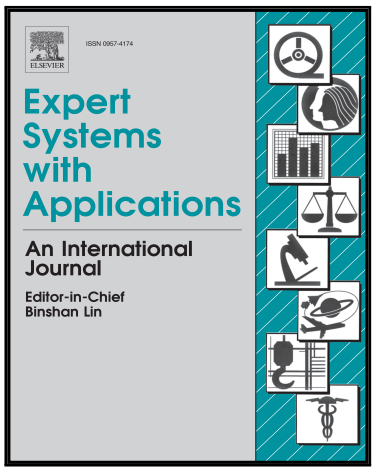


Automatic Driver Stress Level Classification Using Multimodal Deep Learning

Mohammad Naim Rastgoo , Bahareh Nakisa , Frederic Maire , Andry Rakotonirainy , Vinod Chandran

PII: S0957-4174(19)30489-0
DOI: https://doi.org/10.1016/j.eswa.2019.07.010
Reference: ESWA 12793

To appear in: *Expert Systems With Applications*

Received date: 26 February 2019
Revised date: 30 June 2019
Accepted date: 5 July 2019

Please cite this article as: Mohammad Naim Rastgoo , Bahareh Nakisa , Frederic Maire , Andry Rakotonirainy , Vinod Chandran , Automatic Driver Stress Level Classification Using Multimodal Deep Learning, *Expert Systems With Applications* (2019), doi: https://doi.org/10.1016/j.eswa.2019.07.010

Highlights

- Multimodal fusion model based on CNN-LSTM network to recognize driver stress level.
- First Deep learning approach applied to ECG, vehicle data and environmental data.
- Multimodal deep learning approach is effective in detecting driver stress level.
- Fusion approach using CNN-LSTM performs better than handcrafted feature extraction.

Automatic Driver Stress Level Classification Using Multimodal Deep Learning

Mohammad Naim Rastgoo, Bahareh Nakisa, Frederic Maire, Andry Rakotonirainy, Vinod Chandran

[a] School of Electrical Engineering and computer Science, Queensland University of Technology, Brisbane, QLD, Australia
[b] Centre for Accident Research & Road Safety-Queensland, Queensland University of Technology, Brisbane, QLD, Australia

Emails :{ mohammadnaim.rastgoo, Bahareh.nakisa, r.andry, f.maire, v.chandran} @qut.edu.au

ABSTRACT Stress has been identified as one of the contributing factors to vehicle crashes which create a significant cost in terms of loss of life and productivity for governments and societies. Motivated by the need to address the significant costs of driver stress, it is essential to build a practical system that can detect drivers' stress levels in real time with high accuracy. A driver stress detection model often requires data from different modalities, including ECG signals, vehicle data (e.g. steering wheel, brake pedal) and contextual data (e.g. weather conditions and other ambient factors). Most of the current works use traditional machine learning techniques to fuse multimodal data at different levels (e.g. feature level) to classify drivers' stress levels. Although traditional multimodal fusion models are beneficial for driver stress detection, they inherently have some critical limitations (e.g. ignore non-linear correlation across modalities) that may hinder the development of a reliable and accurate model. To overcome the limitations of traditional multimodal fusion, this paper proposes a framework based on adopting deep learning techniques for driver stress classification captured by multimodal data. Specifically, we propose a multimodal fusion model based on convolutional neural networks (CNN) and long short-term memory (LSTM) to fuse the ECG, vehicle data and contextual data to jointly learn the highly correlated representation across modalities, after learning each modality, with a single deep network. To validate the effectiveness of the proposed model, we perform experiments on our dataset collected using an advanced driving simulator. In this paper, we present a multi-modal system based on the adoption of deep learning techniques to improve the performance of driver stress classification. The results show that the proposed model outperforms model built using the traditional machine learning techniques based on handcrafted features (average accuracy: 92.8%, sensitivity: 94.13%, specificity: 97.37% and precision: 95.00%).

Keywords: Deep Learning; Driver stress detection; Convolutional Neural Network; Long short Term Memory; ECG signal; Vehicle data

1. INTRODUCTION

Stress is an unpleasant mental state can contribute to road traffic crashes and lead to a large numbers of injuries and fatalities every year. Stress can increase crash risk nearly tenfold, according to Virginia Tech Transportation Institute (Brown et al., 2016). Australian national crash reports also show that feeling stressed is a contributing factor to fatal crashes (Beanland, Fitzharris, Young, & Lenné, 2013). Reducing the numbers of traffic accidents and enhancing public road safety is becoming an urgent issue for governments and industries. Therefore, an accurate drivers' stress level detection system is needed to reduce the risk of crashes and to increase driver safety.

Recent research into automatic driver stress level recognition has shown that the use of multimodal data can substantially improve the classification performance (Healey & Picard, 2005; Katsis, Katertsidis, Ganiatsas, Fotiadis,

& others, 2008; Lanatà et al., 2015; Rigas, Goletsis, Fotiadis, & others, 2012). The multimodal data, which can monitor a driver's stress, includes the driver's physiological signals, physical responses and ambient parameters (Rastgoo, Nakisa, Rakotonirainy, Chandran, & Tjondronegoro, 2018). There are multitude studies that have used physiological signals to detect driver stress levels (Healey & Picard, 2000; Katsis et al., 2008; Singh, Conjeti, & Banerjee, 2013; Urbano, Alam, Ferreira, Fonseca, & Simíões, 2017). Among the various physiological responses, electrocardiogram (ECG) signals which are influenced largely by momentary Autonomic nervous system (ANS) activities are known to be an important stress indicator (Lee et al., 2007). The ECG signal represents the heart's electrical activity over time. In a stressful driving situation, the sympathetic part of the ANS will be activated instantaneously, which results in an increase in heart activity (Rastgoo et al., 2018). In addition to physiological responses to stress, drivers' stress responses can be also detected through physical or behavioural reactions. In a stressful situation, drivers physically react to control vehicles and avoid crashes. The reaction time is fast and depending on the type of reaction can vary from milliseconds to seconds (Green, 2000). These physical reactions of drivers can be monitored using vehicle data such as steering wheel, vehicle acceleration and deceleration data (Bořil, Sadjadi, Kleinschmidt, & Hansen, 2010; Lanatà et al., 2015; D. S. Lee et al., 2017; Rigas et al., 2012).

Apart from a driver's physical responses, contextual data contains important information for detecting drivers' stress level. Ambient or environmental data as an important contextual data can affect stress level of drivers (Hill & Boyle, 2007). The environmental data include different ambient information such as weather, visibility level, time of the day, driving routes, and other drivers' behaviours.

Fusing environmental data with physiological and physical stress responses can be beneficial for building an accurate driver's stress level detection model. This is due to that fact that data from different modalities can describe the same stressful event and each modality carries stress related information. Moreover, there is some stress related information across the modalities which can provide the complementary information to improve the performance of driver stress detection. In order to extract such information, it is important to capture the correlation between modalities with a compact set of latent variables.

To date, most of the studies have applied traditional multimodal fusion models like sensor-level, feature level or decision level fusion (Healey & Picard, 2005; Katsis et al., 2008; Lanatà et al., 2015; Rigas et al., 2012), whereby all the extracted features from each modality are combined into one vector to use in a next data analysis step. However, these approaches are not able to capture the non-linear correlation across data modalities, as the correlation between features in each modality is stronger (Ngiam et al., 2011). This is because, these sort of approaches focus on learning the patterns within each modality separately, rather than learning patterns that occurs simultaneously across multiple data modalities.

Moreover, most of the existing studies rely solely on extracting handcrafted features from different modalities to build multimodal fusion models. However, handcrafted feature extraction techniques are time-consuming, ad-hoc and require an in-depth knowledge and expertise (Pourbabaee, Roshtkhari, & Khorasani, 2017).

In recent years, deep learning (DL) techniques have been successfully applied in different contexts to build multimodal fusion models (Hu & Li, 2016; Huang & Kingsbury, 2013; Kanjo, Younis, & Ang, 2019; Karpathy et al., 2014; Y. Liu, Chen, Peng, & Wang, 2017; Ngiam et al., 2011; Yang et al., 2017). Deep learning approaches relieve the burden of manually extracting handcrafted features. Multimodal fusion models based on deep learning methods have been proposed to jointly learn and explore the highly correlated representation across modalities after learning each modality data,

using a single deep network. Moreover, deep learning approaches remove the need for expert input and are able to construct the salient features from each modality.

In the context of driving, multimodal deep learning approaches have been applied for path prediction using a fusion of LiDAR and GPS (Aranjuelo, Unzueta, Arganda-Carreras, & Otaegui, 2018; Virdi, 2017), pedestrian detection for autonomous vehicles (Liu, Zhang, Wang, & Metaxas, 2016), predicting driver action via the fusion of image, speed and angular velocity (Chi & Mu, 2017), and action prediction using steering action and vehicle status (Xu, Gao, Yu, & Darrell, 2017). However, fusion of multi-model data based on deep learning approaches has not previously been investigated for driver stress detection.

In this study, we propose a multimodal deep learning model with the aim of fusing ECG signals, vehicle dynamics data and environmental data. The proposed model is based on convolutional neural networks and long-short term memory networks (CNN-LSTM). The multimodal deep learning model based on CNN-LSTM networks helps in fusing the ECG signals, vehicle dynamics data and environmental data to capture the driver stress related information within and across the modalities.

In this study, drivers' stress is categorized into three stress levels: low, medium and high. The proposed model is evaluated on our dataset collected from an advanced driving simulator. This dataset contains ECG signals, vehicle dynamics data and environmental data from 27 participants captured from multiple scenarios made up of driving situations designed to induce different levels of stress. The performance of the proposed model is also compared with models based on handcrafted feature extraction methods.

This study's contributions are:

- To propose a framework based on a deep learning technique (CNN-LSTM) to accurately build a drivers' stress level detection model based on the fusion of ECG, vehicle data and contextual data. To the best of our knowledge, no other work has applied deep learning to a combination of modalities like physiological signal, physical and environmental data for stress detection.
- To conduct various experiments to compare the performance of the proposed multimodal deep learning technique with frameworks based on traditional multimodal fusion models (handcrafted feature extraction method with feature level fusion). In addition, the analysis and fusion of drivers' physiological, physical and environmental data to explore their significance in stress detection is presented.

2. RELATED WORK

Due to the multimodal nature of stress, interpreting and analysing the multimodal data together is recommended to build a robust and reliable stress detection model. Different modalities can be used to measure a driver's stress including driving stressors, the driver's ambient and individual parameters (contextual data), and the driver's psychological, physiological and physical responses to stress. Given the relative ease of collecting physiological signals and vehicle dynamics data, researchers have investigated the correlation between these data and drivers' stress levels (Lanatà et al., 2015; Lee et al., 2017; Rigas et al., 2012). The majority of the works in the literature employ the traditional statistical analysis methods (tradition machine learning techniques), whereby, a number of handcrafted features are extracted from the data, and then computational models are built to classify driver stress levels. Although handcrafted features yield promising results, the quality of the selected features significantly affects the classification performance. Therefore, extracting the most important representative and critical features is always a challenging problem (Bahareh Nakisa, Rastgoo, Tjondronegoro, & Chandran, 2017). Moreover, extracting salient features using expert

knowledge is time-consuming and ad-hoc and the extracted features are not always robust to variations such as noise and scaling.

Several studies have taken advantage of combining different modalities to build an accurate driver stress level detection models with acceptable detection speed. Urbano et al. (2017) proposed a model using a fusion of ECG and electrodermal activity (EDA) signals to detect two stress levels of drivers using a linear discriminant analysis classifier. The built models achieved 81–97% accuracy. In another study (Singh, Conjeti, & Banerjee, 2013), a three-level stress level detection model based on the fusion of photoplethysmogram (PPG), EDA and respiration (RSP) signals was built using a recurrent neural network classifier. The precision, sensitivity and specificity of the proposed model for discriminating three stress levels of drivers (low, medium, and high) were reported to be 89.23%, 88.83% and 94.92%, respectively. Rigas et al. (2012) fused several features extracted (feature level fusion) from ECG, electrodermal activity (EDA), respiration signals (RSP), and vehicle dynamics data to detect two stress levels of drivers. The features were fed to a naïve Bayes classifier and achieved 96% accuracy. In another study (Lanatà et al., 2015), 42 features extracted from ECG, EDA, RSP and vehicle dynamics data (steering wheel, car velocity and driver response time) and then fused (feature level fusion) to detect three stress levels of drivers using a nearest mean classifier. The built model achieved over 90% accuracy. Rigas, Goletsis, Bougia, & Fotiadis (2011) extracted a number of features from physiological signals (ECG, EDA and RSP), physical data (head movement) and contextual data (environmental data) and fused them to build a driver stress detector. The built model using an SVM classifier achieved 86% accuracy.

The fusion of the data in most of these studies is performed at the feature level. Using this approach, feature data from each modality is concatenated to form one feature vector which is used to solve classification problems. However, this approach is not able to capture the non-linear correlation in multimodal time-series data, as the correlation between features in each modality is stronger (Ngiam et al., 2011). This is because, these sorts of approaches focus on learning the patterns within each modality separately rather than learning patterns that occur simultaneously across multiple data modalities.

To overcome the difficulties in obtaining effective and robust features and capturing the non-linear correlation across the modalities, deep learning (DL) techniques have been proposed. DL techniques have the ability to learn the features from the raw data and can be actively applied to multidimensional signal processing due to their state-of-the-art performance and strong capabilities in constructing reliable features in different fields such as speech recognition (Hinton et al., 2012) and time-series data analysis (Y. Liu et al., 2017; Zheng, Liu, Chen, Ge, & Zhao, 2014).

Among the DL methods, convolutional neural networks (CNNs) have been successfully used for constructing strong and suitable features for different problems (Burkert, Trier, Afzal, Dengel, & Liwicki, 2015; Dwivedi, Biswaranjan, & Sethi, 2014; Hajinoroozi, Mao, Jung, Lin, & Huang, 2016; Yan, Teng, Smith, & Zhang, 2016). The strong feature learning capabilities of CNNs make them an ideal choice for multidimensional signal processing applications. CNNs are artificial neural networks that can capture the local dependencies and invariant features in the data. CNNs with different layers can first extract local, low-level features from the raw input and then increasingly more global and high level features in deeper layers. Experimental results confirm that CNNs surpass traditional machine learning approaches (Hajinoroozi et al., 2016; Zhu et al., 2014). Zhu et al. (2014) proposed a CNN network using electrooculography (EOG) signal to detect driver drowsiness and the obtained result using deep learning is proven to be more efficient in drowsiness compared to manual ad-hoc feature extraction. Acharya, Oh, Hagiwara, Tan, & Adeli (2018) proposed a 13-layer deep convolutional neural network (CNN)

is implemented to detect normal, preictal, and seizure classes. Using a CNN model, different studies could model complex models to classify different emotions using physiological signals (Chen & Jin, 2015; Martinez, Bengio, & Yannakakis, 2013).

Long Short-Term Memory (LSTM) networks, a special type of Recurrent Neural networks (RNNs), have gained popularity due to their ability to exploit the temporal dependencies in time-series data. Recently, LSTM networks achieved state-of-the-art performance in different domains such as machine translation, voice recognition and emotion recognition (Hinton et al., 2012; Nakisa, Rastgoo, Rakotonirainy, Maire, & Chandran, 2018; Neverova et al., 2016; F. Yan & Mikolajczyk, 2015). Recently, LSTM network is applied to driving domain and obtained a promising performance. Wollmer et al. (2011) proposed a novel algorithm for driver's distraction and showed that LSTM enable a reliable subject independent distraction model with the accuracy of 96%. More modern approaches are proposed based on the combination of CNN and LSTM networks across a wide variety of domains to understand long-term context (Valiente, Zaman, Ozer, & Fallah, 2019; Zhang, Chan, & Jaitly, 2017). For example, Donahue et al. (2015) proposed a new model that combines LSTM with CNN networks for visual recognition problem. Zhang et al. (2017) explored a very deep Convolutional LSTMs for speech recognition.

Moreover, deep learning techniques have been shown to be effective methods in generating a joint representation across modalities in different domains such as audio-visual speech recognition (Huang & Kingsbury, 2013), predicting driver actions using image and speed and angular velocity (Xu et al., 2017), and action prediction using steering action and vehicle status (Chi & Mu, 2017). However, the temporal information is not considered in these deep learning networks, which deviate from the natural properties of time-series data.

Recently, some studies attempted to model temporal multimodal data to capture the temporal information about the multimodal sequences (Karpathy et al., 2014; Liu et al., 2016; Liu et al., 2017; Yang et al., 2017) and the proposed models were evaluated using image and audio-visual data. However, temporal multimodal fusion based on deep learning techniques (CNN-LSTM network) has not been exploited in the domain of driving.

3. METHODOLOGY

In this section, a multimodal deep learning model based on CNN-LSTM networks to build a drivers' stress level detection model is presented. The proposed model is evaluated on our dataset collected from an advanced driving simulator. A description of dataset is provided in Section 3.1. Then, the proposed multimodal fusion framework is presented in Section 3.2.

3.1 Dataset Description

The dataset used in this study was recorded in response to different stressful situations in the context of driving. The experiment was conducted in an advanced driving simulator (Figure 1). Consisting of a car with an automatic transmission, front view (180-degree), rear-view mirrors, audio system, a hydraulic system to simulate vehicle motion and SCANeR™ studio software. The surrounding sounds such as engine, road noise and other traffic interactions sounds is simulated accurately by the audio system. Further information on the driving simulator can be found at https://research.qut.edu.au/carrsq/services/advanced-driving-simulator/

The physiological signals such as ECG signals, vehicle data and environmental data are acquired using the software: SCANeR™ studio and BIOPAC MP150. BIOPAC MP150 was used to acquire physiological signals such as ECG signals with a sampling rate of 1000 Hz which was down sampled to 200 Hz. In addition, SCANeR™ studio was used to acquire vehicle data such as brake pedal, gas pedal, steering wheel data and environmental data such as distance to next vehicle, number of

lanes, lane width, weather related data, visibility related data and time of the day.

Figure 1. CARRS-Q advanced simulator used for our data collection. The simulator consists of a car with an automatic transmission, front view (180-degree), rear-view mirrors, audio system, and hydraulic system to simulate vehicle motion, and a BIOPAC system to obtain physiological signals.

The sampling rate of SCANeR™ studio is set to 60 Hz. It should be noted that all the data from the SCANeR™ and BIOPAC software was synchronised. In this study, data were collected from 27 participants aged 21–40 years (55% male). All participants were required to have a valid driver's license in Australia, and to regularly drive for a total of at least one hour every week. The experiment for each participant took about one hour on average.

3.2 Data collection Scenarios

The experimental protocol is structured into two phases: pre-experiment, and driving scenario experiment.

Prior to the commencement of experiments, all participants were instructed via email about the experiment (task details, wearable sensing, data acquisition, driving routes, and safety instructions). Some restrictions such as to avoid drinking caffeine and alcohol prior to the data collection were applied to the participants.

Before starting driving, each participant was asked to relax for 2–3 minutes in order to record their physiological baseline. Then, the participant drove through six driving scenarios. Along with data from the experimental scenarios, driver's data (ECG signal, vehicle dynamic data and environmental data) were continuously acquired. In the first driving scenario (practice drive), the participant was asked to drive on a simple route to become familiar with simulator environment and how to control the car. After the practice drive, the participant drove on the next five driving scenarios: Urban1, Urban2, Highway, City1 and City2 landmarks. It should be noted that the scenario order was randomised across participants to avoid learning effects.

Each scenario contains several designed stressors to induce different stress levels into the driver. During each scenario, the drivers were asked every two minutes to provide their responses (verbally) to a short questionnaire about their stress levels (low, medium and high). The applied stressors in this study are derived from different studies (Hill & Boyle, 2007; Lee et al., 2017; Rodrigues, Kaiseler, Aguiar, Cunha, & Barros, 2015) and categorised into four groups.

3.2.1 Traffic

Traffic congestion is a resource of stress (Rastgoo et al., 2018). Several traffic densities were designed in this study in order to simulate driving in different traffic levels. Table 1 shows

the vehicle densities per km in the different driving scenarios.

Table1: Traffic densities in different driving scenarios

Scenarios	Number of Vehicles per Kilometre
Urban 1	0
Urban 2	30
Highway	50
CBD 1	50
CBD 2	60

Table 2: driving road situations for different scenarios

Scenarios	Narrow roads	Curve road	Tight corner
Urban 1	-	-	-
Urban 2	X	X	-
Highway	-	X	-
CBD1	X	X	X
CBD 2	-	X	X

Table3: Simulator parameters set for other vehicles to simulate bad driving behaviors

Simulator parameters
Stay on lane
Sign observing
Priority observing
Safety time
Speed limit risk
Overtake risk

Table 4: Weather-related condition for the driving scenarios.

Scenarios	Rain density (0-1)	Foggy weather	Night-time driving
Urban 1	-	-	-
Urban 2	-	-	X
Highway	(0.2-1)	X	X
CBD 1	(0.3-0.6)	X	-
CBD 2	-	-	-

3.2.2 Driving road situations

Several driving road situations such as driving in narrow roads, curved roads and tight corners are applied to induce different stress levels in drivers. Table 2 presents the used driving road situations in each scenario. It should be noted the range of curve radius in the proposed scenarios are varied between 17m and 1200m.

3.2.3 Other drivers' behaviours.

Other drivers' behaviours such as changing lane, dangerous overtaking, travelling at speed greater than allocated signage, tailgating can induce different stress levels into the driver. In this study, we have set several parameters in the simulator, listed in Table 3, to simulate the mentioned behaviors into some other drivers. It should be noted that the reported behaviors are set for the Highway and CBD1 and CBD 2 driving scenarios.

3.2.4 Weather and visibility related conditions

In this study, several weather-related conditions like rain, fog and driving at night are applied to the driving scenarios as stressor. Table 4 presents the list of weather conditions in each driving scenario.

3.3 Multimodal Fusion Architecture

This section presents a temporal multimodal deep learning model architecture. The proposed model aims to fuse ECG signals, vehicle dynamics data and environmental data over time into a joint representation to capture the stress related information within and across modalities over time and improve driver stress level classification. The proposed model consists of four different steps: pre-processed input data, feature learning, fusion and classification (see Figure 2).

In this study, the *pre-processed input data* are collected from ECG signals, vehicle dynamics data and environmental modalities. The raw ECG signals were used for this study.

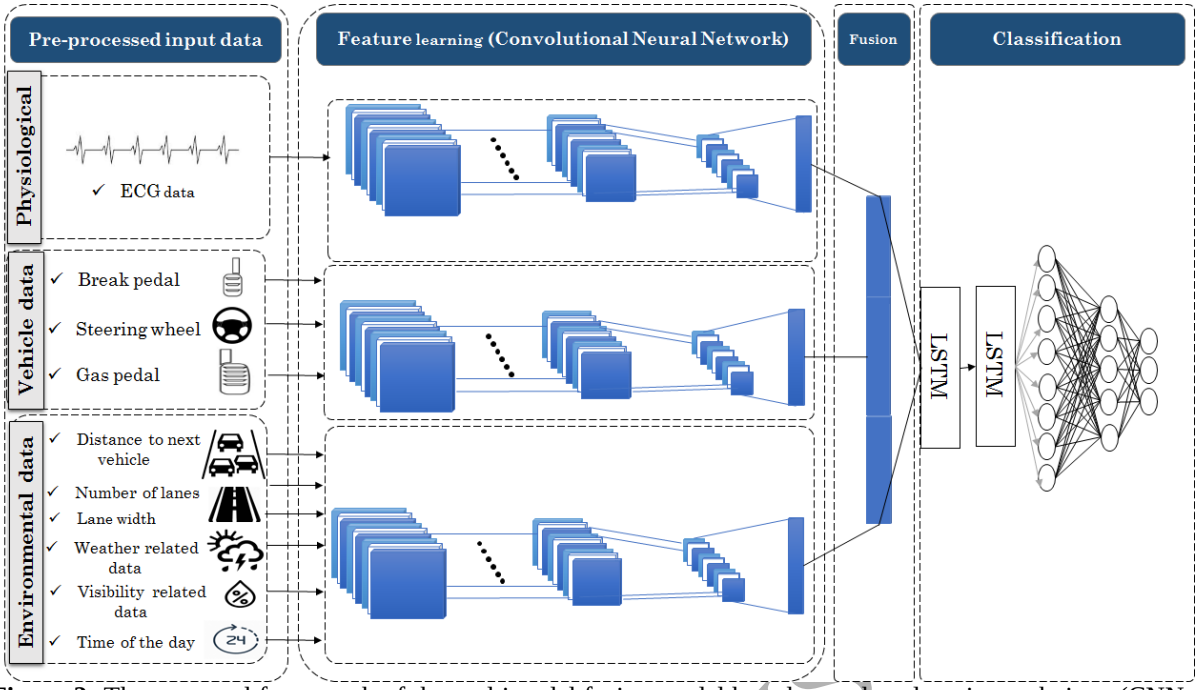


Figure 2: The proposed framework of the multimodal fusion model based on a deep learning technique (CNN-LSTM) to classify the stress level of drivers into three levels. The input of this system is the ECG signals, vehicle data (brake pedal, steering wheel, gas pedal) and contextual data (distance to next vehicle, lane width, number of lanes, time of the day, weather related conditions and visibility. The output of this model is the driver's stress level, classified into low, medium or high stress. Each modality is fed into individual CNN to extract feature maps (Feature learning step). The output of the feature maps from the ECG signals, vehicle data and contextual data is concatenated to form the joint representation (fusion step). The joint representation is fed into a two-layer LSTM followed by a dense layer (classification step). The output of the dense layer from the modalities is combined to create a joint representation and then is fed into a two-layer dense followed by a Softmax layer for stress level classification.

From the vehicle dynamics data, steering wheel, gas pedal, and brake pedal data are used in this study. From the contextual data, distance to next vehicle, lane width, number of lanes, time of the day, weather-related conditions data and visibility-related conditions data are used.

To prepare the data, it is assumed that we are given five drives per subject as each subject drove in five different stressful situations. Every two minutes in each drive, the driver's stress level is classified as low, medium or high stress and recorded. To remove subject specific physiological signal baseline and life-style factors influencing it, a min-max normalisation is applied. The normalised ECG signals are passed to a Butterworth band-pass filter (5–15 Hz) to reduce the noise from muscle noise and baseline wander. Then, the normalised and filtered ECG signals, vehicle dynamics data and contextual data are fed into the CNN architecture for feature learning.

To prepare the data for the CNN architecture, the sliding window strategy was applied to each low level feature from each modality. Using sliding window strategy, the signals are segmented into a fixed window size and degree of overlap. The segmented windows are the new training data and will get the same labels as the original trials.

After that, each new training data (segmented window at time t) from each modality is fed into the CNN model to learn features. In this step, the CNN architecture is applied to the segmented window (e.g. window t) from the raw ECG signals as well as low level features from vehicle and contextual data to construct feature maps and learn hierarchical features. The CNN presented in this study is composed of two convolutional max-pooling blocks (see

Table 5.), each of which is constituted by a convolutional layer, an Exponential Linear Unit (ELU), a batch normalisation layer and a max pooling layer. Since in this study, all the used modalities/ low features from each modality are time-series, the 1D convolution layer was used. The convolution layer convolves the low-level features of each modality (window t) or the previous layer's output with the set of filters to be learned.

It captures the temporal information using trainable filters with the fixed size. The second layer of the CNN architecture, the activation function, Exponential Linear Units (ELU), maps the output of previous layer. In the next layer, a batch normalisation layer normalises the value of different feature maps in the previous layer. Finally, the last layer, a max-pooling layer, finds the maximum feature maps over a range of local neighbourhood.

It should be noted that during the process of feature learning from each modality by the CNN architecture, both the current input data (window t) and its history are considered. Specifically, at window t, the recent per-modality history (CNN^{t-1}) is appended to the current window to obtain the current feature maps representation. The proposed number of filters, window size and strides in the CNN architecture are selected by trial and error to achieve high detection accuracy.

In the *fusion* step, the generated feature maps from ECG signals, vehicle dynamics data and contextual data at time t are concatenated to create one vector feature map (joint representation) from all the modalities. The generated joint representation at time t is fed into the *classification* step, a two-layer LSTM network followed by a two-layer dense and a Softmax layer to model the overall multimodal feature representation.

It should be noted that an LSTM network consists of hidden state or memory, which can help to store its previous hidden layers and learn the stress related information over time. Thus, the output of the $LSTM^t$ is computed using the current state as well as the previous hidden states (*t-1*), which can capture the temporal pattern of the previous joint representations as well. We should emphasise that this architecture not only tries to learn the patterns within each modality individually, but it also learns the patterns across modalities using the joint representations. This architecture is fully trained in an end-to-end manner and does not require any explicit feature extraction.

4. RESULTS

Extensive experiments were conducted to determine if the proposed multimodal fusion model (CNN-LSTM) can be used as an effective method for an accurate driver stress level detection model using the combination of ECG signals, vehicle dynamics data and contextual data. The performance of the proposed deep learning model is also compared with a handcrafted features approach (refer to Section 4.1 for details of the handcrafted feature approach). The ECG signals, vehicle data and contextual data were segmented into consecutive windows of a fixed size and degree of overlap. In this study, we evaluated the performance of the proposed temporal models using different window sizes. However, the degree of window overlap in this study is fixed, the raw physiological signals were segmented with 90% overlap.

The performance of the multimodal fusion models (handcrafted features and deep learning model) are evaluated and compared with different window sizes (Section 4.3.1).

Since the goal of this study is to build an accurate driver stress level detection model with the potential to be applied to real-time applications, the performance of the multimodal fusion models are evaluated on small window sizes.

In order to train our models, learning batches of 10 sequences were used. The early stopping on validation set is also applied.

Table 5: The CNN architecture

CNN	
Convolutional Layer	Filter= 20, Kernel size=(10, 1), Stride=2
Exponential Linear Units (ELU)	Alpha= 0.1
Batch Normalisation + Dropout (0.15)	
Max-Pooling	Pool-size= (2, 1), Stride=2
Convolutional Layer	Filter=20, Kernel size= (10, 1), Stride=2
Exponential Linear Units(ELU)	Alpha=0.1
Batch Normalisation+ Dropout (0.15)	
Max-pooling	Pool-size= (2, 1), Stride=2

4.1 Handcrafted features approach for comparison with the proposed model

To evaluate the performance of the proposed model based on a CNN-LSTM network and to present the effectiveness of this model, we provided handcrafted feature methods for comparison. This section presents the framework of driver stress level detection model based on feature extraction methods.

The framework based on feature extraction from each modality (ECG, vehicle data and contextual data) contains three main steps: pre-processing, feature extraction, and classification.

In the first stage, a Butterworth band-pass filter (5–15 Hz) is applied to the ECG signals to reduce the noise from muscle noise and baseline wander. The two other modalities, vehicle data and contextual data are normalised with a zero mean and unit variance.

To extract features from the ECG signal, Heart Rate Variability (HRV) parameter is used. This parameter is considered to be a low-level ECG feature. HRV is defined as the time fluctuations between sequences of successive heart beats. To measure HRV, first R-peaks are extracted from the ECG signal using the Pan-Tompkins algorithm (Pan & Tompkins, 1985), and then based on the extracted peaks, the HRV is measured. Afterwards, the most common time-domain features from HRV are extracted. HRV time-domain features are influenced largely by momentary ANS activities; thus, HRV time-domain analysis can be used to measure instantaneous driver stress responses (Lee et al., 2007). Statistical features such as mean and standard deviation are commonly used to detect drivers' stress levels (Rastgoo et al., 2018). The mean of the first difference of the HRV data, average normal-to-normal (NN) and intervals, standard deviation of normal-to-normal intervals (SDNN), square root of the mean squared difference of successive normal-to-normal intervals (RMSSD), and number of pairs of successive normal-to-normal intervals that differ by more than 50 ms (PNN50) are the most common HRV time-domain features extracted in relation to driver stress detection (Guo, Brennan, & Blythe, 2013; Healey & Picard, 2000; Katsis et al., 2008; Lanatà et al., 2015; Wang, Lin, & Yang, 2013) . The list of extracted features in this study is tabulated in Table 6.

Extracted features from vehicle dynamics data are mean of steering wheel, mean of gas pedal data, and mean of brake pedal data. These features shows the instantaneous body reactions to control the car. Therefore, we can use the short window sizes to capture their information related to stress.

Generally, environmental features can be divided into four main categories: (1) weather-related conditions, (2) visibility-related conditions, (3) driver environment interactions, and (4) driving routes data. In our previous work, the common environmental features extracted from these categories to detect driver stress level was reviewed (Rastgoo et al., 2018).

The advanced driving simulator used in this study is able to collect a group of these features such as distance to next vehicle, lane width, number of lanes, time of the day, weather info (sun/ low rain, medium rain, high rain) and fog (Lee et al., 2017; Lanatà et al., 2015, Rigas et al., 2012).

Table 6. Handcrafted features from ECG signals, vehicle dynamics data and environmental data

	Handcrafted Features
ECG features (HRV time-domain)	mean
	Standard deviation
	Mean of the first difference of HRV
	average normal-to-normal (NN) and intervals
	standard deviation of normal-to-normal intervals (SDNN)
	square root of the mean squared difference of successive normal-to-normal intervals (RMSSD)
	number of pairs of successive normal-to-normal intervals that differ by more than 50 ms (PNN50)
Vehicle dynamics data	Mean of steering wheel angle
	Mean of brake pedal data
	Mean of gas pedal data
Environmental data	distance to next vehicle
	lane width
	number of lanes
	time of the day
	weather info (sun/ low rain, medium rain, high rain, fog)

In the next step the extracted features from each modality are concatenated to create one single vector of features which is fed into a classifier to classify driver stress into three levels. In this study, the LSTM network with two stacked cells is used to classify drivers' stress levels.

It should be noted that all the extracted features from the various modalities are shown to be effective in detecting driver stress levels, and these features are used to compare the proposed framework based on CNN-LSTM with other works in the literature.

4.2 Experimental results

4.2.1 Fusion of modalities and comparison with handcrafted features model

In this section, we assess the performance of the proposed driver stress level detection model using a deep learning (CNN-LSTM) approach against the model using a handcrafted feature approach. The effectiveness of the models is evaluated based on different window sizes: 30 seconds, 10 seconds, and 5 seconds. Since the goal of this study is to build an automatic stress classification model with the potential to be applied to real-time applications, small window sizes were selected.

Table 7 presents the relative performance of the deep learning model and the handcrafted features model based on ECG signal, vehicle, and contextual data individually, and the fusion of the modalities using multimodal fusion models. The average accuracies of multimodal fusion models over 10 runs is calculated.

Based on Table 7, the multimodal fusion model with the CNN-LSTM network using the fusion of modalities over different window sizes outperformed the same model with a handcrafted features approach. The performance of each modality using handcrafted features and CNN-LSTM is also presented in the table. The false positive for three modalities as well as fusion of them are also reported. Of the three modalities, vehicle dynamics data followed by ECG signals, over different window sizes, performed better than the contextual data. When the window size was reduced, the average accuracies of vehicle dynamics data and ECG signal decreased.

While their false positives with smaller window sizes is higher than bigger window sizes. The average accuracies of vehicle

Table 7. The average performance (accuracy and false positive) of multimodal fusion models using CNN-LSTM and handcrafted features methods with different window sizes.

	Modalities	**T1= 30sec**		**T2= 10 sec**		**T3= 5 sec**	
		Accuracy	False Positive	Accuracy	False Positive	Accuracy	False Positive
Handcrafted features	ECG	74.3%	9.5%	73.8%	10.3%	70.4%	11.1%
	Vehicle dynamics data	78.1%	8.9%	77.4%	9.4%	72.5%	9.8%
	Environmental data	55.5%	16.5%	54.6%	14.4%	51.2%	15.5%
	Fusion	92.1%	4.19%	86.6%	13.7%	85.3%	14.1%
Automatic features extracted by CNN	ECG	**81.4%**	**8.6%**	**79.5%**	**8.3%**	**76.5%**	**8.3%**
	Vehicle dynamics data	**77.2%**	**9.1%**	**76.8%**	**8.8%**	**70.7%**	**8.7%**
	Environmental data	**58.2%**	**15.2%**	**55.4%**	**11.4%**	**53.6%**	**12.5%**
	Fusion	**96.3%**	**7.5%**	**93.3%**	**8.2%**	**92.8%**	**9.1%**

dynamics data and ECG signals using CNN-LSTM with a small window size (T3=5 sec) are still higher than the average accuracy of the handcrafted features model. The reported false positives using CNN-LSTM is better than using handcrafted features model.

The results also show that the fusion of modalities over different window sizes outperformed a single modality. Over the fusion of modalities, the average accuracy and false positive of CNN-LSTM with a short window size is better than the model with the handcrafted features approach.

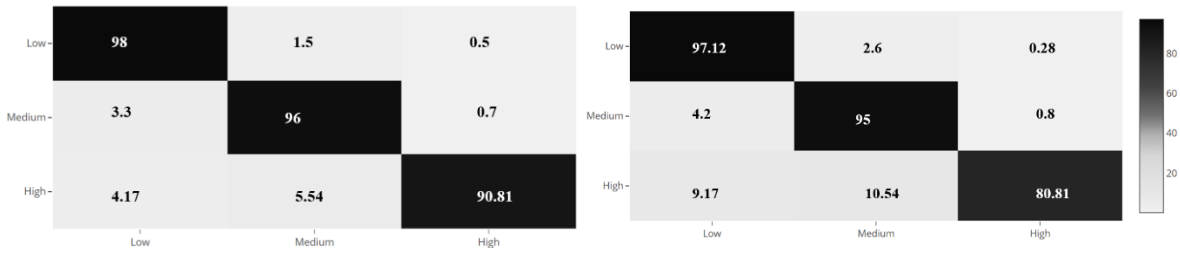
The average accuracy and false positive for deep learning model confirm the outstanding performance over the handcrafted feature model for driver stress levels recognition model

4.2.2 Confusion Matrices of Multimodal Fusion Model Based on CNN-LSTM and Handcrafted Features over Different Window Sizes

In this section, the best performance achieved using the multimodal fusion models (handcrafted features and CNN-LSTM) for different window sizes is investigated (Figures 3–5).

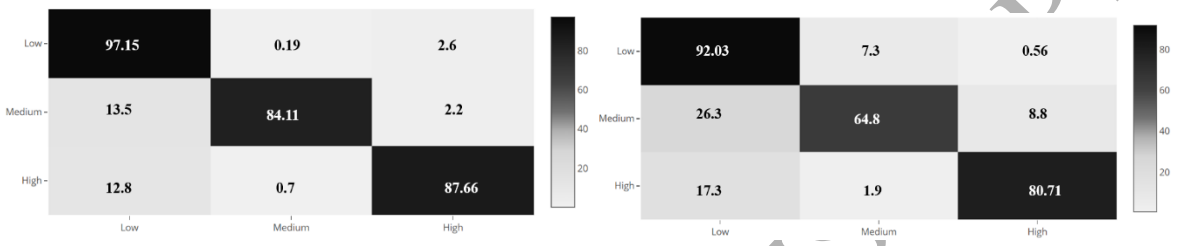
From all the figures, we can generally observe that recognising a high stress level is more difficult than recognising low stress or medium stress. In fact, a high stress level is mostly misclassified as medium stress, due to the fact that the number of samples for high stress are less than for the other two levels of stress.

Figure 3 shows the confusion matrices of driver stress level detection based on the CNN-LSTM network and handcrafted features with 30-second window respectively. Based on the figure, recognising high stress level using both models are more difficult than other stress levels. Although the performance of both models in detecting driver stress levels with a 30-second window is high, detecting high stress level using handcrafted features is more confusing than medium and low stress levels. Similarly, the performance of the CNN-LSTM network in detecting three stress levels with a 10-second window is better than the handcrafted features model (Figure 4). It has been also shown that high stress is often confused with medium stress. It can be seen from figure 5 that as the window size decreased, the performance of the two multimodal fusion models also decreased.



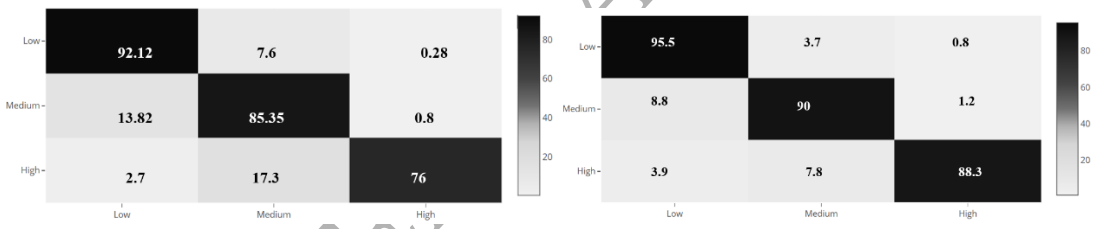
(a) (b)

Figure 3. The confusion matrix of the best model achieved using: (a) CNN-LSTM network, (b) handcrafted features, with a 30-second window.



(a) (b)

Figure 4. The confusion matrix of the best model achieved using: (a) handcrafted features, (b) CNN-LSTM network, with a 10-second window.



(a) (b)

Figure 5. The confusion matrix of the best model achieved using: (a) handcrafted features, (b) CNN-LSTM network, with a 5-second window size.

Based on figure 5, the accuracy of detecting low and medium stress levels using CNN-LSTM is higher than using the handcrafted features approach. The CNN-LSTM network can detect a high stress level better than the handcrafted features model.

4.2.3 Comparison of the proposed framework with other works

The final experiment compares the best achieved result (highest accuracy) from the multimodal fusion model using the CNN-LSTM network against some other recent works with state-of-the-art performance. Experimental results are shown in Table 8, indicating the classification accuracies of different driver stress level detection methods.

The comparison shows that most of the studies in the literature focused on the fusion of physiological modalities, which may not be practical for continuous stress monitoring of drivers and, therefore, adversely affect the acceptability of the system.

These compared studies in the literature categorized driver's stress into different stress classes.

Table 8. The comparison of the best performance achieved using the proposed model based on the CNN-LSTM network with other recent works.

Reference	No. of Subjects	Method	Used modalities			Classifier	Window size (second)	Performance	No. Classes
			Physiological	Physical	Context				
(Healey & Picard, 2000)	10	Handcrafted features	ECG, EDA and RSP signals	-	-	KNN	60	Accuracy: 88.6%	4 stress class (low, medium, high, very high)
(R. R. Singh et al., 2013)	19	Handcrafted features	PPG, EDA, HRV, RSP signals	-	-	LRNN	10	Specificity: 94.92% Sensitivity: 88.83% Precision : 89.23%	3 stress class (low, medium, high)
(Lanatà et al., 2015)	14	Handcrafted features	ECG, EDA and RSP	Vehicle dynamics data	-	(NMC)	5	Accuracy: 91.33% Sensitivity: 92.33% Specificity: 96.00%	3 stress class (no stress, stress level1, stress level 2)
(Rigas et al., 2011)	1	Handcrafted features	ECG, EDA, RSP	Eye, head movement	Environmental parameters	SVM	10	Accuracy: 86%	2 stress class (no stress-stress)
(Rigas et al., 2012)	13	Handcrafted features	ECG, EDA, RSP	Vehicle dynamic data	-	Bayesian Network	10	Accuracy: 96%	2-stress class (no stress-stress)
Our work	27	Deep learning (CNN-LSTM) network	ECG	Vehicle dynamic data	Environmental parameters	LSTM	5	Accuracy: 92.8% Sensitivity: 94.13% Specificity: 97.37% Precision : 95.00%	3 stress class (low, medium, high)

Based on the table, Healey and Picard (2000) categorized stress classes into four stress classes (low, medium, high and very high stress classes. Singh et.al (2013) proposed a model based on three stress classes (low, medium and high). Lanata et.al (2015) proposed three stress classes: normal (no-stress), stress level 1 (low and medium) and stress level 2 (high). Rigas et.al (2011) only focused on two stress classes, no stress versus stress. In our study, we focused on three stress classes, low, medium and high stress levels. It should be noted that high stress level in our study is equal to high and very high stress classes in Healey and Picard (2000).

Although, Rigas et.al (2012) proposed a model to fuse physiological signals with other modalities like vehicle data and contextual data, the applied methods were used to detect only two stress classes (stress versus non-stress) and the accuracy achieved was lower than that achieved using our model. Moreover, they utilised different physiological signals that may affect the acceptability of their system by drivers for real-time applications. In addition, most of these studies are not appropriate to rapidly detect driver stress, as they use long window sizes (longer than 5 seconds), whereas, we achieved state-of-the-art performance using a 5-second window.

Although Lanatà et al. (2015) classified three stress levels (normal, stress levels 1 and stress level 2) of drivers using a handcrafted features model based on a 5-second window and achieved a high accuracy, the achieved performance is not as high as that of our proposed model using the CNN-LSTM network.

In addition, the proposed model fused different physiological signals which is more obtrusive and invasive than our proposed model. In this study, we only fused ECG signals with vehicle dynamics data and contextual data, which can result in higher system acceptability compared to Lanatà et al. (2015).

In comparison with other recent works, our proposed model has shown state-of-the-art performance in building an accurate driver stress level detection model. There are some advantages in using our proposed model based on the CNN-LSTM model. The first advantage is that the proposed model using ECG signals, vehicle dynamics data and contextual data is less invasive and obtrusive than many of the other recent models. Therefore, it can be used for real-time applications. In addition, we

showed that by using a small window size, the proposed multimodal model can achieve a more accurate result. Lastly, the use of the CNN-LSTM network to build such a system, removes the need for expert knowledge to extract features, as required by many of the other recent models.

4.3 CONCLUSION

In this paper, we proposed an accurate driver stress level detection model using the multimodal fusion model based on deep learning techniques. Specifically, we used the CNN-LSTM network to automatically fuse ECG signals, vehicle dynamics data and contextual data to find a joint feature representation across multimodal data and enhance the detection performance. The performance of the proposed model with different window sizes was evaluated on our dataset collected from an advanced driving simulator. The results showed that using a multimodal fusion model based on the CNN-LSTM network with a small window size using ECG signals, vehicle dynamics data and contextual data, increased the accuracy of drivers' stress detection compared to the handcrafted features model. This is due to the fact that the multimodal fusion model based on deep learning (CNN-LSTM) is able to efficiently combine the complementary information across and within ECG signals, vehicle dynamics data and contextual data during the feature representation process. Also, the results showed that deep learning can be a promising approach for the study of driver's stress classification.

Authorship Statement:
All persons who meet authorship criteria are listed as authors, and all authors certify that they have participated sufficiently in the work to take public responsibility for the content, including participation in the concept, design, analysis, writing, or revision of the manuscript. Furthermore, each author certifies that this material or similar material has not been and will not be submitted to or published in any other publication before its appearance in the *Hong Kong Journal of Occupational Therapy*.

Conflict of interest
Author confirms there is no conflict of interest.

4.4 ACKNOWLEDGEMENTS

This Work is supported by QUT Postgraduate Research Award (QUTPRA).

REFERENCES:

Acharya, U. R., Oh, S. L., Hagiwara, Y., Tan, J. H., & Adeli, H. (2018). Deep convolutional neural network for the automated detection and diagnosis of seizure using EEG signals. *Computers in Biology and Medicine*, *100*, 270–278.

Aranjuelo, N., Unzueta, L., Arganda-Carreras, I., & Otaegui, O. (2018). Multimodal Deep Learning for Advanced Driving Systems. In *International Conference on Articulated Motion and Deformable Objects* (pp. 95–105). Springer.

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.

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

Brown, T. G., Ouimet, M. C., Eldeb, M., Tremblay, J., Vingilis, E., Nadeau, L. Bechara, A. (2016). Personality, executive control, and neurobiological characteristics associated with different forms of risky driving. *PLoS One*, *11*(2), e0150227.

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*.

Chen, S., & Jin, Q. (2015). Multi-modal dimensional emotion recognition using recurrent neural networks. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge* (pp. 49–56). ACM.

Chi, L., & Mu, Y. (2017). Deep steering: Learning end-to-end driving model from spatial and temporal visual cues. *ArXiv Preprint ArXiv:1708. 03798*.

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 *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2625–2634).

Dwivedi, K., Biswaranjan, K., & Sethi, A. (2014). Drowsy driver detection using representation learning. In *2014 IEEE International Advance Computing Conference (IACC)* (pp. 995–999). IEEE.

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

Guo, W., Brennan, D., & Blythe, P. (2013). Detecting Older Drivers' Stress Level during Real-World Driving Tasks. In *Proceedings of World Academy of Science, Engineering and Technology* (p. 1773). World Academy of Science, Engineering and Technology (WASET). Retrieved from http://search.proquest.com/openview/bbc6155bee386d6a2424b0747a13d70a/1?pq-origsite=gscholar

Hajinoroozi, M., Mao, Z., Jung, T.-P., Lin, C.-T., & Huang, Y. (2016). EEG-based prediction of driver's cognitive performance by deep convolutional neural network. *Signal Processing: Image Communication*, *47*, 549–555.

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

Healey, J., & Picard, R. (2000). SmartCar: detecting driver stress. In *15th International Conference on Pattern Recognition, 2000. Proceedings* (Vol. 4, pp. 218–221 vol.4). https://doi.org/10.1109/ICPR.2000.902898

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., Yu, D., Dahl, G. E., Mohamed, A., Jaitly, N., … Sainath, T. N. (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.

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

Huang, J., & Kingsbury, B. (2013). Audio-visual deep learning for noise robust speech recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on* (pp. 7596–7599). IEEE.

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

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., & Fei-Fei, L. (2014). Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition* (pp. 1725–1732).

Katsis, C. D., Katertsidis, N., Ganiatsas, G., Fotiadis, D., & others. (2008). Toward emotion recognition in car-racing drivers: A biosignal processing approach. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions On*, *38*(3), 502–512.

Lanatà, 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

Lee, H. B., Kim, J. S., Kim, Y. S., Baek, H. J., Ryu, M. S., & Park, K. S. (2007). The relationship between HRV parameters and stressful driving situation in the real road. In *2007 6th International Special Topic Conference on Information Technology Applications in Biomedicine* (pp. 198–200). https://doi.org/10.1109/ITAB.2007.4407380

Liu, J., Zhang, S., Wang, S., & Metaxas, D. N. (2016). Multispectral deep neural networks for pedestrian detection. *ArXiv Preprint ArXiv:1611.02644*.

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

Martinez, H. P., Bengio, Y., & Yannakakis, G. N. (2013). Learning deep physiological models of affect. *IEEE Computational Intelligence Magazine*, *8*(2), 20–33.

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*, 1–1. https://doi.org/10.1109/ACCESS.2018.2868361

Nakisa, Bahareh, 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*. 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).

Pan, J., & Tompkins, W. J. (1985). A real-time QRS detection algorithm. *Biomedical Engineering, IEEE Transactions On*, (3), 230–236.

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., Rakotonirainy, A., Chandran, V., & Tjondronegoro, D. (2018). A Critical Review of Proactive Detection of Driver Stress Levels Based on Multimodal Measurements. *ACM Comput. Surv*, *51*(5), 88:1–88:35. 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. *International Journal of Vehicular Technology*, *2011*. Retrieved from http://www.hindawi.com/journals/ijvt/2011/617210/abs/

Rigas, G., Goletsis, Y., Fotiadis, D., & others. (2012). Real-time driver's stress event detection. *Intelligent Transportation Systems, IEEE Transactions On*, *13*(1), 221–234.

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

Singh, R. R., Conjeti, S., & Banerjee, R. (2013). A comparative evaluation of neural network classifiers for stress level analysis of automotive drivers using physiological signals. *Biomedical Signal Processing and Control*, *8*(6), 740–754. https://doi.org/10.1016/j.bspc.2013.06.014

Urbano, M., Alam, M., Ferreira, J., Fonseca, J., & Simíões, P. (2017). Cooperative driver stress sensing integration with eCall system for improved road safety. In *IEEE EUROCON 2017 -17th International Conference on Smart Technologies* (pp. 883–888). https://doi.org/10.1109/EUROCON.2017.8011238

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. *ArXiv Preprint ArXiv:1904.04375*.

Virdi, J. (2017). *Using Deep Learning to Predict Obstacle Trajectories for Collision Avoidance in Autonomous Vehicles* (PhD Thesis). UC San Diego.

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.

Wollmer, M., Blaschke, C., Schindl, T., Schuller, B., Farber, B., Mayer, S., & Trefflich, B. (2011). Online driver distraction detection using long short-term memory. *IEEE Transactions on Intelligent Transportation Systems*, *12*(2), 574–582.

Xu, H., Gao, Y., Yu, F., & Darrell, T. (2017). End-to-end learning of driving models from large-scale video datasets. *ArXiv Preprint*.

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

Yan, S., Teng, Y., Smith, J. S., & Zhang, B. (2016). Driver behavior recognition based on deep convolutional neural networks. In *2016 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)* (pp. 636–641). IEEE.

Yang, X., Ramesh, P., Chitta, R., Madhvanath, S., Bernal, E. A., & Luo, J. (2017). Deep multimodal representation learning from temporal data. *CoRR, Abs/1704.03152*.

Zhang, Y., Chan, W., & Jaitly, N. (2017). Very deep convolutional networks for end-to-end speech recognition. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4845–4849). IEEE.

Zheng, Y., Liu, Q., Chen, E., Ge, Y., & Zhao, J. L. (2014). Time series classification using multi-channels deep convolutional neural networks. In *International Conference on Web-Age Information Management* (pp. 298–310). Springer.

Zhu, X., Zheng, W.-L., Lu, B.-L., Chen, X., Chen, S., & Wang, C. (2014). EOG-based drowsiness detection using convolutional neural networks. In *2014 International Joint Conference on Neural Networks (IJCNN)* (pp. 128–134). IEEE.