IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Ingénierie des Systèmes d'Information
Vol. 29, No. 5, October, 2024, pp. 1779-1788
Journal homepage: http://iieta.org/journals/isi

A Deep Learning-Based System for Driver Fatigue Detection

Abderrahim Benmohamed, Hafed Zarzour*

Department of Mathematics and Computer Science, University of Souk Ahras, Souk Ahras 41000, Algeria

Corresponding Author Email: hafed.zarzour@univ-soukahras.dz

https://doi.org/10.18280/isi.290511

Received: 14 February 2024
Revised: 10 September 2024
Accepted: 21 September 2024
Available online: 24 October 2024

Keywords:
deep learning, structural features, drowsiness detection, driver fatigue detection

ABSTRACT

Driver fatigue is still a principal cause of traffic accidents. While many ways allowing fatigue detection, a diversity of obstacles such as head position, luminosity, and facial expressions make it a very challenging problem. In this paper, we propose a hybrid approach using deep learning techniques to detect driver drowsiness by combining between structural and global classification methods. The structural method tracks eyes, eyebrows, and mouth movements to assess blink and yawning, for this purpose we calculate eye-opening and mouth-opening ratios relative to their width. Five parameters are extracted LEM (left eye movement), REM (right eye movement), LEB M (left eyebrow movement), REBM (right eyebrow movement), and MM (mouth movement), whereas the global method is based on Convolutional Neural Network (CNN) to describe the whole face. Eight-layer pre-trained Alexnet network is used to extract features and make classification of each frame. To do video classification, the five structural parameters, along with the global classification decision, are combined into a single vector to be input into Long-Short-Term Memory (LSTM) networks that is an improved version of Recurrent Networks. LSTM decision score is determined after running 150 steps, providing information about driver state Extensive Experiments are performed on a Driver Drowsiness Detection Dataset that contains subjects of different ethnicities. The experimental results show that the proposed method with the combined features improves drowsiness detection significantly as well as outperforms the state-of-the-art models in terms of drowsiness scores.

1. INTRODUCTION

There are approximately 1.2 billion vehicles around the world and the rate of road accidents continues to increase [1]. Statistics show that 1.25 million persons die each year [2], and 20-30% of road accidents are due to drowsiness. According to the World Health Organization [3], the fatality rate linked to road accidents stands at 26.60 deaths per 100,000 people. Excessive sleepiness or driver fatigue which may be mental or physical are crucial problems occurred during the long trips.

Recently, the evolution of new technologies allowed vehicles manufacturers to fit out cars by technology-based safety systems. The latter could detect the driver drowsiness in order to reduce the number of accidents by intervening automatically at the right time (warn the driver, slow down a car speed, etc.). Many technologies are proposed in which some systems are designed to operate as monitor for driver attention, and others are based on the evaluation of steering movements. All of them have to alert drivers rapidly and with high reliability when the latter are tired while avoiding as much as possible the false alerts. In the literature, the fatigue detection methods are divided into three classes: vehicle driving parameters based method [4], driver physiological signal-based method [5], and facial features-based method [6-8]. Figure 1 gives more clarification about the different fatigue detection methods.

Driving parameters-based methods take parameters from vehicle travel path, speed, lateral acceleration, etc. They are a non-contact and can give a high accuracy, but results can be affected by external factors like bad weather and road conditions. On the other hand, physiological signal-based methods are simple to carry out and to operate, but the driver could be disturbed because of direct contact of the sensors. Finally, driver facial features-based detection methods are more suitable because they are simple, reliable, and characterized by their low cost. The most promising among them are based on blink frequency and yawning detection. The blink frequency-based methods are real-time and give high accuracy, but non-effective when the driver wears glasses. Otherwise, yawning methods are accurate, but return a false positive result, that is the driver can open his or her mouth for laugh or when he or she is surprised.

In the Table 1, we give summaries of the three methods and show the advantages and inconvenient of each one:

This table shows that methods based on facial features are more appropriate due to their cost-effectiveness and non-intrusive nature. However, these methods can be further classified into two categories: those based on global features and those based on local features.

All non-intrusive methods rely on the tracking of eye and mouth movements. The variations among these methods lie in minor details, specifically in terms of eye and mouth configuration or the material used for classification. The challenge arises from the fact that the mouth and eyes may not consistently convey the true state, often due to issues such as insufficient detection caused by variations in luminosity, unexpected obstacles, or driver-related tics. Therefore, in this paper, we suggest a hybrid approach that merges both global and local features. Specifically, we extract CNN features from the entire face, and concurrently compute certain dimensions derived from the movements of the eyes and mouth. LSTM recurrent algorithm is employed to precisely determine the moment of drowsiness. The rest of this paper is organized as follows: in Section 2, we present some important related work. Then, we explain in detail both used features extraction methods and classification process in Section 3. In Section 5, we discuss the experiment results. Finally, we conclude the paper in Section 6.

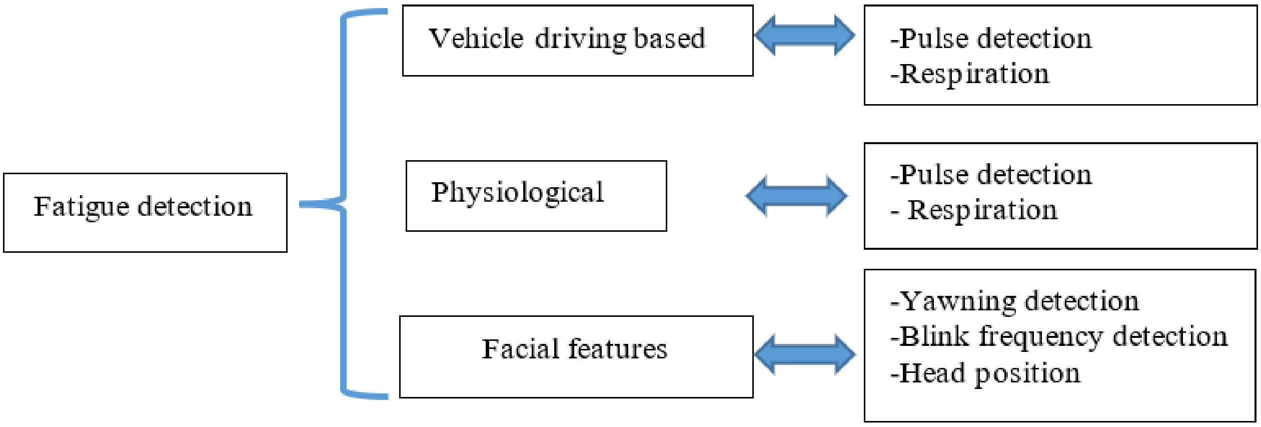


Figure 1. Different fatigue detection methods

Table 1. Comparative table between methods

Method	Advantages	Limitations
driving parameters-based	- high accuracy - parameters obtained from vehicle behavior (non-contact)	-sensitive to external factors (road, weather, …etc.) -need of expensive material
physiological signal-based	-high accuracy -related to conductor behavior	-need of expensive material - direct contact of the sensors disturb the conductor
driver facial features-based (global features)	-Simple and reliable - characterized by their low cost - non-intrusive	-sensitive to illumination variation
driver facial features-based (structural features)	-characterized by their low cost -non-Sensitive to illumination variation -non-intrusive	-non-effective when the driver wears glasses. -Sensitive to orientation -yawning methods are accurate, but return a false positive

2. RELATED WORK

In this section, we introduce some promising methods from fatigue detection classes by making a short comparison between them.

2.1 Vehicle driving-based methods

One of the solutions proposed to reduce the road accidents is the design of smart vehicle that can alert driver when a danger will happen, contact other vehicles in its neighborhood, and in extreme case take over driving [2, 9] proposed a method, which processes front view road images to evaluate the vehicle position and then advises drivers when they will across road side. The system used a 3D model of the road sides and all information about vehicle position, speed, and direction. The study [4] proposed a system that detects driver drowsiness and controls a vehicle speed. Practically, the vehicle control position detection requires expensive materials and hard computing. In addition, it is sensitive to the weather conditions and road state. To remedy, [1, 10] proposed a steering angle analysis method that uses an artificial neural network to classify driver states.

2.2 Physiological signal-based methods

Systems based-physiological parameters are more accurate and reliable. Numerous parameters are used among them are the following:

2.2.1 System using electro cardiogram (ECG)

ECG are based on heart rate variability, most of them need a very expensive material [11]. ECG methods use the electrode and attach them to the left leg and both arms to measure the voltage and produce an electro cardiogram [12]. The study [2] combined ECG with facial expression to improve results. Otherwise, the study [13] combined between the ECG and PPG (Photoplethysmogram), the latter uses lights to capture heart motion. Also, the study [14] fused ECG features, time domain, and frequency domain to detect driver fatigue.

2.2.2 System using electro encephalogram (EEG)

EEG signal is also reliable and trustworthy. It is associated with both brain and physical activities, that is, for each activity EEG recording changes in terms of frequency and magnitude. In the literature, several methods have used EEG to detect drowsiness. The study [15] used five types of entropies with

EEG signal to estimate fatigue level. The study [16] improved detection accuracy by analyzing EEG signal in a clustering brain network in order to extract more sensitive features from spatial and temporal dimension. Otherwise, the study [17] developed a detection indicators based on EEG signal and tested the system in low-voltage. The study [18] used a deep convolutional network to detect driver's states from EEG signal. The study [19] used a convolutional auto-encoder to combine Electro Encephalogram EEG and electrooculography EOG.

2.2.3 System using electromyograph (EMG)

EMG signal is also used to detect fatigue because it contains many transient components and it is related to muscle activity. For example, in the driving case we can capture both triceps and biceps movements. Many works have used EMG signal [20, 21]. The study [22] proposed a method based on time domain and frequency domain. First, they removed noise by using Butterworth filter and then they used amplitude, phase, and frequency to evaluate the muscle fatigue. The study [23] proposed a non-contact method by installing recording electrodes into car seat. They combined EMG signal, ECG signal, and performed analysis with Fast Independent Component Analysis (ICA) and digital filter [24]. The problem encountered with EMG signal is related to the noisy signal, all methods that use it must go first through a refinement phase.

2.3 Facial features-based methods

Facial expressions are used in emotion detection applications [25-28]. To predict the driver state, facial features methods typically rely on the detection and monitoring of movements in facial features.

While all non-intrusive methods utilize eye and mouth movement data [29], the distinguishing factors lie in the granularities of analysis, such as eye/mouth feature extraction or classification algorithms.

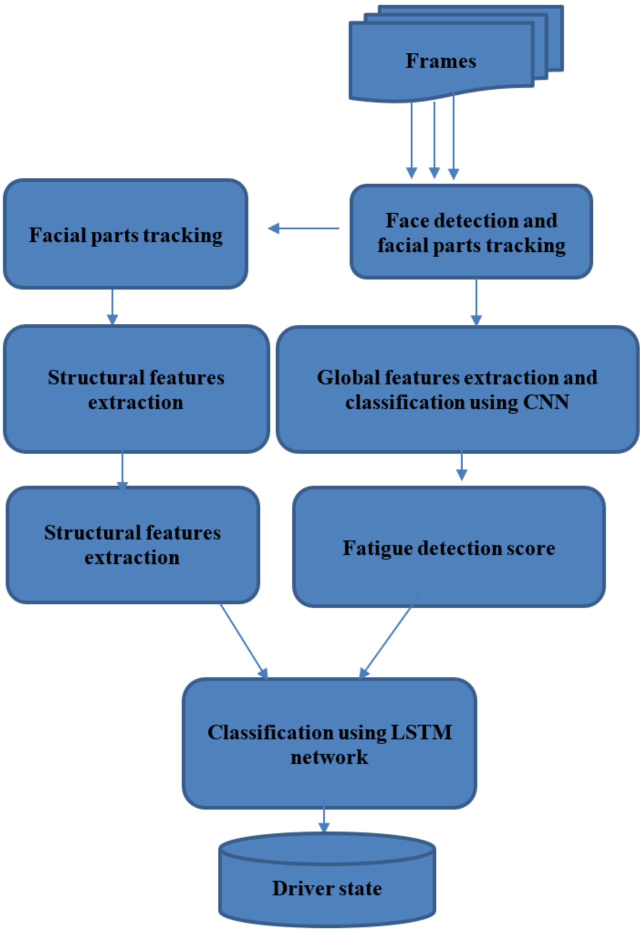
The study [6] presents a prominent approach utilizing three parameters. It partitions the driver's face into regions from which they extract descriptors such as HOG (Histogram of Oriented Gradients), covariance, and LBP (Local Binary Pattern). Subsequently, the extracted features undergo reduction through PCA (Principal Component Analysis) and Fisher score. To accomplish classification, it employs Support Vector Machines (SVM). Otherwise, the study [30] employs a deep learning approach for facial region detection, subsequently tracking eye closure (PERCLOS) and mouth aspect ratio (MAR). Driver fatigue is estimated by leveraging the selected features from the eyes and mouth. Also, the study [31] presents a system designed to identify drowsiness in drivers by employing Convolutional Neural Networks (CNNs) to extract features from the eyes and mouth. Its objective is to enhance existing systems that struggle to detect alcohol-consumed drivers, even when equipped with sensors. Another methodology adheres to a similar principle for driver state detection [32], focusing on features related to the mouth and eyes. The concept introduced in this study involves preprocessing all frames to enhance their quality. Subsequently, pixels' gradient orientation, along with CNN features, are extracted and integrated to form a feature space. Experiments conducted on the CELEB and YAW datasets yielded an accuracy score of 95%, achieved at a processing speed of 25 frames per second.

Deep learning methods frequently rise to the top due to their robust performance and ability to achieve high accuracy. A study presents a comparison between machine learning and deep learning approaches [33]. During the machine learning phase, they utilized EEG signals with various classifiers, including SVM, KNN, Gaussian NB, MLP, QDA, RF, and LR. Their comprehensive experiments revealed that SVM achieved the highest accuracy score. In the deep learning phase, three models were employed: CNN, 2D recurrent network, and a hybrid model combining both. The results indicated that CNN yielded the most effective solution. In their work [34], the authors introduced an innovative fatigue detection algorithm that utilizes integrated facial features and a Gate Recurrent Unit (GRU) judgment neural network. This algorithm was designed to effectively analyze contextual information spanning multiple image frames arranged chronologically. The extraction of features from facial feature points was employed, and a judgment network was developed by inputting change curves of six features over 20 consecutive frames. This approach enables real-time detection and provides output indicating the driver's fatigue status. Recently, the study [35] employs a shallow CNN architecture with a reduced number of layers to detect driver fatigue. Eye regions are first identified using a 68-point detection method, followed by the extraction of CNN features from these regions. Also, the study [30] examines the effectiveness of CNNs for yawning detection, they use five architectures DenseNet201, AlexNet, MobileNetv2, ResNet50, and VGG16 trained with yaw-DD dataset and they achieve that DenseNet201 and ResNet50 give the highest accuracy.

Facial features systems are non-contact and allow more comfort to the driver. Also, they are less expensive and easy to be implemented. However, these systems are not credible in real applications due to occlusion (driver wearing glasses), illumination, and head position. To remedy, the study [23] proposed 3D head motion estimation method that combines RGB data and depth data to decide on drive state. In the same way, the study [36] provided a review of recent investigations of the impact of occlusion for performance in facial expression analysis.

3. PROPOSED ARCHITECTURE

Several symptoms indicate fatigue, including the face becomes pale and looks tired, eyes almost closed, mouth often very open for yawning, and head is tilted. In this approach, we begin by face detection and facial parts extraction using Dlib library [37]. Next, to make a robust and reliable system, we combine two types of features. First, we use Alexnet for global features extraction and classification. Second, we extract structural features from face parts. The local features, combined with the AlexNet [38] network decision, are integrated into a unified vector. This vector is then input into the LSTM networks to predict the driver's state. The overview of the proposed architecture is shown in Figure 2. All execution steps and choice justifications are illustrated below. Then, we obtain 68 landmarks coordinates delimiting the face, eyes, eyebrows, and mouth. From these lasts, we can frame the face and track mouth, eyes, and eyebrows movements. We do not use all landmarks points, but only we use those that can help us to frame the face, detect the head pose, and track parts movements.



Frames
Face detection and facial parts tracking
Facial parts tracking
Structural features extraction
Global features extraction and classification using CNN
Structural features extraction
Fatigue detection score
Classification using LSTM network
Driver state

Figure 2. System architecture

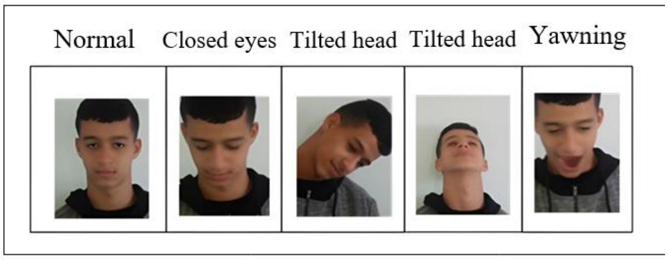


Figure 3. Facial expression and head position indicating fatigue or drowsiness

3.1 Face detection and facial parts tracking

Figure 3 shows the different head positions and facial expressions that indicate fatigue situation examples. Face detection is the first step in which we use the Dlib library, which is based on Support Vector Machine (SVM) and Histograms of Oriented Gradients. Dlib is widely utilized for detecting and tracking facial features, as well as for face detection and recognition. It is renowned for its robustness in handling variations in lighting conditions. Then, we obtain 68 landmarks coordinates delimiting the face, eyes, eyebrows, and mouth. From these lasts, we can frame the face and track mouth, eyes, and eyebrows movements. We do not use all landmarks points, but only we use those that can help us to frame the face, detect the head pose, and track parts movements. We can detect the face by drawing a two-dimensional rectangle, where the width spans from the leftmost to the rightmost point of the face, and the height extends from the lowest point to the upper point (eyebrow level). Additionally, 20% is added to the top to include the forehead in the frame (Figure 4).

In fact, we have to cover all situations that may arise when a driver is tired. The driver is considered tired when eyes are almost closed, frequent yawning occurs, strong tilt of head. Figure 5 shows facial landmarks points detected using Dlib library and those chosen for the next step.

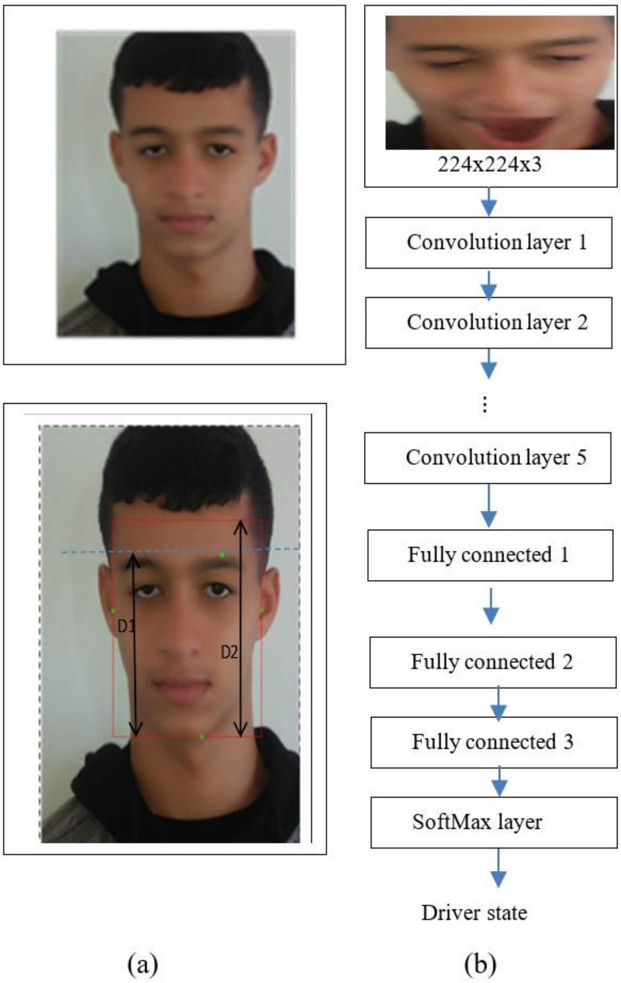


Figure 4. Face detection and global features extraction (a) detecting benchmark points, D_1 is the distance between upper and bottom points D_2=D_1+20% (b) cropped face and Alexnet architecture

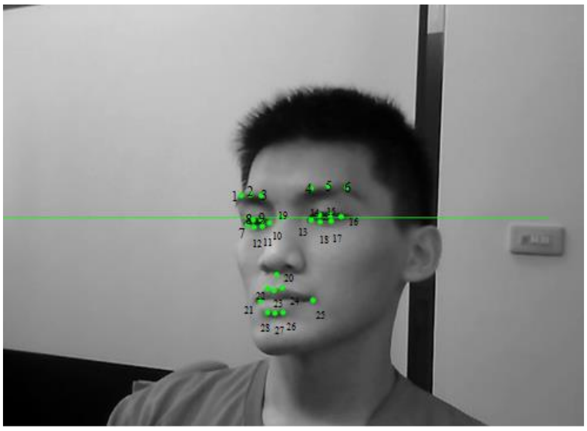


Figure 5. Fiducial points used to extract Structural features from facial parts

We crop the image located between the rightmost, leftmost, the undermost, and the uppermost (by adding 20% to include the forehead and eyebrows) points.

3.2 Features extraction

When a human is being tired, this is easily noticed on the features of his or her face (overall appearance). In addition, a weary person often opens the mouth (yawning), closes eyes, shrinks eyebrows and cannot keep his or her head straight. To take all this into account, we use features extracted from facial parts dimension and we combine them with global features to overcome luck of information due to occlusion wearing glasses, and so on.

Actually, features extraction is performed in two stages. We focus on contextual information, which makes the global features, and parts movement information, which makes the structural features. Then, the combination of both features provides a refined and more precise class prediction. We should also take into consideration the timing to avoid the false positive detection. For example, a laughing person also opens his or her mouth, while others do a frequent wink without being tired. To remedy it, we use the LSTM network for classification.

3.2.1 Global features extraction

After detecting the face, we extract the global features using CNN. CNNs are amply involved in the most recent computer vision applications [39-43]. The deep learning mimics human brain process, in such networks an image is directly used without passing by feature extraction step.

Training CNNs needs a huge amount of data. The reuse of pre-trained models makes the job easier. In fact, in transfer learning we benefit of knowledge obtained from learning massive datasets such as ImageNet [13, 15], Alexnet.

To extract the first features vector, we use Alexnet architecture, which is commonly employed in intricate situations. The AlexNet model consists of a total of eight layers, comprising five convolutional layers and three fully connected layers.

As shown in Figure 5, we use some landmarks point to estimate the face location and then we crop the face between the bottom point and the upper, by adding 20% to include forehead and eyebrows, and the lateral points. The captured facial image is resized to dimensions of 224×224×3 and then inputted into the initial layer, undergoing filtration by 96 kernels sized 11×11×3 with a stride of 4 pixels each. In the second layer we use 256 kernels of size 5×5×48. The outputs of the first and second layers are passed through max pooling layers. The third and fourth layers are equipped with 384 kernels each, and the fifth layer is characterized by 256 kernels. The following three layers, each with a size of 4096, are fully connected. Next, the SoftMax classifier is employed to generate the intended one output. Figure 4(b) provides additional clarification.

3.2.2 Structural features extraction

Second promising features must also be used, they represent facial parts movements. Human emotional state is expressed by facial part state: eyes and mouth (opened, closed, wink, agape), eyebrows (raised, tight). For example, yawning induces mouth opening and eyes closing, also when a person laughs his or her mouth is opened and his or her eyes are closed but in a different manner.

In this work, we track the eyes, mouth, and eyebrows movement as well as head position. We are motivated by those movements because they well express the fatigue state i.e. a tired person same time opens often his or her mouth for yawning, he or she can also close his or her eyes and in the he or she cannot keep his or her head straight, thereby he or she tilts it forward or back. Figure 5 shows all structural features extracted from the face.

As shown in Figure 5, we use twenty-eight points to calculate structural features that capture eyes, eyebrows, mouth movements, and head position. All features are grouped in a vector to be fed to classifier based on LSTM network.

In the following, we give in detail all steps used to extract six features i.e., left/right eyebrow movement LEBM/REBM, left/right eye movement LEM/REM, mouth opening MO, and head position angle.

Closed eye is a first promising factor to detect drowsiness, and by tracking eye movement we can decide if eye is closed or not. We use points, P_7..P_{12} to set up the right eye and P_{13} to P_{18} to set up the left eye. LEM and REM are given as follows:

$$LEM = \frac{|P_8 - P_{12}| + |P_9 - P_{11}|}{|P_7 - P_{10}|} \tag{1}$$

$$REM = \frac{|P_{14} - P_{18}| + |P_{15} - P_{17}|}{|P_{16} - P_{13}|} \tag{2}$$

Eyebrow movement is also a basic factor to describe emotional state. Drowsy and laughing person both open their mouths, but the first one shrinks his eyebrows and the last one stretch them. In this work we propose a metric based on distances between eyebrow detected points and the horizontal line passing through head point P_{19}. REBM (Right Eyebrow Movement) and LEBM (Left Eyebrow Movement) are calculated as follows:

$$LEBM = \frac{|P_1 - P_{19}| + |P_2 - P_{19}| + |P_3 - P_{19}|}{|P_1 - P_3|} \tag{3}$$

$$REBM = \frac{|P_4 - P_{19}| + |P_5 - P_{19}| + |P_6 - P_{19}|}{|P_4 - P_6|} \tag{4}$$

Generally*, a* drowsy person opens his mouth to breath oxygen. He is in yawning state. To track mouth movement, we use eight points P_{21}..P_{28} and the mouth movement is given as follows:

$$MM = \frac{|P_{22} - P_{18}| + |P_{23} - P_{27}| + |P_{24} - P_{26}|}{|P_{21} - P_{25}|} \tag{5}$$

The last parameter is the angle between the horizontal line and a line passing through P19 and P20 points. A high tilt angle means that a person is drowsy, for that, we choose tilt angle as the sixth and last parameters to make the features vector.

3.3 Video classification using LSTM network

Classification using recurrent neural network are widely used in machine translation, speech recognition and video sequences processing [44]. LSTM network is an improved version of them [45] in the fact that it fills the vanishing problem that occurs when treating long sequences. A hidden layer unit in the LSTM network is a bloc of memory cells and three gating units: input gate, output gate, and forget gate.

Each block receives as input the state of the previous frame S_{t-1} and the input at time t, X_t. Also, each gate receives the same features as the block input.

$$g(t)=\delta\left(\left(X\_t+S\_(t-1)\right)\times W\_(g)+b\_g\right) \tag{6}$$

$$Y\_in(t)=\delta\left(\left(X\_t+S\_(t-1)\right)\times W\_(i)+b\_i\right) \tag{7}$$

$$Y\_f(t)=\delta\left(\left(X\_t+S\_(t-1)\right)\times W\_(f)+b\_f\right) \tag{8}$$

$$C\_t=\left(C\_(t-1)\odot f\_(t)+g\odot Y\_in\right) \tag{9}$$

$$Y\_o(t)=\delta\left(\left(X\_t+S\_(t-1)\right)\times W\_(o)+b\_o\right) \tag{10}$$

$$S\_t=\tanh\left(C\_t\right)\odot Y\_in(t)\big) \tag{11}$$

For the classification purpose, a single LSTM network with forget gates is utilized to discard irrelevant features. To train the LSTM network for determining the driver's drowsiness state, the initial step involves extracting structural features and the CNN classification result for each frame. Subsequently, the features and decision score are consolidated into a unified vector, which is then input into the LSTM to predict the drowsiness or non-drowsiness of the entire video sequence.

The feature vector comprises seven parameters extracted from frames, namely (LEM, REM, LEBM, REBM, MM, HeadposeAngle, CNN_decision). We run 150 steps to generate a driver state decision in the output. If the output value is equal to or greater than 0.5, it indicates that the driver is in a drowsy state (refer to Figure 6).

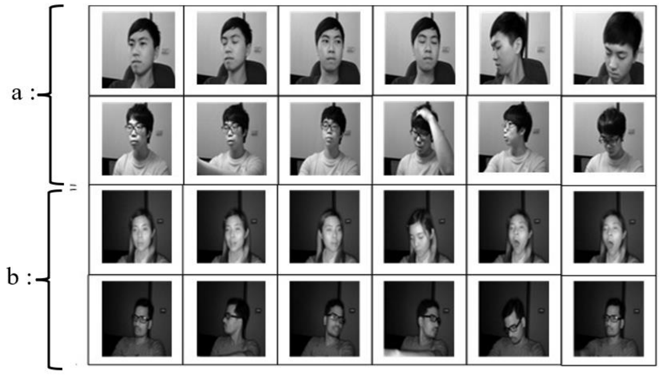


Figure 6. Samples of drowsy and normal frames showing different possible expressions, head positions, mouth and eyes opening (a) day time sequences (b) night-time sequences

4. EXPERIMENTS

4.1 Dataset

To perform experiments, we use Driver drowsiness detection dataset [9]. Video dataset is collected by NTHU Computer Vision Lab. It includes training, evaluation, and testing sets, features recordings of 36 subjects from diverse ethnic backgrounds, both with and without glasses/sunglasses, in various simulated driving scenarios. These scenarios range from normal driving to yawning, slow blinking, falling asleep, and laughing, under both day and night lighting conditions. Subjects were recorded while seated, playing a simple driving game using a simulated steering wheel and pedals, and were guided by an experimenter to perform specific facial expressions. The total duration of the dataset is approximately 9.5 hours.

The training set includes 18 subjects across five scenarios (BareFace, Glasses, Night_BareFace, Night_Glasses, Sunglasses). For each subject, sequences featuring yawning and slow blinking with nodding were recorded for about 1 minute each, while key scenarios involving drowsiness-related behaviors (yawning, nodding, slow blinking) and non-drowsiness-related actions (talking, laughing, looking to the sides) were recorded for 1.5 minutes each. The evaluation and testing sets consist of 90 driving videos from the remaining 18 subjects, with mixed drowsy and non-drowsy states across different scenarios.

An active infrared (IR) illumination was used to capture IR videos. The video resolution is 640x480 in AVI format. Videos in the Night_BareFace and Night_Glasses scenarios were recorded at 15 frames per second, while those in the BareFace, Glasses, and Sunglasses scenarios were recorded at 30 frames per second. The dataset is divided into training, evaluation, and testing sets, with testing videos comprising a mixture of different driving scenarios.

4.2 Proposed LSTM parameters

The experiments are conducted on a machine equipped with an i7 processor and an 8GB NVIDIA RTX GPU. Each video is decomposed into frames, with an average of 300 frames per video. We apply frame skipping, selecting one frame out of every 10. To achieve optimal classification performance, the LSTM network is configured as follows:

- The first LSTM layer consists of 128 units, designed to capture dependencies across time frames.
- The second, third, and fourth LSTM layers each have 64 units, focusing on identifying high-level patterns.
- Dropout is applied to all LSTM layers to prevent overfitting.
- A dense layer with 128 units follows the LSTM layers for further processing.
- A final SoftMax layer with one unit is used for binary classification: a value close to 1 indicates a drowsy driver, while a value close to 0 indicates an alert state.
- We put batch size=32, learning rate=0.001, and we run 150 epochs.

4.3 Evaluation measures

System performance evaluation is not only based on single frame accuracy, because it is not admissible in real application to decide about driver state 30 times per second. In our system, we propose a new routine, which helps us to improve the score by eliminating irrelevant detections. To do that, we consider a new state only when it appears at least in 30 frames. The following algorithm reads score vectors of thirty frames and makes a predominant decision.

Let “M” be the video size, “T” be a vector collecting real scores, and “State” contain the decision score.

```
$S \leftarrow 'Drowsy'; cpt \leftarrow 0;$
$Current_{frame} = 1$
$repeat$

while $(T[i] ==' Drowsy')$
    $cpt \leftarrow cpt + 1;$
    $Current_{frame} \leftarrow Current_{frame} + 1;$
    $if\ (cpt \geq 15)\ State = 'Drowsy'; cpt \leftarrow 0;$
    $else\ State =' NoDrowsy';$
$until\ end\ of\ T$
```

After performing refinement of score vector obtained, we calculate the global quadratic error estimated on all the frames. It is given in Eq. (7).

$$ERR = \frac{\sum_{i=1}^{M}(Out_i - T_i)^2}{M} \qquad (12)$$

Tests are performed on the mentioned dataset. To show the usefulness of our approach, we conduct tests as follows:

(1) We prepare both structural and CNN-decision score of all frames.

(2) First, we train LSTM network with only structural features.

(3) Second, we add CNN-decision to the structural features, then we train again the LSTM network.

(4) Finally, we compare the obtained results with other works that used the same dataset.

4.4 Structural features

First, we test with structural features and use the LSTM network to classify the frames. Table 2 shows the obtained results.

Table 2. Drowsiness detection using structural features

Scenario	Drowsiness Score
No Glasses	93.25%
Glasses	90.47%
Sunglasses	93.33%
Night-No-Glasses	80.00%
Night-Glasses	38.18%
Overall	79.05%

4.5 CNN-features classification

Table 3 presents the drowsiness detection scores achieved using global features.

Table 3. Drowsiness detection using global features

Scenario	Drowsiness Score
No Glasses	63.63%
Glasses	88.89%
Sunglasses	89.38%
Night-No-Glasses	75.00%
Night-Glasses	71.51%
Overall	77.68%

4.6 Merged features

In the third step, we merge both scores of CNN-networks, which get features from the hole face, and the structural obtained futures. Then we train the LSTM network to get a final decision. Results are shown in Table 4.

Table 4. Drowsiness detection using merged structural features with CNN decision

Scenario	Drowsiness Score
No Glasses	96.77%
Glasses	90.57%
Sunglasses	95.74%
Night-No-Glasses	85.71%
Night-Glasses	81.81%
Overall	90.12%

4.7 Result analysis

Structural features alone are not sufficient and give an overall accuracy of 79.05%. We can see in Table 2 that drowsiness detection accuracy with diurnal videos is best than accuracy of nocturnal ones. This is explainable because structural features are extracted by tracking eyes, eyebrow and mouth movements, and in infrared image the poor quality of images hinders this process. We can see clearly in Figure 7 the series of sequences showing a person in drowsiness state, images from 1 to 6 reflect a person in normal state but really it is in drowsiness situation. Here, we deduce that structural features used alone are not efficient.

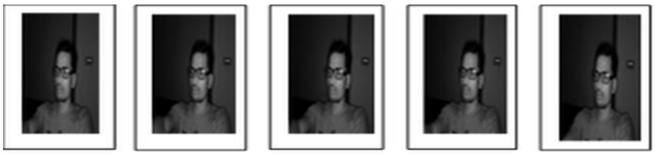


Figure 7. Sequence of infrared images chowing a drowsiness state

Secondly, the global features give different scores between 63.63% and 89.38%, which are satisfactory, except in night glasses case they give a score between 70% and 75%. CNN features have a good ability of expression, even for nocturne videos, and are less computationally expensive.

Finally, by combining both CNN-scores and structural features we obtain considerable results improvement as shown in Table 4. Also, the graph in Figure 8 summarizes all results and allows us to make a reasonable comparison.

The results above indicate that structural features provide a weak classification rate during nighttime, primarily due to the difficulty in accurately detecting features, making it challenging to determine the driver's state. Similarly, during the daytime, the system experiences a few misclassification instances. This is understandable because only a few video frames clearly display expressions associated with the driver’s state, i.e., many frames capture a drowsy person without overt signs such as yawning or closed eyes. The error stems from the algorithm used to determine the driver's state, which classifies the driver as drowsy after detecting 15 instances of drowsiness, i.e., after processing 150 frames. A potential solution could be to make this threshold dynamic, allowing it to adjust based on the situation.

By combining structural features with a global classification approach, we improved the night-time performance and addressed some misclassifications caused by drowsy states without clear facial expressions. See Figure 7.

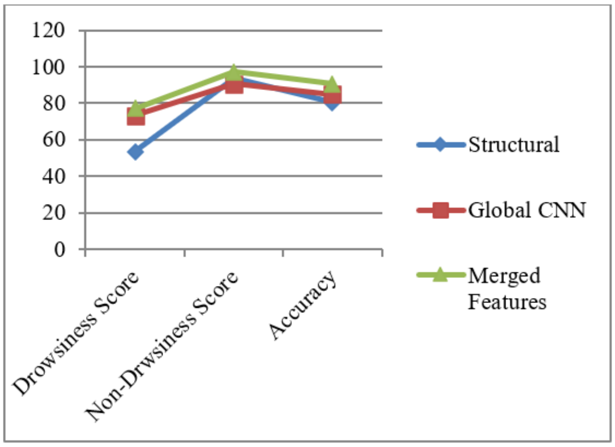


Figure 8. Graph summarizing the three steps results obtained

4.8 Comparison with other works

The proposed approach is compared with models used in [46] experiments. The authors propose a hierarchical temporal Deep Belief Network (HTDBN) that uses three deep networks to extract features from mouth, head, and eyes. Following that, the features vectors are grouped in a single one which, in turn, is regarded as observation vector for two HMMs i.e., drowsiness HMM and non-drowsiness HMM. Then in their experiment, they propose four different scenarios to evaluate their system:

(1) SVM+SVM: in that event, they use one SVM to extract features and other one to do classification

(2) SVM+HMM: here SVM is used to extract features and HMM to classify a sequence

(3) DBN + SVM Deep Belief Network (DBN) is used to extract deep features and a binary class SVM to detect drowsiness state

(4) HTDBN

Table 5 summarizes all results and the graph in Figure 9 allows us to make a comparison between all models.

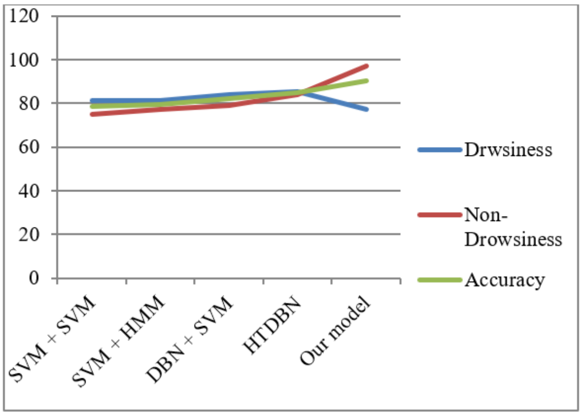


Figure 9. Comparison between our model and other solutions

Table 5. Comparison between our model and other solutions

Model	Drowsiness Score
SVM+SVM	81.16%
SVM+HMM	81.30%
DBN+SVM	84.26%
HTDBN	85.39%
Our model	90.12%

The global accuracy of our model exceeds that of other approaches. The goal of this research is to guide a driver of vehicle based on his or her state, then we have to improve our features detection process that shows weakness with nighttime videos.

5. CONCLUSION

In this work, we proposed a method to guide vehicle drivers throughout of their journey. We developed a hybrid approach based on deep learning and recurrent network that combined structural and global features. We employed an LSTM network that takes structural features, delineating parts movements, as input, and integrated it with a CNN network decision. Combining structural features with CNN decisions can mitigate nighttime misclassification and ensure accurate detection, even when facial expressions do not fully reflect the true state of drowsiness. Our study demonstrates that structural features are robust but not sufficient to decide about driver state. Additionally, the proposed classification algorithm requires further improvements to enhance the accuracy rate.

The findings offer a new idea in driver drowsiness that allow combination between two light systems. The experiments were conducted using each type of feature independently, as well as with both feature types combined. Results showed that combined results give considerable improvements. Also, tests performed on driver drowsiness detection dataset showed satisfactory results compared to other methods.

Although the results are promising, our approach remains inefficient and requires improvements when handling nighttime videos and certain unexpressed situations. Additionally, further research with larger sample sizes is necessary to validate these findings across a broader range of populations.

REFERENCES

[1] Sayed, R., Eskandarian, A. (2001). Unobtrusive drowsiness detection by neural network learning of driver steering. Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering, 215(9): 969-975. https://doi.org/10.1243/0954407011528536

[2] Gromer, M., Salb, D., Walzer, T., Madrid, N.M., Seepold, R. (2019). ECG sensor for detection of driver's drowsiness. Procedia Computer Science, 159: 1938-1946. https://doi.org/10.1016/j.procs.2019.09.366

[3] World Health Organization. (2019). Global status report on road safety 2018. World Health Organization.

[4] Zhang, Z., Zhang, J.S. (2006). Driver fatigue detection based intelligent vehicle control. In 18th International Conference on Pattern Recognition (ICPR'06), Hong Kong, China, 2: 1262-1265. https://doi.org/10.1109/ICPR.2006.462

[5] Bertozzi, M., Broggi, A., Cellario, M., Fascioli, A., Lombardi, P., Porta, M. (2002). Artificial vision in road vehicles. Proceedings of the IEEE, 90(7): 1258-1271. https://doi.org/10.1109/JPROC.2002.801444

[6] Moujahid, A., Dornaika, F., Arganda-Carreras, I., Reta, J. (2021). Efficient and compact face descriptor for driver drowsiness detection. Expert Systems with

Applications, 168: 114334. https://doi.org/10.1016/j.eswa.2020.114334

[7] Doudou, M., Bouabdallah, A., Berge-Cherfaoui, V. (2020). Driver drowsiness measurement technologies: Current research, market solutions, and challenges. International Journal of Intelligent Transportation Systems Research, 18: 297-319. https://doi.org/10.1007/s13177-019-00199-w

[8] Zhang, L., Verma, B., Tjondronegoro, D., Chandran, V. (2018). Facial expression analysis under partial occlusion: A survey. ACM Computing Surveys (CSUR), 51(2): 1-49. https://doi.org/10.1145/3158369

[9] Chausse, F., Aufrere, R., Chapuis, R. (2000). Vision based vehicle trajectory supervision. In ITSC2000. 2000 IEEE Intelligent Transportation Systems. Proceedings (Cat. No.00TH8493), Dearborn, MI, USA, pp. 143-148. https://doi.org/10.1109/ITSC.2000.881039

[10] Coetzer, R.C., Hancke, G.P. (2009). Driver fatigue detection: A survey. In AFRICON 2009, Nairobi, Kenya, pp. 1-6. https://doi.org/10.1109/AFRCON.2009.5308101

[11] Shi, S.Y., Tang, W.Z., Wang, Y.Y. (2017). A review on fatigue driving detection. In ITM Web of Conferences. EDP Sciences, 12: 01019. https://doi.org/10.1051/itmconf/20171201019

[12] Vicente, J., Laguna, P., Bartra, A., Bailón, R. (2016). Drowsiness detection using heart rate variability. Medical & Biological Engineering & Computing, 54: 927-937. https://doi.org/10.1007/s11517-015-1448-7

[13] Dai, J., Li, Y., He, K., Sun, J. (2016). R-FCN: Object detection via region-based fully convolutional networks. In Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS'16). Curran Associates Inc., Red Hook, NY, USA, pp. 379-387.

[14] Zhu, T., Zhang, C., Wu, T., Ouyang, Z., Li, H., Na, X., Liang, J., Li, W. (2022). Research on a real-time driver fatigue detection algorithm based on facial video sequences. Applied Sciences, 12(4): 2224. https://doi.org/10.3390/app12042224

[15] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L. (2015). ImageNet large scale visual recognition challenge. International Journal of Computer Vision, 115: 211-252. https://doi.org/10.1007/s11263-015-0816-y

[16] Zhang, C., Sun, L., Cong, F., Kujala, T., Ristaniemi, T., Parviainen, T. (2020). Optimal imaging of multi-channel EEG features based on a novel clustering technique for driver fatigue detection. Biomedical Signal Processing and Control, 62: 102103. https://doi.org/10.1016/j.bspc.2020.102103

[17] Balasubramanian, K., Ramya, K., Devi, K.G. (2022). Improved swarm optimization of deep features for glaucoma classification using SEGSO and VGGNet. Biomedical Signal Processing and Control, 77: 103845. https://doi.org/10.1016/j.bspc.2022.103845

[18] Zeng, H., Yang, C., Dai, G., Qin, F., Zhang, J., Kong, W. (2018). EEG classification of driver mental states by deep learning. Cognitive Neurodynamics, 12: 597-606. https://doi.org/10.1007/s11571-018-9496-y

[19] Jing, D., Liu, D., Zhang, S., Guo, Z. (2020). Fatigue driving detection method based on EEG analysis in low-voltage and hypoxia plateau environment. International Journal of Transportation Science and Technology, 9(4): 366-376. https://doi.org/10.1016/j.ijtst.2020.03.008

[20] Moshou, D., Hostens, I., Papaioannou, G., Ramon, H. (2005). Dynamic muscle fatigue detection using self-organizing maps. Applied Soft Computing, 5(4): 391-398. https://doi.org/10.1016/j.asoc.2004.09.001

[21] Boon-Leng, L., Dae-Seok, L., Boon-Giin, L. (2015). Mobile-based wearable-type of driver fatigue detection by GSR and EMG. In TENCON 2015-2015 IEEE Region 10 Conference, Macao, China, pp. 1-4. https://doi.org/10.1109/TENCON.2015.7372932

[22] Mohd Azli, M.A.S., Mustafa, M., Abdubrani, R., Abdul Hadi, A., Syed Ahmad, S.N.A., Zahari, Z.L. (2019). Electromyograph (EMG) signal analysis to predict muscle fatigue during driving. In Proceedings of the 10th National Technical Seminar on Underwater System Technology 2018: NUSYS'18, Springer Singapore, pp. 405-420. https://doi.org/10.1007/978-981-13-3708-6_35

[23] Fu, R., Wang, H. (2014). Detection of driving fatigue by using noncontact EMG and ECG signals measurement system. International Journal of Neural Systems, 24(03): 1450006. https://doi.org/10.1142/S0129065714500063

[24] Comon, P. (1994). Independent component analysis, A new concept? Signal Processing, 36(3): 287-314. https://doi.org/10.1016/0165-1684(94)90029-9

[25] Benmohamed, A., Neji, M., Ramdani, M., Wali, A., Alimi, A.M. (2015). Feast: Face and emotion analysis system for smart tablets. Multimedia Tools and Applications, 74: 9297-9322. https://doi.org/10.1007/s11042-014-2082-3

[26] Zhang, Y., Guo, H., Zhou, Y., Xu, C., Liao, Y. (2023). Recognising drivers' mental fatigue based on EEG multi-dimensional feature selection and fusion. Biomedical Signal Processing and Control, 79: 104237. https://doi.org/10.1016/j.bspc.2022.104237

[27] Kim, C.L., Kim, B.G. (2023). Few-shot learning for facial expression recognition: A comprehensive survey. Journal of Real-Time Image Processing, 20(3): 52. https://doi.org/10.1007/s11554-023-01310-x

[28] Zhang, L., Liu, F.A.N., Tang, J. (2015). Real-time system for driver fatigue detection by RGB-D camera. ACM Transactions on Intelligent Systems and Technology (TIST), 6(2): 1-17. https://doi.org/10.1145/2629482

[29] Chhimpa, G.R., Kumar, A., Garhwal, S., Dhiraj. (2023). Development of a real-time eye movement-based computer interface for communication with improved accuracy for disabled people under natural head movements. Journal of Real-Time Image Processing, 20(4): 81. https://doi.org/10.1007/s11554-023-01336-1

[30] Rahmawati, Y., Ardiyanto, I., Nugroho, H.A. (2024). Comparative study of yawn classification on CNN architectures. In 2024 IEEE International Conference on Artificial Intelligence and Mechatronics Systems (AIMS), Bandung, Indonesia, pp. 1-6. https://doi.org/10.1109/AIMS61812.2024.10513022

[31] Ashlin Deepa, R.N., Sai Rakesh Reddy, D., Milind, K., Vijayalata, Y., Rahul, K. (2023). Drowsiness detection using IoT and facial expression. In Proceedings of the International Conference on Cognitive and Intelligent Computing: ICCIC 2021, Springer, Singapore, 2: 679-692. https://doi.org/10.1007/978-981-19-2358-6_61

[32] Deng, W., Wu, R. (2019). Real-time driver-drowsiness detection system using facial features. IEEE Access, 7:

118727-118738. https://doi.org/10.1109/ACCESS.2019.2936663

[33] Alharbey, R., Dessouky, M.M., Sedik, A., Siam, A.I., Elaskily, M.A. (2022). Fatigue state detection for tired persons in presence of driving periods. IEEE Access, 10:79403-79418. https://doi.org/10.1109/ACCESS.2022.3185251

[34] Li, D., Zhang, X., Liu, X., Ma, Z., Zhang, B. (2023). Driver fatigue detection based on comprehensive facial features and gated recurrent unit. Journal of Real-Time Image Processing, 20(2): 19. https://doi.org/10.1007/s11554-023-01260-4

[35] Venkateswarlu, M., Ch, V.R.R. (2024). DrowsyDetectNet: Driver drowsiness detection using lightweight CNN with limited training data. IEEE Access, 12: 110476-110491. https://doi.org/10.1109/ACCESS.2024.3440585

[36] Bhardwaj, R., Natrajan, P., Balasubramanian, V. (2018). Study to determine the effectiveness of deep learning classifiers for ECG based driver fatigue classification. In 2018 IEEE 13th International Conference on Industrial and Information Systems (ICIIS), Rupnagar, India, pp. 98-102. https://doi.org/10.1109/ICIINFS.2018.8721391

[37] Kazemi, V., Sullivan, J. (2014). One millisecond face alignment with an ensemble of regression trees. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1867-1874. https://doi.org/10.1109/CVPR.2014.241

[38] Dua, M., Shakshi, Singla, R., Raj, S., Jangra, A. (2021). Deep CNN models-based ensemble approach to driver drowsiness detection. Neural Computing and Applications, 33: 3155-3168. https://doi.org/10.1007/s00521-020-05209-7

[39] Yang, Y., Sun, Y. (2017). Facial expression recognition based on arousal-valence emotion model and deep learning method. In 2017 International Conference on Computer Technology, Electronics and Communication (ICCTEC), Dalian, China, pp. 59-62. https://doi.org/10.1109/ICCTEC.2017.00022

[40] Ekman, P., Friesen, W.V. (1971). Constants across cultures in the face and emotion. Journal of Personality and Social Psychology, 17(2): 124-129. https://psycnet.apa.org/doi/10.1037/h0030377

[41] Russell, J.A. (1980). A circumplex model of affect. Journal of Personality and Social Psychology, 39(6): 1161-1178. https://psycnet.apa.org/doi/10.1037/h0077714

[42] Chatfield, K., Simonyan, K., Vedaldi, A., Zisserman, A. (2014). Return of the devil in the details: Delving deep into convolutional nets. Computer Science. arXiv Preprint arXiv: 1405.3531. https://doi.org/10.48550/arXiv.1405.3531

[43] Leonardi, G., Montani, S., Striani, M. (2022). Novel deep learning architectures for haemodialysis time series classification. International Journal of Knowledge-based and Intelligent Engineering Systems, 26(2): 91-99. https://doi.org/10.3233/KES220010

[44] Guo, Z., Yang, C., Wang, D., Liu, H. (2023). A novel deep learning model integrating CNN and GRU to predict particulate matter concentrations. Process Safety and Environmental Protection, 173: 604-613. https://doi.org/10.1016/j.psep.2023.03.052

[45] Ramezanpanah, Z., Mallem, M., Davesne, F. (2022). Autonomous gesture recognition using multi-layer LSTM networks and laban movement analysis. International Journal of Knowledge-based and Intelligent Engineering Systems, 26(4): 289-297. https://doi.org/10.3233/KES-208195

[46] Weng, C.H., Lai, Y.H., Lai, S.H. (2017). Driver drowsiness detection via a hierarchical temporal deep belief network. In Computer Vision – ACCV 2016 Workshops. ACCV 2016. Lecture Notes in Computer Science. https://doi.org/10.1007/978-3-319-54526-4_9