Classification of Driver Cognitive Load in Conditionally Automated Driving: Utilizing ECG-based Spectrogram with Light-weight Neural Network

Wenxin Shi
Thrust of Robotics and Autonomous Systems, Systems Hub,
The Hong Kong University of Science and Technology (Guangzhou), China
Email: wshi559@connect.hkust-gz.edu.cn

Zuyuan Wang
Thrust of Robotics and Autonomous Systems, Systems Hub,
The Hong Kong University of Science and Technology (Guangzhou), China
Email: zwang534@connect.hkust-gz.edu.cn

Ange Wang
Intelligent Transportation Thrust, Systems Hub,
The Hong Kong University of Science and Technology (Guangzhou), China
Email: awang324@connect.hkust-gz.edu.cn

Dengbo He, Corresponding Author
Thrust of Intelligent Transportation, Systems Hub
The Hong Kong University of Science and Technology (Guangzhou), China
Department of Civil and Environmental Engineering
The Hong Kong University of Science and Technology, Hong Kong SAR, China
Email: dengbohe@hkust-gz.edu.cn

Word Count: 4,627 words + 3 table (250 words per table) = 5,377 words

Submission Date: November 22, 2023

ABSTRACT

With the development of conditionally automated driving, drivers will be allowed to perform non-driving-related tasks. Under such circumstances, continuous monitoring of driver cognitive load plays an increasingly important role in ensuring that drivers have sufficient mental resources to take over control of the vehicle when the driving automation fails. However, cognitive load estimation is a challenging task due to the difficulties in identifying high-level feature representation and accounting for inter-individual differences. The physiological measure is believed to be a promising candidate for cognitive load estimation in partially automated vehicles. However, the existing estimation methods are mainly based on feature manual extraction of time-domain or frequency-domain indicators from physiological signals, which may not adapt to dynamic driving conditions. With the development of deep learning, the neural network has shown good performance in capturing high-level features from input data automatically. Inspired by this, we adopt a novel approach to classify driver cognitive load based on electrocardiogram (ECG) spectrograms, in which the driver's ECG signal is collected and transformed into a 2D spectrogram by a short-time Fourier Transform. A SENet-based deep learning framework that can capture high-level features and pay more attention to the cognition-related features from the spectrogram is proposed for classification. Experiments in a publicly available dataset demonstrate that our method achieves an accuracy of 96.76% in differentiating two levels of cognitive load in the within-subject evaluation and 71.50% accuracy in the across-subjects evaluation. The results demonstrate the feasibility of detecting drivers' cognitive load through deep learning using ECG Spectrogram alone.

Keywords: Automated Driving, Cognitive Load, Driver State Estimation ECG signal, Spectrogram

INTRODUCTION

The development of autonomous driving technology makes it possible for vehicles with conditionally automated driving systems (SAE L3, (*1*)) or higher to enter the consumer market within the foreseeable future. According to the latest McKinsey Company forecast (*2*), 4% of newly sold vehicles are expected to be equipped with an L3-level conditionally automated driving system by 2030; and in 2035, this number may increase to 14%. No doubt conditionally automated driving can create huge value for drivers and society. With conditional automation, vehicles can control the speed and direction by themselves and monitor the surrounding traffic environment on roads with good traffic conditions. This allows drivers to shift their attention away from the driving tasks to non-driving tasks and make better use of their commute time. However, L3 automation is not fully autonomous and drivers may still need to take over control of the vehicle when necessary (*3*). Drivers' performance in regaining control of their vehicles can be affected by several factors and the cognitive load of drivers is one of them.

Cognitive load is defined as "*a multidimensional construct representing the load that is imposed on the cognitive system while performing a particular task*" (*4*). Previous research found that high cognitive load among drivers can impair their capability to anticipate hazards in non-automated vehicles (*5*). With driving automation, the cognitive load has been found to impair the takeover performance (*6*). Previous research claimed that although driving automation can reduce drivers' perceived cognitive load, it can cause higher cognitive load, given that drivers may still need to monitor not just the traffic, but also the states of the driving automation (*7*). What is more alerting is that drivers are more willing to engage in non-driving-related tasks (NDRTs) with driving automation, which may further occupy their cognitive resources, leading to a high cognitive load. For example, a survey study found that about 45% of US drivers and 32% of UK drivers are willing to engage in secondary tasks with higher cognitive demands in autonomous vehicles (*8*). Given the detrimental effects of high cognitive load in conditionally automated vehicles and drivers' willingness to engage in NDRTs, it is necessary to continuously monitor the driver's cognitive load and take actions to ensure drivers' remaining cognitive resources are within a reasonable range to guarantee driving safety in such vehicles.

In the past few years, many algorithms have been proposed to estimate drivers' cognitive load in non-automated vehicles. Most of these previous works utilized three types of measures, i.e., driving performance measures (e.g.,(*9*, *10*)), eye-tracking measures (e.g.,(*11–13*)), and physiological measures (e.g.,(*9*, *10*, *14*)). In recent years, deep learning has also been applied to estimate driver cognitive load. For example, Prithila Angkan et al. extracted 21 features from Electroencephalogram (EEG), Electrodermal activity (EDA), and eye movement data, and utilized VGG Net (*15*) and ResNet (*16*) networks to train a cognitive load classification model (*17*). Hamidu Rahman et al. presented a vision-based method to extract useful features from driver's eye movement based on domain knowledge and achieved 91% through the convolutional neural network (*11*). Haohan Yan et al. proposed an attention neural network with decision-level fusion to learn features from EEG, eye movements, and vehicle states for classification (*18*). Table 1 summarizes previous research on driver cognitive load estimation in non-automated vehicles. However, although the driving performance data has been widely adopted in cognitive load detection in non-automated vehicles (*19*, *20*), they become unavailable in conditionally automated driving, as the driver is not controlling the vehicle most of the time. Similarly, as drivers may engage in NDRTs, the eye-tracking measures that were found to be effective in non-automated vehicles may also become invalid in conditionally automated vehicles. Thus, the physiological measures seemed to be a more promising option for real-time cognitive estimation in vehicles with driving automation. Previous studies found that various indicators extracted from physiological signals are highly related to cognitive load in drivers. For example, heart rate (HR) and HR variability (HRV) extracted from electrocardiogram (ECG) signals are closely related to the activity of the autonomic nervous system, and thus they can explain the mental states of individuals such as fatigue and cognition (*21*). In addition to the ECG signal, other indicators extracted from EEG, Galvanic Skin Response (GSR), and Respiration are also sensitive to changes in driver cognitive load (*22*, *23*).

However, it is worth mentioning that, although previous research has developed a variety of algorithms to differentiate cognitive load based on physiological measures, to the best of our knowledge, all of them relied on hand-crafted feature extraction and feature selection, which is time-consuming and may lead to loss of information, as feature extraction and selection are processes of information reduction. For example, researchers (*10*, *14*, *24*) extracted SDNN (the standard deviation of normal-to-normal intervals) and pNN50 (the percentage of adjacent NN intervals differing by more than 50 ms) from HRV as indicators to predict cognitive load. These features can be categorized as low-level features, which are not discriminative enough to differentiate the driver's cognitive states. In contrast, the features that are automatically learned from the deep learning models are called high-level features, which can keep more information compared to low-level features (*25*, *26*). For example, very rich information can be extracted from the ECG data. Even for HRV alone, tens of features can be extracted in both the time domain and frequency domain (*27*). Traditionally, only a few features are selected and fed into the machine learning models, based on domain knowledge or data-driven selection (e.g., principle component analysis (*28*)). However, information might be lost in the process. With deep learning, it is possible to realize automatic feature selection and reduce information loss in the process of extracting hand-crafted low-level features.

TABLE 1 Research on Cognitive Load Estimation in Drivers

Reference	Selected Features	Task	Time Window	Best Classifier	Performance	Evaluation
Darzi et al. (*29*)	Driver characteristics, vehicle kinematics, physiological measures	Cell phone use	240s	LR (2 classes)	82.3%	Across-subjects
Le et al. (*30*)	Hemodynamic data (oxyhemoglobin at the 35mm, DoxHb35, ToxHb35, etc.)	N-back	114s	Not mentioned (3 classes)	89.91%	Across-subjects
Rahman et al. (*10*)	Physiological signal, Facial video and Driving performance	N-back	180s	LR (2 classes)	90%	Not mentioned
Barua et al. (*31*)	ECG, EEG, EOG, GSR, Respiration and Driving performance features	N-back	50s	RF (2 classes)	79%	Not mentioned
Prabhakar et al. (*32*)	Pupil-based and Gaze-based metrics	Common Tasks in Vehicle	1s	NN (2 classes)	75%	Not mentioned
He et al. (*33*)	Eye-tracking, HR and GSR	N-back	5s	RF (3 classes)	97.8%	Within-subject
Hamidu Rahman et al. (*11*)	Eye-tracking data	N-back	30s	CNN	91%	Not Mentioned
Prithila Angkan et al. (*17*)	ECG, EDA, eye tracking data	Reading and arithmetic tasks	10s	VGG Net (3 classes)	74.54%	Not Mentioned
Haohan Yan et al.(*18*)	EEG, eye-movements, vehicle data	A visual-auditory mixed n-back task	60s	HyperLSTM-based module (2 classes)	95.0%	With-in subject

Thus, in this work, we evaluated the feasibility of estimating drivers' cognitive load using an ECG spectrogram instead of handcrafted features. We chose the ECG measure as the sole measure as vehicle manufacturers are already considering deploying ECG sensors on the steering wheel (*34*). Further, the spectrogram of ECG (instead of the ECG signal in the time domain) was selected because a lot of ECG features are in the frequency domain. The performance of several deep learning algorithms was compared when spectrogram versus handcrafted features were used as inputs. In addition to the classical deep learning models, we also designed a new model named SE-ECG Net, which uses a convolution module and an attention layer, the Squeeze-and-Excitation Network (SENet (*35*)), to automatically learn high-level features from ECG-based spectrograms to estimate drivers' cognitive loads.

The major contributions of this paper are summarized as follows. First, being different from previous studies that emphasized data fusion (*33*) and hand-crafted feature extraction, we used a single measure that is practically feasible in vehicles to estimate driver cognitive load. To explore the cognitive-load-related information in the ECG, we creatively transferred the ECG signals into spectrograms and designed a SENet-based deep neural network to keep high-level features in the image-like spectrograms. Given that the ECG sensor has the potential to be integrated into the steering wheel, the algorithms developed in this study have the potential to be deployed in mass-produced vehicles. Further, to the best of our knowledge, this is among the first works (in addition to Meterier et al. (*14*, *24*)) that focused on driver cognitive load estimation in SAE L3 vehicles.

CLASSIFICATION TASK AND SIGNAL PREPROCESSING

Dataset and Classification Task

Two datasets were used in this study. The first one (*Dataset 1*) is a publicly available dataset by Meteier et al. (*36*). In this dataset, 90 subjects (Mean age: 24.2, standard deviation: 6.0 years, 40 males, 49 females, and 1 other) finished conditionally automated driving (defined as SAE Level 3 by the authors). Half of the participants were asked to perform a verbal-cognitive task, which required them to count backward from 3645 by a step of 2 throughout the whole drive and they were assumed to be under high cognitive load; while the other half of participants did only the driving task, so that they had low cognitive load. The driving session for each participant lasted 20 minutes, and three physiological signals including ECG, Electrodermal activity (EDA), and Respiration were collected during the whole period, with a sampling frequency of 1000 Hz, leading to 120,000 samples (20 minutes * 60 seconds * 1000 Hz) per physiological signal channel for each participant. The experiment was conducted in a fixed-base driving simulator and all physiological data was collected using the sensors by BioPac. After experiments, the NASA-Task Load Index (NASA-TLX) ratings regarding the drive were collected. Then, statistical analysis was conducted to verify the induced high cognitive load in drivers. As a result, data from 87 subjects were deemed valid in the dataset. More details regarding the data can be found in Meteier et al. (*36*) and downloaded following the link: https://doi.org/10.5281/zenodo.7214953. The task of our study is to develop deep learning models to estimate the level of cognitive load of drivers (i.e., high versus low) based on the dataset using the ECG measures alone.

To further verify the generalizability of our proposed method, we evaluated our models in an unpublished dataset collected by our team. In this dataset, 42 participants (21 males and 21 females) were recruited for a conditionally automated driving study. Each participant was required to perform or not perform a verbal 2-back cognitive task while driving with SAE Level-3 driving automation. The ECG data was collected using a 3-lead ECG sensor at 100 Hz frequency, leading to 252 minutes of drive for each task level and in total 1,512,000 samples for each task level (6 minutes * 60 seconds * 100 Hz * 42 participants). This dataset will be made publicly available when fully prepared (*Dataset 2*).

Signal Preprocessing

As has been stated in the introduction, although the datasets contain multiple physiological measures, we selected only the ECG measure. As shown in **Figure 1**, four signal pre-processing procedures were conducted for the ECG signal, i.e., down sampling, denoising, R peak detection, and segmentation.

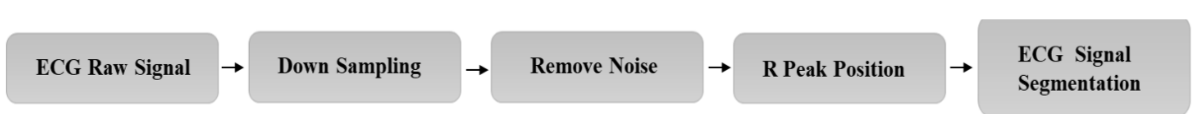


FIGURE 1 The process of ECG signal

We first down-sampled the original 1000 Hz sampling frequency to 360 Hz to reduce the computational loads of the algorithm and make real-time cognitive load detection possible. Then, given that the ECG signals are easily affected by different types of noise, including baseline drift, power frequency interference, and EMG interference, denoising is an important pre-processing step to reduce noise and enhance ECG morphology (*37*). Thus, we denoised the signal using Butterworth filer, followed by a powerline filtering (*38*, *39*). Specifically, a 0.5 Hz high-pass Butterworth filter was utilized to remove low-frequency noise and baseline drift. Then, powerline filtering was also applied to remove noise caused by electrical equipment or nearby power lines. **Figure 2** compares an example segment of the ECG signal before and after the denoise procedure. It can be observed that the filtered ECG signal exhibits a smoother waveform and has a lower baseline drift.

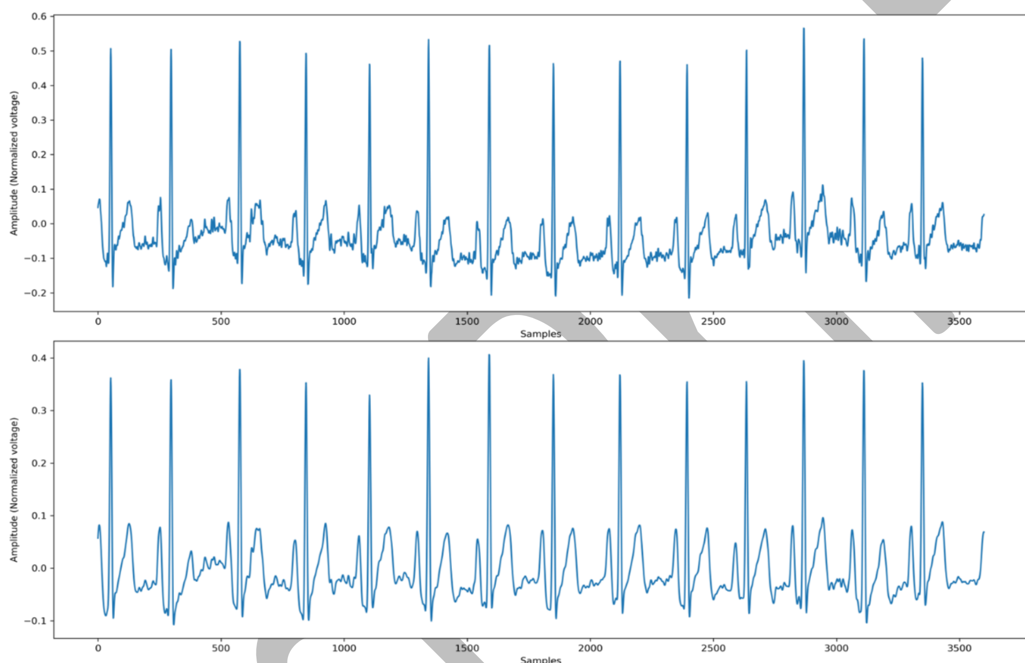


FIGURE 2 ECG raw signal (top) and denoised signal (bottom)

Next, R peaks were detected based on the local maxima in the QRS complexes (*40*), which are based on the steepness of the absolute gradient of the ECG signal. The R peak is defined as the highest amplitude point in the QRS wave morphology. The detection of the R peak was done with the Neurokit Python toolkit (*41*).

After detecting the R peak, the filtered ECG signal was segmented into 60-second time windows. We chose this window size because although we did not extract HRV features in our SE-ECG Net, other baseline models relying on HRV features require at least 60 seconds of the window for reliable HRV measurement (*42*, *43*). Specifically, to capture the entire cardiac cycle (including a series of P-QRS-T waves), each time window includes 30 seconds of the data in front of a specific R peak and 30 seconds of the data after that R peak, leading to a total of data 21600 points (i.e., 60 seconds × 360 Hz) in a window. The sliding window was set to 10 R-peaks, which means that the position of the next window is 10 R peaks away from the last window. **Figure 3** visualizes our signal segmentation process.

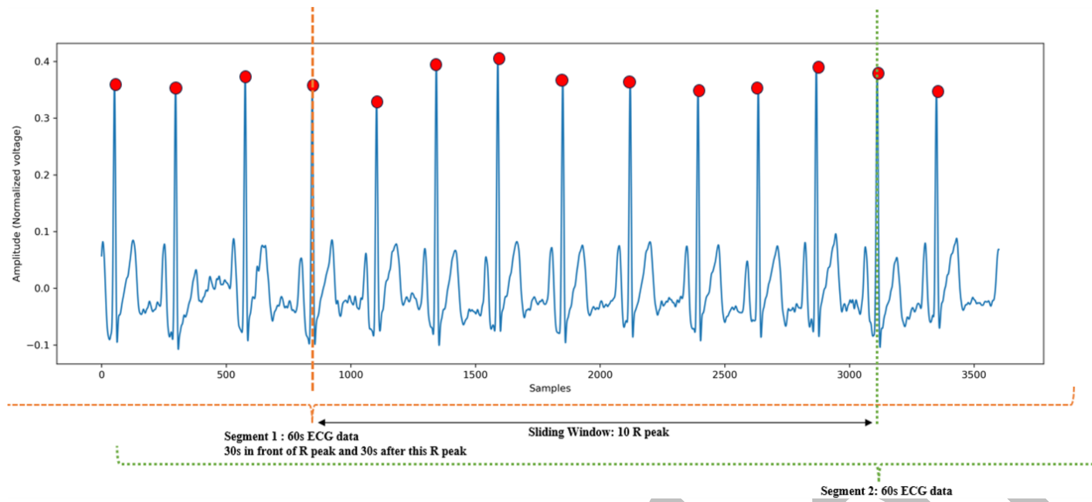


FIGURE 3 R-peak detection and signal segmentation process

Generation of ECG Spectrogram

The 60-second ECG segments from signal pre-processing (See Figure 2) were transformed into a spectrogram, which is a visual representation of the spectrum of frequencies of a signal in a time window. Specifically, given the 60s ECG segment, the data was split into short frames (each frame was 350 ms) with Hamming windows. Then, we calculate the power spectrum from each frame by using a short-time Fourier Transform (STFT) to produce a spectrogram. The horizontal axis of the spectrogram represents time, and the vertical axis represents frequency. Thus, the spectrogram contains information on how the frequency content of the ECG signal changes over time. **Figure 4** presents the time-domain (left) and spectrograms (right) of the filtered ECG signals under high cognitive load and low cognitive load, respectively.

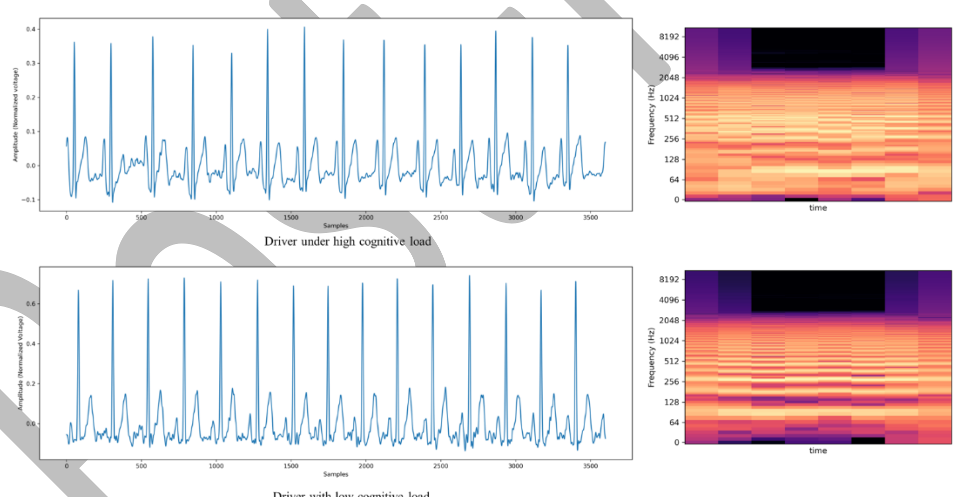


FIGURE 4 ECG signal and spectrogram of drivers under high and low cognitive load

Extraction of handcrafted ECG Features

For handcrafted HRV features, a long enough time window is needed for feature extraction. Given that a minimum of 1-minute time window was recommended for reliable HRV feature extraction (*42*, *43*), in our study, 70 ECG features were extracted from time-domain, frequency-domain and non-linear domain with using a 60-second time window of the signal (See **Table 2**).

TABLE 2 Hand-crafted HRV features extracted from ECG signal

Domain	Indicators
Time Domain	MeanNN, SDNN, RMSSD, SDSD, CVNN, CVSD, MedianNN, MadNN, MCVNN, IQRNN, SDRMSSD, Prc20NN, Prc80NN, pNN50, pNN20, MinNN, MaxNN, HTI, TINN
Frequency Domain	LF, HF, VHF, TP, LFHF, LFn, HFn, LnHF
Non-linear Domain	SD1, SD2 , SD1SD2, S, CSI, CVI, CSI_Modified, PIP, IALS, PSS, PAS, GI, SI, AI, PI, C1d, C1a, SD1d, SD1a, C2d, C2a, SD2d, SD2a, Cd, Ca, SDNNd, SDNNa, DFA_alpha1, MFDFA_alpha1_width, Peak, Mean, Max, Delta, Asymmetry, Fluctuation, Increment, ApEn, ShanEn, FuzzyEn, CD, HFD, KFD, LZC

Note: All hand-crafted features are extracted with Neurokit Python toolkit (*41*). Please check the official website for the explanations of each indicator.

CANDIDATE MODELS

SE-ECG Net

A schematic demonstration of the SE-ECG net we proposed for the classification of driver cognitive load is presented in **Figure 5**. The developed SE-ECG Net mainly contains two modules, i.e., the Convolution Module and the SE-Inception Module. Then, a fully connected layer (FC layer) and a SoftMax layer were used to generate the output (i.e., estimation of the driver's cognitive load level)(*44*).

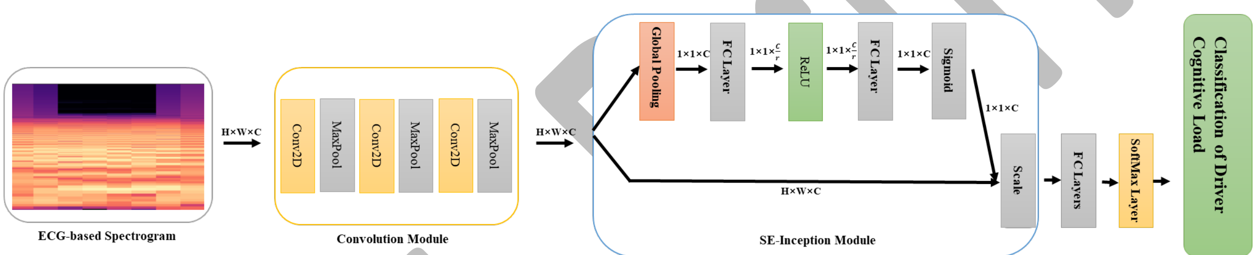


FIGURE 5 Framework of SE-ECG Net

Specifically, three convolutional layers and three max-pooling layers of the convolution module were utilized to extract local features from the spectrogram. Given that not all frame-level features from the ECG spectrogram make the same contribution to the classification of cognitive load, a channel attention mechanism called SE-Inspection Module (*35*) was added to the architecture, which showed great advantages in image recognition tasks by learning the importance weights of each channel, enhancing, or reducing their contribution to the classification. SE-Interception can assign more weights to important features automatically. Thus, the SE-ECG Net can automatically learn high-level feature representations from the spectrogram for cognitive load classification.

Figure 6 shows the structural details of a SE-Inception module. The squares of different colors in the figure demonstrate that SE-Inception can automatically learn the weight information of different channels of the input spectrogram. Given the input ECG spectrogram (H ×W×C), Convolutional Neural Network (CNN) modules mapped the ECG spectrogram into $H_i \times W_{\ i} \times C_i$, which then passed through a squeeze operation. In this operation, for each channel in the feature map, the average value of all pixel values in each channel was calculated, thereby compressing the information of each channel into a scalar value, so the output of the squeeze operation was $1 \times 1 \times C_i$. The squeeze operation was followed by an excitation operation, which involved two fully connected layers to learn the weight of each channel based on its importance. The first fully connected layer reduced the number of channels to a smaller number $1 \times 1 \times \frac{C_i}{r}$, while the second fully connected layer mapped the reduced channels back to the original number of channels $1 \times 1 \times C_i$. The excitation output was then multiplied with the original input feature map $H_i \times W_{\ i} \times C_i$ to produce the final output of the SE-Inception module. The 2D output was then flattened into a 1D vector and was fed into a fully connected layer, followed by a SoftMax layer. The SoftMax layer is a type of activation function to convert the output of the fully connected layer into a probability distribution over predicted output classes.

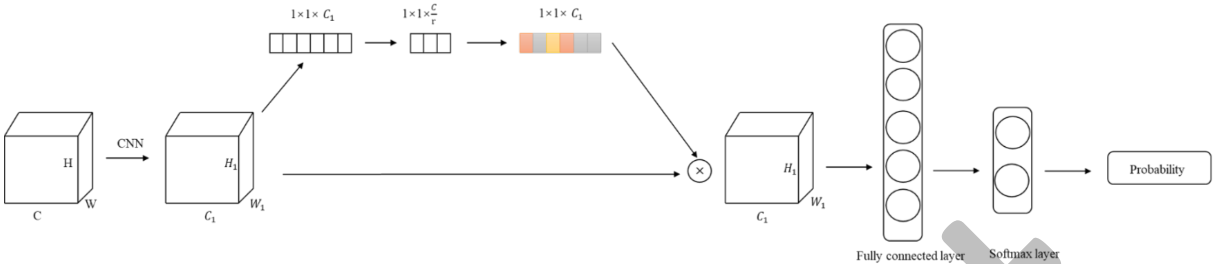


FIGURE 6 Structural details of SE-Inception module

Other Deep Learning Models

To better evaluate the potential of estimating drivers' cognitive load with ECG-based spectrograms, we selected two classic network structures as baselines, including a VGG net (Figure 7a), and a ResNet (Figure 7b). The VGG net was selected as it can capture complex spatial structures, while the ResNet was selected as it can mitigate the vanishing gradient problem through residual connections (*16*). Further, A SE-ECG net without SENet was compared to our proposed SE-ECG Net, to show the contribution of SENet to model performance. It should be noted that when a SENet is removed, the SE-ECG Net is reduced to a convolutional neural network (CNN).

To evaluate the performance of spectrograms, all models were trained and tested twice, once with ECG spectrograms as inputs (i.e., spectrogram-based models) and once with hand-crafted features as inputs (i.e., feature-based models). Since the shape of the spectrogram is two-dimensional and the handcrafted features are one-dimensional, the two-dimensional convolution module (Conv2D) and one-dimensional convolution module (Conv1D) were used, when the models were trained on different inputs, respectively.

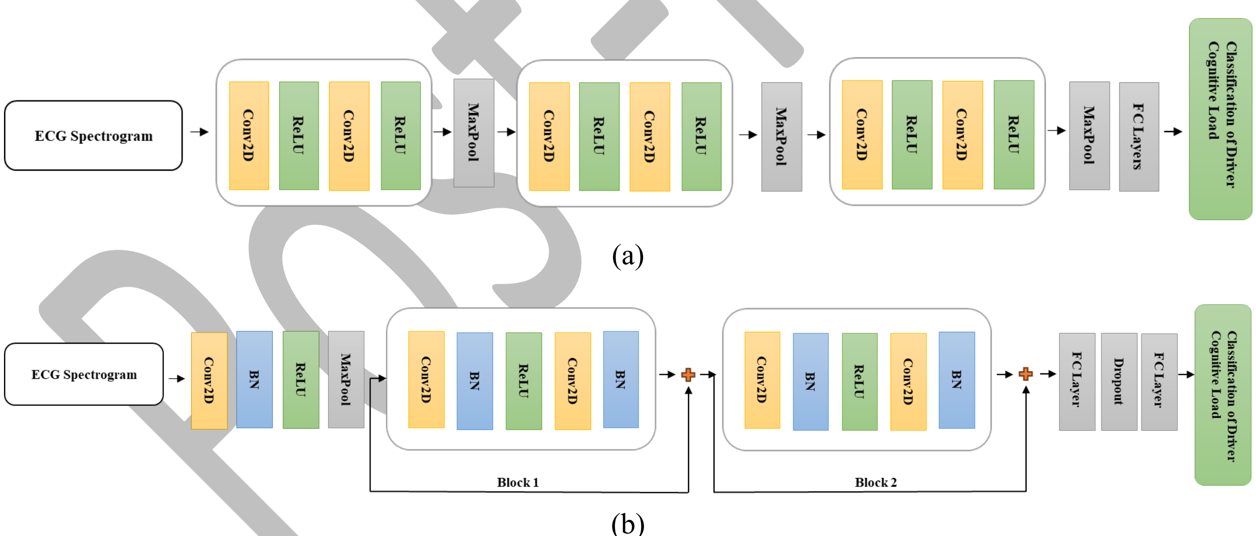


FIGURE 7 Structure of a) VGG Net b) and ResNet. VGG Net adopts a series of stacked convolutional and pooling layers to effectively extract features from images. ResNet network alleviates the vanishing gradient problem using a residual structure (block1 and block2).

EXPERIMENT SETUP

All models were trained and tested using both within-subject and across-subjects data partitions. Considering that K-fold cross-validation has the advantages of avoiding bias, over-fitting, and ensuring the reliability and accuracy of experimental results, a 10-fold cross-validation was employed. Specifically, for within-subject data partition, the ECG spectrograms computed from all subjects are combined. Then, the dataset was divided into 10 continuous equal parts, using 9 parts for training and 1 part for testing. This procedure was repeated 10 times with each part used once for testing. For the across-subjects data partition, the data from 8 or 9 participants were used for testing, and the rest 79 or 78 participants were used as the training dataset. Again, the data from each participant was used for testing once in the 10-fold cross-validation process.

All deep learning models were implemented with the PyTorch toolkit, and the parameters of the model were optimized by minimizing the cross-entropy loss, with a minibatch of 32 samples. The initial learning rate was set to 0.001 and became 0.0001 after 5 epochs to prevent overfitting. The optimizer was Adam (*45*). All model training and testing processes were conducted on NVIDIA GeForce RTX 2080 Ti.

RESULTS

To illustrate the feasibility of using a spectrogram for driver's cognitive estimation, we first compared our spectrogram-based models to feature-based models. Then, we compared our proposed SE-ECG Net with other deep-learning baseline models on spectrograms. Lastly, to verify the generalization of our proposed method, the SENet was re-trained and evaluated on a second dataset (i.e., Dataset 2).

As shown in Table 3 and Figure 8, the comparison between spectrogram-based models to feature-based models shows that the process of generating hand-crafted features may result in information loss. Specifically, for within-subject experiments, the spectrogram-based methods all achieved 96.76% and 91.97% on two datasets, respectively. For across-subjects experiments, although the accuracies of all models have all dropped, the spectrogram-based models still exhibited higher performance compared to the models based on hand-crafted ECG features. We assumed that the spectrogram-based models were able to keep more individual-independent information compared to models based on hand-crafted features. These experiments show the advantages of spectrogram for drivers' cognitive estimation.

Table 3 and Figure 8 further demonstrate the superior performance of SE-ECG Net when compared to other deep-learning neural networks and the CNNs (i.e., SE-ECG Net without SE-Inception Module). Specifically, the SE-ECG Net achieved high accuracies on two datasets with both within-subject and across-subjects data partitions, exceeding the performance of CNNs, VGG Net, and ResNet. This indicates that learning the importance weights of the features by the SE-Inception module can increase the model performance.

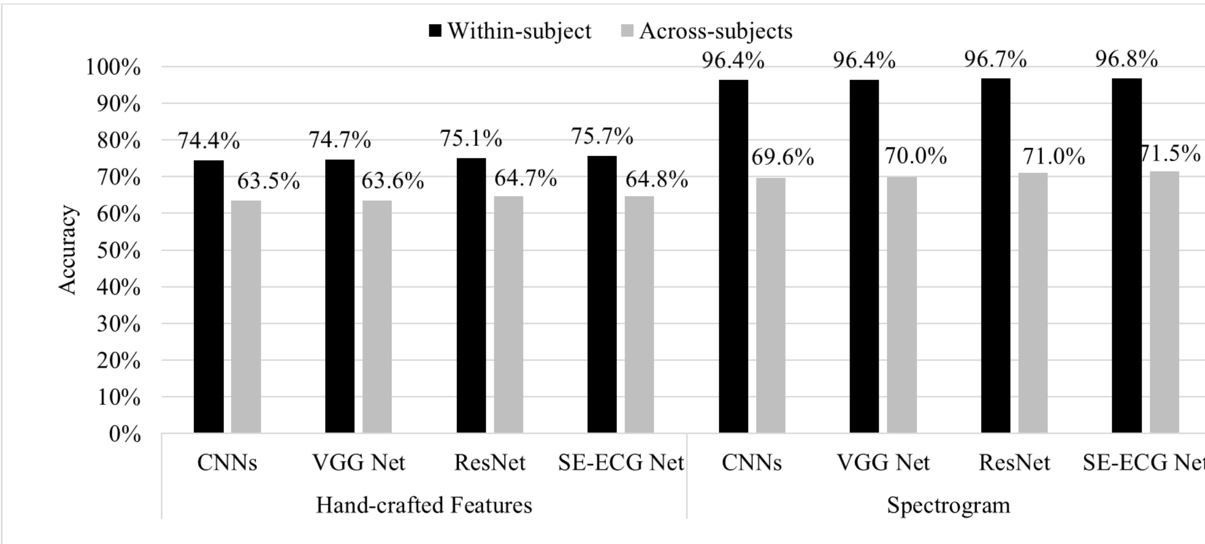
Finally, it should be noted that compared with within-subject experiments, the accuracies of all classifiers have all dropped significantly with an across-subjects data partition, indicating that the deep-learning neural networks in our study may still not be able to fully learn individual-independent features from ECG-based spectrograms.

TABLE 3 Model Results

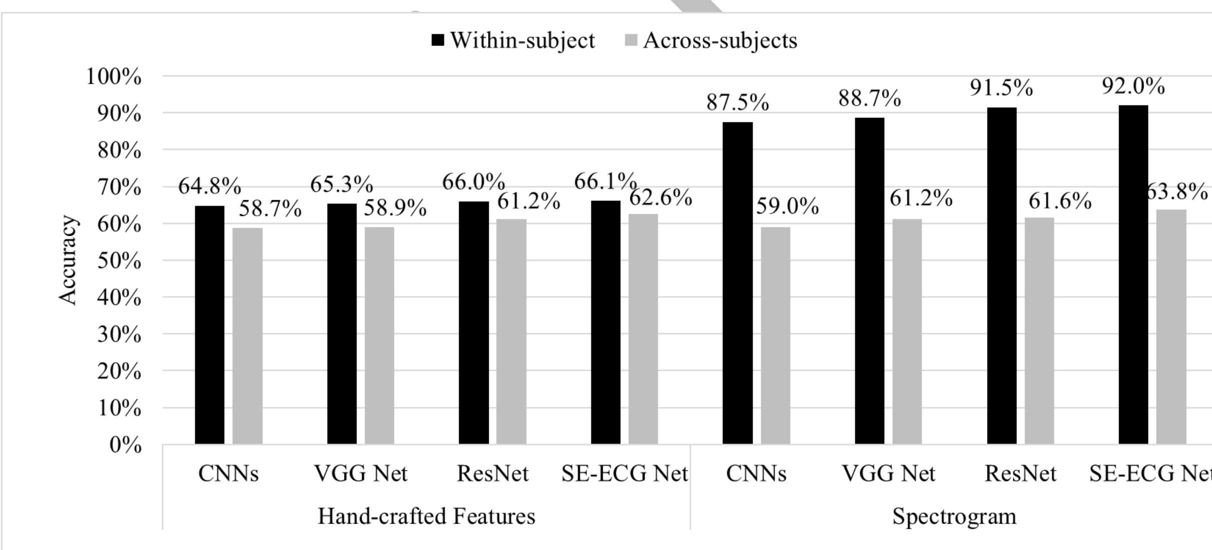
Data partitions	Features	Model	Accuracy (%) Dataset 1	Accuracy (%) Dataset 2
Within-subject	Hand-crafted Features	CNNs	74.44%	64.80%
	Hand-crafted Features	VGG Net	74.70%	65.26%
	Hand-crafted Features	ResNet	75.10%	65.95%
	Hand-crafted Features	SE-ECG Net	75.70%	66.09%
	Spectrogram	CNNs	96.35%	87.51%
	Spectrogram	VGG Net	96.40%	88.65%
	Spectrogram	ResNet	96.70%	91.50%
	Spectrogram	SE-ECG Net	**96.76%**	**91.97%**
Across-subjects	Hand-crafted Features	CNNs	63.50%	58.70%
	Hand-crafted Features	VGG Net	63.57%	58.90%
	Hand-crafted Features	ResNet	64.70%	61.15%

	Hand-crafted Features	SE-ECG Net	64.76%	62.55%
	Spectrogram	CNNs	69.60%	58.95%
	Spectrogram	VGG Net	69.95%	61.17%
	Spectrogram	ResNet	71.00%	61.60%
	Spectrogram	SE-ECG Net	**71.50%**	**63.82%**

Note: the bolded texts highlight the models with the highest accuracies.



(a)



(b)

FIGURE 8 Accuracies of models with different data partitions on a) Dataset 1; b) Dataset 2

To further compare the performance of the spectrogram-based methods, the confusion matrices are provided in Figure 9. It was found that in general, the models performed better in identifying the low workload versus the high workload (3 out of 32 cases), potentially because of the ceiling effect of the ECG signals during high cognitive load (*46*). Such a trend is more obvious with the within-subject data partition compared to the across-subjects data partition; and more obvious in spectrogram-based models compared to the feature-based models. Thus, future research may consider fine-tuning the model structures to strengthen the model performance based on the task needs.

(a) Within-subject partitions method

Input	Target	SE-ECG Net Predicted Low Workload	SE-ECG Net Predicted High Workload	ResNet Predicted Low Workload	ResNet Predicted High Workload	VGG Net Predicted Low Workload	VGG Net Predicted High Workload	CNNs Predicted Low Workload	CNNs Predicted High Workload
Spectrogram	Low Workload	97.9%	2.1%	96.8%	2.2%	97.1%	2.3%	96.2%	3.8%
Spectrogram	High Workload	4.2%	95.8%	3.5%	96.5%	3.9%	96.1%	4.6%	95.4%
HRV Features	Low Workload	72.2%	27.8%	69.4%	30.6%	71.1%	28.9%	69.9%	30.1%
HRV Features	High Workload	37.5%	65.3%	33.5%	66.5%	32.5%	67.5%	34.2%	65.8%

(a) Across-subjects partitions method

Input	Target	SE-ECG Net Predicted Low Workload	SE-ECG Net Predicted High Workload	ResNet Predicted Low Workload	ResNet Predicted High Workload	VGG Net Predicted Low Workload	VGG Net Predicted High Workload	CNNs Predicted Low Workload	CNNs Predicted High Workload
Spectrogram	Low Workload	78.9%	21.1%	77.9%	22.1%	77.7%	22.3%	76.9%	23.1%
Spectrogram	High Workload	35.6%	64.4%	35.5%	64.5%	36.1%	63.9%	36.6%	63.4%
HRV Features	Low Workload	64.4%	35.6%	63.7%	36.3%	62.8%	37.2%	62.2%	37.8%
HRV Features	High Workload	33.9%	66.1%	35.3%	64.7%	38.8%	61.2%	39.2%	60.8%

(a)

(b) Within-subject partitions method

Input	Target	SE-ECG Net Predicted Low Workload	SE-ECG Net Predicted High Workload	ResNet Predicted Low Workload	ResNet Predicted High Workload	VGG Net Predicted Low Workload	VGG Net Predicted High Workload	CNNs Predicted Low Workload	CNNs Predicted High Workload
Spectrogram	Low Workload	87.1%	12.9%	90.6%	9.4%	89.1%	10.9%	88.3%	11.7%
Spectrogram	High Workload	9.2%	90.8%	13.1%	86.9%	12.3%	87.7%	14.1%	85.9%
HRV Features	Low Workload	65.7%	34.3%	66.1%	33.3%	65.8%	34.2%	65.5%	34.5%
HRV Features	High Workload	34.7%	65.3%	34.2%	65.8%	34.6%	65.4%	34.9%	65.1%

(b) Across-subjects partitions method

Input	Target	SE-ECG Net Predicted Low Workload	SE-ECG Net Predicted High Workload	ResNet Predicted Low Workload	ResNet Predicted High Workload	VGG Net Predicted Low Workload	VGG Net Predicted High Workload	CNNs Predicted Low Workload	CNNs Predicted High Workload
Spectrogram	Low Workload	67.4%	32.6%	66.9%	33.1%	66.5%	33.5%	76.9%	23.1%
Spectrogram	High Workload	33.9%	66.1%	34.5%	65.5%	33.7%	66.3%	36.6%	63.4%
HRV Features	Low Workload	65.4%	34.6%	65.7%	34.3%	64.2%	35.8%	63.9%	36.1%
HRV Features	High Workload	36.2%	63.8%	35.9%	64.1%	35.9%	64.1%	37.2%	62.8%

(b)

FIGURE 9 Confusion matrices of models on a) Dataset 1; b) Dataset 2.

DISCUSSION

In this study, we proposed a novel approach to assess driver cognitive load using ECG spectrograms and we designed an SE-ECG Net architecture to identify high-level features in the spectrogram automatically and estimate the cognitive load in drivers. Experimental results show that the ECG spectrogram performed better than machine learning models that relied on hand-crafted features extracted from ECG signals when both within-subject and across-subjects data partitions were used. At the same time, the comparison experiment shows that the SENet Module contributes to the estimation accuracy, increasing the estimation accuracy when both within-subject and across-subjects data partitions were used.

Although hand-crafted features can be designed to capture specific characteristics of the data (e.g., the RR interval will become narrower when there is a high cognitive load), extracting and selecting hand-crafted features can lead to a loss of information. In contrast, the deep learning model can perform a non-linear transformation on the input spectrogram. When the input reaches the final layer of the network, the input can be transformed into a high-level feature representation that is well-suited for the classification of cognitive load, which explains the good performance of the deep learning models, including CNNs, VGG Net, ResNet, and SE-ECG Net. On the other hand, the superior performance of SE-ECG net as compared to other deep learning baseline models can be explained by the capability of the SE-Inception Module to pay more attention to cognition-related features in the spectrogram. However, it should be noted that, from within-subject data partition to across-subjects data partition, the model performance still dropped significantly, although the performance of the spectrogram-based deep learning models was still better than the models that relied on hand-crafted features. This might be because the deep learning models used in our study were still not able to capture all individual-independent features of high cognitive load.

Overall, our proposed method is practically valuable for real-world applications as it can utilize ECG data only to detect high cognitive load among drivers, which can be measured using steering-wheel-integrated sensors. From the practical perspective of view, collecting eye-tracking data that is accurate enough for driver state detection can be expensive and unreliable (due to complex lighting conditions and eye occlusion by sunglasses or hair) while other physiological signals such as EEG and respiration are intrusive to be collected. The single physiological signal was believed to contain too little information for accurate driver state estimation. The results of our study indicate that a single signal, such as ECG may contain more information in high-level features, and satisfactory models may be built with few numbers of physiological signals if the information in the signals can be fully exploited. However, some practical issues need to be considered before the actual deployment of the ECG sensors in vehicles, such as noise reduction and sensor reliability. In both datasets we used, the ECG signals were collected in labs and thus are relatively clean compared to ECG signals collected using wearable sensors or embedded sensors in real vehicles. Therefore, future data cleaning and noise filtering techniques may need to be developed before the algorithms can be deployed in smart cabins. In addition, automakers also need to consider the reliability of the ECG sensor to ensure stable and reliable performance over an extended period. For example, if the sensors are embedded in the steering wheel, similar to what has been planned by Toyota ((*34*) ensuring that the drivers can steadily hold the steering wheel is a prerequisite of the ECG-based models. Hence, future work may also need to consider model switching or fusion when different signals are available.

As far as we know, this is an early work using a spectrogram for driver cognitive load estimation. Despite the promising results, our study still has some limitations. First, the dataset utilized in this research was limited to a fixed simulator, which might not fully replicate real-world driving conditions. Thus, the performance of the models in estimating the cognitive load imposed by different cognitive tasks, especially real in-vehicle tasks, should be further explored. Future research could also benefit from the inclusion of data collected in actual driving environments to further validate the findings. Finally, so far, the spectrogram-based method is still not able to handle heterogeneity between individuals. The performance of the proposed model dropped significantly with an across-subjects data partition. Future studies may need to consider other approaches such as data fusion (*47*), pre-training, and domain generalization to improve the model performance in estimating the cognitive load levels across subjects. For example, (*48*) proposed that two sets of ECG data can be used, i.e., the local ECG data collected from each individual, which reflects their unique characteristics and individual differences; and the global ECG data collected from a large population, which can be used for comparison and analysis to better understand the commonalities and differences of ECG signals. Therefore, global physiological data can be used for pre-training and the local physiological data can be used to customize the models to improve the accuracy of cognitive load estimation. Future research should explore this approach when a larger physiological dataset becomes available. In the computer vision domain (*49*), domain generalization has demonstrated its advantages in solving generalization capabilities, but it has not been explored in the field of cognitive load estimation.

CONCLUSIONS

In this paper, the 2-D spectrogram was first extracted from the ECG signal and the superior performance of this approach was validated on a novel SE-ECG Net and two classical deep learning benchmark models (i.e., VGG Net and ResNet). Experiments on a publicly available dataset (i.e., Dataset 1) achieve an accuracy of 96.76% in the within-subject evaluation and 71.50 % in the across-subjects evaluation. The results highlight the superiority of the spectrogram-based approach compared to the approaches based on handcrafted features for driver high cognitive load detection in SAE L3 vehicles. This success can be attributed to the implementation of deep learning with attention mechanisms and using an ECG spectrogram as input. Therefore, the findings demonstrate the feasibility of utilizing these techniques for cognitive load detection in drivers based on a few mensurable or even single physiological signals, which can facilitate the application of non-intrusive physiological sensors to improve the safety of SAE L3 vehicles, potentially alleviating the privacy concerns of the driver monitoring systems when computer-vision-based approach was adopted.

ACKNOWLEDGMENTS

This work was supported by the Guangzhou Municipal Science and Technology Project (No. 2023A03J0011), and the Guangzhou Science and Technology Program City-University Joint Funding Project (No. 2023A03J0001).

AUTHOR CONTRIBUTIONS

The authors confirm their contribution to the paper as follows: Wenxin Shi: Conceptualization, Data analysis, model training, and original draft. Zuyuan Wang: Hyper-parameter tuning and model validation process and draft. Ange Wang: Validation and revising. Dengbo He: Methodology, supervision, Validation, review editing.

REFERENCES

1. J3016_202104: Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles - SAE International. https://www.sae.org/standards/content/j3016_202104/. Accessed Nov. 24, 2023.
2. The Future of Autonomous Vehicles (AV) | McKinsey. https://www.mckinsey.com/industries/automotive-and-assembly/our-insights/autonomous-drivings-future-convenient-and-connected. Accessed Nov. 24, 2023.
3. J3018_201503: Guidelines for Safe On-Road Testing of SAE Level 3, 4, and 5 Prototype Automated Driving Systems (ADS) - SAE International. https://www.sae.org/standards/content/j3018_201503/. Accessed Nov. 24, 2023.
4. Paas, F. G. W. C., and J. J. G. Van Merriënboer. Instructional Control of Cognitive Load in the Training of Complex Cognitive Tasks. *Educational Psychology Review*, Vol. 6, No. 4, 1994, pp. 351–371. https://doi.org/10.1007/BF02213420.
5. Savage, S., D. Potter, and B. Tatler. The Effects of Cognitive Distraction on Behavioural, Oculomotor and Electrophysiological Metrics during a Driving Hazard Perception Task. *Accident Analysis & Prevention*, Vol. 138, 2020, p. 105469. https://doi.org/10.1016/j.aap.2020.105469.
6. Du, N., E. Pulver, L. Robert, A. Pradhan, X. J. Yang, N. Du, J. Kim, F. Zhou, D. Tilbury, and X. Jessie. *Evaluating Effects of Cognitive Load, Takeover Request Lead Time, and Traffic Density on Drivers' Takeover Performance in Conditionally Automated Driving*. 2020.
7. Stapel, J., freddy a mullakkal babu, and R. Happee. Automated Driving Reduces Perceived Workload, but Monitoring Causes Higher Cognitive Load than Manual Driving. *Transportation Research Part F: Traffic Psychology and Behaviour*, Vol. 60, 2019, pp. 590–605. https://doi.org/10.1016/j.trf.2018.11.006.
8. Public Opinion about Self-Driving Vehicles in China, India, Japan, the US, the UK, and Australia. , 2014.
9. Solovey, E., M. Zec, E. Garcia Perez, B. Reimer, and B. Mehler. Classifying Driver Workload Using Physiological and Driving Performance Data: Two Field Studies. *Conference on Human Factors in Computing Systems - Proceedings*, 2014. https://doi.org/10.1145/2556288.2557068.
10. Rahman, H., M. U. Ahmed, S. Barua, and S. Begum. Non-Contact-Based Driver's Cognitive Load Classification Using Physiological and Vehicular Parameters. *Biomedical Signal Processing and Control*, Vol. 55, 2020, p. 101634. https://doi.org/10.1016/j.bspc.2019.101634.
11. Rahman, H., M. U. Ahmed, S. Barua, P. Funk, and S. Begum. Vision-Based Driver's Cognitive Load Classification Considering Eye Movement Using Machine Learning and Deep Learning. *Sensors*, Vol. 21, No. 23, 2021, p. 8019. https://doi.org/10.3390/s21238019.
12. Liu, T., Y. Yang, G.-B. Huang, Y. K. Yeo, and Z. Lin. Driver Distraction Detection Using Semi-Supervised Machine Learning. *IEEE Transactions on Intelligent Transportation Systems*, Vol. 17, 2015, pp. 1–13. https://doi.org/10.1109/TITS.2015.2496157.
13. N. Li and C. Busso. Predicting Perceived Visual and Cognitive Distractions of Drivers With Multimodal Features. *IEEE Transactions on Intelligent Transportation Systems*, Vol. 16, No. 1, 2015, pp. 51–65. https://doi.org/10.1109/TITS.2014.2324414.
14. Meteier, Q., E. De Salis, M. Capallera, M. Widmer, L. Angelini, O. Abou Khaled, A. Sonderegger, and E. Mugellini. Relevant Physiological Indicators for Assessing Workload in Conditionally Automated Driving, Through Three-Class Classification and Regression. *Frontiers in Computer Science*, Vol. 3, 2022, p. 775282. https://doi.org/10.3389/fcomp.2021.775282.

15. Simonyan, K., and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. http://arxiv.org/abs/1409.1556. Accessed Mar. 12, 2024.
16. He, K., X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. http://arxiv.org/abs/1512.03385. Accessed Mar. 12, 2024.
17. Angkan, P., B. Behinaein, Z. Mahmud, A. Bhatti, D. Rodenburg, P. Hungler, and A. Etemad. Multimodal Brain-Computer Interface for In-Vehicle Driver Cognitive Load Measurement: Dataset and Baselines. *IEEE Transactions on Intelligent Transportation Systems*, 2024, pp. 1–16. https://doi.org/10.1109/TITS.2023.3345846.
18. Liang, Y., M. L. Reyes, and J. D. Lee. Real-Time Detection of Driver Cognitive Distraction Using Support Vector Machines. *IEEE Transactions on Intelligent Transportation Systems*, Vol. 8, No. 2, 2007, pp. 340–350. https://doi.org/10.1109/TITS.2007.895298.
19. Mehler, B., B. Reimer, J. Coughlin, and J. Dusek. The Impact of Incremental Increases in Cognitive Workload on Physiological Arousal and Performance in Young Adult Drivers. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2138, 2009, pp. 6–12. https://doi.org/10.3141/2138-02.
20. Brookhuis, K. A., G. de Vries, and D. de Waard. The Effects of Mobile Telephoning on Driving Performance. *Accident Analysis & Prevention*, Vol. 23, No. 4, 1991, pp. 309–316. https://doi.org/10.1016/0001-4575(91)90008-S.
21. Bush, N. R., Z. K. Caron, K. S. Blackburn, and A. Alkon. Measuring Cardiac Autonomic Nervous System (ANS) Activity in Toddlers - Resting and Developmental Challenges. *Journal of Visualized Experiments : JoVE*, No. 108, 2016, p. 53652. https://doi.org/10.3791/53652.
22. Shimomura, Y., T. Yoda, K. Sugiura, A. Horiguchi, K. Iwanaga, and T. Katsuura. Use of Frequency Domain Analysis of Skin Conductance for Evaluation of Mental Workload. *Journal of physiological anthropology*, Vol. 27, 2008, pp. 173–7. https://doi.org/10.2114/jpa2.27.173.
23. Zihisire Muke, P., Z. Telec, and B. Trawiński. Cognitive Load Measurement Using Arithmetic and Graphical Tasks and Galvanic Skin Response. Cham, 2022.
24. Meteier, Q., M. Capallera, S. Ruffieux, L. Angelini, O. Abou Khaled, E. Mugellini, M. Widmer, and A. Sonderegger. Classification of Drivers' Workload Using Physiological Signals in Conditional Automation. *Frontiers in Psychology*, Vol. 12, 2021, p. 596038. https://doi.org/10.3389/fpsyg.2021.596038.
25. Molnar, C. *10.1 Learned Features | Interpretable Machine Learning*.
26. What Is Deep Learning and How Does It Work? | Definition from TechTarget. https://www.techtarget.com/searchenterpriseai/definition/deep-learning-deep-neural-network. Accessed Nov. 24, 2023.
27. Fujiwara, K., E. Abe, K. Kamata, C. Nakayama, Y. Suzuki, T. Yamakawa, T. Hiraoka, M. Kano, Y. Sumi, F. Masuda, M. Matsuo, and H. Kadotani. Heart Rate Variability-Based Driver Drowsiness Detection and Its Validation With EEG. *IEEE Transactions on Biomedical Engineering*, Vol. 66, No. 6, 2019, pp. 1769–1778. https://doi.org/10.1109/TBME.2018.2879346.
28. Song, F., Z. Guo, and D. Mei. Feature Selection Using Principal Component Analysis. In *Engineering Design and Manufacturing Informatization 2010 International Conference on System Science*, No. 1, 2010, pp. 27–30.
29. Darzi, A., S. M. Gaweesh, M. M. Ahmed, and D. Novak. Identifying the Causes of Drivers' Hazardous States Using Driver Characteristics, Vehicle Kinematics, and Physiological Measurements. *Frontiers in Neuroscience*, Vol. 12, 2018, p. 568. https://doi.org/10.3389/fnins.2018.00568.
30. Le, A. S., H. Aoki, F. Murase, and K. Ishida. A Novel Method for Classifying Driver Mental Workload Under Naturalistic Conditions With Information From Near-Infrared Spectroscopy. *Frontiers in Human Neuroscience*, Vol. 12, 2018, p. 431. https://doi.org/10.3389/fnhum.2018.00431.
31. Barua, S., M. U. Ahmed, and S. Begum. Towards Intelligent Data Analytics: A Case Study in Driver Cognitive Load Classification. *Brain Sciences*, Vol. 10, No. 8, 2020, p. 526. https://doi.org/10.3390/brainsci10080526.

32. Prabhakar, G., A. Mukhopadhyay, L. Murthy, M. Modiksha, D. Sachin, and P. Biswas. Cognitive Load Estimation Using Ocular Parameters in Automotive. *Transportation Engineering*, Vol. 2, 2020, p. 100008. https://doi.org/10.1016/j.treng.2020.100008.
33. He, D., Z. Wang, E. B. Khalil, B. Donmez, G. Qiao, and S. Kumar. Classification of Driver Cognitive Load: Exploring the Benefits of Fusing Eye-Tracking and Physiological Measures. *Transportation Research Record*, Vol. 2676, No. 10, 2022, pp. 670–681. https://doi.org/10.1177/03611981221090937.
34. Toyota Demos ECG Sensing Steering Wheel | MobiHealthNews. https://www.mobihealthnews.com/12128/toyota-demos-ecg-sensing-steering-wheel. Accessed Nov. 24, 2023.
35. Hu, J., L. Shen, S. Albanie, G. Sun, and E. Wu. Squeeze-and-Excitation Networks. http://arxiv.org/abs/1709.01507. Accessed Nov. 24, 2023.
36. Meteier, Q., M. Capallera, E. de Salis, L. Angelini, S. Carrino, M. Widmer, O. Abou Khaled, E. Mugellini, and A. Sonderegger. A Dataset on the Physiological State and Behavior of Drivers in Conditionally Automated Driving. *Data in Brief*, Vol. 47, 2023, p. 109027. https://doi.org/10.1016/j.dib.2023.109027.
37. Shi, F. A Review of Noise Removal Techniques in ECG Signals. Presented at the 2022 IEEE Conference on Telecommunications, Optics and Computer Science (TOCS), 2022.
38. Elgendi, M., M. Jonkman, and F. De Boer. *Frequency Bands Effects on QRS Detection.* 2010.
39. Hamilton, P. Open Source ECG Analysis. Presented at the Computers in Cardiology, 2002.
40. Kohler, B.-U., C. Hennig, and R. Orglmeister. The Principles of Software QRS Detection. *IEEE Engineering in Medicine and Biology Magazine*, Vol. 21, No. 1, 2002, pp. 42–57. https://doi.org/10.1109/51.993193.
41. Makowski, D., T. Pham, Z. J. Lau, J. C. Brammer, F. Lespinasse, H. Pham, C. Schölzel, and S. H. A. Chen. NeuroKit2: A Python Toolbox for Neurophysiological Signal Processing. *Behavior Research Methods*, Vol. 53, No. 4, 2021, pp. 1689–1696. https://doi.org/10.3758/s13428-020-01516-y.
42. Patel, M., S. K. L. Lal, D. Kavanagh, and P. Rossiter. Applying Neural Network Analysis on Heart Rate Variability Data to Assess Driver Fatigue. *Expert Systems with Applications*, Vol. 38, No. 6, 2011, pp. 7235–7242. https://doi.org/10.1016/j.eswa.2010.12.028.
43. Begum, S., S. Barua, R. Filla, and M. U. Ahmed. Classification of Physiological Signals for Wheel Loader Operators Using Multi-Scale Entropy Analysis and Case-Based Reasoning. *Expert Systems with Applications*, Vol. 41, No. 2, 2014, pp. 295–305. https://doi.org/10.1016/j.eswa.2013.05.068.
44. Softmax Function - Wikipedia. https://en.wikipedia.org/wiki/Softmax_function. Accessed Nov. 24, 2023.
45. Kingma, D. P., and J. Ba. Adam: A Method for Stochastic Optimization. http://arxiv.org/abs/1412.6980. Accessed Nov. 24, 2023.
46. Chikhi, S., N. Matton, and S. Blanchet. EEG Power Spectral Measures of Cognitive Workload: A Meta-Analysis. *Psychophysiology*, Vol. 59, No. 6, 2022, p. e14009. https://doi.org/10.1111/psyp.14009.
47. Mou, L., C. Zhou, P. Zhao, B. Nakisa, M. N. Rastgoo, R. Jain, and W. Gao. Driver Stress Detection via Multimodal Fusion Using Attention-Based CNN-LSTM. *Expert Systems with Applications*, Vol. 173, 2021, p. 114693. https://doi.org/10.1016/j.eswa.2021.114693.
48. Saadatnejad, S., M. Oveisi, and M. Hashemi. LSTM-Based ECG Classification for Continuous Monitoring on Personal Wearable Devices. *IEEE Journal of Biomedical and Health Informatics*, Vol. 24, No. 2, 2020, pp. 515–523. https://doi.org/10.1109/JBHI.2019.2911367.
49. Zhou, K., Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy. Domain Generalization: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022, pp. 1–20. https://doi.org/10.1109/TPAMI.2022.3195549.