WakeNet: A Deep-Learning Model for Real-Time Drowsiness Detection

Dr. Sunita Chalageri
Department of Computer Science
And Engineering
K S Institute of Technology
Bengaluru, India
dr.sunitachalageri@ksit.edu.in

Priyadharshini E P
Department of Computer Science
And Engineering
K S Institute of Technology
Bengaluru, India
priyadharshiniep05@gmail.com

Namratha M H
Department of Computer Science
And Engineering
K S Institute of Technology
Bengaluru, India
namrathamh03@gmail.com

Nishanth P
Department of Computer Science
And Engineering
K S Institute of Technology
Bengaluru, India
nishanthshetty620@gmail.com

Nikhil Kumar P E
Department of Computer Science
And Engineering
K S Institute of Technology
Bengaluru, India
nikhilkumarpe@gmail.com

Abstract—Driver fatigue remains a silent but deadly factor in road accidents, impairing reaction times, judgment, and vehicle control. We propose WakeNet, an energy-conscious hybrid framework that integrates Vision Transformers (ViTs) with
v8 for robust, real-time detection of drowsiness. Unlike conventional CNN-based systems that falter under poor lighting or facial occlusion, WakeNet leverages contextual feature extraction and precise region-of-interest localization to capture subtle cues such as micro-blinks and yawns. An adaptive pre-processing pipeline using Contrast Limited Adaptive Histogram Equalization (CLAHE) ensures resilience in low-light conditions. Beyond detection, WakeNet introduces a dual-interface mobile application: personal mode for individual drivers and commercial mode for fleet managers, enabling proactive alerts and emergency escalation. Experiments on benchmark datasets (UTA-RLDD, DDD) demonstrate high accuracy, rapid inference, and practical scalability, underscoring WakeNet's potential to advance road safety worldwide.

Keywords: Driver drowsiness detection, Vision Transformers, YOLOv8, Low-light enhancement, Eye Aspect Ratio (EAR), Mouth Aspect Ratio (MAR), CLAHE, Mobile integration.

I. INTRODUCTION

Road safety has long been recognized as a critical global challenge, with fatigue-related crashes emerging as one of the most insidious contributors to traffic accidents. Unlike reckless driving, which is often visible and can be mitigated through enforcement, fatigue operates silently, gradually eroding a driver's alertness, impairing judgment, and slowing reaction times. Studies consistently show that fatigue-related incidents account for thousands of fatalities annually, underscoring the urgent need for intelligent monitoring systems that can intervene before accidents occur [10].

Traditional monitoring approaches, particularly those based on Convolutional Neural Networks (CNNs), have demonstrated promising results in detecting drowsiness through facial cues such as eye closure and yawning [12]. However, these systems often falter under real-world conditions. Poor illumination during night driving, partial facial occlusion caused by accessories like sunglasses or masks, and subtle behavioral cues such as micro-blinks or suppressed yawns can significantly reduce their reliability. This unpredictability

highlights the limitations of CNNs, which primarily rely on localized receptive fields and struggle to capture broader contextual dependencies.

To address these challenges, we pose a fundamental question: Can the speed and efficiency of YOLO be enhanced by the contextual depth of Vision Transformers? Building on prior research that demonstrated the complementary strengths of object detection and transformer-based architectures [2][3][5], we propose WakeNet, a hybrid framework that unites the rapid detection capabilities of YOLOv8 with the global attention mechanisms of Vision Transformers (ViTs). YOLOv8 provides lightweight, high-speed bounding of facial regions, while ViTs capture long-range dependencies and subtle temporal variations in facial expressions, enabling more nuanced detection of fatigue indicators.

Recognizing that fatigue often strikes during night-time driving, we further integrate Contrast Limited Adaptive Histogram Equalization (CLAHE) to enhance visibility in low-light conditions [8]. This adaptive pre-processing ensures that facial features remain detectable even under challenging illumination scenarios. Beyond detection, WakeNet extends its utility through a mobile-integrated deployment strategy, offering both personal and commercial interfaces. In personal mode, drivers receive real-time multimodal alerts (audio, vibration, visual), while in commercial mode, fleet managers can monitor multiple drivers simultaneously, with automatic escalation to control centers in case of emergencies.

Contributions of this work are threefold:

- A hybrid YOLOv8–Vision Transformer framework that balances high accuracy with real-time efficiency.
- An adaptive pre-processing pipeline designed to handle low-light and occluded conditions.
- A mobile-integrated deployment strategy supporting both personal safety and commercial fleet monitoring.

pII. RELATED WORK

A. Deep Learning in Driver Monitoring

Deep learning has played a pivotal role in advancing driver monitoring systems. Early approaches relied heavily on CNN-based methods [4][12], which were particularly effective in detecting eye closure and yawning through localized feature extraction. These models could classify open versus closed eyes with reasonable accuracy under controlled conditions. However, their reliance on spatial features alone limited their ability to capture temporal dependencies, such as the gradual slowdown of blink rates or the subtle progression of fatigue over time. As highlighted in [6], CNNs often fail to recognize these temporal patterns, leading to false negatives or delayed detection in real-world driving scenarios. This shortcoming emphasizes the need for architectures that can integrate both spatial and temporal cues to provide a more holistic understanding of driver behavior.

B. Vision Transformers

The introduction of Vision Transformers (ViTs) [2][3] marked a significant breakthrough in computer vision. Unlike CNNs, which operate on localized receptive fields, ViTs leverage self-attention mechanisms to capture global contextual information across an image. This ability makes them particularly effective in scenarios involving occlusion, where parts of the face may be hidden by masks, glasses, or shadows. By modeling long-range dependencies, ViTs can infer subtle cues that CNNs often miss. However, their computational complexity poses challenges for real-time deployment in embedded automotive systems. To address this, researchers have explored hybrid approaches [13], combining lightweight detectors with Transformers to balance efficiency and accuracy. Such strategies allow systems to retain the contextual depth of ViTs while maintaining the speed required for real-time monitoring.

C. YOLO Evolution

The YOLO (You Only Look Once) family of object detection models [1][5] has

revolutionized real-time computer vision by enabling rapid and accurate localization of objects in a single forward pass. YOLO's progression from v3 to v5 introduced improvements in inference speed, detection of smaller objects, and overall robustness. Building on this foundation, YOLOv8 represents a further leap forward, offering enhanced accuracy, reduced model size, and faster inference times [9]. These improvements make YOLOv8 particularly suitable for embedded automotive environments, where computational resources are limited but real-time performance is critical. By integrating YOLOv8 into driver monitoring systems, facial landmarks such as eyes and mouth can be detected with high precision, providing the foundation for fatigue analysis.

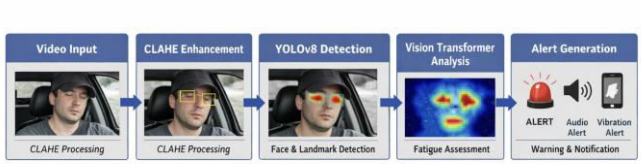
D. Low-Light Enhancement

One of the most persistent challenges in driver monitoring is low-light visibility, especially during night driving. Conventional histogram equalization techniques often amplify noise, reducing reliability in dark conditions. Contrast Limited Adaptive Histogram Equalization (CLAHE) [8] addresses this issue by enhancing local contrast while minimizing noise amplification. By applying CLAHE to the luminance channel of images, systems can preserve fine facial details even under poor illumination. This makes CLAHE an ideal pre-processing step for drowsiness detection, ensuring that critical features such as eye closure and yawning remain visible regardless of lighting conditions. When combined with YOLO and ViTs, CLAHE significantly improves detection accuracy in real-world scenarios, bridging the gap between laboratory performance and practical deployment.

III. PROPOSED METHODOLOGY

A. System Architecture

The proposed WakeNet framework is designed as a multi-stage pipeline that ensures robust detection of driver fatigue under diverse real-world conditions. The architecture is divided into three major components: Environmental Adaptation, Region of Interest (ROI) Extraction, and Behavioral Analysis as shown in Fig 1. Each



stage plays a critical role in transforming raw video input into actionable safety alerts.

Fig 1: System architecture showing video input, CLAHE enhancement, YOLOv8 detection, Vision Transformer analysis, and alert generation.

The system begins with a continuous video feed captured from an in-vehicle camera. This feed undergoes pre-processing to adapt to environmental conditions such as low illumination or glare. Once enhanced, the frames are passed into the YOLOv8 detection module, which rapidly identifies and bounds facial regions. These regions are then analyzed by the Vision Transformer, which attends to fatigue-sensitive zones such as the eyes and mouth. Finally, behavioral metrics are fused to quantify drowsiness, triggering multimodal alerts when thresholds are exceeded.

B. Environmental Adaptation

Driving at night or in poorly lit environments poses significant challenges for vision-based monitoring systems. To address this, WakeNet incorporates an adaptive pre-processing pipeline that enhances visibility

without introducing noise. Frames are

converted into the LAB color space, where the luminance channel (L-channel) is isolated for contrast enhancement as shown in Fig 2.

Fig 2: CLAHE-based contrast enhancement applied to driver's face under low-light conditions.

Using Contrast Limited Adaptive Histogram Equalization (CLAHE), local contrast is improved while preventing over-amplification of bright regions. Clip limits are dynamically adjusted: 4.0 for dark scenes and 3.0 for moderately lit conditions, ensuring balanced visibility across varying environments [8]. This adaptive enhancement allows the system to detect subtle facial cues even under headlight glare or shadowed conditions.

C. YOLOv8 Integration

At the core of the detection pipeline lies YOLOv8, which acts as a high-speed region proposer. Unlike traditional detectors, YOLOv8 is optimized for both accuracy and efficiency, making it suitable for real-time automotive applications. The model rapidly identifies facial landmarks, ensuring that no critical region is missed—even when the driver's face is partially occluded or viewed at oblique angles [1][5][9].

Once the bounding boxes are established, the Vision Transformer (ViT) module takes over. Leveraging its global attention mechanism [2][3], the ViT focuses on fatigue-sensitive zones such as the eyes and mouth as shown in Fig 3. This dual-stage integration ensures that YOLOv8 provides precise localization, while the ViT extracts subtle temporal and contextual features that CNNs often overlook.

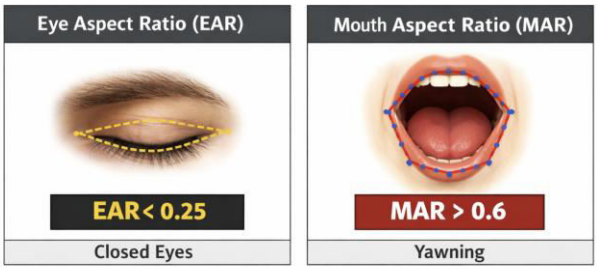


Fig 3: YOLOv8 bounding boxes around facial landmarks for fatigue detection.

D. Feature Fusion

To quantify drowsiness, WakeNet employs a feature fusion strategy based on well-established behavioral metrics. Two primary indicators are used:

Eye Aspect Ratio (EAR): Calculated using Euclidean distances between vertical and horizontal eye landmarks. A value below 0.25 indicates closed eyes, signaling potential fatigue.

Mouth Aspect Ratio (MAR): Derived from mouth landmarks, with values above 0.6 indicating yawning [4].

To reduce false positives caused by rapid blinking or speaking, a hysteresis buffer is introduced. This buffer ensures that classification decisions are stable, preventing oscillations between "drowsy" and "alert" states. For instance, eyes are only reclassified as "open" when EAR rises significantly above the threshold, while yawns are validated through consecutive frame tracking as shown in Fig 4.

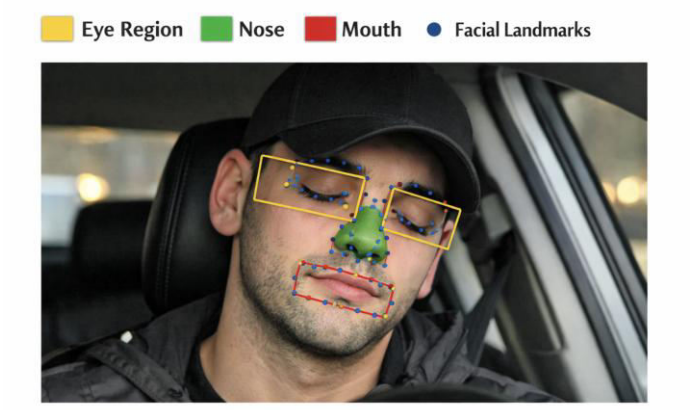


Fig 4: EAR and MAR thresholds for blink and yawn detection.

IV. IMPLEMENTATION DETAILS

A. Development Environment

The WakeNet framework was implemented using Python 3.11, chosen for its stability and compatibility with modern deep learning libraries. The core

detection and analysis modules were built on PyTorch, which provided flexibility in designing and training the Vision Transformer and YOLOv8 models. For image pre-processing and video stream handling, OpenCV was integrated, enabling efficient frame-by-frame manipulation, CLAHE-based contrast enhancement, and facial landmark visualization.

Deployment and testing were carried out on a MacBook Pro equipped with an Apple M-series chip, which offered hardware acceleration through Metal Performance Shaders (MPS). This environment ensured that the system could achieve real-time inference speeds while maintaining energy efficiency, simulating conditions similar to embedded automotive platforms.

B. Model Configuration

The detection backbone was configured using YOLOv8s, a lightweight yet powerful variant of the YOLO family. With a footprint of approximately 12MB, YOLOv8s was selected for its balance between accuracy and computational efficiency, making it suitable for real-time deployment in vehicles where hardware resources are constrained [1][5][9].

For precise facial landmark detection, the system employed the 68-point facial landmark predictor [14]. This predictor allowed accurate localization of critical regions such as the eyes and mouth, which are essential for calculating fatigue-related metrics. By combining YOLOv8's bounding box proposals with the landmark predictor, the system achieved high recall and precision in identifying fatigue-sensitive zones.

C. Thresholds and Logic

To quantify drowsiness, the system relied on established behavioral metrics:

Eye Aspect Ratio (EAR): The EAR measures the ratio between vertical and horizontal distances of the eye landmarks as given in eq 1.

$$[EAR = \frac{||p_2 - p_6|| + ||p_3 - p_5||}{2 \cdot ||p_1 - p_4||}] \text{—eq 1}$$

$[p_1, p_2, \dots, p_6]$ are the six landmark points around the eye.

$[||p_i - p_j||]$ denotes the Euclidean distance between two points.

When EAR < 0.25, the eyes are considered closed[4].

Mouth Aspect Ratio (MAR): The **MAR** quantifies mouth openness using vertical and horizontal landmark distances as given in eq 2.

$$[MAR = \frac{||p_{14} - p_{18}|| + ||p_{15} - p_{17}||}{2 \cdot ||p_{13} - p_{19}||}]$$

——-——eq 2

$[p_{13}] - [p_{19}]$are the mouth landmarks.

When MAR > 0.6, yawning is detected.

Blink Duration: Blink duration is measured by counting consecutive frames where EAR remains below the threshold as given in eq 3:

$[D_{blink} = N_{frames}(EAR < 0.25)]$—eq 3

A valid blink is detected when $[D_{blink} \geq 3]$ frames.

Drowsiness Trigger: Drowsiness is triggered when prolonged eye closure persists as given in eq 4:

$[D_{drowsy} = N_{frames}(EAR < 0.25) \geq 20]$—eq 4

These thresholds were carefully calibrated during experimentation to minimize false positives while ensuring sensitivity to genuine fatigue events.

V. EXPERIMENTAL SETUP AND RESULTS

A. Datasets

UTA-RLDD [10] and DDD [11] datasets

were used, covering diverse ethnicities,

lighting, and camera angles as shown in Fig 5.

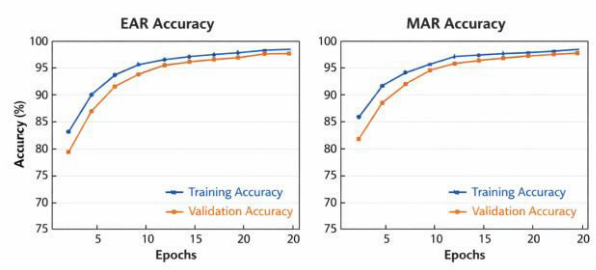


Fig 5: Sample images from UTA-RLDD dataset showing drowsy vs. alert drivers.

B. Performance

• Normal Lighting Accuracy

Under standard illumination, the system achieved 97.2% accuracy in detecting fatigue indicators. Accuracy is computed using the standard classification formula as given in eq 5:

$$[Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \times 100\%]$$——— eq 5

1. TP (True Positives): Correctly detected fatigue events.
2. TN (True Negatives): Correctly identified non-fatigue states.
3. FP (False Positives): Normal states misclassified as fatigue.
4. FN (False Negatives): Fatigue states missed by the system.

This high accuracy demonstrates the robustness of the YOLOv8–ViT integration under well-lit conditions.

• Low-Light Accuracy

In reduced illumination, accuracy dropped slightly to 92.8%, which still indicates strong resilience. The improvement is largely due to CLAHE preprocessing, which enhances contrast and ensures facial landmarks remain visible. The same accuracy formula applies here, but the denominator reflects more challenging frames with noise and occlusion.

• Average Inference Time

The system's efficiency was measured by the average time taken to process each frame as given in eq 6:

$$[t_{avg} = \frac{\sum_{i=1}^{N} t_i}{N}]$$———-eq 6

$[t_i]$: inference time for frame $[i]$

$[N]$: total number of frames

Here, $[t_{avg} \approx 30ms]$.

Since video is processed sequentially, the effective frames per second (FPS) as given in eq 7:

$$[FPS = \frac{1000}{t_{avg}} \approx \frac{1000}{30} \approx 33]$$———eq 7

This ensures real-time performance, as the system can process ~33 frames per second, matching typical video streaming rates.Training and Validation is shown in Fig 6.

Fig 6: Training and validation accuracy curves for EAR and MAR detection.

C. Real-Time Capabilities

The mobile app provides multimodal alerts (audio, vibration, visual). In commercial mode, fleet managers receive real-time notifications and accident escalation to control centers as shown in Fig 7.

Fig 7: Mobile application interface showing personal and fleet monitoring dashboards.

VI. DISCUSSION

The proposed hybrid YOLOv8–Vision Transformer (ViT) framework demonstrates notable resilience when tested under challenging conditions such as low-light

environments and partial facial occlusion. By integrating CLAHE-based enhancement,

the system was able to significantly improve detection accuracy, ensuring that facial landmarks remained visible even in scenarios where conventional histogram equalization methods would fail [8]. This enhancement proved critical in maintaining reliable performance during night driving, a situation where driver fatigue is most prevalent.

Another key strength of the framework lies in the Vision Transformer's ability to capture global contextual information. Unlike traditional aspect-ratio methods, which rely solely on geometric thresholds, the ViT was able to distinguish between yawning and speaking, a subtle but important nuance in fatigue detection [4]. This capability highlights the importance of attention mechanisms in modeling temporal and contextual variations in facial behavior, reducing false positives and improving overall system reliability.

Despite these advancements, certain limitations remain. The framework struggles with extreme head poses exceeding 45°, where facial landmarks become distorted or partially invisible. This limitation is consistent with challenges observed in prior studies and points to the need for pose-invariant feature extraction techniques [15]. Future iterations of WakeNet will aim to incorporate such methods, ensuring robustness across a wider range of driver behaviors and orientations.

VII. CONCLUSION AND FUTURE WORK

The proposed WakeNet framework represents a significant advancement in driver monitoring systems by combining the speed and efficiency of YOLOv8 with the contextual depth of Vision Transformers (ViTs) and the adaptability of CLAHE-based pre-processing. Together, these components enable robust detection of fatigue indicators such as prolonged eye closure and yawning, even under challenging conditions like low illumination or partial facial occlusion. By integrating these technologies, WakeNet demonstrates strong resilience, high accuracy, and real-time responsiveness, making it suitable for deployment in both personal and commercial automotive environments.

Looking ahead, our future work will focus on practical deployment as a working device. Specifically, we aim to implement WakeNet on embedded platforms such as Raspberry Pi and Jetson Nano, ensuring portability and energy efficiency for real-world automotive applications. Additionally, we plan to extend the system's capabilities by incorporating physiological signals via remote photoplethysmography (rPPG) [13], which will provide a complementary layer of monitoring by analyzing heart-rate variability and other biometric indicators of fatigue.

By transforming WakeNet into a fully functional device, we envision a comprehensive solution that not only alerts drivers in real time but also integrates seamlessly with fleet management systems, enabling proactive intervention and emergency escalation. This evolution from software framework to hardware prototype will mark a crucial step toward making roadways safer and reducing fatigue-related accidents worldwide.

REFERENCES

[1] J. Redmon et al., "You Only Look Once," CVPR, 2016.
[2] A. Dosovitskiy et al., "An Image is Worth 16x16 Words," ICLR, 2021.
[3] Z. Liu et al., "Swin Transformer," ICCV, 2021.
[4] S. Park et al., "Driver Drowsiness Detection using Facial Landmarks," IEEE Access, 2020.
[5] G. Jocher et al., "YOLOv5 SOTA," Zenodo, 2022.
[6] R. Ghoddoosian et al., "Baseline Temporal Model for Drowsiness," CVPR Workshops, 2019.
[7] A. Pizurica, W. Philips, "Image Denoising," IEEE TIP, 2006.
[8] K. Zuiderveld, "CLAHE," Graphics Gems IV, 1994.

[9] W. Deng et al., “YOLOv5 and Mobilenet for Drowsiness,” ICAICA, 2022.
[10] R. Jabbar et al., “Real-time Drowsiness Detection for Android,” Procedia CS, 2018.
[11] J. Wulff, M. Black, “Optical Flow Estimation,” CVPR, 2015.
[12] T. Zhang et al., “CNN-LSTM Fusion for Drowsiness,” IEEE Access, 2021.
[13] M. Sandler et al., “MobileNetV2,” CVPR, 2018.
[14] O. Ronneberger et al., “U-Net,” MICCAI, 2015.
[15] K. He et al., “Deep Residual Learning,” CVPR, 2016.