Journal of Artificial Intelligence and Data Mining (JAIDM), Vol. 13, No. 4, 2025, 469-467.

Shahrood University of Technology

Journal of Artificial Intelligence and Data Mining (JAIDM)
Journal homepage: http://jad.shahroodut.ac.ir

Research paper

# Accuracy Improvement of Real-Time Driver Drowsiness Detection Using Transformer Model

Havva Askari, Razieh Rastgoo*, and Kourosh Kiani

*Electrical and Computer Engineering Department, Semnan University, Semnan, Iran.*

**Article Info**

**Article History:**
*Received 07 May 2025*
*Revised 01 June 2025*
*Accepted 08 July 2025*

*DOI:10.22044/jadm.2025.16191.2743*

**Keywords:**
*Transformer Network, Driver Drowsiness Detection, Facial Keypoints, Mediapipe, Real-Time.*

**Corresponding author: rrastgoo@semnan.ac.ir (R. Rastgoo).*

**Abstract**

Drowsiness remains a significant challenge for drivers, often resulting from extended working hours, inadequate sleep, and accumulated fatigue. This condition impairs reaction time and decision-making and contributes to a substantial number of road accidents globally. Therefore, reliable and timely detection of driver drowsiness is essential for enhancing transportation safety and reducing the risk of traffic-related fatalities. With the rapid progress in deep learning, numerous models have been developed to detect driver drowsiness with high accuracy. However, the real-world performance of these models can deteriorate under varying environmental conditions, such as changes in cabin illumination, facial occlusions, and dynamic shadows on the driver's face. To address these limitations, this paper proposes a robust, real-time driver drowsiness detection model that leverages facial behavioral features and a Transformer-based neural network architecture. The MediaPipe framework is utilized to extract a comprehensive set of facial keypoints, capturing subtle facial movements and expressions indicative of drowsiness. These keypoints are then encoded to form feature vectors that serve as input to the Transformer network, enabling effective temporal modeling of facial dynamics. The proposed model is trained and evaluated on the National Tsing Hua University (NTHU) Driver Drowsiness Detection dataset, achieving a state-of-the-art accuracy of 99.71%, demonstrating its potential for deployment in real-world in-vehicle systems.

## 1. Introduction

Drowsiness is generally defined as the abnormal feeling of sleepiness or tiredness during the day that can lead to distraction and falling asleep at inappropriate times, especially while driving [1]. According to a 2017 report from the National Highway Traffic Safety Administration (NHTSA), 91,000 police-reported accidents involved drowsy drivers, resulting in approximately 50,000 injuries and nearly 800 fatalities [2]. Additionally, the Foundation for Traffic Safety reported that 21% of all fatal crashes involved a drowsy driver [3]. These accidents often lead to financial losses, injuries, and fatalities [4]. These alarming statistics underscore the need for a system to monitor and alert drivers to drowsiness, aiming to prevent traffic accidents.

Recently, various models have been developed for automatic drowsiness detection systems. Advanced Driver Assistance Systems (ADAS) offer features such as parking sensors, lane departure warnings, traffic sign recognition, and blind-spot monitoring systems [3-6]. However, traditional ADAS systems lack drowsiness and fatigue detection. Recognizing this challenge, this paper proposes a driver drowsiness detection model to automatically alert drivers before irreparable damage occurs.

As mentioned, driver drowsiness is a perilous condition that occurs when an individual becomes

sleepy or fatigued while driving, often resulting from factors such as sleep deprivation, long hours of driving, medication use, and certain medical conditions. Drowsiness reduces reaction time and attention span, impairing decision-making abilities. Warning signs include yawning, difficulty keeping the eyes open, trouble focusing, drifting in and out of lanes, and nodding off. Therefore, a key challenge for fatigue detection systems is accurately detecting drowsiness in real-time using efficient features. This paper employs lip and eye movement features during yawning and keeping the eyes open to propose a model for accurate driver drowsiness detection.

Recent advances in Artificial Intelligence and deep learning models have encouraged researchers to propose various models in different research areas [7-27], especially for diverse research tasks [28-30]. Within this field, numerous models have been proposed for driver drowsiness detection, employing different features [31-33]. However, real-time and accurate detection of driver drowsiness remains challenging. Some models focused solely on accuracy for offline applications [34], while others have been developed for real-time applications [35]. This work proposes a novel real-time driver drowsiness detection model that achieves high accuracy while maintaining robustness under challenging real-world conditions. The model leverages a Transformer-based architecture [36] to effectively capture temporal dependencies in facial behavioral patterns and integrates Mediapipe-extracted facial keypoints to enhance sensitivity to subtle drowsiness-related cues such as prolonged blinking and yawning.

The remainder of the paper is organized as follows. Section 2 provides a literature review on drowsiness detection with deep learning networks. Section 3 details the proposed model. The experimental results of the proposed model are presented and discussed in Section 4. Finally, Section 5 concludes the work and provides a roadmap for future research.

## 2. Literature Review

In this section, recent works in driver drowsiness detection are reviewed in three categories: Mouth Feature Extraction, Eye Feature Extraction, and Face Feature Extraction. Mouth feature extraction includes models that utilize mouth features, such as calculating the mouth's aspect ratio or mouth frequency, for drowsiness detection. Eye Feature Extraction involves models that consider the aspect ratio of the eyes or the percentage of eye closure, often using facial landmarks for these models. Face Feature Extraction encompasses models that learn most facial features of the driver (without extracting facial landmarks) to detect drowsiness based on the features of the eyes, mouth, and subsequently, the face.

### 2.1. Mouth Feature Extraction

This category primarily relies on mouth features for accurately detecting driver drowsiness. To this end, Anber et al. [34] proposed and compared two AlexNet-based Convolutional Neural Network (CNN) models [37] for detecting driver fatigue behaviors, focusing on head position and mouth movements as behavioral criteria. They employed two different approaches. The first approach involves transfer learning, specifically fine-tuning the final layers of AlexNet to adapt the model to the specific task of detecting driver drowsiness. The second approach involved using AlexNet to extract features by fine-tuning the upper layers of the network. These features were then reduced using Non-Negative Matrix Factorization (NMF) and classified using a Support Vector Machine (SVM) classifier. Results on the NTHU dataset show 95.7% accuracy for AlexNet and 99.65% for SVM.

In another study, Yogesh et al. [38] introduced a module for detecting driver fatigue to mitigate accidents caused by drowsy drivers. They utilized the Dlib algorithm, facial landmarks, and coordinates, such as the Facial Feature Triangle (FFT), representing a geometric feature area. They created a Facial Feature Vector (FFV) containing all the data related to the region and center of each FFT. They used FFV as a criterion for detecting driver fatigue or alertness. Finally, a sliding window was created to calculate facial data entropy, including the Eye Aspect Ratio (EAR) and Mouth Aspect Ratio (MAR), over sequential frames displaying the facial layout. These FFVs are then summed to provide facial information needed for final detection in the model. Results on their private dataset show 94.66% accuracy for the MAR-based model and 95.99% for the EAR-based model. However, this model cannot detect drowsiness from both eye and mouth features simultaneously.

### 2.2. Eye Feature Extraction

The prominent features utilized in this category for accurately detecting driver drowsiness are mouth features. In this context, Flores-Monroy et al. [3] proposed a real-time driver drowsiness detection system in which the driver's facial region is extracted and fed into a specially designed Shallow Convolutional Neural Network (SS-CNN). For

face detection, they employed the Viola-Jones face detector, extracting Haar-like features from a unified image and applying a cascading Adaboost classifier to quickly eliminate non-face objects. Subsequently, the SS-CNN was employed for detecting driver drowsiness. When the system detects driver drowsiness, it activates an alert sound. The proposed system's performance was evaluated using the NTHU dataset, achieving a recognition accuracy of 98.95%. However, it's worth noting that this model does not utilize mouth features.

Jahan et al. [35] proposed a deep neural network-based approach, namely the Deep Driver Drowsiness Detector (4D), to detect driver drowsiness. The 4D model is designed to detect sleepiness based on eye position. They used the Dlib face detection algorithm to detect face and eye keypoints to calculate the Eye Aspect Ratio (EAR). Results on the MRL Eye dataset, comprising images of open and closed eyes under different lighting conditions and environmental factors, showed an accuracy of 97.53% in predicting eye position. However, it's important to consider that the Dlib algorithm may be sensitive to variations in lighting conditions.

### 2.3. Face Feature Extraction

This category primarily employs facial features for the accurate detection of driver drowsiness. In this context, Bai et al. [39] proposed a Two-Stream Spatiotemporal Graph Convolutional Network (2S-STGCN) for detecting driver drowsiness. Their approach uses video as the processing unit instead of individual video frames. This method employs a two-stream network for modeling both spatial and temporal features, simultaneously handling first-order and second-order information. Evaluation results on the NTHU and YawDD datasets show recognition accuracy of 92.7% and 93.4%, respectively [30,31]. However, the model complexity of this method is high. Bekhouche et al. [42] developed a computer vision-based framework for detecting driver drowsiness using the input video. This framework consists of three main stages: Face Detection, Feature Extraction, and Classification. In the first stage, the driver's facial region is identified within a sliding time window using a pre-trained YOLO model. Bounding boxes and class predictions are obtained after evaluating the input image. Deep features from each detected face are extracted in the second stage, using a pre-trained ResNet-50 model trained on the VGGFace2 dataset. Statistical operators are used to compute a sequence of feature vectors. Finally, the obtained feature vectors are fed into an SVM classifier to determine the presence or absence of driver drowsiness. The model achieved a recognition accuracy of 86.74%.

Aytekin and Mençik [43] proposed a CNN-based model for driver drowsiness detection using the position of eyelids. In this way, the pre-trained VGG16 model is used for feature extraction. Results on the YawDD dataset, including images of drivers of different genders and ages obtained using in-car cameras during driving, show a recognition accuracy of 91%. However, the model complexity can be decreased. Jia et al. [44] designed a system, namely Facial Feature Detection (FFD-System), as well as an algorithm, entitled Facial Multi-Factor Fusion Algorithm (MF-Algorithm), for detecting driver facial features and assessing driver fatigue status, respectively. In the FFD-System, they designed three networks: M1-FDNet, M2-PENet, and M3-SJNet for detecting the driver's face, estimating head position, and recognizing the status of the eyes and mouth, respectively. Three datasets, WIDER FACE, BIWI, and a self-built dataset, are used to train the M1-FDNet, M2-PENet, and M3-SJNet models, respectively. The MF-Algorithm uses the output from the FFD-System to calculate some parameters, such as eye blink duration, blink frequency, eye closure rate, mouth opening frequency, and head asynchrony time. These parameters are combined to provide a comprehensive recognition of the driver's fatigue status, obtaining a recognition accuracy of 97.8%. While this model benefits from simultaneously using the eyes, mouth, and head features, the model complexity is high due to using three different networks for each part of the model. Krishna et al. [45] proposed a framework for driver drowsiness detection using the YOLOv5 and ViT (Vision Transformer) models. After extracting the frames from the input videos, each frame serves as input to the framework for final drowsiness detection. They used two video datasets, The University of Texas at Arlington Real-Life Drowsiness Detection Dataset (UTARLDD) [46] for training their model and a custom-collected dataset for testing the model, obtaining detection accuracies of 97.4% and 95.5%, respectively. However, the proposed model needs a large amount of annotated data for training in different scene conditions.

Aiming for performance enhancement, we propose an accurate model using recent advancements in deep learning models, especially Transformer models, for real-time driver drowsiness detection. Using the keypoints of the face, the proposed model outperforms state-of-the-art models in the

field. Details of the proposed model are explained in the following section.

## 3. Proposed Method

Here, we present the details of the proposed model. Since most previous works have primarily relied on image features, we introduce a novel model that has significantly improved accuracy in detecting driver drowsiness. Previous works have utilized the Vision Transformer (ViT) for drowsiness detection by converting the input image into smaller patches and feeding them into the network [3]. In this work, we employ the standard Transformer network for driver drowsiness detection using facial keypoints. In the standard Transformer, mainly employed in Natural Language Processing (NLP) tasks, the network comprises two parts: an Encoder and a Decoder. However, in our case, we exclusively utilize the Encoder part. Unlike traditional recurrent models (e.g., LSTMs or GRUs), Transformers are specifically designed to model long-range temporal dependencies using self-attention mechanisms, which allow the network to weigh the importance of each time step in the input sequence relative to others. This is particularly advantageous in the context of driver drowsiness detection, where subtle behavioral patterns—such as delayed blinks, prolonged eye closure, or gradual yawning—unfold over several frames and require global temporal context to be accurately identified. Moreover, the parallel processing capability of Transformers offers improved efficiency over sequential models, enabling real-time performance, which is critical for in-vehicle deployment. The self-attention mechanism also makes the model more robust to irrelevant or noisy frames, as it can selectively attend to frames containing key behavioral cues. These characteristics make Transformers especially well-suited for modeling the temporal dynamics of facial keypoint sequences.

An overview of the proposed model can be found in Figure 1. Details of the proposed model are explained below.

### 3.1. Video Pre-processing

In this method, as the network's input is a sequence, videos were pre-processed to create suitable inputs for the model. To achieve this, given the varying number of frames in each video, the average number of frames across all videos was used to determine the input size for the network. Additionally, the input videos were annotated with one of two possible labels: drowsy or non-drowsy.

### 3.2. Facial Keypoint Extraction

In a typical Transformer used in NLP, the input consists of sequential words. Since we are also using a standard Transformer, the input for this network is a sequence of facial keypoints. To prepare the input for the Transformer network, 52 facial keypoints were extracted from shortened videos captured at a rate of 30 frames per second using the Mediapipe face mesh library. Mediapipe is a real-time, multi-person library for human pose estimation, hand tracking, face analysis, and more. It has the capability to detect body, hand, foot, and key facial points in images [47]. Mediapipe Face Mesh is a solution that estimates 468 3D facial landmarks in real-time, even on mobile devices. This algorithm employs machine learning to infer the 3D facial structure and requires only a camera, without the need for specialized depth sensors. Figure 2 shows the 468 facial keypoints (a) and the selected 52 facial keypoints (b) in the proposed model.

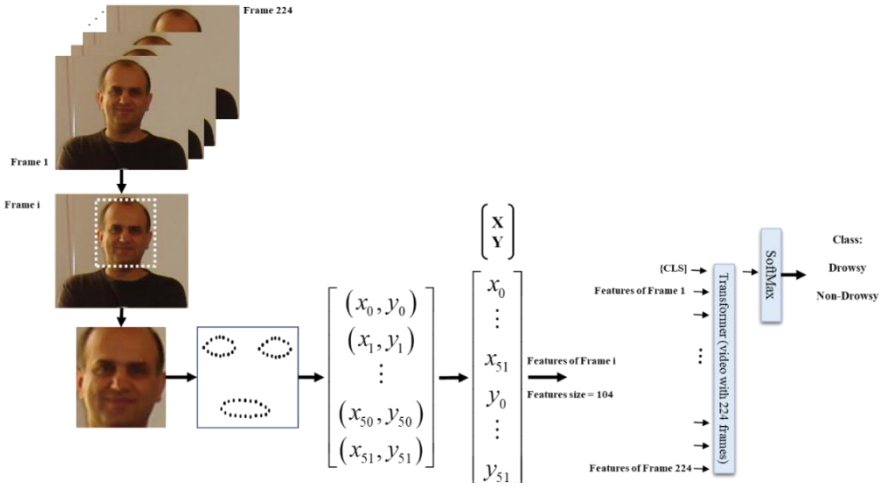


**Figure 1. An overview of the proposed model.**

The obtained keypoints consist of 16 keypoints for the left eye, 16 keypoints for the right eye, and 20 keypoints for the mouth. Aiming to decrease model complexity while maintaining the capability to work in a real-time environment, only 52 facial landmarks were used in the proposed model.

### 3.3. Network Implementation

In the implementation of the proposed model, facial keypoints (eyes and mouth) are first extracted from video frames and used as input to a Transformer network. The output vector obtained from the Transformer network is then fed into a Fully Connected (FC) layer with a Softmax activation function to recognize whether the driver is drowsy or not.

**Figure 2. Face and keypoint detection on one frame: a) 468 points, b) Selected 52 points.**

## 4. Experiments and Results Evaluation

In this section, the dataset, implementation details, ablation analysis, and results comparing the proposed model to state-of-the-art models are presented.

### 4.1. Dataset

The video dataset for driver drowsiness detection was obtained from the National Tsing Hua University (NTHU-DDD), encompassing various scenarios captured under both daytime and nighttime lighting conditions [41]. These scenarios include regular driving, yawning, slow blinking, deliberate smiling, chatting, and dizziness. The dataset includes male and female drivers from diverse ethnic backgrounds. Each frame in the videos is labeled as "fatigue" or "non-fatigue." The videos depict various activities related to "fatigue" and "non-fatigue." Figure 3 shows some examples from the NTHU dataset. Subsequently, an experiment was conducted using infrared (IR) lighting to obtain the IR videos available in the dataset. These videos were recorded at a resolution of 480x640 pixels, with a frame rate of 30 frames per second in mp4 format and without audio [48]. All videos in this dataset required initial preprocessing before serving as input for the network. Since each video contains various scenarios, including drowsy and normal, we separated the drowsy and normal frame segments within each video. The training dataset consists of 1,860 videos, including 1,151 labeled as "drowsy" and 709 labeled as "alert." For validation, we used 253 videos, including 149 "drowsy" and 104 "alert" labeled videos. As the number of frames in each video varies, we used the average number of frames across all videos as the input size for the network. Considering the mean value, we ultimately selected 224 frames as the input size for the model. Subsequently, all videos were down-sampled and up-sampled to 224 frames. Yawning typically lasts between 2–5 seconds, while fatigue-related blinks (e.g., micro-sleeps or slow eye closures) often range from 0.5 to 2 seconds. At a frame rate of 30 FPS, 224 frames represent approximately 7.5 seconds of video—sufficient to capture multiple occurrences and temporal dynamics of these drowsiness indicators.

**Figure 3. Some samples from the NTHU-DDD dataset.**

### 4.2. Implementation details

The model was implemented in the Google Colab environment, equipped with a GPU and approximately 7GB of available RAM. The total training time for the model was approximately 150 minutes, involving 150 epochs with early stopping. The input shape of the Transformer network in the proposed model is set to (224x104). The values of hyperparameters, including the learning rate, batch size, the number of heads in the Transformer network, and the number of layers in the Transformer network, were set to 0.0001, 16, 16, and 12, respectively.

### 4.3. Results

Here, we present various analyses performed on the proposed model. Different numbers of self-attention heads and layers in the Transformer

model were experimented with, as shown in Table 1. Initially, we used 4 self-attention heads and 2 Transformer layers. Subsequently, we increased the Transformer layers to 4, 6, and 8. We then adjusted the head numbers to 6, 8, 12, and 16. Initially, the batch size and learning rate were set to 64 and 0.0001, respectively. After several runs, these parameters were updated based on the model's behavior. The highest performance was achieved with a learning rate of 0.0001 and a batch size of 32. Simultaneously, we adjusted the number of heads and Transformer layers, and the highest performance was obtained by setting these parameters to 16 and 12, respectively. Additionally, we conducted ablation experiments to evaluate the impact of using eye-only, mouth-only, and combined eye–mouth features in Table 2. The results of Table 2 consistently showed that using both sets of features significantly improved the model's performance. This improvement is intuitive, as drowsiness manifests through multiple facial cues—for instance, eye-related behaviors such as prolonged blinking or eye closure, and mouth-related actions such as yawning. These cues often occur together or sequentially, and relying on only one modality may overlook important indicators. The combined use enables the model to capture a more holistic temporal pattern of drowsy behavior.

Upon reaching the highest accuracy, we selected these parameters as superior parameters. To ensure robustness and mitigate the effects of any potential bias in model initialization or data ordering, we conducted 10 independent runs of the experiment using the same training/validation split, but with different random seeds for model initialization and data shuffling. The results of the proposed model during different epochs are shown in Table 3 and Figure 4. While the accuracy obtained fluctuates between 99% and 100%, the best result is obtained in epoch 150 with a value of 99.74%. After averaging the results of 10 runs, the accuracy and loss are equal to 99.71 and 0.83, respectively. Moreover, the standard deviations (std) for both accuracy and loss have been included in this table.

Finally, the model complexity, as well as the detection time of the model, is shown in Table 4. The number of parameters is an important factor in understanding the complexity and capacity of a neural network. Larger models with more parameters may have a greater capacity to learn intricate patterns in the data, but they also require more computational resources for training and inference. In the proposed model, using 52 facial keypoints, 12 Transformer layers, and 16 self-attention heads, the total number of network parameters was 9,331,090, and the model complexity was 35.60 MB. We also calculated the complexity and the number of parameters for the proposed model using 468 facial keypoints and found that choosing 52 facial keypoints reduces the network complexity (the total number of network parameters for 468 facial keypoints was 750,591,506, and the model complexity was 2.80 GB). In addition, the confusion matrix of the proposed model has been shown in Table 5. This confusion matrix summarizes the classification performance of the proposed drowsiness detection model on the validation set of 253 videos. Out of 149 videos labeled as drowsy, the model correctly identified 148 and misclassified 1 as alert. All 104 alert videos were correctly classified. This results in a high overall accuracy of 99.71%, demonstrating the model's strong ability to distinguish between drowsy and alert states, with minimal misclassification.

Finally, the proposed model is evaluated against state-of-the-art methods on the NTHU dataset. As shown in Table 6, the proposed model outperforms all comparative models in terms of accuracy Furthermore, the table also presents model complexity, confirming the suitability of the proposed approach for real-time drowsiness detection.

### 4.4. Discussion

This section discusses the strengths and limitations of the proposed model. The proposed model utilizes facial keypoints, specifically from the eyes and mouth, for accurate driver drowsiness detection. The model's reliance on these key features contributes to its success, as demonstrated by its superior performance compared to state-of-the-art models.

Additionally, the model's efficient use of 52 facial landmarks reduces complexity while maintaining high accuracy, making it suitable for real-time applications. However, it's essential to acknowledge certain limitations. The dataset used for training and evaluation might not capture the full spectrum of real-world scenarios, potentially limiting the model's generalization to diverse driving conditions. Moreover, the proposed model focuses on facial keypoints, neglecting other potential indicators of drowsiness, such as head position or external environmental factors. Expanding the model's capabilities to encompass a broader range of features could enhance its robustness. Despite these limitations, the proposed model represents a significant advancement in driver drowsiness detection, showcasing the potential of leveraging facial keypoints for accurate

and real-time monitoring. Further research and refinement could address the discussed limitations and contribute to the continuous improvement of such models.

## 5. Conclusion

In this study, a deep learning-based model was introduced to enhance the accuracy of driver drowsiness detection. The Mediapipe algorithm was employed to extract keypoints of the eyes and mouth due to its high accuracy and robustness, capable of extracting 468 points from the face. To capture essential drowsiness-related features in images, positional encoding of facial keypoints was utilized. Through analyzing the positions of the eyes and mouth keypoints, the proposed model demonstrated effectiveness in detecting driver drowsiness or alertness. With the utilization of only 52 facial keypoints, additional information in the images is unnecessary and can be masked. Overall, considering the achieved results and comparing them with other methods, this approach significantly improved the accuracy of drowsy driver detection.

.

**Table 1. Accuracy and loss for executing the model with different parameters.**

| Test Accuracy | Test Loss | Learning Rate | Batch size | Transformer Layers | Heads |
|---|---|---|---|---|---|
| 99.41 % | 1.48 | 0.0001 | 16 | 10 | 20 |
| 98.23 % | 4.67 | | 32 | | |
| **99.95 %** | **0.25** | | **16** | **12** | |
| 99.25 % | 2.22 | | 16 | 14 | |
| 98.6 % | 3.47 | | 32 | | |
| 97.74 % | 5.03 | | 64 | 2 | |
| 98.33 % | 3.98 | | | 4 | |
| 97.15 % | 7.22 | | | 6 | |
| 93.71 % | 14.9 | | | 8 | |
| 98.92 % | 3.8 | | 16 | 4 | 22 |
| 99.25 % | 2.26 | | 32 | | |

**Table 2. Results of the ablation study on two types of features in the proposed model.**

| Model | Accuracy | Loss |
|---|---|---|
| Only Eyes features | 91.15 | 1.84 |
| Only Mouth features | 88.16 | 2.12 |
| **Both Eyes and Mouth features** | **99.95** | **0.25** |

**Table 3. Accuracy and error of the network over 10 runs with consistent parameters.**

| **Run #** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **mean** | **std** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Accuracy** | 99.95 | 99.73 | 99.73 | 99.62 | 99.95 | 99.57 | 99.89 | 99.78 | 99.46 | 99.46 | 99.71 | 0.17 |
| **Loss** | 0.25 | 0.77 | 0.77 | 1.09 | 0.25 | 1.19 | 0.55 | 0.50 | 1.48 | 1.46 | 0.83 | 0.43 |

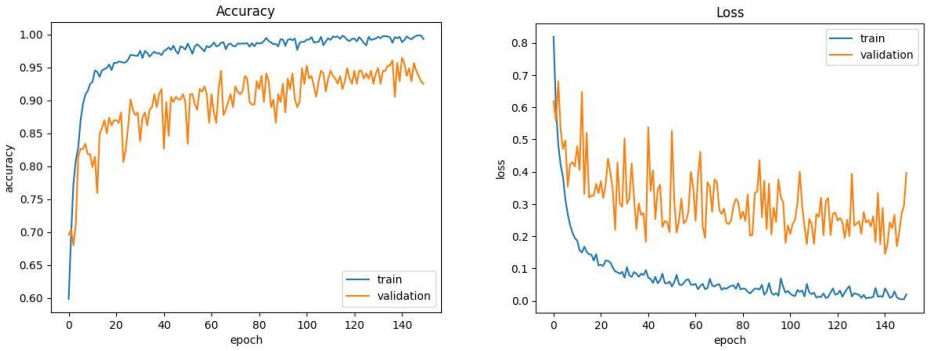


**Figure 4. The accuracy and error of the proposed model. (Left) The accuracy during training and validation for 150 epochs. (Right) The training and validation losses for 150 epochs.**

**Table 4. The model complexity and detection time of the proposed model.**

| Proposed model | Model complexity | Detection time | |
|---|---|---|---|
| | | CPU | GPU |
| | 35.60 MB | 0.550s | 0.135s |

**Table 5. Confusion matrix of the proposed model.**

| | **Predicted Drowsy** | **Predicted Alert** | **Total** |
|---|---|---|---|
| Actual Drowsy (149) | 148 (TP) | 1 (FN) | 149 |
| Actual Alert (104) | 0 (FP) | 104 (TN) | 104 |
| Total | 148 | 105 | 253 |

**Table 6. Comparison of the proposed method with the state-of-the-art models on NTHU dataset.**

| **Ref.** | **Model** | **Feature** | **Accuracy** | **Real-time** | **Model complexity** |
|---|---|---|---|---|---|
| [43] | SS-CNN | Eyes | 98.95 | **Yes** | - |
| [46] | 2sSTGCN | Eyes and Mouth | 92.7 | No | $1.6\times10^9$(FLOPs) |
| [32] | AlexNet | Mouth | 95.7 | No | - |
| [32] | AlexNet+SVM | Mouth | 99.65 | No | - |
| [40] | ResNet-50 | Eyes | 86.74 | No | - |
| **Proposed** | **Transformer** | **Eyes and Mouth** | **99.71** | **Yes** | 35.60 MB |

The outcomes of this research not only contribute to the early detection of driver drowsiness but also have potential applications in traffic or security cameras, security guard rooms, and healthcare settings for professional guidance and monitoring. As part of future work, considering neck angle and head dropping time as behavioral features could further enhance detection accuracy, particularly in scenarios where the driver is wearing a mask and glasses.

## References

[1] S.A. El-Nabi, W. El-Shafai, ES.M. El-Rabaie, et al., "Machine learning and deep learning techniques for driver fatigue and drowsiness detection: a review," Multimed Tools Appl, 2023. https://doi.org/10.1007/s11042-023-15054-0

[2] "NHTSA report (accessed on Nov 29, 2023)." [Online]. Available: *https://www.nhtsa.gov*.

[3] J. Flores-Monroy, M. Nakano-Miyatake, G. Sanchez-Perez, and H. Perez-Meana, "Visual-based real time driver drowsiness detection system using CNN," in *2021 18th International Conference on Electrical Engineering, Computing Science and Automatic Control (CCE)*, IEEE, 2021, pp. 1–5.

[4] P. Mate, N. Apte, M. Parate, et al., "Detection of driver drowsiness using transfer learning techniques," Multimed Tools Appl, 2023.

[5] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv Prepr. ArXiv201011929*, 2020.

[6] E. Magán, M. P. Sesmero, J. M. Alonso-Weber, and A. Sanchis, "Driver Drowsiness Detection by Applying Deep Learning Techniques to Sequences of Images," *Appl. Sci.*, vol. 12, no. 3, Art. no. 3, Jan. 2022, doi: 10.3390/app12031145.

[7] R. Rastgoo, K. Kiani, S. Escalera, "Diffusion-Based Continuous Sign Language Generation with Cluster-Specific Fine-Tuning and Motion-Adapted Transformer," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops (CVPRW)*, pp. 4088-4097, 2025.

[8] K. Kiani, R. Rastgoo, A. Chaji, S. Escalera, "Image Inpainting Enhancement by Replacing the Original Mask with a Self-attended Region from the Input Image," *Journal of AI and Data Mining*, vol. 13, no. 3, pp. 379-391, 2025.

[9] N. Esfandiari, K. Kiani, R. Rastgoo, "Development of a Persian Mobile Sales Chatbot based on LLMs and Transformer," *Journal of AI and Data Mining*, vol. 12, no. 4, pp. 465-472, 2024.

[10] N. Esfandiari, K. Kiani, R. Rastgoo, "Transformer-based Generative Chatbot Using Reinforcement Learning," *Journal of AI and Data Mining*, vol. 12, no. 3, pp. 349-358, 2024.

[11] A.M. Ahmadi, K. Kiani, R. Rastgoo, "A Transformer-based model for abnormal activity recognition in video," *Journal of Modeling in Engineering*, vol. 22, no. 76, pp. 213-221, 2024.

[12] F. Bagherzadeh, R. Rastgoo, "Deepfake image detection using a deep hybrid convolutional neural network," *Journal of Modeling in Engineering*, vol. 21, no. 75, pp. 19-28, 2023.

[13] M. Talebian, K. Kiani, R. Rastgoo, "A Deep Learning-based Model for Fingerprint Verification," *Journal of AI and Data Mining*, vol. 12, no. 2, pp. 241-248, 2024.

[14] H. Zaferani, K. Kiani, R. Rastgoo, "Real-time face verification on mobile devices using margin distillation," *Multimedia Tools and Applications*, vol. 82, no. 28, pp. 44155-44173, 2023.

[15] S. Zarbafi, K. Kiani, R. Rastgoo, "Spoken Persian digits recognition using deep learning," *Journal of Modeling in Engineering*, vol. 21, no. 74, pp. 163-172, 2023.

[16] N. Esfandiari, K. Kiani, R. Rastgoo, "A conditional generative chatbot using transformer model," *arXiv:2306.02074*, 2023.

[17] N. Majidi, K. Kiani, R. Rastgoo, "A deep model for super-resolution enhancement from a single image," *Journal of AI and Data Mining*, vol. 8, no. 4, pp. 451-460, 2020.

[18] R. Rastgoo, K. Kiani, "Face recognition using fine-tuning of Deep Convolutional Neural Network and transfer learning," *Journal of Modeling in Engineering*, vol. 17, no. 58, pp. 103-111, 2019.

[19] R. Rastgoo, V. Sattari-Naeini, "Gsomcr: Multi-constraint genetic-optimized qos-aware routing protocol for smart grids," *Iranian Journal of Science and Technology, Transactions of Electrical, Engineering*, vol. 42, no. 2, pp. 185-194, 2018.

[20] R. Rastgoo, V. Sattari-Naeini, "Tuning parameters of the QoS-aware routing protocol for smart grids using genetic algorithm," *Applied Artificial Intelligence*, vol. 30, no. 1, pp. 52-76, 2016.

[21] R. Rastgoo, V. Sattari Naeini, "A neurofuzzy QoS-aware routing protocol for smart grids," *22nd Iranian Conference on Electrical Engineering (ICEE)*, pp. 1080-1084, 2014.

[22] F. Alinezhad, K. Kiani, R. Rastgoo, "A Deep Learning-based Model for Gender Recognition in Mobile Devices," *Journal of AI and Data Mining*, vol. 11, no. 2, pp. 229-236, 2023.

[23] S. Shekarizadeh, R. Rastgoo, S. Al-Kuwari, M. Sabokrou, "Deep-disaster: unsupervised disaster detection and localization using visual data," *26th International Conference on Pattern Recognition (ICPR)*, pp. 2814-2821, 2022.

[24] R. Rastgoo, K. Kiani, S. Escalera, "ZS-GR: zero-shot gesture recognition from RGB-D videos," *Multimedia Tools and Applications*, vol. 82, no. 28, pp. 43781-43796, 2023.

[25] R. Rastgoo, K. Kiani, S. Escalera, "A deep co-attentive hand-based video question answering framework using multi-view skeleton," *Multimedia Tools and Applications*, vol. 82, no. 1, pp. 1401-1429, 2023.

[26] A. Pourreza, K. Kiani, "A partial-duplicate image retrieval method using color-based SIFT," *24th Iranian Conference on Electrical Engineering (ICEE)*, pp. 1410-1415, 2016.

[27] A. Fakhari, K. Kiani, "A new restricted boltzmann machine training algorithm for image restoration," *Multimedia Tools and Applications*, vol. 80, no. 2, pp. 2047-2062, 2021.

[28] A. Alsayat, "Improving Sentiment Analysis for Social Media Applications Using an Ensemble Deep Learning Language Model," Arabian Journal for Science and Engineering, vol. 47, pp. 2499–2511, 2022.

[29] A. Mukhamadiyev, L. Khujayarov, O. Djuraev, J. Cho, "Automatic Speech Recognition Method Based on Deep Learning Approaches for Uzbek Language," Sensors, vol. 22, no. 10, 2023.

[30] T. D. Pereira *et al.*, "SLEAP: A deep learning system for multi-animal pose tracking," *Nat. Methods*, vol. 19, no. 4, 2022, doi: 10.1038/s41592-022-01426-1.

[31] J. Cui *et al.*, "A compact and interpretable convolutional neural network for cross-subject driver drowsiness detection from single-channel EEG," *Methods*, vol. 202, pp. 173–184, 2022.

[32] M. H. Z. M. Fodli, F. H. K. Zaman, N. K. Mun, and L. Mazalan, "Driving Behavior Recognition using Multiple Deep Learning Models," in *2022 IEEE 18th International Colloquium on Signal Processing & Applications (CSPA)*, IEEE, 2022, pp. 138–143.

[33] A. Quddus, A. S. Zandi, L. Prest, and F. J. Comeau, "Using long short-term memory and convolutional neural networks for driver drowsiness detection," *Accid. Anal. Prev.*, vol. 156, pp. 106107, 2021.

[34] S. Anber, W. Alsaggaf, and W. Shalash, "A Hybrid Driver Fatigue and Distraction Detection Model Using AlexNet Based on Facial Features," *Electronics*, vol. 11, no. 2, p. 285, 2022.

[35] I. Jahan *et al.*, "4D: A Real-Time Driver Drowsiness Detector Using Deep Learning," *Electronics*, vol. 12, no. 1, 2023, doi: 10.3390/electronics12010235.

[36] A. Vaswani *et al.*, "Attention is all you need," *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[37] Y. LeCun and Y. Bengio, "Convolutional networks for images, speech, and time series," *Handb. Brain Theory Neural Netw*, vol. 3361, no. 10, p. 1995, 1995.

[38] R. Yogesh, V. Ritheesh, S. Reddy, and R. G. Rajan, "Driver Drowsiness Detection and Alert System using YOLO," in *2022 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES)*, IEEE, 2022, pp. 1–6, 2022.

[39] J. Bai *et al.*, "Two-stream spatial-temporal graph convolutional networks for driver drowsiness detection," *IEEE Trans. Cybern.*, 2021.

[40] M. Omidyeganeh et al., "YawDD: Yawning Detection Dataset." IEEE Transactions on Instrumentation and Measurement, vol. 65, no. 3, 2016. https://ieee-dataport.org/open-access/yawdd-yawning-detection-dataset.

[41] Ch.H. Weng, Y.H. Lai, Sh.H. Lai, "Driver Drowsiness Detection via a Hierarchical Temporal Deep Belief Network," In Asian Conference on Computer Vision Workshop on Driver Drowsiness Detection from Video, Taipei, Taiwan, Nov. 2016

[42] S. E. Bekhouche, Y. Ruichek, and F. Dornaika, "Driver drowsiness detection in video sequences using hybrid selection of deep features," Knowledge-Based Systems, vol. 252, pp. 109436, 2022.

[43] A. Aytekin and V. Mençik, "Detection of Driver Dynamics with VGG16 Model," *Appl. Comput. Syst.*, vol. 27, no. 1, pp. 83–88, 2022.

[44] H. Ja, Z. Xiao, and P. Ji, "Real-time fatigue driving detection system based on multi-module fusion," *Comput. Graph*, vol. 108, pp. 22–33, 2022.

[45] G. S. Krishna, K. Supriya, and J. Vardhan, "Vision Transformers and YoloV5 based driver drowsiness detection framework," *arXiv Prepr. ArXiv220901401*, 2022.

[46] R. Ghoddoosian, M. Galib, and V. Athitsos, "A realistic dataset and baseline temporal model for early drowsiness detection," in *Proceedings of the IEEE conference on computer* vision *and pattern recognition workshops (CVPRW)*, 2019, pp. 1–10.

[47] C. Lugaresi et al., "Mediapipe: A framework for building perception pipelines," ArXiv190608172, 2019.

[48] R. Jabbar, K. Al-Khalifa, M. Kharbeche, W. Alhajyaseen, M. Jafari, and S. Jiang, "Real-time driver drowsiness detection for android application using deep neural networks techniques," *Procedia Comput. Sci.*, vol. 130, pp. 400–407, 2018.

# بهبود دقت شناسایی بلادرنگ خواب‌آلودگی راننده با استفاده از مدل ترنسفورمر

**حوا عسکری، راضیه راستگو* و کورش کیانی**

**دانشکده برق و کامپیوتر، دانشگاه سمنان، سمنان، ایران.**

ارسال ۲۰۲۵/۰۵/۰۷؛ بازنگری ۲۰۲۵/۰۶/۰۱؛ پذیرش ۲۰۲۵/۰۷/۰۸

**چکیده:** کاهش هوشیاری یکی از چالش‌های مهم برای رانندگان به‌شمار می‌رود که اغلب ناشی از ساعات کاری طولانی، خواب ناکافی و خستگی تجمع‌یافته است. این وضعیت زمان واکنش و تصمیم‌گیری را مختل می‌کند و نقش قابل‌توجهی در بروز تصادفات جاده‌ای در سراسر جهان دارد. بنابراین، شناسایی قابل‌اعتماد و به‌موقع خواب‌آلودگی راننده برای ارتقای ایمنی حمل‌ونقل و کاهش خطر مرگ‌ومیر ناشی از تصادفات ترافیکی ضروری است. با پیشرفت سریع یادگیری عمیق، مدل‌های متعددی برای شناسایی خواب‌آلودگی راننده با دقت بالا توسعه یافته‌اند. با این حال، عملکرد این مدل‌ها در شرایط محیطی واقعی ممکن است کاهش یابد؛ به‌ویژه در مواجهه با تغییرات نور داخل کابین، انسدادهای صورت و سایه‌های پویا روی چهره راننده. برای غلبه بر این محدودیت‌ها، این مقاله مدلی مقاوم و بلادرنگ برای شناسایی خواب‌آلودگی راننده ارائه می‌دهد که از ویژگی‌های رفتاری چهره و معماری شبکه عصبی مبتنی بر ترنسفورمر بهره می‌برد. در این مدل، از چارچوب MediaPipe برای استخراج مجموعه‌ای جامع از نقاط کلیدی چهره استفاده می‌شود که حرکات و حالات ظریف صورت مرتبط با خواب‌آلودگی را ثبت می‌کند. این نقاط کلیدی به بردارهای ویژگی رمزگذاری شده و به عنوان ورودی به شبکه ترنسفورمر داده می‌شوند تا مدل‌سازی زمانی مؤثر از پویایی‌های چهره فراهم گردد. مدل پیشنهادی با استفاده از دیتاست تشخیص خواب‌آلودگی راننده دانشگاه تسینگ‌هوا (NTHU) آموزش داده شده و ارزیابی می‌گردد. این مدل به دقتی برابر با ۹۹/۷۱٪ دست یافته است که نشان‌دهنده پتانسیل بالای آن برای استفاده در سامانه‌های واقعی داخل خودرو می‌باشد.

**کلمات کلیدی:** شبکه ترنسفورمر، شناسایی خواب‌آلودگی راننده، نقاط کلیدی چهره، مدیاپایپ، بلادرنگ.