Expert Systems With Applications 173 (2021) 114693

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Driver stress detection via multimodal fusion using attention-based CNN-LSTM

Luntian Mou [a,*], Chao Zhou [a], Pengfei Zhao [a], Bahareh Nakisa [b], Mohammad Naim Rastgoo [c], Ramesh Jain [d], Wen Gao [e]

[a] Beijing Key Laboratory of Multimedia and Intelligent Software Technology, Faculty of Information Technology, Beijing University of Technology, Beijing, China
[b] School of Information Technology, Faculty of Science Engineering and Built Environment, Deakin University, VIC, Australia
[c] School of Electrical Engineering and Computer Science, Queensland University of Technology, Brisbane, QLD, Australia
[d] Institute for Future Health, Bren School of Information and Computer Sciences, University of California, Irvine, CA, USA
[e] Institute of Digital Media, School of Electronic Engineering and Computer Science, Peking University, Beijing, China

ARTICLE INFO

Keywords:
Driver stress detection
Convolutional neural network
Long short-term memory
Eye data
Vehicle data
Attention mechanism

ABSTRACT

Stress has been identified as one of major contributing factors in car crashes due to its negative impact on driving performance. It is in urgent need that the stress levels of drivers can be detected in real time with high accuracy so that intervening or navigating measures can be taken in time to mitigate the situation. Existing driver stress detection models mainly rely on traditional machine learning techniques to fuse multimodal data. However, due to the non-linear correlations among modalities, it is still challenging for traditional multimodal fusion methods to handle the real-time influx of complex multimodal and high dimensional data, and report drivers' stress levels accurately. To solve this issue, a framework of driver stress detection through multimodal fusion using attention based deep learning techniques is proposed in this paper. Specifically, an attention based convolutional neural networks (CNN) and long short-term memory (LSTM) model is proposed to fuse non-invasive data, including eye data, vehicle data, and environmental data. Then, the proposed model can automatically extract features separately from each modality and give different levels of attention to features from different modalities through self-attention mechanism. To verify the validity of the proposed method, extensive experiments have been carried out on our dataset collected using an advanced driving simulator. Experimental results demonstrate that the performance of the proposed method on driver stress detection outperforms the state-of-the-art models with an average accuracy of 95.5%.

1. Introduction

One of the contributing factors to road traffic crashes which lead to a large number of injuries and fatalities, is being stressed while driving. Stress can be defined as a nonspecific bodily response to a combination of external demands and internal concerns. Stress can be distinguished in the concept of eustress (positive stress) and distress (negative stress) (Selye, 1974). Generally, there are two types of stress: eustress and distress, by which eustress refers to positive correlation with life satisfaction while distress is the opposite mental state. In daily life, we often use the term "stress" to describe negative stress rather than positive stress. In this study, the term "stress" also refers only to negative stress.

According to the World Health Organization (WHO)'s report on road safety, the total number of deaths caused by various traffic accidents has reached 1.3 million each year (Sauerzapf, 2012). The European Commission estimated that the cost of car accidents in Europe reached 160 billion euros, of which 60%–80% came from the drivers' psychophysical condition (Vivoli, Bergomi, Rovesti, Bussetti, & Guaitoli, 2006). Stress often leads to poor psychophysical condition that can increase the risk of crash almost tenfold, according to Virginia Tech Transportation Institute (Brown et al., 2016). National crash reports in Australia also show that feeling stressed is a critical factor in fatal crashes (Beanland, Fitzharris, Young, & Lenné, 2013). Stress increases the risk of crash by weakening the cognitive ability of drivers, which would result in undermined driving performance (Useche, Ortiz, & Cendales, 2017). Therefore, in order to reduce the risk of crashes and improve driving safety, it is essential to build a system which can detect drivers' stress levels accurately.

* Corresponding author.
E-mail address: ltmou@bjut.edu.cn (L. Mou).

https://doi.org/10.1016/j.eswa.2021.114693
Received 15 July 2020; Received in revised form 18 November 2020; Accepted 5 February 2021
Available online 11 February 2021
0957-4174/© 2021 Elsevier Ltd. All rights reserved.

It has been shown that the use of multimodal data can substantially improve driver stress classification performance (Healey & Picard, 2005; Katsis, Katertsidis, Ganiatsas, & Fotiadis, 2008; Rigas, Goletsis, & Fotiadis, 2012; Lanata et al., 2015). Different types of data such as eye data, vehicle data and environmental parameters are used to monitor driver stress (Rastgoo, Nakisa, Rakotonirainy, Chandran, & Tjondronegoro, 2018). Several studies have shown that there is strong correlation between eye data and driver's behaviour (Haak, Bos, Panic, & Rothkrantz, 2009; Palinko, Kun, Shyrokov, & Heeman, 2010; Wu, Zhao, Rong, & Ma, 2013; Zhang, Liu, & Tang, 2015). Pupillary response from eye data has been shown to be a potential physiological data for detecting driver stress (Pedrotti et al., 2014). The autonomic nervous system (ANS) can continuously regulate pupil size. When a driver is under stress, the pupils are dilated due to sympathetic nervous system (SNS) stimulation. Therefore, pupil diameter (PD) can be used in driver stress detection. Baltaci and Gokcay (2016) combined the features of PD and face temperature to detect driver stress. In addition to PD, blink and gaze of eye are also concerned by researchers (Hansen & Ji, 2010; Kübler et al., 2014). Some researchers even combined PD, blink and gaze data of eye and electroencephalogram (EEG) to perform emotion recognition on drivers (Soleymani, Pantic, & Pun, 2012; Liu, Zheng, & Lu, 2016).

In stressful situations, drivers reacts physically to control the vehicle to avoid collisions. Depending on the types of reaction, the reaction time can range from milliseconds to seconds (Green, 2000). The drivers' physical reactions can be monitored with vehicle data such as steering wheel, acceleration and deceleration data (Bořil, Sadjadi, Kleinschmidt, & Hansen, 2010; Rigas et al., 2012; Lanata et al., 2015; Lee, Chong, & Lee, 2017). In addition, environmental data have also been shown to be helpful in detecting driver stress levels (Hill & Boyle, 2007). The environmental data include different information affecting drivers, for instance, weather conditions, visibility, time of day, road situations, and other driver behaviors.

Building a multimodal fusion model based on eye data, vehicle data and environmental data can improve the performance of driver stress detection. The strategies for fusion mainly include sensor-level, feature-level and decision-level fusion. At present, feature-level strategies are the main fusion approach, which uses handcrafted features or deep learning features to build multimodal models (Hu & Li, 2016; Pourbabaee, Roshtkhari, & Khorasani, 2017; Nakisa, Rastgoo, Rakotonirainy, Maire, & Chandran, 2018, 2020). Compared to handcrafted features methods, deep learning methods can automatically extract features without expertise. Although existing deep learning models perform relatively well in driver stress detection (Ngiam, Khosla, Kim, Nam, Lee, & Ng, 2011; Kanjo, Younis, & Ang, 2019; Rastgoo, Nakisa, Maire, Rakotonirainy, & Chandran, 2019), there is still space for performance improvement. This is due to the fact that these methods just learn the relationships of features within a single modality and lack proper mechanisms to handle the non-linearity across modalities.

To effectively fuse features from different modalities, attention mechanism is introduced in this paper to integrate eye data, vehicle data and environmental data. Originally, the attention mechanism was used in machine translation, which can quickly extract sparse features for natural language processing tasks (Bahdanau et al., 2015). Self-attention is an improvement of the attention mechanism, which can reduce the dependence on external data and capture the internal relationship of longer data or features (Lin et al., 2017). The self-attention mechanism can be used to process the hidden states of an LSTM or Gated Recurrent Unit (GRU) for classification tasks. For example, an attention-based LSTM network was proposed to classify aspect-level sentiment (Wang, Huang, Zhao, & Zhu, 2016). To classify documents, a hierarchical attentional network was constructed using GRU and attention mechanisms (Yang et al., 2016). In recent years, researchers have begun to focus on combining CNN-LSTM network (Sainath, Vinyals, Senior, & Sak, 2015) with attention mechanism in a variety of areas. In the script recognition problem of scene text images and video scripts, an attention-based CNN-LSTM framework was proposed to extract local and global features and dynamically weight them (Bhunia et al., 2019). On the issue of electrocardiogram (ECG) based arrhythmia classification, Liu et al. (2019) proposed an attention-based hybrid LSTM-CNN model to extract overall variation trends and local features of ECG.

In this study, we propose a multimodal fusion model based on an attentional CNN-LSTM network to fuse eye data, vehicle data, and environmental data. The proposed model first uses convolutional neural networks (CNN) and long short-term memory networks (LSTM) to extract features, and then allocates different levels of attention to features with different modalities using a self-attention mechanism. Thus, it can capture the relationship between multimodal data and driver stress levels. Here, three driver stress levels are currently considered, namely low, medium and high. We have collected a dataset for this study from an advanced driving simulator. This dataset contains eye data, vehicle dynamics data and environmental data. And the data is sampled from 22 participants from multiple driving situations designed to induce different levels of stress.

The main contributions of this study are as follows:

- A framework based on attentional CNN-LSTM network is proposed to build an accurate driver stress detection system. This attention-based multimodal fusion model can not only automatically extract features but also weigh the features from different modalities to improve performance on driver stress level classification.
- A non-invasive multimodal data combination is proposed for driver stress detection, which includes eye data, vehicle dynamics data and environmental data. The non-invasive characteristics makes it more suitable for practical deployment.
- Extensive experiments have been conducted to validate the proposed model. Experimental results show that eye data is a promising data for driver stress detection and attention-based multimodal fusion model is superior to other non-attention models.

2. Related work

Since stress detection tends to utilize multimodal data, a stable and reliable stress detection model can be established by analysing and fusing multimodal data. There are different modalities that can be used to measure driver stress, including stressors that stimulate drivers, the driving environment, personal parameters, and the physiological, psychological, and physical responses of drivers under stressors. Since eye data and vehicle dynamics data are easy to obtain, researchers have studied the relationship between these data and stress (Bořil et al., 2010; Pedrotti et al., 2014; Lanata et al., 2015). The studies in the literature have mainly used traditional machine learning methods to extract features manually from data, and then combine the features to create stress detection models. Although the handcrafted features approach has yielded positive results, it is always a challenge to accurately extract important and representative features (Nakisa, Rastgoo, Tjondronegoro, & Chandran, 2017). In addition, the handcrafted features approach requires specialized knowledge and is less robust to noise and data variations.

There are several researchers who have built stress level detection models using multimodal data. Benoit et al. (2009) proposed a driver simulator that uses multimodal data to monitor stress states, including video data (facial activity) and physiological data. The model used facial activities such as blinking, yawning and head turning, as well as ECG and electrical skin responses, to assess the driver's stress levels. Rigas, Goletsis, Bougia, and Fotiadis (2011) proposed a model to detect driver stress levels and predict driving performance. The multimodal data include physiological signals, video recordings (eye data, head movement), and environmental data. The model used a support vector machine (SVM) classifier to distinguish between two stress levels (no stress, stress) with an accuracy of 86%.

Most of the above studies fuse multimodal data at feature-level. The traditional feature-level fusion method combines feature data from each

modality into a feature vector, which is fed into a classifier. However, this approach lacks the ability to handle complex multimodal and high dimensional data (Ngiam et al., 2011). To improve the model's robustness and data processing ability, deep learning techniques are used to process multimodal data. Deep learning techniques can directly process the original data and are widely used in high dimensional signals processing, such as speech recognition (Hinton et al., 2012) and time series data analysis (Liu, Chen, Peng, & Wang, 2017).

In deep learning methods, CNN has shown strong performance in the field of image recognition (George & Routray, 2016). One-dimensional convolutional neural network (1D-CNN) is a type of CNN that is primarily used in sequence modeling and natural language processing (Burkert, Trier, Afzal, Dengel, & Liwicki, 2015; Ordóñez & Roggen, 2016). CNN can extract locally dependent and invariant features of the data. Moreover, the deep CNN can extract local features from the original data, and then extract the global high dimensional feature representation in deeper layers. Some studies proved that CNN surpasses the traditional handcrafted features approach in feature extraction (Haidar, Koprinska, & Jeffries, 2017; Urtnasan, Park, Joo, & Lee, 2018). He, Li, Liao, Zhang, and Jiang (2019) proposed to use CNN to process the heart rate variability (HRV) of ECG signals in order to achieve rapid detection of acute cognitive stress.

As an improvement to recurrent neural network (RNN), LSTM network can learn long-term dependence information and avoid gradient explosion. LSTM has excellent performance in dealing with time series problems such as machine translation, emotion recognition and speech recognition (Hinton et al., 2012; Yan & Mikolajczyk, 2015; Neverova et al., 2016). In addition, some researchers have proposed the CNN-LSTM model, which uses CNN for feature extraction on input data and LSTM to perform sequence prediction on the feature vectors (Zhang, Chan, & Jaitly, 2017; Valiente, Zaman, Ozer, & Fallah, 2019). Donahue et al. (2015) combined LSTM and CNN to solve visual recognition problems. Rastgoo et al. (2019) proposed a multimodal fusion model based on CNN-LSTM network to detect the driver's stress level. Arefnezhad et al. (2020) proposed a CNN-LSTM deep network structure using vehicle data on driver drowsiness detection that significantly outperforms traditional machine learning methods.

Although LSTM can handle long time series data, it is still an RNN structure, which focuses on time step relationships and lacks extraction of global information. Thus, some researchers have begun to focus on attention mechanism (Wang et al., 2016; Winata, Kampman, & Fung, 2018). Attention mechanism was first proposed in machine translation and then widely used in text classification and representation learning (Bahdanau, Cho, & Bengio, 2015; Yang et al., 2016). Self-attention is an improvement on attention mechanism with better performance in capturing the internal correlation of data and features. Winata et al. (2018) proposed a bidirectional LSTM model based on self-attention mechanism using recorded texts to classify psychological stress.

Recently, the combination of CNN-LSTM network and attention mechanism has been concerned by researchers. Peng, Tian, Yu, Lv, and Wang (2019) proposed a CNN-LSTM network based on attention, which has a good effect on identifying and detecting malicious uniform resource locator (URL). Li et al. (2020) proposed an attention-based CNN-LSTM model to predict urban PM2.5 concentration. The model used the CNN-LSTM network to learn the correlation of multivariate time series data related to air quality, and used the attention mechanism to weigh the past features to improve the prediction accuracy. Therefore, for the first time, this paper applies the deep learning model of CNN-LSTM combined with the self-attention mechanism in the field of driver stress classification. Specifically, the CNN-LSTM model is used to extract features from non-invasive multimodal time series data, and the self-attention mechanism is used to fuse features of eye, vehicle and environment, and weigh features from different modalities to effectively detect drivers' stress levels.

3. Methodology

In this section, a multimodal fusion model based on attentional CNN-LSTM network is proposed for driver stress detection. The dataset was measured in a simulated driving environment. In the following subsections, the dataset acquisition, the experimental design, and the multimodal fusion architecture are described one by one.

3.1. Dataset acquisition

The dataset was measured by simulating different stressful environments using an advanced driving simulator (see Fig. 1). The driving simulator consists of SCANeR™ studio software, computers, projectors, a real cabin and a six degree of freedom (6DOF) motion platform. The driving simulator can move and twist in three-dimensional space to approximate the actual driving environment. The simulated environment includes 180 degree front view, rear view images, engine and various environment sounds and vehicle movements. Therefore, the driving simulator can simulate real visual scenes, surrounding environmental sounds, and vehicle motion feedback, making the driver immersed in a virtual environment that is close to real vehicle. More information on the advanced driving simulator is available at https://research.qut.edu.au/carrsq/wp-content/uploads/sites/45/2019/02/Simulator.pdf.

The experiment used SCANeR™ studio software (Scanerstudio, 2020) and FaceLAB™ (Funke et al., 2016) remote video eye tracker to acquire data. FaceLAB™ was used to obtain eye data such as pupil diameter, gaze dispersion in X and Y axes and blink frequency, sampled at 60 Hz. Vehicle dynamics and environmental data were collected by SCANeR™ studio, also sampled at 60 Hz. Vehicle data include steering wheel angle, brake pedal and gas pedal. Environmental data include the distance to the preceding vehicle, lane width, number of lanes, time of day, weather conditions (sunny, rain density) and visibility (fog). In particular, all data from FaceLAB™ and SCANeR™ studio were collected synchronously. In this study, there were 22 participants involved in data collection, aged 21–40 years (55% male). All participants were required to have two years of driving experience and qualified driving license.

3.2. Experimental design

Before the start of the experiment, participants were asked to get familiar with the details and precautions of the experiment, such as avoiding alcohol and caffeinated beverages for a week before the experiment.

In this experiment, six driving scenarios were built to collect participants' eye data, vehicle dynamics data and environmental data. Firstly, participants were asked to drive on a simple road to become familiar with the driving environment and operating methods. After that, each participant's driving data was collected in five driving scenarios (Urban1, Urban2, Highway, CBD1 and CBD2) with a random order. These operations and the advanced driving simulator can help avoid simulator sickness while ensuring high realism rating of the volunteers. And each driving scenario contains several different stressors to induce different levels of stress in participants. The data labels were obtained through a verbal question and answer. During each scenario, the participants were asked every two minutes to provide their responses (verbally) to a short questionnaire about their average stress levels. They were asked to express their stress levels between 0 and 3 (0- No stress to 3- High stress) during each scenario. These numbers are then mapped into three different stress levels (0.1–1 = Low, 1.1–2 = Medium, 2.1–3 = High). These mappings allow us to compare our work with other existing works (Pedrotti et al., 2014; Wang, Lin, & Yang, 2013; Baltaci & Gokcay, 2016).

In this study, the stressors were designed with reference to different studies (Hill & Boyle, 2007; Rodrigues, Kaiseler, Aguiar, Cunha, & Barros, 2015; Lee et al., 2017; Rastgoo et al., 2018). There are five main

Fig. 1. CARRS-Q's driving simulator was used to collect data. The simulator consists of a front view (180 degrees), rear view mirror image, real cabin, audio system to simulate the driving environment and a six degree of freedom motion platform, as well as the SCANeR™ studio and FaceLAB™ remote video eye tracker.

stressors: traffic congestion, driving road situations, the behavior of other drivers, weather, and time of day. Table 1 shows the different stressors in different driving scenarios. Since traffic congestion creates stress (Rastgoo et al., 2018), this study designed different vehicle densities per kilometer under different driving scenarios. Then, narrow roads, curvy roads, and sharp bends were used to induce different stress levels in drivers. The behaviors of other drivers can also cause varying degrees of stress, such as overtaking, lane changing, speeding and tailgating. In the Highway, CBD1 and CBD2 scenarios, we set several parameters (stay on lane, sign observing, priority observing, safety time, speed limit risk, and overtake risk) in the simulator to simulate the above behaviors. This study designed rain density (0–1) and foggy under different driving scenarios, and simulated the driving environment during the day and night.

3.3. Multimodal fusion architecture

This subsection presents a framework for a multimodal fusion model based on attentional CNN-LSTM network. The proposed model extracts stress related information by fusing eye data, vehicle dynamics data and environmental data to classify drivers' stress levels. The proposed framework consists of four steps: preprocessing, feature extraction, feature fusion and classification (see Fig. 2).

In the first step, to synchronize eye data, vehicle dynamics data and environmental data, missing eye data and existing anomalies in the data are removed. To reduce the differences in data among participants, all data are normalized to zero mean and unit variance. Finally, we use the sliding window method to divide each feature of each modality into time window with fixed window size and degree of overlap. A new training dataset is produced by consisting of generated time windows, of which each label is the same as the original dataset.

Table 1
Different stressors in different driving scenarios.

Scenarios	Number of Vehicles	Road situations	Simulator parameters	Weather	Time
Urban 1	0	–	–	–	Daytime
Urban 2	30	Narrow, Curve	–	–	Night
Highway	50	Curve	Stay on lane, etc.	Rain density (0.2–1), Foggy	Night
CBD 1	50	Narrow, Curve, Tight corner	Stay on lane, etc.	Rain density (0.3–0.6), Foggy	Daytime
CBD 2	60	Curve, Tight corner	Stay on lane, etc.	–	Daytime

Next, the new training dataset for each modality is fed into the 1D-CNN and LSTM framework to extract features. Specifically, the segmented time window (e.g., window t) from training dataset is first fed into the 1D-CNN to automatically learn features. Since the time window is time series, one-dimensional convolutional layer is used. This feature extraction framework consists of three 1D convolutional layers, three max pooling layers, and two-layer LSTMs. The detailed parameter settings are shown in Table 2. The parameter combination and model framework with the best detection accuracy are selected through trial and error. The convolutional layer uses sliding filters to extract effective features. The activation function of the convolution layer is exponential linear units (ELU), which can accelerate the convergence speed and improve the robustness of the model. Each layer of convolution is followed by a max pooling layer. In order to reduce data complexity, the max pooling layer reduces the amount of data to half of the original. The dropout layer is adopted after the pooling layer to avoid overfitting. In each training epoch, a portion of the neurons of dropout layer are randomly selected and are not allowed to participate in weight optimization. After three layers of convolution and pooling, the input data are transformed into high dimensional feature maps. Since the feature maps are extracted from the time window and the operations of convolution and pooling do not change their temporal order, the feature maps are fed directly into the two-layer LSTMs. The LSTM network processes time series by a gating mechanism, which includes forget gate, input gate and output gate. They can control the discarding or adding of information to enable forgetting and remembering. The feature maps are converted to the corresponding hidden states through the LSTM network.

In the fusion step, the generated hidden states from the eye data, vehicle dynamics data and environmental data are integrated to create a new feature map. There are n hidden states h_t in this feature map. The feature map is denoted as H:

$$H = (h_1, h_2, \ldots, h_n). \tag{1}$$

Since different hidden states have different degrees of impact on stress detection, we introduce a self-attention mechanism to weigh all hidden states. These hidden states are aggregated into a vector representation s by the attention layer, of which the formulas are as follows:

$$u_t = tanh(Wh_t + b), \tag{2}$$

$$\alpha_t = \frac{exp(u_t^T u)}{\sum_{t=1}^{n} exp(u_t^T u)}, \tag{3}$$

$$s = \sum_{t=1}^{n} \alpha_t h_t. \tag{4}$$

The hidden state h_t is first fed into a fully connected layer with an

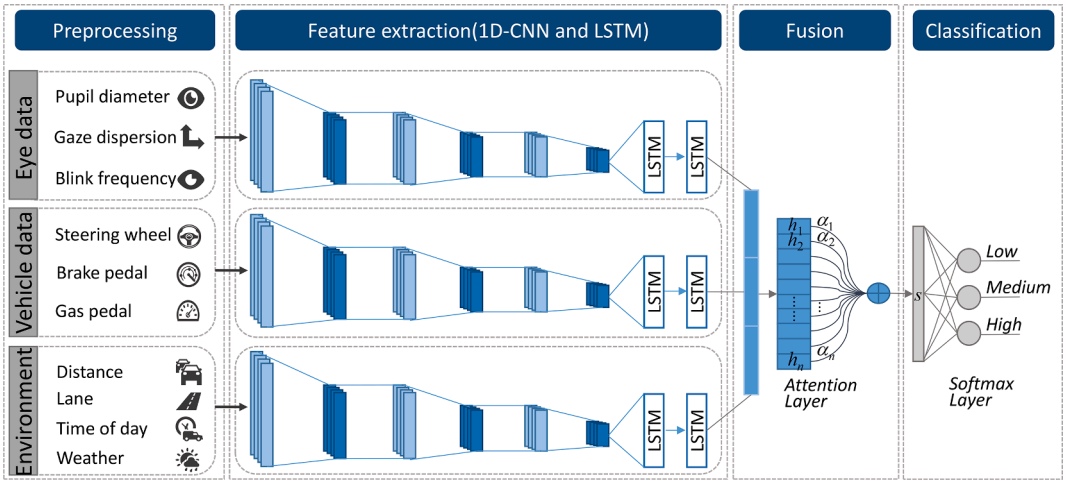

Fig. 2. The multimodal fusion model based on attentional CNN-LSTM network is used to detect driver stress levels. The inputs of the model are eye data, vehicle data and environmental data. The outputs of the model are three levels of driver stress: low, medium and high. The preprocessed data for each modality are fed into the corresponding 1D-CNN and LSTM to extract features (Feature extraction step). The output of hidden states from the three modalities are concatenated to obtain a feature map $(h_1, h_2, ..., h_n)$. The vector representation s is calculated as multiple weighted sums of hidden states from the feature map, where the summation weights $(\alpha_1, \alpha_2, ..., \alpha_n)$ are calculated in formulas (2) (3) (Fusion step). Finally, the s is fed into the Softmax layer for stress level classification (Classification step).

Table 2
The 1D-CNN and LSTM model parameters.

1D-CNN and LSTM	
Convolutional layer	Filter = 20, Kernel size = (10, 1), Stride = 1
Max-pooling + Dropout (0.15)	Pool-size = (2, 1), Stride = 2
Convolutional layer	Filter = 40, Kernel size = (5, 1), Stride = 1
Max-pooling + Dropout (0.15)	Pool-size = (2, 1), Stride = 2
Convolutional layer	Filter = 80, Kernel size = (3, 1), Stride = 1
Max-pooling + Dropout (0.15)	Pool-size = (2, 1), Stride = 2
LSTM	Hidden-size = 64
LSTM	Hidden-size = 64

activation function of tanh to get u_t as a hidden representation of h_t. The transpose of the output u_t is multiplied by the trainable parameter vector u to get the alignment coefficient of attention. Then, the softmax function is used to normalize the alignment coefficient to obtain the summation weight α_t. After that, we compute the vector representation s as a weighted sum of hidden states. In the last step of classification, the vector representation s can be used as a feature vector to be fed into the Softmax layer for stress level classification. The W in the formula (2) is a weight matrix. The b is a bias vector of the fully connected layer in the attention layer, its dimensions is d_a. And the u is a trainable parameter vector used to represent context information and its dimensions is also d_a. Here, d_a is an important hyper-parameter. After extensive experimental verification, it has been shown that the larger the dimension of d_a the better the performance (accuracy and false positive) of the model is in the 5 s window (Fig. 3). Finally, in order to make the model reach a balance between model performance and computational complexity, the best dimension of d_a is set to 64. During the training process, the weight matrix W, bias vector b, and the parameter vector u are initialized randomly.

This subsection proposes a new framework to multimodal data fusion. Specifically, a self-attention mechanism is used to fuse the hidden states of the LSTM output and give different attention to each hidden state. It is worth noting that the proposed framework can not only fuse eye data, vehicle dynamics data and environment data, but also be suitable for processing other multimodal time series data. The proposed framework uses an end-to-end approach to training without the need for

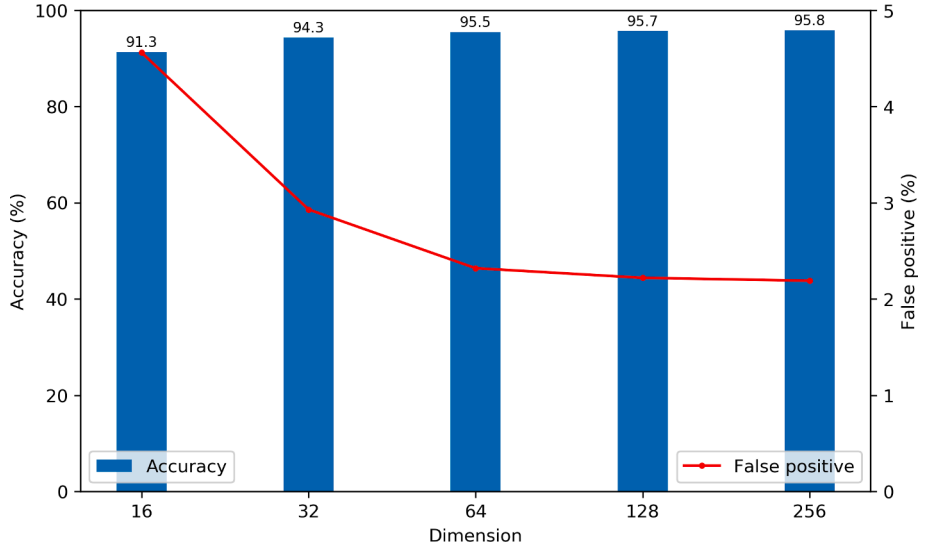

Fig. 3. The average performance (accuracy and false positive) against hyper-parameter d_a in different dimensions.

handcrafted features.

4. Results

To validate the proposed multimodal fusion method based on attentional CNN-LSTM network, we have conducted comparative experiments in this section. This section also evaluates the performance of eye data on stress detection (refer to Section 4.1 for details of eye feature analysis).

In this study, eye data, vehicle dynamics data, and environmental data were segmented and synchronized using sliding windows. We used different window sizes to verify the performance of the proposed method. And, the overlapping degree of sliding windows in this study is 90%. Since the research goal is to establish a real-time driver stress detection system, the 5 s window size is currently selected (Section 4.2).

For the division of the dataset, we used a 10-fold cross validation method to verify the performance of our model. Specifically, the dataset is randomly shuffled and divided into 10 parts, and then one part is used as the test set and the others as the training set in turn. Finally, the experimental results are the average of ten test results.

4.1. Eye data-based stress detection

4.1.1. The framework of handcrafted eye features

The feature extraction framework of eye data mainly includes three steps: preprocessing, feature extraction, and classification.

First of all, maximum-minimum normalization was adopted to normalize all data. Since blinking can result in the loss of pupil diameter and gaze dispersion data, linear interpolation was used to fill in the missing sample data (Soleymani et al., 2012). Then, the average pupil diameter of left and right eyes was used as the time series of pupil diameter.

After preprocessing, the power spectrum and time domain features of the pupil diameter were analyzed. The mean, standard deviation and power spectral density (PSD) were extracted from the pupil diameter. The Hippus effect refers to a small amplitude oscillation in the frequency domain of the pupil diameter between 0.05 and 0.3 Hz with an amplitude of 1 mm (Pamplona, Oliveira, & Baranoski, 2009; Bouma & Baghuis, 1971). It has been shown that the Hippus effect occurs when a person is in a relaxed or passive state. As long as there is mental activity and psychological stress, this effect will disappear and the pupil diameter will expand. Moreover, the PSD features of pupil diameter were computed using the Welch's method in four frequency bands (0–0.2 Hz, 0.2–0.4 Hz, 0.4–0.6 Hz, and 0.6–1 Hz) (Soleymani et al., 2012). In addition to pupil diameter, the dispersion (mean and standard deviation) of eye gaze in X and Y axes were extracted to detect drivers' stress levels (Zheng, Dong, & Lu, 2014). This eye gaze mainly refers to fixation which is a slight deviation of the fixation point. It usually occurs within 2–5 degrees of central vision, and lasts about 80–100 ms (Hansen & Ji, 2010). Blink frequency has also been shown to be correlated to anxiety and stress (Kanfer, 1960). Ultimately, 11 kinds of eye features were extracted from the eye data. A summary of the extracted eye features is shown in Table 3.

Last, all eye features were formed into a feature vector, which was fed into a classifier. The LSTM network was adopted as a classifier for handcrafted eye features.

4.1.2. Comparing the performance of handcrafted features and automatically extracted features

In this subsection, we analyze eye features associated with stress levels and compare the performance of handcrafted features approach with the proposed automatic feature extraction model. The accuracy (ACC) and false positive (FP) of pupil diameter, gaze dispersion (X and Y), and blink frequency in the 5 s window are shown in Table 4. The experimental results are averaged values of ten times. It can be seen that the accuracies of all feature groups are higher than the 33% random level. Therefore, all feature groups are related to stress. The classification accuracies of pupil diameter and gaze dispersion are higher than that of the blink frequency. The average accuracy of 11-dimensional features reaches 74.3% (FP = 13.3%). This result shows that these eye features can effectively distinguish stress levels.

Table 3
Handcrafted features from eye data, vehicle dynamics data and environmental data.

	Handcrafted Features
Eye data	Pupil diameter (Mean, standard deviation) Pupil diameter (PSD in four bands:0–0.2 Hz, 0.2–0.4 Hz, 0.4–0.6 Hz, 0.6–1 Hz) Dispersion (X and Y) (Mean, standard deviation) Blink frequency
Vehicle dynamics data	Steering wheel angle (Mean, standard deviation) Brake pedal data (Mean, standard deviation) Gas pedal data (Mean, standard deviation)
Environmental data	Distance to preceding vehicle Time of day Road situations (Lane width, Number of lanes) Visibility (fog) Weather conditions (sun, low rain, medium rain, high rain)

Based on Table 4, the proposed automatic feature extraction model has better performance than handcrafted features approach. Compared with the handcrafted features approach, the accuracy of pupil diameter, gaze dispersion (X and Y), and blink frequency under the proposed model are improved by 24.6%, 34.4%, and 3.5%, respectively. And their false positives have dropped significantly. Since the blink frequency is calculated at a certain time and is not a continuous time series, its improvement is smaller than the other two features. Finally, all eye features form a modality that is fed into the proposed model. The average accuracy of the proposed model based on all eye features reaches 92.9%, which is nearly twenty percent higher than that of the handcrafted features approach. The results show that the proposed model is better in identifying driver stress levels in a short window than the handcrafted features approach.

4.2. Performance of multimodal fusion model

To verify the performance of the proposed multimodal fusion model, we collected vehicle dynamics data and environmental data besides eye data. For vehicle dynamics data, steering wheel angle, brake pedal, and gas pedal indirectly reflect the driver's instantaneous reaction to stress. For environmental data, the distance to the preceding vehicle, lane width, number of lanes, time of day, weather conditions (sunny, rain density), and visibility (fog) were selected from the advanced driving simulator (see Table 3). These features have been verified to be related to driver stress (Lee et al., 2017; Lanata et al., 2015; Rigas et al., 2012).

4.2.1. Attention based fusion model and comparison with other multimodal fusion models

In this subsection, we evaluate the performance of the proposed multimodal fusion model of attentional CNN-LSTM in detecting driver stress levels against other multimodal fusion models. Meanwhile, the performance of the proposed model is validated under different window sizes: 5 s, 10 s, and 15 s.

Similar to handcrafted features approach in subsection 4.1, the handcrafted features approach in this subsection combines the features from eye data, vehicle dynamics data and environmental data into a feature vector, which is fed into the LSTM network. Meanwhile, we use the LSTM model with direct input of raw data as a comparison model. The multimodal fusion model (CNN-LSTM) adopted by Rastgoo et al. (2019) is also compared with the proposed model.

The comparison results of the proposed model with other fusion models are shown in Fig. 4, which indicates that the proposed model has

Table 4
The average performance (accuracy and false positive) of handcrafted features and automatically extracted features.

Approach	Pupil diameter		Gaze dispersion		Blink frequency		All eye features	
	ACC (%)	FP (%)	ACC (%)	FP (%)	ACC (%)	FP (%)	ACC (%)	FP (%)
Handcrafted	54.7	23.9	56.5	23.4	48.9	27.9	74.3	13.3
Automatic	**79.3**	**10.7**	**90.9**	**4.6**	**52.4**	**25.7**	**92.9**	**3.6**

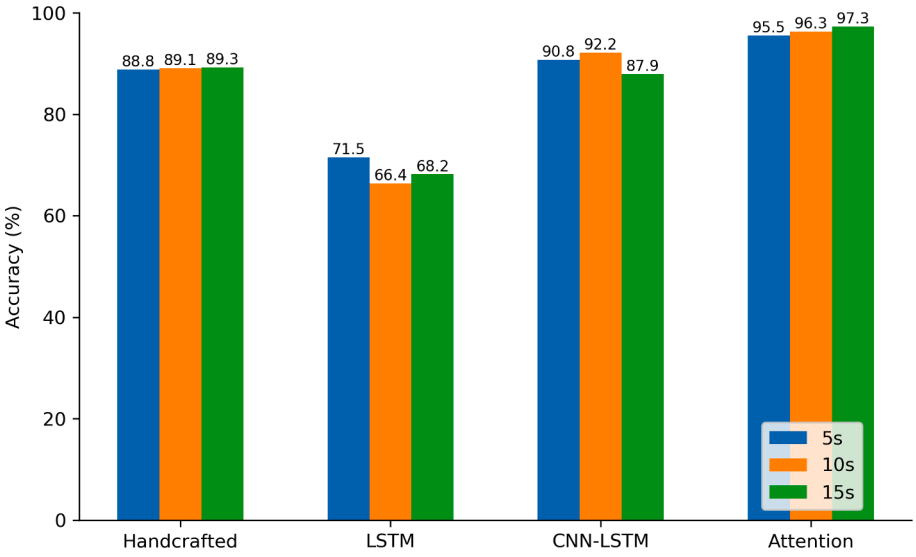

Fig. 4. The average accuracy of different fusion models with different window sizes.

the highest accuracy rate under different window sizes. The accuracies of the multimodal fusion models were obtained through averaging on 10 times. As the window size increases, the accuracy of the proposed model is improved by 0.8% and 1.8% under the 10 s and 15 s windows, respectively. The accuracy of LSTM model is not improved with increased window size. Although handcrafted features approach and the CNN-LSTM model get good results in the 10 s window, their accuracies are still lower than the proposed model. Particularly, the average accuracy of the proposed model in the 15 s window reaches 97.3%, which is improved by 8%, 29.1%, and 9.4% respectively compared to handcrafted features approach, LSTM model, and CNN-LSTM model. Although the long window sizes can improve the accuracy of the model, we finally chose the 5 s window size data in order to build a real-time driver stress detection system.

Table 5 shows the performance of different fusion models under single-modal data and multimodal data in the 5 s window. Obviously, the accuracies of the multimodal fusion models are superior to those of the single-modal models, which means that the fusion models can complement the information of each modality. Since the data of different modalities are very different, eye data and vehicle dynamics data have a greater influence on the detection results of the stress level, while the influence of environmental data is smaller. The self-attention mechanism was introduced to deal with features with different degrees of influence. Thus, the average accuracy of the proposed model reaches 95.5%, which is 4.7 percentage points higher than the traditional CNN-LSTM model. In multimodal fusion, the average accuracy and false positives of multimodal fusion model based on attentional CNN-LSTM network have verified its superior performance in driver stress detection.

4.2.2. *Comparison of confusion matrices for multimodal fusion models*

In this subsection, the confusion matrix of each multimodal fusion model in the 5 s window is shown in Fig. 5, which details the advantages and disadvantages of each model at three stress levels. Obviously, the accuracy of each model for detecting high stress levels is lower than that of low and medium stress levels. This is because the number of high stress samples is less than that of low or medium stress, which leads the model to misclassify high stress as low or medium stress.

Among the confusion matrices of the four multimodal fusion models, the LSTM model has the worst performance (Fig. 5b). The accuracy of the handcrafted features approach is significantly improved, but it is only 76% accurate for high stress level (Fig. 5a). The handcrafted features approach is less effective than the CNN-LSTM model in detecting high stress level (Fig. 5c). Finally, the accuracies of the proposed model have been significantly improved at low, medium, and high stress levels (Fig. 5d). It can be seen that the introduction of the self-attention mechanism strengthens the ability of model to detect three stress levels, and significantly improves the detection accuracy of high stress

Table 5
The average performance (accuracy and false positive) of handcrafted features approach, LSTM model, CNN-LSTM model, and attention-based CNN-LSTM model in different modalities.

Approach	Environmental data		Vehicle data		Eye data		Fusion	
	ACC (%)	FP (%)	ACC (%)	FP (%)	ACC (%)	FP (%)	ACC (%)	FP (%)
Handcrafted	49.9	27.0	49.2	27.7	74.3	13.3	88.8	5.9
LSTM	50.6	26.6	44.9	30.1	58.1	22.1	71.5	14.7
CNN-LSTM	51.0	26.4	73.6	13.6	85.8	7.4	90.8	4.8
Attention	**52.6**	**25.1**	**85.1**	**7.8**	**92.9**	**3.6**	**95.5**	**2.3**

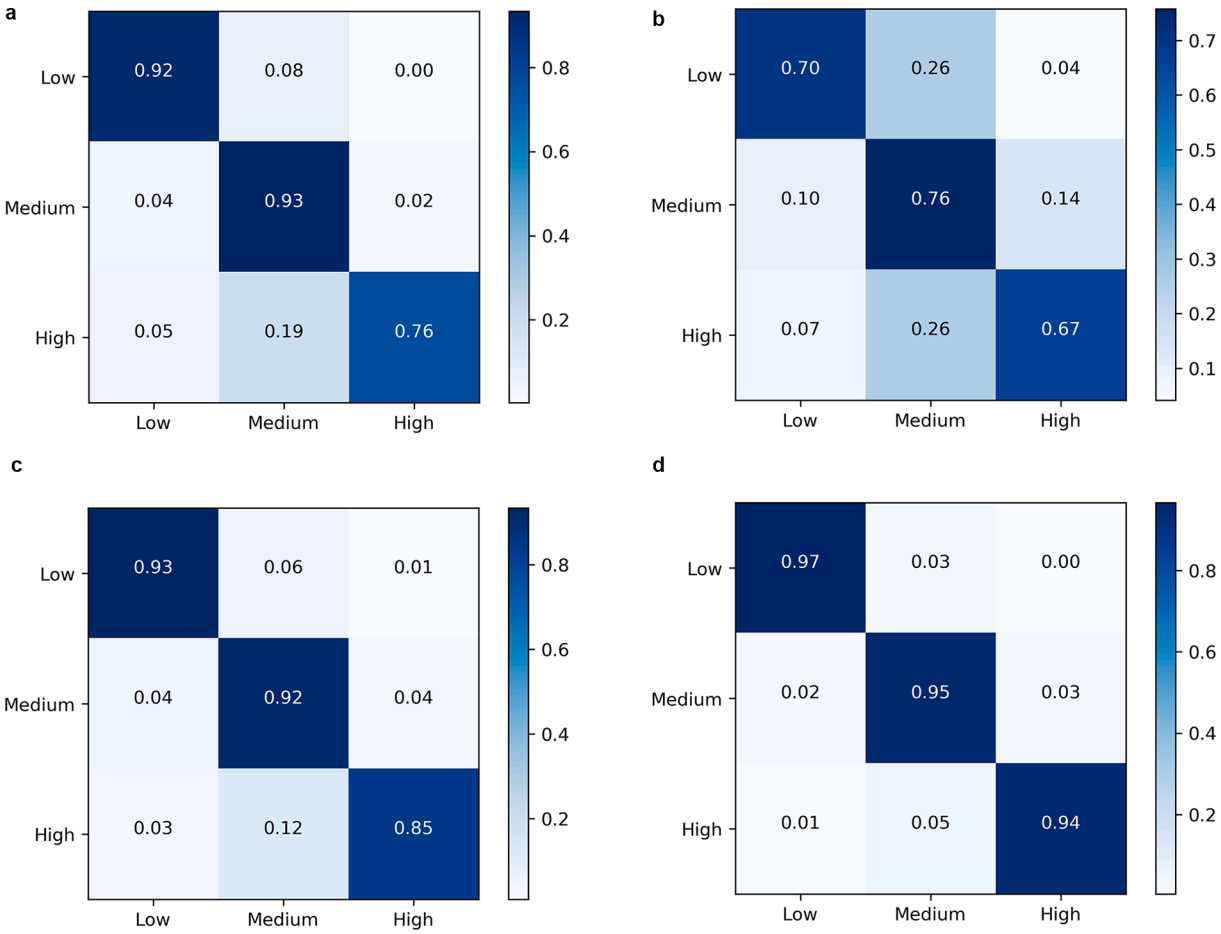

Fig. 5. The confusion matrix of multimodal fusion models: (a) handcrafted features approach, (b) LSTM model, (c) CNN-LSTM model, (d) attention-based CNN-LSTM model.

level. As a result, the attention-based CNN-LSTM model has the best capability for driver stress levels among all multimodal fusion models.

4.2.3. Comparison of attention based fusion model with other works

In this subsection, the experimental results of the multimodal model based on attentional CNN-LSTM network are compared with other recent works. Table 6 shows some works using eye data and multimodal fusion methods.

As can be seen, Pedrotti et al. (2014), Wang et al. (2013), and Baltaci and Gokcay (2016) mainly used physiological modalities to detect the driver's stress level. Among them, pupil diameter (PD) obtained without contact was concerned by researchers. Rigas et al. (2011) combined physiological signals (ECG; electrodermal activity, EDA; respiration, RSP), video data, and environmental data to classify driver stress into

Table 6
The comparison of the best performance of the multimodal model based on attentional CNN-LSTM network with other recent works.

Reference	No. Subjects	of Method	Physiological	Used modalities		Window size	Performance	No. Classes
				Physical	Context			
(Pedrotti et al., 2014)	33	Handcrafted features	PD, EDA	–	–	80 s	Accuracy: 79.2%	4 stress class (low, medium, high, very high)
(Wang et al., 2013)	17	Handcrafted features	ECG	–	–	300 s	Accuracy: 80%	3 stress class (low, medium, high)
(Baltaci et al., 2016)	11	Handcrafted features	PD, Face temperature	–	–	18 s	Accuracy: 83.8% Sensitivity: 83.9% Specificity: 83.8%	2 stress class (no stress, stress)
(Rigas et al., 2011)	1	Handcrafted features	ECG, EDA, RSP	Eye, Head movement	Environmental data	10 s	Accuracy: 86%	2 stress class (no stress, stress)
(Rastgoo et al., 2019)	27	Deep learning (CNN-LSTM) network	ECG	Vehicle dynamic data	Environmental data	5 s	Accuracy: 92.8% Sensitivity: 94.13% Specificity: 97.37% Precision: 95.00%	3 stress class (low, medium, high)
Our work	22	Attention based CNN-LSTM network	PD	Eye movement, Vehicle dynamic data	Environmental data	5 s	Accuracy: 95.5% Sensitivity: 95.31% Specificity: 97.67% Precision: 95.34%	3 stress class (low, medium, high)

two stress levels (no stress, stress). Compared with the work using only one modality, the performance of stress detection by multimodalities has been improved. The first three works used handcrafted features approach, which required a large window size of data, making it difficult to monitor driver stress status in real time. Rastgoo et al. (2019) proposed a multimodal deep learning model to fuse physiological signals (ECG), vehicle dynamics data and environmental data and achieved good stress detection performance, but the applied physiological signals were invasive. Our proposed model is non-invasive, and achieves better accuracy, and the sensitivity, specificity, and precision of the proposed model are improved as well.

In summary, the comprehensive performance of our proposed model outperforms other recent works on driver stress level detection. Specifically, there are some advantages of the proposed model. Firstly, the proposed model uses eye data, vehicle dynamics data and environmental data, which is non-invasive compared with other works using traditional physiological data (ECG, EDA, and EEG). In addition, the proposed model using a small window size (5 s) of data can detect the driver stress in real time. Finally, the self-attention mechanism can effectively fuse data from different modalities by assigning different weights to different features.

5. Conclusion

In this paper, we propose a multimodal fusion model based on attentional CNN-LSTM network for driver stress detection. It is characterized by being non-invasive, accurate and real-time. On the dataset collected by the advanced driving simulator, extensive experiments were carried out to verify the performance of the proposed model. Experimental results demonstrate that eye data is highly correlated to stress levels and very effective in stress detection. Furthermore, the performance of the proposed model is verified as superior to other state-of-the-art models under different window sizes. Additionally, the proposed model can effectively complement and weigh the information from eye data, vehicle dynamics data and environmental data. Therefore, the attention-based CNN-LSTM network is a promising method for driver stress detection.

Though the non-invasive multimodal data combination and feature-level multimodal fusion model have good application prospects, concerns of its application in real vehicles could be raised. First, limitations could be induced by the number and the selection method of participants. Specifically, the sample size of 22 participants is relatively small and the experiment did not consider people over 40 years old. Second, the ground truth of driver stress is determined by the participant's subjective assessment. This is completely subjective, even if the participants are already familiar with the evaluation method of the stress level before the experiment and the subjective assessment is the most reliable technique to annotate data. Yet, it may fall short of objectiveness.

For future work, we will adhere to multimodal fusion method and introduce non-invasive physiological data. Our research has shown that different modalities can compensate for each other, thus the choice of appropriate fusion levels (sensor, feature, score, or decision) could be the key to building an optimal model. Next, we will increase the sample size and age range of participants to improve the generalization ability of the model. Due to the labeling of data relies entirely on subjective assessment, we need to improve the ground truth of driver stress through combining subjective assessment with expert assessment by video or physiological data. Recently, unsupervised learning has made great strides in other areas, and has even surpassed supervised learning in some areas (He, Fan, Wu, Xie, & Girshick, 2020). Unsupervised learning can learn features related to driver stress on unlabeled data, which can solve the problem of laborious, tedious, and inaccurate labeled data.

CRediT authorship contribution statement

Luntian Mou: Investigation, Methodology, Project administration, Writing - original draft, Writing - review & editing. **Chao Zhou:** Methodology, Validation, Writing - original draft, Writing - review & editing. **Pengfei Zhao:** Methodology, Writing - review & editing. **Bahareh Nakisa:** Writing - original draft, Writing - review & editing. **Mohammad Naim Rastgoo:** Writing - review & editing. **Ramesh Jain:** Supervision, Writing - review & editing. **Wen Gao:** Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This work was supported by the National Natural Science Foundation of China under grant 61672068.

References

Arefnezhad, S., Samiee, S., Eichberger, A., Frühwirth, M., Kaufmann, C., & Klotz, E. (2020). Applying deep neural networks for multi-level classification of driver drowsiness using vehicle-based measures. *Expert Systems with Applications, 162*, 113778. https://doi.org/10.1016/j.eswa.2020.113778

Bahdanau, D., Cho, K., & Bengio, Y. (2015, May). Neural machine translation by jointly learning to align and translate. 3rd International Conference on Learning Representations, ICLR 2015, San Diego, SD.

Baltaci, S., & Gokcay, D. (2016). Stress detection in human-computer interaction: Fusion of pupil dilation and facial temperature features. *International Journal of Human-Computer Interaction, 32*(12), 956–966. https://doi.org/10.1080/10447318.2016.1220069

Beanland, V., Fitzharris, M., Young, K. L., & Lenné, M. G. (2013). Driver inattention and driver distraction in serious casualty crashes: Data from the Australian National Crash In-depth Study. *Accident Analysis & Prevention, 54*, 99–107. https://doi.org/10.1016/j.aap.2012.12.043

Benoit, A., Bonnaud, L., Caplier, A., Ngo, P., Lawson, L., Trevisan, D. G., … Chanel, G. (2009). Multimodal focus attention and stress detection and feedback in an augmented driver simulator. *Personal and Ubiquitous Computing, 13*(1), 33–41. https://doi.org/10.1007/s00779-007-0173-0

Bhunia, A. K., Konwer, A., Bhunia, A. K., Bhowmick, A., Roy, P. P., & Pal, U. (2019). Script identification in natural scene image and video frames using an attention based Convolutional-LSTM network. *Pattern Recognition, 85*, 172–184. https://doi.org/10.1016/j.patcog.2018.07.034

Bořil, H., Sadjadi, S. O., Kleinschmidt, T., & Hansen, J. H. (2010). Analysis and detection of cognitive load and frustration in drivers' speech. In *In Eleventh Annual Conference of the International Speech Communication Association* (pp. 502–505).

Bouma, H., & Baghuis, L. C. J. (1971). Hippus of the pupil: Periods of slow oscillations of unknown origin. *Vision Research, 11*(11), 1345–1351. https://doi.org/10.1016/0042-6989(71)90016-2

Brown, T. G., Ouimet, M. C., Eldeb, M., Tremblay, J., Vingilis, E., Nadeau, L., … Yechiam, E. (2016). Personality, executive control, and neurobiological characteristics associated with different forms of risky driving. *PLoS One, 11*(2), e0150227. https://doi.org/10.1371/journal.pone.015022710.1371/journal.pone.0150227.t00110.1371/journal.pone.0150227.t00210.1371/journal.pone.0150227.t003

Burkert, P., Trier, F., Afzal, M. Z., Dengel, A., & Liwicki, M. (2015). Dexpression: Deep convolutional neural network for expression recognition. ArXiv Preprint ArXiv:1509.05371.

Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., & Darrell, T. (2015). Long-term recurrent convolutional networks for visual recognition and description. In *In Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2625–2634).

Funke, G., Greenlee, E., Carter, M., Dukes, A., Brown, R., & Menke, L. (2016). Which eye tracker is right for your research? Performance evaluation of several cost variant eye trackers. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 60*(1), 1240–1244. https://doi.org/10.1177/1541931213601289

George, A., & Routray, A. (2016, June). Real-time eye gaze direction classification using convolutional neural network. 2016 International Conference on Signal Processing and Communications (SPCOM), Bangalore. https://doi.org/10.1109/spcom.2016.7746701.

Green, M. (2000). " How long does it take to stop?" Methodological analysis of driver perception-brake times. *Transportation Human Factors, 2*(3), 195–216.

Haak, M., Bos, S., Panic, S., & Rothkrantz, L. J. M. (2009). Detecting stress using eye blinks and brain activity from EEG signals. Proceeding of the 1st Driver Car Interaction and Interface (DCII 2008), 35–60.

Haidar, R., Koprinska, I., & Jeffries, B. (2017). Sleep Apnea event detection from nasal airflow using convolutional neural networks. *Lecture Notes in Computer Science*, 819–827.

Hansen, D. W., & Ji, Q. (2010). In the eye of the beholder: A survey of models for eyes and gaze. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 32*(3), 478–500. https://doi.org/10.1109/tpami.2009.30

He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. (2020). Momentum Contrast for Unsupervised Visual Representation Learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.1109/cvpr42600.2020.00975

He, J., Li, K.e., Liao, X., Zhang, P., & Jiang, N. (2019). Real-time detection of acute cognitive stress using a convolutional neural network from electrocardiographic signal. *IEEE Access, 7*, 42710–42717. https://doi.org/10.1109/Access.628763910.1109/ACCESS.2019.2907076

Healey, J. A., & Picard, R. W. (2005). Detecting stress during real-world driving tasks using physiological sensors. *IEEE Transactions on Intelligent Transportation Systems, 6*(2), 156–166. https://doi.org/10.1109/TITS.2005.848368

Hill, J. D., & Boyle, L. N. (2007). Driver stress as influenced by driving maneuvers and roadway conditions. *Transportation Research Part F: Traffic Psychology and Behaviour, 10*(3), 177–186.

Hinton, G., Deng, L.i., Yu, D., Dahl, G., Mohamed, A.-R., Jaitly, N., … Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine, 29*(6), 82–97. https://doi.org/10.1109/MSP.2012.2205597

Hu, D., & Li, X. (2016). Temporal multimodal learning in audiovisual speech recognition. In *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3574–3582).

Kanfer, F. H. (1960). Verbal rate, eyeblink, and content in structured psychiatric interviews. *The Journal of Abnormal and Social Psychology, 61*(3), 341–347. https://doi.org/10.1037/h0038933

Kanjo, E., Younis, E. M. G., & Ang, C. S. (2019). Deep learning analysis of mobile physiological, environmental and location sensor data for emotion detection. *Information Fusion, 49*, 46–56.

Katsis, C. D., Katertsidis, N., Ganiatsas, G., & Fotiadis, D. I. (2008). Toward emotion recognition in car-racing drivers: A biosignal processing approach. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 38*(3), 502–512. https://doi.org/10.1109/tsmca.2008.918624

Kübler, T. C., Kasneci, E., Rosenstiel, W., Schiefer, U., Nagel, K., & Papageorgiou, E. (2014). Stress-indicators and exploratory gaze for the analysis of hazard perception in patients with visual field loss. *Transportation Research Part F: Traffic Psychology and Behaviour, 24*, 231–243. https://doi.org/10.1016/j.trf.2014.04.016

Lanata, A., Valenza, G., Greco, A., Gentili, C., Bartolozzi, R., Bucchi, F., … Scilingo, E. P. (2015). How the autonomic nervous system and driving style change with incremental stressing conditions during simulated driving. *IEEE Transactions on Intelligent Transportation Systems, 16*(3), 1505–1517. https://doi.org/10.1109/TITS.2014.2365681

Lee, D. S., Chong, T. W., & Lee, B. G. (2017). Stress events detection of driver by wearable glove system. *IEEE Sensors Journal, 17*(1), 194–204. https://doi.org/10.1109/JSEN.2016.2625323

Li, S. Z., Xie, G., Ren, J. C., Guo, L., Yang, Y. Y., & Xu, X. Y. (2020). Urban PM2.5 concentration prediction via attention-based CNN-LSTM. *Applied Sciences (Switzerland), 10*(6), Article e1953. https://doi.org/10.3390/app10061953

Lin, Z., Feng, M., Dos Santos, C. N., Yu, M., Xiang, B., Zhou, B., & Bengio, Y. (2017). April). A structured self-attentive sentence embedding. *5th International Conference on Learning Representations, ICLR 2017*.

Liu, Y.u., Chen, X., Peng, H.u., & Wang, Z. (2017). Multi-focus image fusion with a deep convolutional neural network. *Information Fusion, 36*, 191–207.

Liu, W., Zheng, W.-L., & Lu, B.-L. (2016). Emotion recognition using multimodal deep learning. *Lecture Notes in Computer Science*, 521–529.

Liu, F., Zhou, X., Wang, T., Cao, J., Wang, Z., Wang, H., & Zhang, Y. (2019). July). An Attention-based Hybrid LSTM-CNN Model for Arrhythmias Classification. In *2019 International Joint Conference on Neural Networks (IJCNN)*. https://doi.org/10.1109/ijcnn.2019.8852037

Nakisa, B., Rastgoo, M. N., Rakotonirainy, A., Maire, F., & Chandran, V. (2018). Long short term memory hyperparameter optimization for a neural network based emotion recognition framework. *IEEE Access, 6*, 49325–49338. https://doi.org/10.1109/ACCESS.2018.2868361

Nakisa, B., Rastgoo, M. N., Rakotonirainy, A., Maire, F., & Chandran, V. (2020). Automatic emotion recognition using temporal multimodal deep learning. *IEEE Access*. https://doi.org/10.1109/ACCESS.2020.3027026

Nakisa, B., Rastgoo, M. N., Tjondronegoro, D., & Chandran, V. (2017). Evolutionary computation algorithms for feature selection of EEG-based emotion recognition using mobile sensors. *Expert Systems with Applications, 93*, 143–155. https://doi.org/10.1016/j.eswa.2017.09.062

Neverova, N., Wolf, C., Lacey, G., Fridman, L., Chandra, D., Barbello, B., & Taylor, G. (2016). Learning human identity from motion patterns. *IEEE Access, 4*, 1810–1820.

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., & Ng, A. Y. (2011). Multimodal deep learning. In Proceedings of the 28th international conference on machine learning (ICML-11) (pp. 689–696).

Ordóñez, F., & Roggen, D. (2016). Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition. *Sensors, 16*(1), 115. https://doi.org/10.3390/s16010115

Palinko, O., Kun, A. L., Shyrokov, A., & Heeman, P. (2010). Estimating cognitive load using remote eye tracking in a driving simulator. In *In Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications* (pp. 141–144).

Pamplona, V. F., Oliveira, M. M., & Baranoski, G. V. G. (2009). Photorealistic models for pupil light reflex and iridal pattern deformation. *ACM Transactions on Graphics, 28*(4), 1–12. https://doi.org/10.1145/1559755.1559763

Pedrotti, M., Mirzaei, M. A., Tedesco, A., Chardonnet, J.-R., Mérienne, F., Benedetto, S., & Baccino, T. (2014). Automatic stress classification with pupil diameter analysis. *International Journal of Human-Computer Interaction, 30*(3), 220–236. https://doi.org/10.1080/10447318.2013.848320

Peng, Y., Tian, S., Yu, L., Lv, Y., & Wang, R. (2019). Malicious uniform resource locator attention-based CNN-LSTM. *KSII Transactions on Internet and Information Systems, 13*(11), 5580–5593. https://doi.org/10.3837/tiis.2019.11.017

Pourbabaee, B., Roshtkhari, M. J., & Khorasani, K. (2017). Deep convolutional neural networks and learning ecg features for screening paroxysmal atrial fibrillation patients. *IEEE Transactions on Systems, Man, and Cybernetics: Systems, 99*, 1–10.

Rastgoo, M. N., Nakisa, B., Maire, F., Rakotonirainy, A., & Chandran, V. (2019). Automatic driver stress level classification using multimodal deep learning. *Expert Systems with Applications, 138*, 112793. https://doi.org/10.1016/j.eswa.2019.07.010

Rastgoo, M. N., Nakisa, B., Rakotonirainy, A., Chandran, V., & Tjondronegoro, D. (2018). A critical review of proactive detection of driver stress levels based on multimodal measurements. *ACM Computing Surveys, 51*(5), 1–35. https://doi.org/10.1145/3186585

Rigas, G., Goletsis, Y., Bougia, P., & Fotiadis, D. I. (2011). Towards Driver's State Recognition on Real Driving Conditions. Retrieved from *International Journal of Vehicular Technology, 2011*, 1–14 http://www.hindawi.com/journals/ijvt/2011/617210/abs/.

Rigas, G., Goletsis, Y., & Fotiadis, D. I. (2012). Real-Time Driver's Stress Event Detection. *IEEE Transactions on Intelligent Transportation Systems, 13*(1), 221–234. https://doi.org/10.1109/tits.2011.2168215

Rodrigues, J. G. P., Kaiseler, M., Aguiar, A., Cunha, J. P. S., & Barros, J. (2015). A mobile sensing approach to stress detection and memory activation for public bus drivers. *IEEE Transactions on Intelligent Transportation Systems, 16*(6), 3294–3303. https://doi.org/10.1109/TITS.2015.2445314

Sainath, T. N., Vinyals, O., Senior, A., & Sak, H. (2015). Convolutional, Long Short-Term Memory, fully connected Deep Neural Networks. In *In 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4580–4584). https://doi.org/10.1109/icassp.2015.7178838

Sauerzapf, V. A. (2012). *Road Traffic Crash Fatalities: An Examination of National Fatality Rates and Factors Associated with the Variation in Fatality Rates between Nations with Reference to the World Health Organisation Decade of Action for Road Safety 2011–2020*. Retrieved from: University of East Anglia. https://ueaeprints.uea.ac.uk/id/eprint/46589.

Scanerstudio. (2020). https://www.avsimulation.com/scanerstudio/.

Selye, H. (1974). Stress without distress. *New York*, 26–39.

Soleymani, M., Pantic, M., & Pun, T. (2012). Multimodal emotion recognition in response to videos. *IEEE Transactions on Affective Computing, 3*(2), 211–223. https://doi.org/10.1109/t-affc.2011.37

Urtnasan, E., Park, J.-U., Joo, E.-Y., & Lee, K.-J. (2018). Automated detection of obstructive sleep apnea events from a single-lead electrocardiogram using a convolutional neural network. *Journal of Medical Systems, 42*(6). https://doi.org/10.1007/s10916-018-0963-0

Useche, S. A., Ortiz, V. G., & Cendales, B. E. (2017). Stress-related psychosocial factors at work, fatigue, and risky driving behavior in bus rapid transport (BRT) drivers. *Accident Analysis & Prevention, 104*, 106–114. https://doi.org/10.1016/j.aap.2017.04.023

Valiente, R., Zaman, M., Ozer, S., & Fallah, Y. P. (2019). Controlling Steering Angle for Cooperative Self-driving Vehicles utilizing CNN and LSTM-based Deep Networks. In *In 2019 IEEE Intelligent Vehicles Symposium (IV)* (pp. 2423–2428). https://doi.org/10.1109/ivs.2019.8814260

Vivoli, R., Bergomi, M., Rovesti, S., Bussetti, P., & Guaitoli, G. M. (2006). Biological and Behavioral Factors Affecting Driving Safety. Journal of Preventive Medicine and Hygiene 2006, 47, (pp. 69–73).

Wang, Y., Huang, M., Zhao, L., & Zhu, X. (2016). Attention-based LSTM for aspect-level sentiment classification. In 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016 (pp. 606–615).

Wang, J.-S., Lin, C.-W., & Yang, Y.-T.-C. (2013). A k-nearest-neighbor classifier with heart rate variability feature-based transformation algorithm for driving stress recognition. *Neurocomputing, 116*, 136–143. https://doi.org/10.1016/j.neucom.2011.10.047

Winata, G. I., Kampman, O. P., & Fung, P. (2018). Attention-Based LSTM for Psychological Stress Detection from Spoken Language Using Distant Supervision. In *In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6204–6208). https://doi.org/10.1109/icassp.2018.8461990

Wu, Y., Zhao, X., Rong, J., & Ma, J. (2013). Effects of chevron alignment signs on driver eye movements, driving performance, and stress. *Transportation Research Record, 2365*(1), 10–16. https://doi.org/10.3141/2365-02

Yan, F., & Mikolajczyk, K. (2015). Deep correlation for matching images and text. In *In Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3441–3450). https://doi.org/10.1109/CVPR.2015.7298966

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., & Hovy, E. (2016). Hierarchical attention networks for document classification. In *In 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 1480–1489). NAACL HLT 2016.

Zhang, Y., Chan, W., & Jaitly, N. (2017). In *Very deep convolutional networks for end-to-end speech recognition* (pp. 4845–4849). IEEE.
Zhang, L., Liu, F., & Tang, J. (2015). Real-time system for driver fatigue detection by RGB-D Camera. *ACM Transactions on Intelligent Systems and Technology, 6*(2), 1–17.
Zheng, W. L., Dong, B. N., & Lu, B. L. (2014). Multimodal emotion recognition using EEG and eye tracking data. In *In 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (pp. 5040–5043). https://doi.org/10.1109/embc.2014.6944757