A Complex CNN&LSTM Neural Network Model for Determining a Person's Drowsiness State by Their Face

Anatolii Nikolenko[1,*,†], Olena Arsirii[1,†], Svitlana Antoshchuk and Oksana Babilunha[1,†]

[1]*Odesa Polytechnic National University, Shevchenko Ave. 1, 65044, Odesa, Ukraines*

Abstract

The article analyzes the capabilities of modern Advanced Driver Assistance Systems (ADAS) from well-known car manufacturers. It is shown that the main drawback of ADAS is the rather low accuracy of recognizing a person's drowsiness from frames and video sequences of facial images. In addition, the shortcomings of computer vision systems related to industrial safety are shown, when a tired worker is a source of increased risk in such areas as metallurgy, the chemical industry, energy, etc. On the other hand, the capabilities of neural network models of classical convolutional neural networks and their hybrids with SVM and Random Forest layers for processing additional sensory signals for binary classification of drowsiness are analyzed. The prospects for using recurrent neural networks RNN and LSTM for recognizing a person's drowsiness from video sequences of human images are shown. In the work, the architecture of a complex neural network model CNN&LSTM is proposed for determining the state of human drowsiness, which combines a multilayer convolutional neural network (CNN) for analyzing the context of images and a model with long short-term memory (LSTM) for analyzing sequential data on the state of human drowsiness over time. The combination of multilayer CNN and LSTM allows combining the advantages of both approaches. It is proposed to increase the accuracy of determining the state of human drowsiness at the LSTM input, in addition to the data obtained for each frame of the video sequence as a result of its processing by CNN, to additionally transmit the values of the EAR, MAR, PUC, MOE indicators calculated by MediaPipe Face Mesh. The complex CNN&LSTM model is implemented in the Python language using Keras/TensorFlow in the Google Colab environment. CNN&LSTM research using Drowsiness Detection Dataset version 1 and video sequences of EAR, MAR, PUC, MOE values showed an increase in the accuracy of determining drowsiness using the Accuracy and F1 metrics to 96-97%.

Keywords

Advanced Driver Assistance Systems, CNN, LSTM, drowsiness state, MediaPipe Face Mesh, classification

1. Introduction

Solving the problem of recognizing drowsiness from a human face image is relevant in computer vision systems related to transport safety, which depends, among other things, on timely determination of the degree of driver fatigue; with industrial safety, when a tired worker is a source of increased risk in such areas as metallurgy, chemical industry, energy, etc. Analysis of the state of the eyes may be the only available biomarker in patient condition monitoring systems in intensive care, as well as in user condition monitoring systems in smartphones, laptops and VR/AR headsets for automatic blocking or pausing in case of fatigue while viewing content.

It is known that the use of modern neural network models with deep learning (Deep Learning) opens up new opportunities for increasing the accuracy of detecting the state of the driver, operator, patient and other users of such computer vision systems. Therefore, the current task, which is solved within the framework of this work, is the analysis of the capabilities of recognizing human face images using the architecture of a classical convolutional neural network and its hybrid with additional layers of SVM and Random Forest for processing additional sensory signals for binary classification of the state of drowsiness.

ICST-2025: Information Control Systems & Technologies, September 24-26, 2025, Odesa, Ukraine

*Corresponding author.

†These authors contributed equally.

anatolyn@ukr.net(A. Nikolenko); e.arsiriy@gmail.com(O. Arsirii); asgonpu@gmail.com(S. Antoshchuk); babilunga.onpu@gmail.com(O. Babilunha)

0000-0002-9849-1797(A. Nikolenko); 0000-0001-8130-9613 (O. Arsirii); 0000−0002−9346−145X (S. Antoshchuk); 0000−0001−6431−3557 (O. Babilunha)

As well as the development of a complex neural network model that combines the advantages of convolutional networks (CNN) with long short-term memory (LSTM), which allows detecting the process of slow eye closure over several frames.

The creation of such a complex CNN&LSTM model will allow to increase the accuracy of determining the state of drowsiness from images and video sequences of a person's face by identifying long-term dependencies, which is important in preventing false positives of the classifier.

2. Literature overview

Recognition of human drowsiness is important in computer vision systems aimed at driver safety, because driver drowsiness is one of the main causes of road accidents worldwide. Eye monitoring systems using video analysis of a person's face allow them to recognize closed or half-closed eyes, determine the degree of fatigue, and generate alarms in real time. Such technologies are actively implemented in ADAS (Advanced Driver Assistance Systems) [1, 2, 3, 4, 5].

One of the well-known ADAS is the Audi Rest Recommendation System, which is an innovative feature introduced by Audi to improve the comfort and well-being of drivers and passengers during long journeys. The system uses various technologies, including sensors and artificial intelligence algorithms, to analyze driver behavior, external factors such as traffic and road conditions, and vehicle status data [1]. Unfortunately, the system overloads the driver with unnecessary information and the accuracy of driver status recognition is 92%.

BMW Active Driving Assistant with Attention Assistant analyses driving behaviour and, if necessary, advises the driver to take a break. Advice on taking a break is provided in the form of graphic symbols on the control display [2]. The system is sensitive to the angle of the driver's face in relation to the control panel and to adverse weather conditions, with an accuracy of 90% in detecting driver drowsiness.

The Bosch Driver Drowsiness Detection system [3] is a proactive approach to reducing the risks of drowsy driving using advanced sensors, algorithms and real-time analysis of driver behavior and intervention if necessary. The system has a disadvantage of false positives. The accuracy of the system is 91%.

Ford Driver Alert is a driver assistance feature designed to help prevent accidents caused by driver fatigue or inattention [4]. Ford Driver Alert is specifically designed to detect signs of driver fatigue or drowsiness. The system uses sensors and cameras to monitor the driver's behavior while driving. They track various parameters such as steering, lane departure, vehicle speed and even time of day to assess the driver's level of attentiveness. The system looks for patterns in the driver's behavior that may indicate reduced alertness, such as uneven steering movements, delayed reactions or lane departure. One of the system's shortcomings is that the sensors do not read fatigue as a condition well, focusing only on driving patterns. The system has an accuracy of 86%.

Honda's Driver Alert System uses Eye-Tracking technology to track the driver's eye movements and eyelid movements to detect signs of fatigue. By analyzing patterns such as prolonged eye closure or frequent blinking, the system can detect signs of drowsiness or distraction [5]. Honda can use machine learning algorithms to analyze data from various sensors and cameras in real time. These algorithms can learn patterns of driver behavior and identify deviations that may indicate fatigue or decreased attention. The system's shortcomings include the fact that the system can distract the driver with an overloaded interface and can block the car if there is no response to the warning. The system has an accuracy of 93%.

In computer vision systems aimed at industrial safety in metallurgical, chemical, and energy environments, worker state recognition modules operate using cameras in rooms or in helmets, analyzing eyes in real time. The CORE for Tech software product [6] detects states of lethargy, drowsiness, and fatigue in high-risk environments based on biometric indicators of heart rate variability (HRV), skin conductance, eye movements, and algorithms that allow early detection and intervention. The disadvantage is the use of inconvenient sensors that affect user comfort. The module's recognition accuracy is 92%.

An example of monitoring the user's state in mobile and consumer applications in smartphones, laptops and automatic blocking in case of fatigue is the use of the Vigo smart Bluetooth headset [7] . Vigo is equipped with sensors that track the user's eye and head movements in real time, recording factors such as blinking, duration of eye closure and head nodding. The disadvantage is the mandatory wearing of a headset. The system's recognition accuracy is 91%.

Analysis of the above computer vision systems related to security shows that the accuracy of recognizing the user's drowsiness depends on the quality of neural network models and machine and deep learning algorithms, which are the main components of such system. Convolutional networks in their modern form appeared in the works of Jan LeCun's group at the end of the 1980s [8, 9, 10], and since then they are quite successfully used for image recognition and many other tasks [11, 12]. An overview of methods and neural networks of deep learning is given in [13].

Combining CNN with other types of neural networks can improve recognition results. In particular, recurrent neural network (RNN) and long short-term memory network (LSTM) are used to solve the task of determining the state of drowsiness behind her face. The biggest advantage of CNN is the extraction of features of the eye image, while RNN can effectively obtain information about the timing of the blinking process [14]. This ability allows RNN to take into account the previous state of the eye in the detection of blinking to better determine whether blinking is currently occurring. And because the act of blinking is sequential, RNN can effectively extract temporal features such as the duration and frequency of blinks. This helps to more accurately distinguish between blinking and other similar actions [15, 16]. However, RNN and LSTM are sequential. Each time step must be calculated sequentially, which leads to low computational efficiency. In particular, this will lead to some delay when working with longer sequences or real-time applications [17].

3. Research aim statement

The aim of the work is to increase the accuracy of determining the state of drowsiness from images and video sequences of a human face with the additional use of EAR, MAR, PUC, MOE indicators calculated using MediaPipe Face Mesh by developing a complex neural network model based on CNN and LSTM.

The development of this complex neural network model consists of the following stages:

- Creating a 3D model of a human face based on "landmarks" using the Mediapipe library to obtain control points for further calculations of EAR, MAR, PUC, MOE.
- Development of the architecture of a complex neural network model CNN+LSTM for combining drowsiness recognition from image and video sequences.
- Conducting an experimental study on the accuracy of binary classification of drowsiness from human facial images with different neural network models based on a classical convolutional neural network.Methodology for developing a CNN&LSTM complex neural network model and studying its accuracy in recognizing drowsiness ront matter

3.1. Developing a 3D model of a human face based on "landmarks" using the MediaPipe library information

The Drowsiness Detection Dataset [18], which is based on the MRL [19] and Closed Eyes in Wild (CEW) [20, 21] datasets was selected as the input data for the study. This large-scale dataset, containing images of both closed and open human eyes, can be used for eye detection on faces, and in addition, for drowsiness detection (Figure 1). The datasets contain images of faces with closed and open eyes without glasses.

Figure 1:Examples of open and closed eyes in human face images [19, 21].

The images for the dataset were acquired at different lighting, distance, resolution, face angle, and eye angle parameters. There are different versions of the dataset [19]. Version 1 contains 10,000 images, divided into 5,000 images for closed and open eyes. The degree of eye openness is determined using the numerical measure of eye aspect ratio (EAR). This measure is key for many computer vision applications, such as detecting blinks, drowsiness, or assessing human attention. The eye is represented by a hexagon (Fig. 2) with values P1–P6, with points P1 and P4 defining the extreme left and right corners of the eye, P2 and P3 defining the two upper corners, and P6 and P5 defining the two lower corners of the eye, respectively. The EAR measure is calculated as

$$EAR = \frac{|P_2 - P_6| + |P_3 - P_5|}{2 \cdot |P_1 - P_4|}. \quad (1)$$

If the EAR value is close to 0, then this is the case of a closed eye [23]. Fig. 2 shows the tracking of the EAR value (ordinate axis) over a certain time, represented by the abscissa axis. We see that when the eye is closed, the EAR value drops sharply, almost to zero.

But for the final determination of the state of drowsiness, it is necessary to distinguish it from ordinary blinking. To do this, the duration of eye closure for blinking is determined. This interval on the abscissa axis is 200–400 msec, and for drowsiness – from 500 msec to several seconds (the state of prolonged eye closure PERCLOS – PERcentage of eye CLOSure). In addition, the state of drowsiness is often accompanied by other visible signs of fatigue, namely yawning (frequent and wide opening of the mouth), head tilts, facial micromovements, instability of posture, lack of focus of gaze.

Drowsiness detection systems use various metrics, algorithms, and technologies to analyze these features. For example, for efficient feature calculation, a 3D model of the face surface is used, created based on a regression approach [24, 25]. Technologically, such a model is created using the open computer vision and machine learning library from Google MediaPipe [26]. To detect drowsiness on a person's face, MediaPipe creates a mesh of 468 landmarks, called the Media Pipe Face Mesh. This comprehensive solution from MediaPipe allows you to accurately determine 468 3D coordinates (or landmarks) on the face surface in real time. These landmarks cover the contour of the face, eyebrows, eyes, nose, and mouth, providing a detailed representation of facial expressions and shape (Figure 3, a).

Determination of drowsiness using the created Media Pipe Face Mesh is based on the EAR, MAR, PUC, and MOE indicators, which indicate fatigue or drowsiness. Let's consider each of these indicators in more detail.

As mentioned, EAR (1) is a key indicator that measures the degree of eye openness. Its value ranges from high for an open eye to low for a closed eye (Fig. 3, b and c). In addition, the duration is the tracking of the number of blinks per minute. An extremely high or, conversely, an extremely low blink rate is an indicator of fatigue.

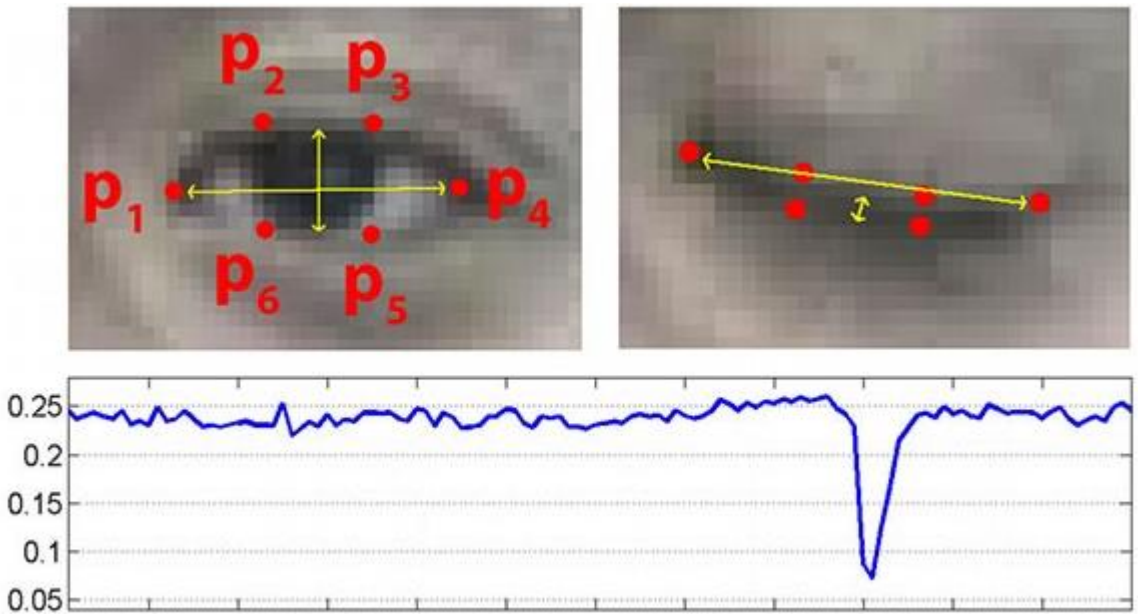
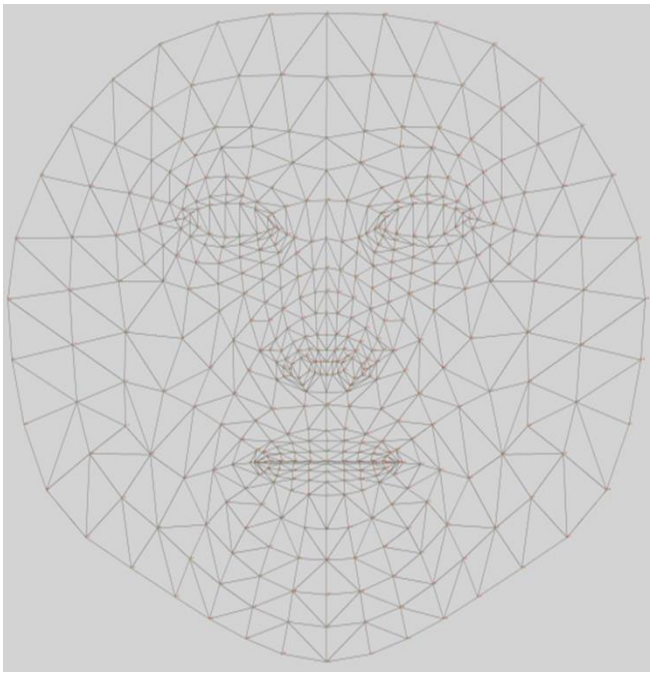
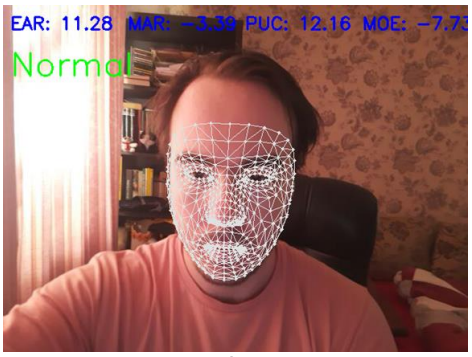


Figure 2: Example of determining drowsiness using the EAR indicator [27].



a



b

c

Figure3: Example of determining drowsiness based on EAR, MAR, PUC, MOE indicators on Media Pipe Face Mesh (a – Media Pipe Face Mesh view, b – normal state, c – drowsiness state).

A significant indicator of sleepiness is the Mouth Aspect Ratio (MAR), which measures the degree of mouth openness and is calculated similarly to EAR, but for landmarks around the mouth (usually 12-20 landmarks are used). At the same time, a high MAR value over a certain period of time indicates yawning (see Fig. 3, b and c), which is one of the most obvious and direct physical signs of sleepiness.

Pupil Circularity (PUC) is a measure of the degree of roundness of the pupil or the overall shape of the pupil and iris. It can be calculated based on the distances from the center of the pupil to points along its contour using the Media Pipe Face Mesh landmarks. The shape of the pupil and its apparent size are change for a sleepy person. A decrease in the PUC (or a decrease in its apparent diameter/area) is an indicator that the eye is not fully open. It complements the EAR by allowing the detection of "heavy" eyelids or a decrease in the pupil aperture, which often occurs with fatigue. It

should be noted that the calculation of PUC is more complex than EAR and MAR, as it requires more accurate detection of the pupil and is sensitive to lighting and image quality.

Mouth Over Eye Ratio (MOE) is a composite measure that combines MAR and EAR. It is calculated as the ratio of MAR to EAR (MOE = MAR / EAR). This measure helps filter out "false positives" from other facial expressions that only affect one of the measures, for example, smiling increases MAR but does not affect EAR. Calculating MOE helps account for the interaction between eye and mouth movements, making it a powerful aggregate measure for drowsiness classifiers.

It should be noted that practical applications of computer vision systems to determine human drowsiness usually use a combination of all the listed metrics and other behavioral features (e.g., head tracking, blink rate, gaze duration, etc.). These indicators are collected during a certain time window and fed as input to machine learning algorithms (e.g., SVM, Random Forest, convolutional and recurrent neural networks, and their combinations), which then classify the human state as normal or drowsy. Such a comprehensive approach significantly increases the accuracy and reliability of the system.

3.2. Development of the architecture of a complex neural network model CNN&LSTM

From the analysis of scientific sources and analogues, it was found that modern systems for determining the state of human drowsiness are implemented based on a neural network approach using the CNN model. But some of them use only the average values of features over a period of time, which leads to the loss of information about the dynamics of events, and others can analyze the state of drowsiness only at certain points in time without taking into account the context and sequence, which leads to the loss of important details about the human state.

In the work, a complex neural network model architecture is proposed for determining the human state, combining a multilayer convolutional neural network (CNN) for analyzing the context of images and a long short-term memory model (LSTM) for analyzing sequential data about the state of human drowsiness over time [17, 28, 29, 30]. The combination of multilayer CNN and LSTM allows combining the advantages of both approaches. It is proposed to increase the accuracy of determining the state of human drowsiness at the LSTM input, in addition to the data obtained for each frame of the video sequence as a result of its processing by CNN, to additionally transmit the values of the EAR, MAR, PUC, MOE indicators calculated by MediaPipe Face Mesh. Thus, when using this model, it is proposed to consider two streams of input data, namely, the stream of images of each frame of the video sequence (224×224×3, scalable RGB images) and the stream of EAR, MAR, PUC, MOE indicators calculated by MediaPipe Face Mesh. In the first stream, each frame of the video sequence is processed, while the EAR, MAR, PUC, MOE indicators are updated only every 10 frames, which allows you to match the duration of data updates with MediaPipe Face Mesh and eye closure during drowsiness.

The main idea of creating a complex neural network model CNN&LSTM is explained by the following algorithm:

- Step 1. Each of the N frames of the first image stream is processed by a CNN to extract spatial features.
- Step 2. A flow of metrics is generated. In the preparatory stage, the EAR, MAR, PUC, MOE metrics are calculated for every 10th frame of the video sequence and repeated for the intermediate 9 frames. This creates a sequence of metrics of the same length - N, as the image sequence.
- Step 3. For each frame, the vector of values obtained at the CNN output and the corresponding (calculated or repeated) EAR, MAR, PUC, MOE indicators are combined.
- Step 4. The combined feature vector is fed to a single LSTM to detect temporal patterns.
- Step 5. Final fully connected layers are added to perform binary classification to obtain a conclusion about the person's sleepiness state.

- Step 5. Final fully connected layers are added to perform binary classification to obtain a conclusion about the person's sleepiness state.

The design of the complex CNN&LSTM neural network model was carried out using the Python programming language.

Input data for the model:

- flow 1 (input_images). Sequence of raw RGB images. (batch_size, sequence_length=64, 224, 224, 3);
- flow 2 (input_mp_features). Sequence of EAR, MAR, PUC, MOE metrics that have already been preprocessed (i.e., values are repeated for intermediate frames to match the length sequence_length = 64). (batch_size, sequence_length=64, 4).

CNN&LSTM architecture for a sequence of 64 frames:

1. architecture CNN(Image Feature Extractor) for flow 1:
   a. input – image (224, 224, 3) (one frame at a time via TimeDistributed);
   b. convolution blocks:
      - Conv2D(32, (3,3), activation='relu'), Conv2D(32, (3,3), activation='relu'), MaxPooling2D((2,2)), Dropout(0.25);
      - Conv2D(64, (3,3), activation='relu'), Conv2D(64, (3,3), activation='relu'), MaxPooling2D((2,2)), Dropout(0.25);
      - Conv2D(128, (3,3), activation='relu'), Conv2D(128, (3,3), activation='relu'), MaxPooling2D((2,2)), Dropout(0.25);
   c. flatten layer – Flatten();
   d. dense layer – Dense(128, activation='relu');
   e. output – 128-dimensional vector for each frame.
2. flow of indicators 2 (Feature Flow). As noted earlier, the indicators have already been calculated at the preparatory stage. The four indicators for frames 1...9, 11...19, etc., repeat the values calculated for frames 0, 10, 20, respectively. A vector (4,) (EAR, MAR, PUC, MOE) is formed for each frame in the sequence;
3. frame-level feature merging (Concatenation per Frame). The output of the 128-dimensional vector from the CNN flow is concatenated with the 4-dimensional vector of metrics (EAR, MAR, PUC, MOE) for each corresponding frame. At the output of the CNN, we obtain the combined vector (132,) for each frame;
4. temporal analysis using LSTM (Temporal Feature Learner):
   a. input – a sequence of 64 concatenated feature vectors (sequence_length=64, features_per_frame=132);
   b. LSTM layer – LSTM(128, return_sequences=False);
   c. dropout layer – Dropout(0.5);
5. final classification (Classification Head):
   a. dense layer– Dense(64, activation='relu').
   b. dropout layer – Dropout(0.25).
   c. output layer – Dense(1, activation='sigmoid')(for binary classification.).

The output layer receives the LSTM outputs and performs binary classification (state "sleepy" / state "normal").

A complex neural network model is trained using human face data, which are labeled in the dataset into two classes: "sleeping" / "awake" [20, 21]. The use of the ReLU activation function helps to prevent the problem of gradient decay and speeds up the training. The Dropout method is used to prevent overtraining of the model and increase its generalization ability.

To optimize the model, backpropagation error methods with the Adam optimization algorithm, training step 0.1 are used.

After training the model, its performance is evaluated on the test dataset [20, 21] to determine the classification accuracy and the F1-Score metric [13].

3.3. Conducting an experimental study on the accuracy of binary classification of drowsiness based on human faces

In order to experimentally prove the correctness of the creation and feasibility of using the proposed complex CNN&LSTM neural network model, it is necessary to conduct experiments on the same data set and compare it with the results of other models according to accuracy criteria (accuracy, F1-Score).

The following systems were considered as analogues:

1. system with a single-layer CNN. The system is implemented as follows:
   a. Conv2D(32, (3, 3), activation='relu', input_shape=(64, 64, 1)):
      - Conv2D – this is a convolutional layer (2D convolution);
      - the number of filters (neurons) in this layer is 32, each of which extracts its own features from the image;
      - size of each filter 3x3 pixels;
      - ReLU (Rectified Linear Unit) activation function activation='relu', which adds nonlinearity to the model, helping it learn more complex patterns;
      - input image shape input_shape=(64, 64, 1). The input data is a 64×64 pixel image with one color channel (e.g., grayscale);
   b. MaxPooling2D((2, 2)):
      - a max pooling sampling layer – MaxPooling2D, which reduces the image size by selecting the maximum value from each 2×2 pixel window.
      - sampling window size (2, 2);
   c. Dense (128, activation='relu'):
      - Dense – a fully connected (or dense) layer in which each neuron is connected to all neurons in the previous layer.
      - number of neurons in this layer – 128.
      - activation function ReLU – activation='relu'.
2. a hybrid system of single-layer CNN, SVM and Random Forest.The difference of the hybrid architecture is the combination of deep learning of the convolutional neural network for feature extraction with traditional machine learning methods of the support vector machine and the random forest method to improve the classification results. That is, the following layers are added to the CNN layer described in the previous model:
   a. SVM training:
      - svm_model = SVC(kernel='linear');
      - svm_model.fit(sensor_data_scaled, labels);
   b. Random Forest training:
      - rf_model = RandomForestClassifier(n_estimators=100);
      - rf_model.fit(sensor_data, labels).
3. system using a single-layer CNN and LSTM model.
   Models using a single-layer CNN with LSTM typically build on the CNN structure found in the previous sections and add an LSTM layer followed by concatenation.:
   a. lstm_out = LSTM(50):
      - LSTM(50).This is an LSTM layer with 50 neurons. LSTM is used to process sequential data, storing information about long-term dependencies in the data;
   b. combined = concatenate([cnn_out, lstm_out]):

- concatenate: This is a pooling function that combines the outputs from two different types of layers (in this case, CNN and LSTM);
- [cnn_out, lstm_out]. This is the list of outputs that we want to merge. cnn_out and lstm_out represent the outputs of the respective previous layers;

c. combined = Dense(50, activation='relu') (combined):
- Dense(50, activation='relu'). This is a fully connected layer with 50 neurons and a ReLU activation function;
- Combined. This layer uses the result of merging the outputs from the previous step as its inp;

d. output = Dense(2, activation='softmax') (combined):
- Dense(2, activation='softmax'. This is the last fully connected layer that produces the final predictions. It has 2 neurons (probabilities for the two classes) and uses a softmax activation function to convert the outputs into probabilities;
- Combined. This layer uses the output of the previous fully connected layer as input.

4. complex multilayer CNN&LSTM model. The architecture of the system based on this model is discussed above in section 4.2.

The results of determining the driver's drowsiness state by systems based on the considered neural network models on a dataset of 10 thousand frames [20, 21] are given in Table 1.

Table 1
Comparative analysis of the results of determining the state of human drowsiness based on different neural network models

Model	Accuracy	F1-Score
Single-layer CNN	0.918	0.954
Hybrid system of single-layer CNN, SVM and Random Forest	0.925	0.960
System using a single-layer CNN and LSTM model.	0.933	0.964
Complex multilayer CNN&LSTM model	0.961	0.979

From the data analysis, it can be concluded that the results improved when using the complex CNN&LSTM model, relative to the best analogue, both in accuracy - by 3%, and in F1-Score - by 1.5%. This is an experimental confirmation of the feasibility of the developed model and its performance and competitiveness.

When implementing a neural network model, an additional increase in the accuracy of its operation can be ensured by the correct selection and adjustment of the network hyperparameters.

During the computer experiment, the following hyperparameters were set for training LSTM:

- Batch Size: 32.
- Number of Epochs: 10.
- Learning Rate: 0.1 (the step at which model parameters are updated during optimization).
- Number of LSTM Layers: 2.
- Number of Hidden Units in LSTM: 64.
- Optimizer: Adam.
- Dropout: 0.25.

When calculating the resulting F1-Score value (see Table 1), the error values were taken into account, namely, False Positive and False Negative:

- the first type of error, False Positive, tells us that a person is considered asleep when they are actually awake;
- a type II error, False Negative, indicates that a person's state is considered normal (active) when they are already falling asleep.

Since the potential risk of type I errors is incomparably smaller than the potential risk of type II errors, it is necessary to minimize type II errors by minimizing the loss of accuracy in the process of determining a person's sleepiness.

Let's conduct an experiment with changing the learning rate, i.e. the step value – Learning rate, with which the model parameters are updated during optimization.

We will sequentially change the value of the Learning rate parameter from the standard 0.1 to 1, 0.01 and 0.001 and evaluate the Accuracy and F1-Score. The results of changing the Learning rate value, evaluating type I and type II errors, and calculating accuracy metrics are given in Table 2.

Table 2
Results of evaluating the accuracy of determining the state of human drowsiness when changing the Learning rate hyperparameter in a complex neural network model of multilayer CNN&LSTM

Learning rate	False Positive	False Negative	Accuracy	F1-Score
0.1	201	181	0.961	0.979
1	309	308	0.941	0.963
0.01	192	287	0.963	0.974
0.001	121	395	0.952	0.973

At Learning rate = 0.001, we see a change in both metrics: Accuracy – 0.952, F1-score – 0.973. The values of these indicators have slightly decreased compared to the standard setting of the Learning rate indicator, however, these indicators are still higher than any proposed analogue, and fulfill the main goal, namely, minimizing type II errors.

4. Conclusions

Thus, the goal of the work on increasing the accuracy of determining the state of drowsiness from images and video sequences of a human face with the additional use of EAR, MAR, PUC, MOE indicators calculated by MediaPipe Face Mesh by developing a complex neural network model based on CNN and LSTM was achieved.

When developing a complex neural network model CNN&LSTM, the following tasks were solved.

It is shown that recognizing the state of drowsiness from an image of a human face is relevant in computer vision systems related to transport safety, which depends, among other things, on the timely determination of the degree of driver fatigue; with industrial safety, when a tired worker is a source of increased risk in such areas as metallurgy, chemical industry, energy, etc. The analysis of such systems showed that their main disadvantages are the rather low accuracy of recognizing the state of drowsiness from frames and video sequences of human facial images, as well as the complexity and inconvenience of additional sensors and overload with unnecessary information.

Analysis of modern computer vision systems related to security has shown that the accuracy of recognizing the user's state of drowsiness depends on the quality of neural network models and machine and deep learning algorithms (Deep Learning), which are the main components of such systems. Combining CNN with other types of neural networks allows you to improve the recognition results. In particular, to solve the problems of determining the state of drowsiness of a person by his face, recurrent neural network (RNN) and long short-term memory network (LSTM) are used. Combining the advantages of convolutional neural networks for recognizing closed eyes in a single frame of a face image with long short-term memory (LSTM), allows us to detect the process of slow eye closure over several frames of a video sequence and increase the accuracy of solving the problem of binary classification of sleepiness.

The calculation of features for the video sequence is implemented using a 3D model of the face surface, created based on the regression approach. For its implementation, the open library of computer vision and machine learning from Google MediaPipe was used. Calculations of EAR, MAR, PUC and MOE indicators are performed using the created Media Pipe Face Mesh.

To determine the human state, architecture of a complex neural network model was created, which combines a multilayer convolutional neural network (CNN) for analyzing the context of images and a model with long short-term memory (LSTM) for analyzing sequential data on the state of human sleepiness over time. It is proposed to increase the accuracy of determining the state of human drowsiness at the LSTM input, in addition to the data obtained for each frame of the video sequence as a result of its processing by CNN, to additionally transmit the values of the calculated indicators EAR, MAR, PUC, MOE. Thus, when using the proposed model, two streams of input data are considered, namely, the stream of images of each frame of the video sequence (224x224x3, scalable RGB images) and the stream of indicators EAR, MAR, PUC, MOE calculated by MediaPipe Face Mesh. In the first stream, each frame of the video sequence is processed, while the indicators EAR, MAR, PUC, MOE are updated only every 10 frames.

The complex CNN&LSTM model is implemented in the Python language using Keras/TensorFlow in the Google Colab environment.

To study the advantages of the developed CNN&LSTM using the Drowsiness Detection Dataset, a comparative experimental analysis of the accuracy of drowsiness recognition of a system with a single-layer CNN, a hybrid system of a single-layer CNN with SVM and Random Forest, as well as a complex CNN&LSTM model was conducted. It was shown that the results of using CNN&LSTM improved in accuracy by 3% and in F1-Score by 1.5%.

It was shown that when implementing the CNN&LSTM neural network model, an additional increase in the accuracy of its operation can be ensured by the correct selection and adjustment of the network hyperparameters. An experiment was conducted with a change in the learning rate, i.e. the value of the Learning rate step. The possibility of significantly reducing the value of the second type of error in drowsiness detection systems due to the rational selection of the Learning rate value was experimentally proven.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] Audi: Assistance Systems, 2023. URL: https://www.audi-mediacenter.com/en/assistance-systems-237.

[2] BMW: Driver assistance. 2024. URL: https://www.press.bmwgroup.com/global/article/search/driver%20drowsiness/topic:5243.

[3] Driver-drowsiness-detection, 2024. URL: https://www.bosch-mobility.com/en/solutions/assistance-systems/driver-drowsiness-detection/.

[4] The importance of Ford's Driver Alert Monitor for Safer Driving. 2023. URL: https://www.uk-car-discount.co.uk/news/fords-driver-alert-monitoring-system.

[5] ACCESS YOUR INFO Find Your Honda, 2025. URL: https://mygarage.honda.com/s/find-honda.

[6] Core for Tech. Develops software technology that analyzes heart rate variability and provides early signs of drowsiness to drivers and vehicles, 2025. URL: https://www.ventureradar.com/organisation/CORE for Tech™/a7d025a6-d217-4ae9-8aa8-9f4805a7c696.

[7] Catherine Shu, Bluetooth Headset Vigo Knows When You Are Tired Before You Do, 2014. URL: https://techcrunch.com/2014/01/17/vigo/.

[8] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, L. D. Jackel, Backpropagation Applied to Handwritten Zip Code Recognition, Neural Computation 1 4 (1989), 541–551. doi: 10.1162/neco.1989.1.4.541.

[9] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner Gradient-Based Learning Applied to Document Recognition, in: Proceedings of the IEEE, vol. 86, no. 11, 1998, pp. 2278-2324, doi:10.1109/5.726791

[10] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, L. D. Jackel, Handwritten Digit Recognition with a Back-Propagation Network, in: Advances in Neural Information Processing Systems 2, Morgan Kaufmann, 1990. — pp. 396–404. URL: https://proceedings.neurips.cc/paper/1989/file/53c3bce66e43be4f209556518c2fcb54-Paper.pdf.

[11] W. Xu, J. He, Y. Shu, and H. Zheng, Advances in Convolutional Neural Networks, Advances and Applications in Deep Learning. IntechOpen, (2020). doi: 10.5772/intechopen.93512.

[12] Mienye, I. D., Swart, T. G., Obaido, G., Jordan, M., & Ilono, P., Deep Convolutional Neural Networks: A Comprehensive Review. Preprints. (2024). URL: https://doi.org/10.20944/preprints202408.1288.v1.

[13] J. Xiong, W. Dai, Q. Wang, X. Dong, B. Ye, J. Yang, A review of deep learning in blink detection. PeerJ Computer Science 11:e2594, 2025. URL: https://doi.org/10.7717/peerj-cs.2594.

[14] A. Fogelton, W. Benesova, Eye blink completeness detection. Computer Vision and Image Understanding 176 11 (2018) 78–85. DOI: 10.1016/j.cviu.2018.09.006.

[15] G. Hu, Y Xiao, Z Gao, L. Meng, Z. Fang, JT. Zhou, J. Yuan, Towards Real-Time Eyeblink Detection in the Wild: Dataset, Theory and Practices, in: IEEE Transactions on Information Forensics and Security, vol. 15, pp. 2194-2208, 2020, doi: 10.1109/TIFS.2019.2959978.

[16] C. Lu, Y. Jiang, K. Fu, Q Zhao, H Yang, LSTPNet: long short-term perception network for dynamic facial expression recognition in the wild, Image and Vision Computing (2024) 142(2):104915 doi: 10.1016/j.imavis.2024.104915.

[17] G. de la Cruz, M. Lira, O. Luaces and B. Remeseiro, Eye-LRCN: A Long-Term Recurrent Convolutional Network for Eye Blink Completeness Detection, in: IEEE Transactions on Neural Networks and Learning Systems, vol. 35, no. 4, pp. 5130-5140, April 2024, doi: 10.1109/TNNLS.2022.3202643 .

[18] P. V. Patil, Drowsiness Detection Dataset, 2020. URL: https://www.kaggle.com/datasets/prasadvpatil/mrl-dataset .

[19] A. Shingha, ·MRL Eye Dataset, 2024. URL: https://www.kaggle.com/datasets/akashshingha850/mrl-eye-dataset/data.

[20] Closed Eyes In The Wild (CEW). 2014. URL: http://parnec.nuaa.edu.cn/_upload/tpl/02/db/731/template731/pages/xtan/ClosedEyeDatabases.html.

[21] F. Song, X.Tan, X.Liu, S.Chen, Eyes Closeness Detection from Still Images with Multi-scale Histograms of Principal Oriented Gradients, Pattern Recognition, Volume 47, Issue 9 (2014) 2825-2838. URL: https://doi.org/10.1016/j.patcog.2014.03.024.

[22] A. Khadraoui, E. Zemmouri, Y. Taki, M. Douimi, Towards a system for real-time prevention of drowsiness-related accidents. IAES International Journal of Artificial Intelligence (IJ-AI), 13 1 (2024) 153-161. http://dx.doi.org/10.11591/ijai.v13.i1.pp153-161.

[23] M. Susman, V. Kimmelman. Eye Blink Detection in Sign Language Data Using CNNs and Rule-Based Methods, in: Proceedings of the LREC-COLING 2024 11th Workshop on the Representation and Processing of Sign Languages: Evaluation of Sign Language Resources, pp. 361–369, Torino, Italia. ELRA and ICCL. URL: https://aclanthology.org/2024.signlang-1.40/ .

[24] F. Sukno, J. Waddington, P. Whelan, 3D Facial Landmark Localization Using Combinatorial Search and Shape Regression. In: A. Fusiello, V. Murino, R. Cucchiara (eds) Computer Vision – ECCV 2012. Workshops and Demonstrations. ECCV 2012. Lecture Notes in Computer Science, vol 7583. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-33863-2_4.

[25] X. Fan, Q. Jia, K. Huyan, X. Gu, Z. Luo, 3D facial landmark localization using texture regression via conformal mapping, Pattern Recognition Letters 83 3 (2016) 395-402, https://doi.org/10.1016/j.patrec.2016.07.005.
[26] MediaPipe solutions guide, 2025. URL: https://ai.google.dev/edge/mediapipe/solutions/guide.
[27] V.T. Sai Sandeep Raju, M. Belwal, Driver Drowsiness Detection. in: Smys, S., Palanisamy, R., Rocha, Á., Beligiannis, G.N. (eds) Computer Networks and Inventive Communication Technologies. Lecture Notes on Data Engineering and Communications Technologies, vol 58. Springer, Singapore, 2021. https://doi.org/10.1007/978-981-15-9647-6_77.
[28] R. Bennett, S. H. Joshi, A CNN and LSTM Network for Eye-Blink Classification from MRI Scanner Monitoring Videos, 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Mexico, 2021, pp. 3463-3466, doi: 10.1109/EMBC46164.2021.9629937.
[29] M. Popat, D. Goyal, V. Raj, N. Jayabalan and C. Hota, Eye Movement Tracking for Computer Vision Syndrome using Deep Learning Techniques, 2024 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), Osaka, Japan, 2024, pp. 317-322, doi: 10.1109/ICAIIC60209.2024.10463437.
[30] Z. Gomolka, E. Zeslawska,B. Czuba and Y. Kondratenko, Diagnosing Dyslexia in Early School-Aged Children Using the LSTM Network and Eye Tracking Technology. Applied Sciences (Switzerland), 14 17 (2024) 8004. https://doi.org/10.3390/app14178004.