

IoT-Enabled Driver Drowsiness Detection Using Facial Dynamics, U-Net Segmentation and CNN-LSTM

Renuka Zende¹, Prof. P. P. Chavan², Prof. A. J. Shaikh³, Dr. H. B. Jadhav⁴

¹Student, Department of Computer Engineering

^{2,3}Assistant Professor, Department of Computer Engineering

⁴Head of Department (HOD), Department of Computer Engineering
Adsul Technical Campus Chas, Ahilyanagar, Maharashtra, India

Abstract: Driver fatigue is a major road-safety concern because a driver may gradually lose attention and enter a drowsy state without fully recognizing the change. This work presents an IoT-oriented approach for automatic drowsiness monitoring using visual information obtained from the driver's face. The framework combines U-Net segmentation with a CNN-LSTM architecture. U-Net first isolates the relevant driver region from the surrounding visual content, reducing the influence of irrelevant background information. The CNN then learns spatial features from the segmented facial images, while the LSTM processes these representations across consecutive video frames. By combining spatial information with temporal behavior, the system distinguishes an alert driver from one exhibiting patterns associated with drowsiness.

Keywords: Driver drowsiness detection; driver fatigue monitoring; Internet of Things (IoT); facial behavior analysis; U-Net segmentation; Convolutional Neural Network (CNN); Long Short-Term Memory (LSTM); Advanced Driver Assistance Systems (ADAS); computer vision; temporal feature analysis.

I. INTRODUCTION

Modern vehicles increasingly use sensing, connectivity, and artificial intelligence to support drivers and improve road safety. Within advanced driver-assistance environments, continuous monitoring of driver alertness is particularly important. Fatigue can reduce attention, slow reaction time, and limit the ability to respond effectively to changing traffic conditions. Therefore, a practical drowsiness-monitoring system should continuously observe the driver without requiring wearable physiological sensors.

A camera-based monitoring approach provides a relatively unobtrusive way to observe facial behavior during driving. Visual indicators such as prolonged eye closure, changes in mouth opening, blinking frequency, and other facial movements can provide useful evidence of reduced alertness. However, drowsiness develops over time, so relying on a single video frame may produce unreliable decisions. The proposed approach therefore combines spatial image analysis with sequential learning. U-Net focuses processing on the relevant driver region, CNN layers learn visual representations from that region, and an LSTM analyzes those representations across consecutive frames.

II. SYSTEM DESIGN

The proposed system consists of a sequence of processing stages that begins with video acquisition and ends with an IoT-based warning decision. Each stage transforms the captured facial information into an estimate of the driver's current alertness state.



A. Video Acquisition and Preprocessing

A driver-facing camera continuously records visual information while the vehicle is operating. The resulting video stream is divided into individual frames and prepared for further processing. These frames provide the input for segmentation, feature extraction, and temporal analysis.

B. U-Net-Based Driver Region Segmentation

For each captured frame, U-Net identifies the region associated with the driver. This segmentation separates relevant facial information from surrounding content, allowing later processing to focus on visual characteristics related to drowsiness.

The focused representation also reduces the influence of unrelated background elements.

C. CNN-Based Facial Feature Extraction

The segmented driver region is provided to a Convolutional Neural Network. Rather than relying entirely on manually defined visual descriptors, the CNN learns spatial patterns directly from the training images. The resulting representations capture facial structures and appearance changes associated with eye, mouth, and other facial movements.

D. LSTM-Based Temporal Modeling

Features obtained from successive frames are processed by a Long Short-Term Memory network. Because fatigue-related behavior usually develops across a sequence rather than at a single instant, the LSTM combines information from previous observations with the current representation. This temporal modeling helps distinguish an isolated facial event from a sustained pattern of drowsiness.

III. EXISTING SYSTEM

Earlier driver-drowsiness detection approaches have largely relied on observable changes in facial behavior. Typical systems examine eye closure, blinking, mouth activity, and related facial movements as indicators of reduced alertness. Video-based monitoring allows these indicators to be observed continuously instead of evaluating only an isolated image.

A. Facial-Based Monitoring

Conventional vision-based methods analyze selected facial regions to estimate whether a driver is alert or fatigued. Eye and mouth behavior are particularly useful visual cues. For example, prolonged eye closure, unusual blinking patterns, and increased mouth activity may indicate drowsiness.

B. Deep Learning-Based Detection

More recent approaches use deep-learning models to learn useful representations directly from driver images and video frames. CNN architectures can extract spatial characteristics from facial imagery, reducing the need for manually engineered features.

C. Existing Classification Models

Various deep-learning architectures have been evaluated for driver-state classification. The source study compares EM-CNN, VGG-16, GoogLeNet, AlexNet, and ResNet50 with a CNN-LSTM-based method. These architectures differ in feature-learning capability and computational requirements.

D. Limitations of Frame-Based Analysis

An individual frame may not adequately represent the gradual development of driver fatigue. Drowsiness is better described by behavioral changes occurring across a sequence of observations. Temporal information is therefore useful for distinguishing a momentary facial event, such as a normal blink, from a continuing pattern of reduced alertness.



E. Need for an Improved Approach

Driver-facing video may contain substantial background information that has little relevance to drowsiness classification. Concentrating the analysis on the appropriate driver region can reduce such distractions. These considerations motivate an integrated framework in which U-Net performs region segmentation, CNN extracts spatial characteristics, and LSTM models temporal changes. The resulting state estimate can then be used to activate an IoT-based warning mechanism.

IV. PROPOSED SYSTEM

The proposed framework analyzes a continuous stream of in-vehicle images rather than treating each frame as an independent sample. The camera provides the raw observations, preprocessing prepares them for analysis, U-Net isolates the relevant driver region, CNN layers learn visual characteristics, and LSTM layers model the resulting feature sequence. A final classifier uses the learned representation to estimate the driver's drowsiness state.

4.1 Overall Architecture

The architecture follows a sequential processing pipeline. An input frame is first normalized and prepared for segmentation. U-Net then produces a segmentation map that identifies the relevant driver region. The selected region is represented through CNN-derived spatial features, which are arranged as a sequence and supplied to the LSTM for temporal analysis.

4.2 U-Net-Based Region Segmentation

U-Net is selected because its encoder-decoder structure combines contextual information with detailed spatial reconstruction. The encoder progressively captures higher-level representations, while the decoder reconstructs the segmentation map. Skip connections transfer spatial information from the encoding path to the corresponding decoding stages, helping preserve useful details during reconstruction. In the proposed application, the resulting mask retains the driver-related region while reducing the influence of reflections, vehicle interiors, surrounding scenery, and other irrelevant visual structures.

4.3 Dataset and Preprocessing

The source study reports the use of a driver-video database associated with I. Nasri et al. and obtained through Kaggle. The dataset is described as approximately 5.3 GB in size and contains around 7.3 hours of video. Frontal driver views were recorded at frame rates between 12 and 27 frames per second with a resolution of 640×480 pixels. The experimental setup included more than 3,144 images, with examples representing open-eye, closed-eye, and open-mouth conditions.

4.4 Feature Extraction with CNN

The CNN stage converts the segmented images into spatial feature representations. Convolutional filters learn local visual patterns from the training data rather than requiring every relevant characteristic to be explicitly programmed. In this application, the learned representations can capture information around the eyes, mouth, facial contour, and other facial structures useful for estimating driver alertness.

4.5 Temporal Modeling with LSTM

The LSTM receives the sequence of CNN-derived representations and retains information from earlier time steps. This capability is important because fatigue-related behavior can develop gradually. A single frame showing closed eyes may correspond to an ordinary blink, whereas prolonged closure or repeated similar observations provide stronger evidence of drowsiness. The LSTM uses gated memory to determine which information should be retained, updated, or passed to the output.

The recurrent stage can be represented in simplified form as:

$$\begin{aligned} i_t &= \sigma(W_i[h_{t-1}, x_t] + b_i) \\ f_t &= \sigma(W_f[h_{t-1}, x_t] + b_f) \\ o_t &= \sigma(W_o[h_{t-1}, x_t] + b_o) \\ c_t &= f_t \odot c_{t-1} + i_t \odot \tanh(W_c[h_{t-1}, x_t] + b_c) \end{aligned}$$



Here, x_t denotes the feature vector at the current time step, h_{t-1} denotes the hidden state from the previous step, and c_t represents the memory state. Through these gates, the recurrent network can preserve informative temporal patterns while limiting the effect of less useful information.

4.6 Drowsiness Decision and PERCLOS

PERCLOS is used in the source study as a behavioral measure for assessing eye closure. It represents the proportion of frames within a specified observation interval in which the eyes are classified as closed. The measure is expressed as:

$$\text{PERCLOS} = (\text{N}_{\text{closed}} / \text{N}_{\text{total}}) \times 100\%$$

For the described experiment, the source reports a threshold of 0.25 or higher for the eye-closure condition.

4.7 Operational Algorithm

- Capture a driver-facing frame using the vehicle camera.
- Normalize and preprocess the captured frame.
- Apply U-Net segmentation to obtain the relevant driver region.
- Use convolutional layers to learn spatial facial features.
- Organize frame-level representations into a temporal sequence.
- Process the sequence through the LSTM to learn temporal behavior.
- Classify the driver state using the learned spatial and temporal representation.
- Generate an alert when the estimated state indicates drowsiness.

V. EXPERIMENTAL SETUP

According to the source study, the model was trained for 100 epochs with a batch size of 32. The reported experimental environment included a six-core CPU, 16 GB of RAM, and an NVIDIA GTX 1650 Ti GPU. The experiments were conducted on Windows 10 using Google Colab Pro Plus, with Python 3.10, Keras 2.4.0, and TensorFlow 2.7.0 reported as the software environment.

VI. RESULTS AND DISCUSSION

The reported evaluation shows that the proposed CNN-LSTM approach achieved the highest testing accuracy among the compared models in the principal accuracy assessment, with a value of 98.80%. A separate precision, recall, and F1-score evaluation reports an accuracy of 98.46%. Because these figures come from different evaluation presentations, they should be interpreted as separate reported results rather than as identical measurements.

Performance Comparison

Model	Precision	Recall	Accuracy (%)
EM-CNN	0.721	0.685	96.62
VGG-16	0.113	0.218	82.98
GoogLeNet	0.708	0.611	94.01
AlexNet	0.461	0.485	91.48
ResNet50	0.677	0.536	93.85
Proposed CNN-LSTM	0.819	0.652	98.46

In the reported comparison, the CNN-LSTM method achieves the highest accuracy and precision among the listed models. These results support the use of CNN-based spatial representation learning together with LSTM-based temporal modeling for driver-drowsiness analysis.

VII. LIMITATIONS AND PRACTICAL CONSTRAINTS

Several practical factors may influence the reliability of the proposed system. The source study notes that binarization-based processing can be sensitive to variations in skin tone and illumination. Differences among subjects and



environmental conditions may therefore affect performance. The available training data may also not fully represent the range of situations encountered during real-world driving.

The experimental environment included a relatively controlled background, including the use of a black sheet during testing to limit unwanted reflections. Rapid head movements were not a major component of the evaluation. Although these conditions simplify experimentation, they may not represent the full range of behavior observed in actual vehicles. Broader testing is therefore required.

Eyewear is another relevant constraint. The source observations indicate stronger classification performance for images captured without glasses, while sunglasses can reduce the visibility of facial characteristics, particularly the eyes. Camera position, reflections, lighting, and eyewear can therefore affect the reliability of eye-based cues.

For greater robustness, the system should not rely exclusively on visible eye information. Additional cues such as mouth movement, head orientation, facial geometry, and temporal behavior can complement eye-related measurements. Changes in illumination, glare, reflections, unusual headlight conditions, and low-light environments should also be considered during evaluation. Broader experiments involving different drivers, vehicles, camera placements, weather conditions, and illumination levels would provide a more realistic assessment.

Computational performance is another practical consideration. The reported CNN-LSTM model has a size of 21.77 MB, with a reported drowsiness-detection time of 26.88 seconds and an overall processing speed of 11.6 fps.

VIII. CONCLUSION AND FUTURE WORK

The proposed framework combines U-Net segmentation, CNN-based spatial feature extraction, and LSTM-based temporal modeling for IoT-oriented driver-drowsiness monitoring. The process first isolates the relevant driver region, learns meaningful facial representations, and then evaluates those representations across consecutive frames. This sequence-oriented strategy allows the system to consider changes in facial behavior over time rather than making a decision from an isolated image.

The reported results show a testing accuracy of 98.80% in the principal accuracy evaluation and 98.46% in the separate precision, recall, and F1-score evaluation. The reported precision is 0.819, while the F1 score is 0.732. The model size is

21.77 MB, with a reported detection time of 26.88 seconds and an overall speed of 11.6 fps. These findings indicate potential for non-contact driver-drowsiness monitoring, although additional validation is needed to assess performance under diverse real-world driving conditions.

Future work may investigate the integration of an attention mechanism to allow the model to assign greater importance to informative facial regions and significant temporal events. Further evaluation should use more diverse data and include variations in illumination, eyewear, head movement, and realistic driving environments. Testing on embedded hardware would also help determine whether the framework is suitable for real-time deployment. Incorporating complementary cues such as lip movement and head orientation may further improve robustness when eye-related information is partially obscured.

REFERENCES

1. S. Das et al., "IoT-Assisted Automatic Driver Drowsiness Detection through Facial Movement Analysis Using Deep Learning and a U-Net-Based Architecture," *Information*, vol. 15, 2024.
2. M. Saradadevi and P. Bajaj, "Driver Fatigue Detection Using Mouth and Yawning Analysis," *International Journal of Computer Science and Network Security*, 2008.
3. M. Dua et al., "Deep CNN Models-Based Ensemble Approach to Driver Drowsiness Detection," *Neural Computing and Applications*, 2021.
4. S. Chaabene et al., "Convolutional Neural Network for Drowsiness Detection Using EEG Signals," *Sensors*, 2021.



5. M. J. Flores et al., "Real-Time Warning System for Driver Drowsiness Detection Using Visual Information," Journal of Intelligent and Robotic Systems, 2010.
6. L. Pauly and D. Sankar, "Detection of Drowsiness Based on HOG Features and SVM Classifiers," IEEE ICRCICN, 2015.
7. B. Reddy et al., "Real-Time Driver Drowsiness Detection for Embedded System Using Model Compression of Deep Neural Networks," CVPR Workshops, 2017.
8. S. E. Bekhouche et al., "Driver Drowsiness Detection in Video Sequences Using Hybrid Selection of Deep Features," Knowledge-Based Systems, 2022.
9. V. T. Teyeb et al., "A Novel Approach for Drowsy Driver Detection Using Head Posture Estimation and Eyes Recognition," IISA, 2014.
10. C. Lea et al.

