribute to the final decision. This reduces the likelihood of incorrect predictions caused by noise or temporary variations.

Temporal smoothing is applied to stabilize the output over consecutive frames. This prevents sudden fluctuations and ensures that alerts are triggered only when drowsiness persists over time.

2.6 Alert Generation Mechanism

The final stage of the system is the alert generation module. Once the system detects sustained drowsiness, it activates warning mechanisms to notify the driver.

The alerts include:

- Audio warnings (alarm sounds)
- Visual notifications (on-screen messages)

The alert is triggered only when the drowsiness condition exceeds a predefined threshold for a continuous duration. This design minimizes unnecessary alerts and enhances user experience.

III. METHODOLOGY

Steps:

The methodology follows a structured pipeline designed for real-time and continuous monitoring:

1. Capture real-time video using a camera
2. Extract frames from the video stream
3. Detect the driver's face using MediaPipe FaceMesh
4. Extract facial landmarks (468 key points)
5. Compute fatigue indicators such as EAR, MAR, and PERCLOS
6. Input facial image patches into the Vision Transformer model
7. Obtain classification results (alert/drowsy)
8. Fuse outputs from feature-based and deep learning models
9. Apply temporal smoothing to avoid sudden fluctuations
10. Generate alerts when drowsiness is consistently detected

This pipeline ensures both **accuracy and robustness** in real-world scenarios.

IV. DETAILED SYSTEM ARCHITECTURE

The proposed system follows a modular architecture designed for real-time performance and explainability. It consists of multiple interconnected components that process video input, extract meaningful features, and generate reliable drowsiness alerts.

4.1 Input Module

The system captures real-time video using a standard RGB camera. The video stream is processed frame-by-frame to ensure continuous monitoring. This approach ensures scalability and compatibility with existing in-vehicle camera systems.

4.2 Face Detection and Landmark Extraction

Face detection is performed using MediaPipe FaceMesh, which detects 468 facial landmarks. These landmarks provide precise spatial information about key facial regions such as eyes, mouth, and jawline.

4.3 Feature Extraction Module

From the extracted landmarks, several fatigue indicators are computed:

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 10, April 2026
Impact Factor: 8.2

- Eye Aspect Ratio (EAR) for detecting eye closure
- Mouth Aspect Ratio (MAR) for yawning detection
- Blink frequency for fatigue patterns
- PERCLOS for prolonged eye closure measurement

These features are interpretable and provide transparency in decision-making.

4.4 Deep Learning Module (Vision Transformer)

The Vision Transformer processes facial images by dividing them into patches and applying self-attention mechanisms. Unlike CNNs, ViT captures global dependencies across facial regions, improving classification performance.

4.5 Decision Fusion Module

A hybrid approach combines:

- Rule-based fatigue indicators
- ViT classification outputs

This fusion improves reliability and reduces false positives.

4.6 Alert Generation Module

The system triggers alerts only when drowsiness is detected consistently over time. This avoids unnecessary disturbances.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 10, April 2026
Impact Factor: 8.2

```
(.venv) PS C:\Users\guguloth venkatesh\Saved Games\Downloads\files (1)> python drowsiness_detection_enhanc
ed.py
>>
pygame 2.5.2 (SDL 2.28.3, Python 3.10.11)
Hello from the pygame community. https://www.pygame.org/contribute.html
INFO: Created TensorFlow Lite XNNPACK delegate for CPU.
============================================================
Enhanced Driver Drowsiness Detection System
Using Vision Transformers (ViT) + Behavioral Indicators
============================================================

🎯 Alert Conditions:
1. 🥱 Yawning - 2 quick beeps
2. 👁️ Excessive Blinking - 3 rapid beeps
3. ⬇️ Head Down/Not Visible - Long continuous beep
4. 😴 Eyes Closed - Urgent beeps
```

V. ALGORITHM DESCRIPTION

- Algorithm: Drowsiness Detection

1. Initialize camera
2. Capture frame
3. Detect face using FaceMesh
4. Extract landmarks
5. Compute EAR, MAR, PERCLOS
6. Input face image into ViT
7. Get classification output
8. Apply fusion rule:
    1. If EAR < threshold AND ViT = drowsy
    2. Then increase drowsiness score
9. Apply temporal smoothing
10. If score exceeds threshold → Trigger alert

VI. MATHEMATICAL MODEL

- Eye Aspect Ratio (EAR)

Used to detect eye closure by measuring vertical and horizontal eye distances.

EAR=‖p2−p6‖+‖p3−p5‖2‖p1−p4‖EAR = \frac{\|p_2-p_6\| + \|p_3-p_5\|}{2\|p_1-p_4\|}EAR=2‖p1−p4‖‖p2−p6‖+‖p3−p5‖

- Low EAR indicates closed eyes
- Threshold-based detection improves accuracy

- Mouth Aspect Ratio (MAR)

Used to detect yawning behavior.

MAR=‖p3−p7‖+‖p4−p6‖2‖p1−p5‖MAR = \frac{\|p_3-p_7\| + \|p_4-p_6\|}{2\|p_1-p_5\|}MAR=2‖p1−p5‖‖p3−p7‖+‖p4−p6‖

- High MAR indicates yawning
- Useful for fatigue detection

DOI: 10.48175/IJARSCT-33444

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 10, April 2026
Impact Factor: 8.2

- PERCLOS

Percentage of eye closure over time.

PERCLOS=Closed Eye FramesTotal FramesPERCLOS = \frac{\text{Closed Eye Frames}}{\text{Total Frames}}PERCLOS=Total FramesClosed Eye Frames

- Strong indicator of drowsiness
- Widely used in real-world systems

VII. EXPERIMENTAL SETUP

The system is evaluated using both controlled datasets and real-time inputs.

- Datasets
- Public drowsiness datasets
- Real-time webcam data
- Hardware
- Standard laptop/PC
- Webcam
- Software
- Python
- OpenCV
- MediaPipe
- TensorFlow / PyTorch
- Evaluation Metrics
- Accuracy
- Precision
- Recall
- F1-score
- Latency

VIII. PERFORMANCE ANALYSIS

The proposed system demonstrates:

- **High Accuracy** due to hybrid model
- **Low Latency** suitable for real-time use
- **Reduced False Alarms** due to temporal smoothing
- **Better Interpretability** compared to CNN models
- Comparison with Existing Systems

Method	Accuracy	Explainability
CNN	High	Low
ViT	High	Medium
Proposed	Very High	High

IX. LIMITATIONS

- Performance may drop in low-light conditions
- Face occlusion (mask, sunglasses) affects accuracy
- Requires proper camera alignment
- Computational cost of Vision Transformers

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 10, April 2026
Impact Factor: 8.2

X. REAL-WORLD APPLICATIONS

- Smart vehicles (ADAS systems)
- Fleet driver monitoring
- Public transport safety
- Mining and industrial vehicles
- Aviation fatigue monitoring

XI. CONCLUSION

The proposed system presents an effective and reliable solution for real-time driver drowsiness detection by integrating Vision Transformers with interpretable behavioral fatigue indicators. By combining deep learning–based classification with feature-based analysis such as Eye Aspect Ratio (EAR), Mouth Aspect Ratio (MAR), blink frequency, and PERCLOS, the system achieves a balanced trade-off between accuracy and explainability.

One of the key strengths of this approach lies in its hybrid decision-making strategy, which reduces dependency on a single model and enhances robustness under varying real-world conditions. The inclusion of temporal smoothing further improves system stability by minimizing false alarms caused by short-term fluctuations in driver behavior. As a result, the system ensures that alerts are triggered only during sustained drowsiness, thereby enhancing usability and driver acceptance.

In addition, the non-intrusive design of the system makes it highly practical for deployment in modern vehicles without requiring specialized hardware or wearable sensors. The use of Vision Transformers enables better feature representation and global context understanding compared to traditional convolutional approaches, while the incorporation of interpretable metrics ensures transparency in decision-making.

Overall, the proposed framework demonstrates strong potential for integration into intelligent transportation systems and Advanced Driver Assistance Systems (ADAS). It contributes not only to improving road safety but also to advancing the development of explainable artificial intelligence in safety-critical applications. With further optimization and real-world validation, the system can serve as a reliable component in next-generation smart vehicle technologies.

XII. FUTURE WORK

- Integration with IoT-enabled vehicles
- Edge AI deployment for faster processing
- Night vision support using IR cameras
- Mobile application integration
- Multi-modal detection (voice + face)

REFERENCES

[1] S. Abtahi, B. Hariri, and S. Shirmohammadi, “Driver drowsiness monitoring based on yawning detection,” in *Proc. IEEE Int. Instrum. Meas. Technol. Conf.*, 2011, pp. 1–4.

[2] L. M. Bergasa, J. Nuevo, M. A. Sotelo, R. Barea, and M. E. Lopez, “Real-time system for monitoring driver vigilance,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 7, no. 1, pp. 63–77, Mar. 2006.

[3] M. Eriksson and N. Papanikolopoulos, “Eye-tracking for detection of driver fatigue,” in *Proc. IEEE Intelligent Transportation Systems Conf.*, 1997, pp. 314–319.

[4] A. Vaswani *et al.*, “Attention is all you need,” in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 5998–6008.

[5] A. Dosovitskiy *et al.*, “An image is worth 16x16 words: Transformers for image recognition at scale,” in *Proc. Int. Conf. Learning Representations (ICLR)*, 2021.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 10, April 2026
Impact Factor: 8.2

[6] S. Khan, M. Naseer, M. Hayat, S. Zamir, F. Khan, and M. Shah, "Transformers in vision: A survey," *ACM Computing Surveys*, vol. 54, no. 10, pp. 1–41, 2022.
[7] T. Soukupová and J. Čech, "Real-time eye blink detection using facial landmarks," in *Proc. 21st Computer Vision Winter Workshop*, 2016.
[8] Google, "MediaPipe FaceMesh," *Google AI Research*, 2020. [Online]. Available: https://mediapipe.dev
[9] R. Ghoddoosian, M. Galib, and V. Athitsos, "A realistic dataset and baseline temporal model for early drowsiness detection," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019, pp. 178–187.
[10] Y. Li, X. Chen, F. Zhao, and J. Liu, "Driver fatigue detection based on deep learning and facial expression analysis," *IEEE Access*, vol. 8, pp. 187914–187924, 2020.
[11] H. Park, S. Pan, and J. Kwak, "Driver drowsiness detection using facial features and deep learning," *Sensors*, vol. 19, no. 21, pp. 1–15, 2019.
[12] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Facial landmark detection by deep multi-task learning," in *Proc. European Conf. Computer Vision (ECCV)*, 2014, pp. 94–108.