Draft Chapter of an Automotive UI Case Study *to appear in Introduction to Digital Signal Processing and Machine Learning for Interactive Systems Developers*

Driver Cognitive Load Classification based on Physiological Data

Dengbo He, Martina Risteska, Birsen Donmez*, Kaiyang Chen

University of Toronto

*Corresponding author.

1.1 Motivation for Driver State Estimation

Technological advancements in the recent past have led to increasingly intelligent vehicle technologies, which aim to track and deduce the driver's state [1-3] as well as the state of the driving environment [4], predict driver intent [5, 6], and warn the driver when certain critical events like vehicle lane departures occur [1, 6]. An example of such technology in production is the Mercedes-Benz Fatigue Detection System [7], which uses sensors to detect drowsiness-related changes in driving style. Although many of the above-mentioned capabilities are still in development, such systems have the potential to enhance safety by detecting driver impairment.

One type of driver impairment that is of major concern for safety is driver distraction [8], which is defined as "the diversion of attention away from activities critical for safe driving toward a competing activity" [9]. In addition to increasingly intelligent vehicle technologies, technological advancements in the recent past have also given rise to infotainment devices that are either built in (e.g., GPS navigation) or carried into (e.g., cell phones) the vehicle. These infotainment devices are potential sources of driver distraction as they can compete with the driving task for drivers' mental resources. In general, driving is a cognitively demanding, mainly visual perception-manual control task that heavily relies on attention. As in-vehicle infotainment systems are becoming more prevalent, the issue of driver distraction becomes of greater concern and the need to estimate driver's state and intervene when necessary also increases.

Although driver distraction caused by visual-manual tasks that are secondary to driving (e.g., texting) is especially detrimental to safety as these tasks divert visual attention away from the roadway [8], tasks that are auditory-verbal in nature are also of concern as they are becoming more common with the rise of voice-command interfaces within the vehicle. In fact, several simulator and on-road studies indicate that high cognitive load caused by auditory-verbal tasks secondary to driving (i.e., secondary or non-driving tasks) impairs drivers' visual scanning behaviors and driving performance [10, 11]. Therefore, it is necessary to consider the effects of auditory-verbal secondary

tasks on drivers' cognitive load, a multidimensional construct representing the load that performing a particular task imposes on the drivers' cognitive system [12]. What the driver experiences, i.e., cognitive load, depends on taskload (based solely on task characteristics), the individual driver characteristics, and the cognitive capacity that the driver allocates to different tasks [12]. Errors increase and performance becomes less efficient when humans are cognitively overloaded [13]. Potential sources of overload or distraction, may be both outside (e.g., high traffic or bad weather) or inside the vehicle (e.g., information from cellphones [13], infotainment devices [14], or the dashboard). Therefore, intelligent vehicle systems should not only strive to detect when visual attention is diverted away from the road, but also strive to detect high levels of cognitive load experienced by drivers that can lead to decrements in visual scanning and driving performance [e.g., 15]. With this information, human-computer interfaces within the vehicle can be adapted to enhance safety.

1.2 Driver Cognitive Load Detection Methods

As stated earlier, driving can be a cognitively demanding task, such as when navigating through high traffic density, poor weather, or unfamiliar driving environments. Interactions with devices that are built into the vehicle (e.g., navigational displays) and are carried-in (e.g., smartphones) can further claim cognitive resources and add to the level of demand experienced by drivers, potentially leading to attention and response impairments. Although drivers can moderate their cognitive load to some extent, such as by reducing their speed, avoiding lane changes, and increasing headway [16, 17], these actions may not be sufficient to fully compensate for the external demands placed on the drivers. By detecting driver cognitive load in real-time, in-vehicle technologies can adapt to the level of load experienced by the driver and, for example, filter information content, delay notifications, and block access to certain actions. Piechulla et al. [18] developed a prototype system to help drivers allocate their attention in difficult driving situations. The authors found a reduction in cognitive load in the adaptive telephone condition where incoming phone calls were automatically filtered when estimated cognitive load exceeded a predetermined threshold. Real-time assessment of cognitive load can also be used to intelligently transfer control to the driver in automated vehicles.

1.2.1 Cognitive Load Measures

Various measures can be used to estimate cognitive load. These measures can be categorized into four groups [19]: a) physiological measures, such as Electroencephalogram (EEG), Electrocardiography (ECG), and galvanic skin response (GSR); b) eye tracking measures, such as gaze position and blink rate; c) performance-based measures, such as vehicle speed; and d) subjective measures, such as NASA Task Load Index [20]. Table 1 provides a summary of cognitive load measures and their response to increased cognitive secondary task difficulty (i.e., cognitive taskload) in driving studies.

Although subjective measures are informative for cognitive load assessment in research settings, they are not appropriate for real-time assessment in the vehicle as their collection would interfere with the performance of the driving task [21]. Physiological,

eye-tracking, and performance-based measures are suitable for real-time cognitive load detection. In particular, previous literature has identified physiological measures, such as EEG, heart rate, and GSR, as valuable predictors for drivers' cognitive load estimation in real-time [22-25]. For example, EEG, a measurement of the electrical activity of the brain [26] provides very high temporal resolution in detecting changes in cognitive activity [27] as it is sampled at high frequencies, e.g., 500 Hz in Wang et al. [22]. This high resolution property makes EEG particularly suitable for driving applications, where the conditions and drivers' tasks can change very rapidly. GSR, a measure of skin conductance, has also been proven to be sensitive to changes in cognitive load, but its response is not as fast as EEG's response to shifts in cognitive load; e.g., 3 seconds latency after stimulus, as suggested by Wagner and Wagner [28]. In general, different measures have varying advantages and disadvantages. Most physiological measures require the drivers to wear sensors on their body at this point in time, which can be intrusive to the drivers; accuracy of eye-tracking is highly affected by lighting conditions [24]; and driving performance measures are affected not only by cognitive load but by several other factors, such as road conditions. Therefore, in practice, multiple measures may need to be combined to boost the performance of cognitive load detection algorithms.

Table 1 Summary of Example Cognitive Load Measures.

Category	Trend with Increased Cognitive Taskload
Physiological	Increase in EEG power of alpha band [26, 29], heart rate [30-32], and GSR [30, 31]
Eye tracking	Decrease in number of glances to periphery, mirror, instrument panel [11], decrease in standard deviation of horizontal and vertical gaze position [31-33], and increase in blink rate [33]
Performance-based	Decrease in average speed and increase in steering wheel reversal rate [31]
Subjective	Increase in NASA Task Load Index [11]

1.2.2 Cognitive Load Detection through Physiological Measures

As stated above, although different types of measures are reactive to changes in cognitive load, physiological measures have been identified to be valuable predictors for driver cognitive load estimation in real-time. Previous literature has used different physiological predictors alone and in combination with other predictors to classify drivers' cognitive load. These studies also explored a variety of machine learning methods including k-nearest neighbors (kNN), artificial neural network (ANN), support vector machine (SVM), Naïve Bayes (NB), decision tree (DT), linear discriminant analysis (LDA), and logistic regression (see Chapter 4: Machine Learning Basics for an overview of these methods).

Solovey et al. [25] used heart rate (HR), galvanic skin response (GSR), and driving performance data from an instrumented vehicle study to classify whether drivers were performing a cognitive task or not. In this study, external cognitive load was imposed on the drivers through an auditory-verbal recall task, i.e., the n-back task, commonly used in working memory literature [34]. In particular, the task difficulty level used in classification was 2-back: participants were required to listen to several pre-recorded series of single-digit numbers and respond verbally with the digit that was presented two positions previously or 2-back from the current number. The authors classified 2-back periods from periods with no secondary task using a variety of machine learning methods and a maximum classification accuracy of 89% was achieved with NB. In Wang et al. [22], a 32-channel EEG device was used for data collection in a driving simulator, and an SVM model with RBF kernel was built to identify whether the driver was performing a yes-no cognitive task (e.g., 24+15=37?). The authors achieved an accuracy of around 85% for classifying the cognitive task and no task periods using the power of four EEG bands: delta (1–3 Hz), theta (4–7 Hz), alpha (8–13 Hz), and low beta (14–20 Hz). Kohlmorgen et al. [23] had their participants drive on a highway in an instrumented vehicle and perform two cognitive tasks: counting down from a three-digit number by steps of 27, and listening to a book in the presence of another audio stream (news read). Through LDA on the power of four EEG bands (i.e., 3–15, 7–15, 10–15, 3–10 Hz) obtained through a 32-channel EEG, the authors achieved classification accuracies of 91.8% and 95.6% in detecting whether the drivers were conducting either task, respectively, vs. not.

1.3 Case Study: Cognitive Load Classification using Driver Physiological Data

The studies cited above demonstrate the potential of physiological measures, especially the EEG signals, in detecting cognitive taskload experienced by drivers. However, although Wang et al. [22] and Kohlmorgen et al. [23] achieved high classification accuracies with EEG signals, they used research-grade EEG systems, which would be too invasive to be implemented in the vehicle. Recent advancements in technology have allowed for the development of less intrusive and much more affordable EEG systems, e.g., 4 channels collected wirelessly through a thin head band. Such a consumer-grade EEG headband has been shown to be reactive to changes in cognitive taskload experienced by drivers [35]. In this chapter, we present a case study classifying cognitive taskload experienced by drivers through the use of EEG signals from this consumer-grade EEG headband, along with HR and GSR measures. The data was obtained in a driving simulator study originally reported in He et al. [35], in which 33 participants completed a drive with no secondary task and two drives with cognitive secondary tasks: one with lower difficulty (1-back) and the other with higher difficulty (2-back). Different machine learning models were built to classify the task condition that was experienced by the drivers (i.e., no task, lower difficulty cognitive task, and higher difficulty cognitive task).

Chapter 1 provided a brief overview of experimental design. Through this case study, we present an example for within-subjects experimental design. We then demonstrate the application of basic signal processing methods discussed in Chapter 3 (e.g., Fast Fourier Transformation, digital filtering) to generate features from physiological data. Next, we use and compare supervised machine learning algorithms

(e.g., Artificial Neural Network, Support Vector Machines) introduced in Chapter 4 to classify drivers' cognitive load as accurately as possible.

1.3.1 Data Source

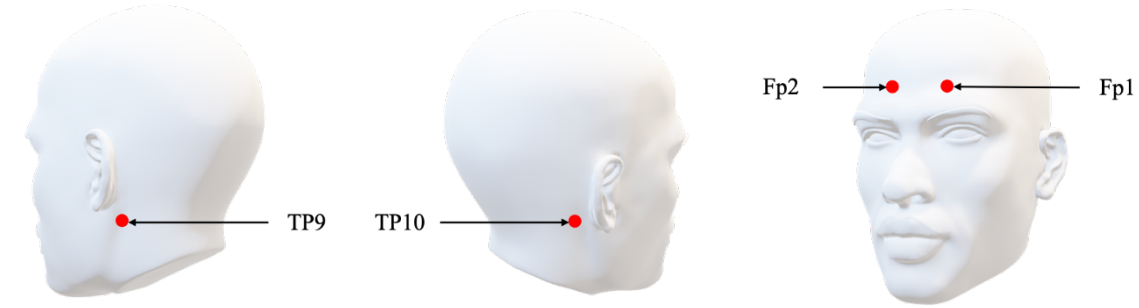
Data was recorded in a driving simulator experiment that had a within-subject design. Each participant completed three drives, each with a different cognitive load condition: no-task (no external secondary task), lower difficulty cognitive task (1-back), and higher difficulty cognitive task (2-back). Each condition was completed by the participants in a separate drive with the order of the three drives counterbalanced across participants. Following sections present the experimental methodology, which is also described in detail in He et al. [35].

Participants In total, 33 drivers (18 males and 15 females), recruited through campus and online posts, participated in this driving simulator study. Participants were required to drive at least several times per month, hold a full driver's license (G license in Ontario, Canada or equivalent) for at least 3 years, and be under 35 years old (average age: 27.6; SD: 4.45). The compensation was C$12 per hour, and participants were told that they would receive a bonus of up to C$14 based on their driving and secondary task performance as motivation for engaging in the secondary task while not prioritizing it over the driving task.

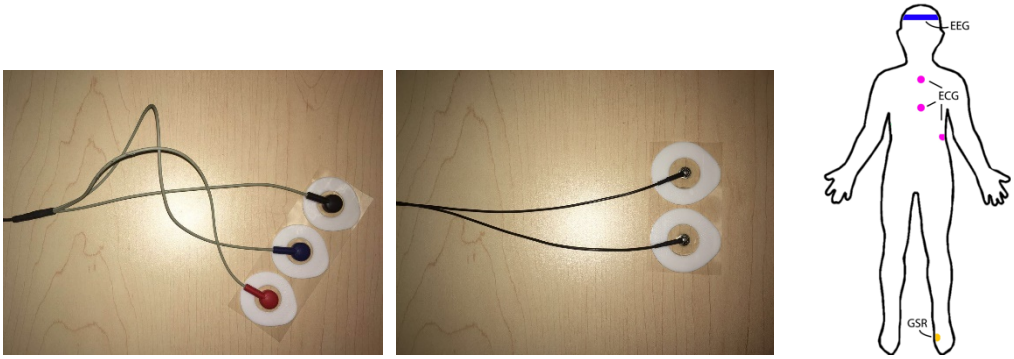
Apparatus A NADS miniSim^{TM} driving simulator was used in the experiment (Figure 1). This fixed-base simulator has three 42-inch screens, creating a 130° horizontal and 24° vertical field at a 48-inch viewing distance. The centre screen displays the left and centre parts of the windshield; the right screen displays the rest of the windshield, the rear-view mirror, and the right-side window and mirror, while the left screen displays the left-side window and mirror. EEG data was collected at 220 Hz using Muse^{TM} by Interaxon (Figure 2a), a wireless consumer-grade headband with 2 dry sensors located at Fp1 and Fp2 positions (on the forehead) and two gel foam electrodes at TP9 and TP10 positions (behind the ears), as shown in Figure 2b. The TP9 and TP10 positions were found to be unreliable due to oversensitivity to head movements, thus were excluded from our models. Heart rate was measured through Electrocardiography (ECG). Both ECG and GSR were collected using sensors by Becker Meditec that captures data at 240 Hz using the D-Lab software developed by Ergoneers. Solid gel foam electrodes were used for the ECG (Figure 2c) and GSR sensors (Figure 2d). ECG was recorded with three electrodes placed on participant's chest (Figure 2e). The GSR sensors were attached beneath the participant's bare left foot with one sensor in the middle and the other under the heel (Figure 2e).

Figure 1. NADS miniSim Driving Simulator.

(a)



(b)



(c) (d) (e)

Figure 2. Physiological sensors. (a) EEG headband. (b) Electrode positions for EEG. (c) ECG sensors. (c) GSR sensors. (d) Sensor placements on the participant.

Experimental tasks A variety of secondary tasks can be used to generate different levels of cognitive load in an experimental setting, including, hands-free phone conversations [32], mental arithmetic [11], auditory-verbal navigation tasks [33], and the n-back task [30, 31]. We chose to use the n-back task given its widespread use in the working memory [34] and the driving literatures (e.g., [30]). In particular, we adopted a modified version of the n-back task; its validation is reported in He et al. [35] who showed that the 2-back version is more difficult than the 1-back version as expected. Participants listened to a pre-recorded series of 10 letters, separated by approximately 2.5 second intervals, for an overall duration of approximately 25 seconds for each n-back task. For the 1-back task, participants were asked to count the number of times two identical letters appeared in pairs in a sequence (e.g., PP). For the 2-back task, participants were asked to count the number of times two identical letters appeared in pairs separated by one letter in between (e.g., DTD). Participants were asked to verbally respond with the total count of n-back instances at the end of each series.

The driving scenarios required the participants to follow a lead vehicle at a speed of 40 mph (around 64.4 km/h) on a 4-lane urban road with light ambient traffic and some vehicles parked on the sides. The scenarios were designed to involve mainly operational driving decisions, with no or minimal strategic or tactical decisions, such as navigation or passing a vehicle. For the 1-back and 2-back drives, the participants were presented with two groups of n-back tasks, each on a straight section of the route. Each group consisted of three n-back tasks (a series of 10 letters each), totalling to six n-back tasks completed within each drive. A notification and a brief reminder of the task was provided to let the participant know that the n-back task was starting. In our machine learning models reported in the next section, we only used four out of the six n-back tasks completed within each drive, as the other two n-back tasks had lead vehicle braking events. The braking events were included as part of a separate research question explored in He et al. [35], that is, to assess driver response to unexpected events. In the n-back drives, the participants spent about 100 seconds performing the four n-back tasks. This 100-second period for each n-back drive and a corresponding 100 second period for the no-task drive were used in model building.

Procedures Participant eligibility was verified and consent form was signed upon arrival. Participants first went through a practice drive in the simulator, on a route identical to the one used in the experimental drives. They were then given written and oral instructions on the n-back task and practiced it without driving to ensure that they fully understood and were capable of doing the task, and then they practiced it while driving. After the preparation and training phase, participants completed the three experimental drives and were given a 5-minute break after each drive. At the end of the experiment, participants were briefed and received their payment.

1.3.2 Machine Learning Models

A variety of machine learning methods, including artificial neural network, k-nearest neighbors, support vector machine (SVM), Naïve Bayes, decision tree, and linear discriminant analysis algorithms were built to classify cognitive taskload (i.e., no task, 1-back task or lower difficulty cognitive task, and 2-back task or higher difficulty cognitive task) using EEG, HR, and GSR data. The basics about these methods are presented in Chapter 4: Machine Learning Basics. A detailed introduction of EEG and its use in brain-computer interfaces can be found in Chapter 6: Brain-Computer Interfacing with Interactive Systems.

Signal processing and feature extraction An overview of our signal processing steps are presented in Figure 3. For more detailed procedures on signal processing, the reader is referred to Chapter 3: DSP Basics.

EEG signals are typically analyzed through power of frequency bands (the amount of activity found in the signal within a particular frequency range, e.g., [22]) or the Event-Related Potential (ERP) (e.g., [36]). Although ERP has shown reactivity to external cognitive taskload, it is not suitable for cognitive load detection in real time in uncontrolled settings, given that it relies on a response to a specific stimulus (e.g., detection response task) that needs to be identified prior to measurement [37]. Therefore, the powers of different frequency bands were used as features in our classification.

EEG signals were processed using the MuseLab Software provided by the device manufacturer. First, a notch filter at 60 Hz was used to reduce the influence of utility frequency in Canada. Following this filter, the power spectrum density (PSD), $S_x(f)$, which describes the distribution of power into frequency components composing a signal, was calculated using Fast Fourier Transformation (FFT) method with a hamming window of 256 samples and overlap of 234 samples ($1/10^{th}$ of a second, 10 Hz, resulting in a 90% overlap between consecutive windows). The resulting resolution after applying FFT was 0.86 Hz/bin ranging from 0 Hz to 110 Hz. Then, the power of five EEG bands (delta: 1-4 Hz; theta: 4-8 Hz; alpha: 7.5-13 Hz; beta: 13-30 Hz; gamma: 30-44 Hz) for each channel (Fp1 and Fp2) was calculated using integration:

$$P = \int_{f_1}^{f_2} S_x(f)df$$

where, P refers to the power of an EEG band; $S_x(f)$ is the PSD of the EEG signal; and f_1 and f_2 are the lower and upper range of the frequency range (e.g., 7.5 to 13 Hz for the alpha

band). Overall, there were 10 features extracted from EEG data: 2 positions × 5 bands per position.

HR was calculated from ECG data using the MATLAB Signal Processing Toolbox; over consecutive 5-second periods, average HR was calculated. A smaller period for the calculation of average HR is less meaningful considering that the intervals between heartbeats are around 1 second. First, the raw ECG data was detrended using the *POLYFIT* function to remove potential artifacts in the signal due to body movement. Then, the function *FINDPEAKS* was used to identify R spikes of the signal. The intervals between the R spikes were used to calculate HR.

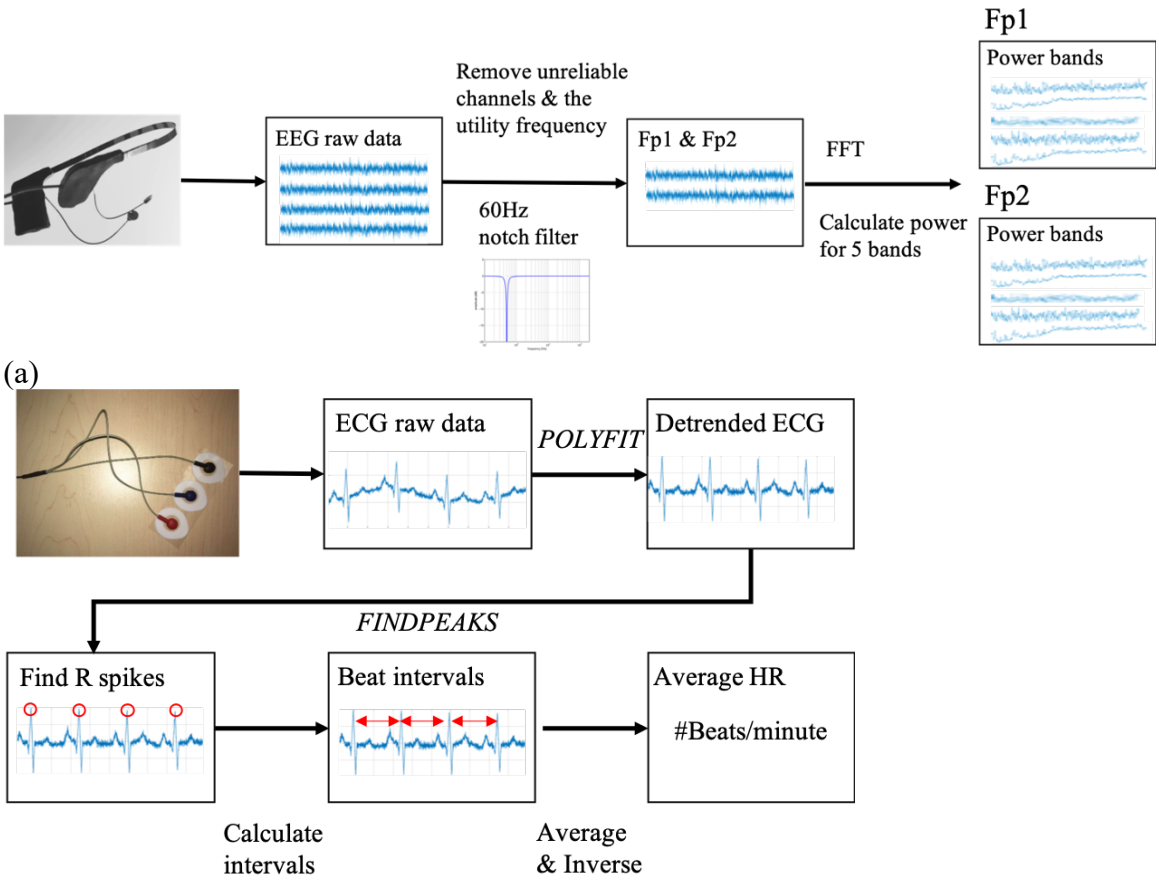


Figure 3. Processing of physiological signals. (a) EEG. (b) ECG. (c) GSR.

The average GSR was calculated every 0.1 second to match the sampling frequency of the EEG features. Overall, there was one feature each obtained from ECG and GSR. For each participant, there were 1000 data points per task condition (i.e., no task, lower difficulty cognitive task, and higher difficulty cognitive task), totalling to 3000 data points per participant. Each data point had 12 features (10 EEG, 1 HR, and 1 GSR).

Data preparation for machine learning The dataset was split into two: 75% for training and 25% for test. However, given that consecutive physiological data points are highly correlated [23], the split was not random. For each drive, the first 75 seconds of the 100 second data period of interest was assigned to the training set and the remaining 25 seconds was assigned to the test set, as shown in Figure 4.

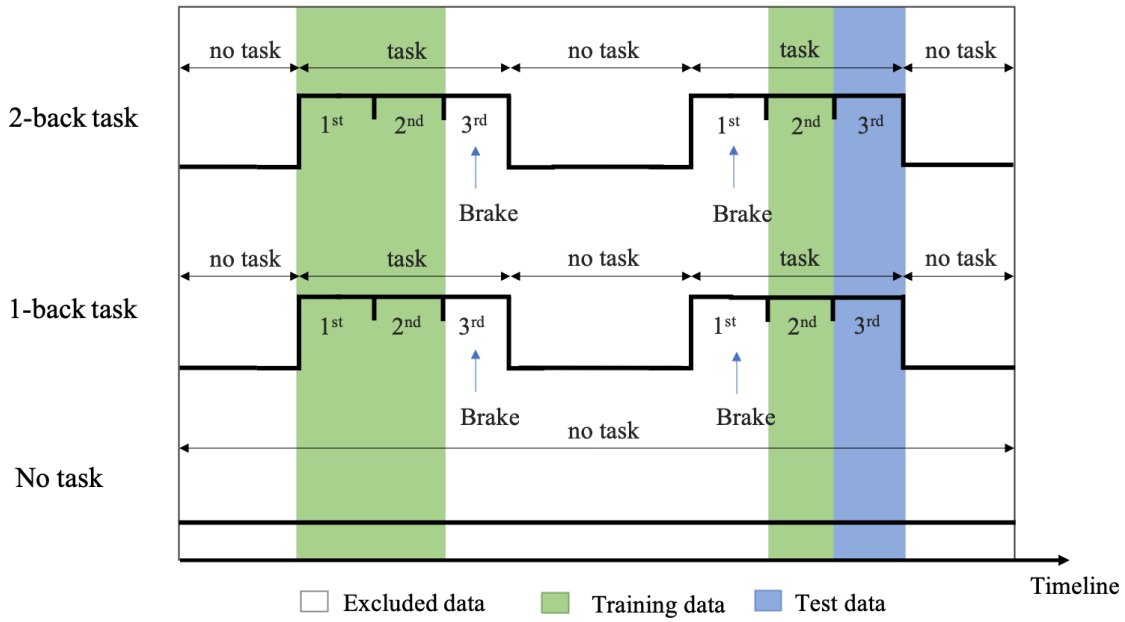


Figure 4. Illustration of the training and the test datasets.

Differences among drivers have been shown to impact the accuracy of driver state classification based on physiological data [38]. In order to minimize the effect of individual differences, each participants' data was normalized with respect to their no-task responses in the training dataset. We used only the training dataset as the reference given that in a real-world application, the training dataset would be available whereas the test dataset would represent future values not yet observed. The normalized feature values were calculated using the following equation for each participant.

$$X_{score} = \frac{X_{raw} - \overline{X}_{no\ task}}{S_{no\ task}}$$

where X_{score} is the normalized feature score, X_{raw} is the raw feature value, $\overline{X}_{no\ task}$ and $S_{no\ task}$ are the mean and the standard deviation of the feature for the no task condition in the training dataset of the participant.

Further, given that some classification algorithms, such as kNN [39], are sensitive to scale differences across features, the scales were also standardized across features using the training dataset, as shown in the equation below. The reason for using the training dataset here as well, was the same as for normalization.

$$X_{scaled} = \frac{X_{score} - \overline{X}_{training}}{S_{training}}$$

where X_{scaled} is the standardized feature score; X_{score} is the normalized feature score; $\overline{X}_{training}$ and $S_{training}$ are the mean and the standard deviation of the feature for the training dataset.

Model Building We utilized six widely-used supervised learning models: kNN, ANN, SVM, NB, decision tree, and LDA. All models were built in MATLAB using the Statistics and Machine Learning Toolbox, except for the SVM classifier. For the multi-class SVM algorithm, the LIBSVM Toolbox was used [40]. For model fit, a 3-fold cross validation was performed for each model. Overall, each type of machine learning model was built twice, once for normalized data and once for non-normalized data, in order to investigate whether normalizing physiological data as suggested by Lin et al. [38] would result in a prediction accuracy boost. Overall, 12 models were fitted. Table 2 summarizes the models, their associated parameters, the values explored in model fitting, and the software used.

Table 2 Summarization of Models and Parameters

Models	Details
kNN	Function: *FITKNN* in MATLAB Euclidian distance with equal weights Number of neighbors: k = 1, 2, 4, 8, 16[#], 32*, 64
ANN	Function: *PATTERNNET* in MATLAB Feedforward neural network Learning rate: 0.001 Architecture (Hidden layers-nodes): 1-16, 1-32, 1-64, 1-128 2-16, 2-32*, 2-64, 2-128 3-16, 3-32, 3-64[#], 3-128
SVM	Function: LIBSVM Toolbox in MATLAB Kernel: radial basis function[*#], sigmoid
NB	Function: *FITCNB* in MATLAB Prior probabilities: 1/3 for each class
Decision tree	Function: *FITCTREE* in MATLAB
LDA	Function: *FITCDISCR* in MATLAB

* denotes best parameters for normalized data; [#] denotes best parameters for non-normalized data

For kNN, the MATLAB function *FITKNN* was used to build the classifier; different k-values (number of neighbors) were explored, and k=32 and k=16 were found to provide

the best classification accuracy for normalized and non-normalized data, respectively. For ANN, a feedforward neural network was built using the MATLAB function *PATTERNNET*, with cross entropy as the loss function. For hidden layers, the activation function *TANSIG* was used, while a *SOFTMAX* function was used for the output layer. We examined different ANN architectures: 1, 2, and 3 hidden layers, with the number of neurons in each layer ranging from 2 to 128. The best performance was obtained with two hidden layers with 32-neurons for normalized data, and with three hidden layers with 64-neurons for non-normalized data. A learning rate of 0.001 was implemented for both normalized and non-normalized data.

As mentioned before, a third-party MATLAB toolbox was used for SVM. The radial basis function (RBF) and sigmoid kernels were compared, and the RBF kernel was found to lead to better accuracy for both normalized and non-normalized data. For NB, the MATLAB function *FITCNB* was used, with flat prior probabilities for the three task conditions (1/3) given that we had a balanced dataset across task conditions (or classes). For the decision tree, the MATLAB function *FITCTREE* was used. Pruning was used to prevent overfitting. Given that LDA has a closed-form solution, it does not require hyperparameters to be tuned in cross-validation. The MATLAB function *FITCDISCR* was used for LDA with a "linear" discriminant type.

Results Figure 5 presents the classification accuracy for the different models both for normalized and non-normalized data. Overall, models performed much better with normalized data, and SVM provided the best accuracy of 79.4% in classifying the three cognitive states. ANN and kNN also performed well at 75.8% accuracy. Confusion matrices were also generated for the models built on the normalized data (Table 2).

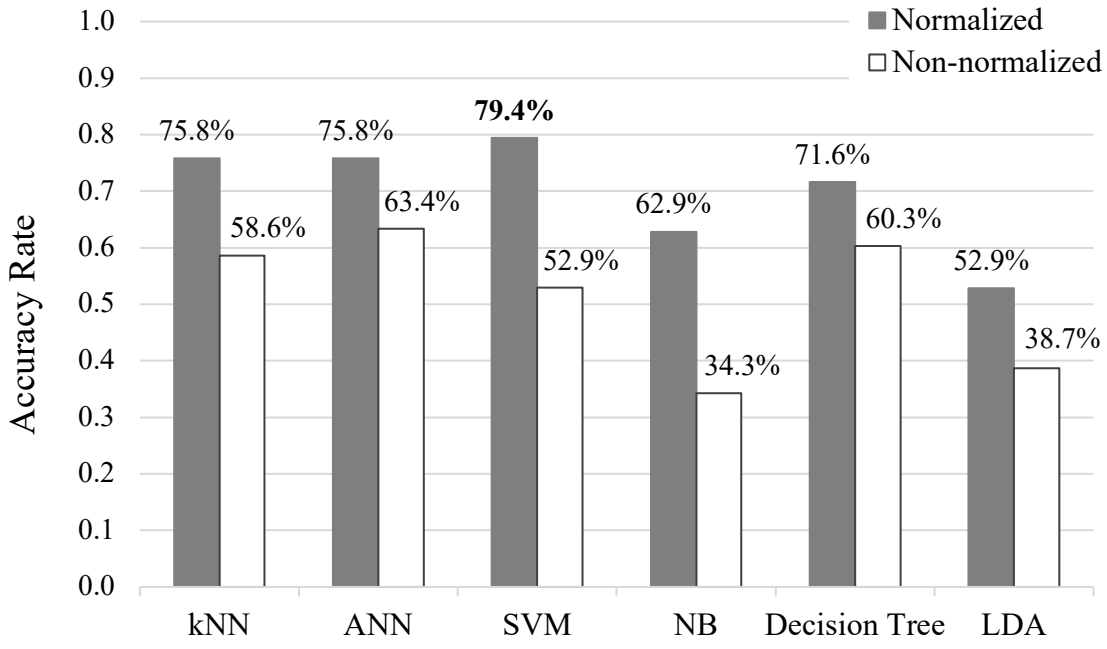


Figure 5. Classification accuracy (three classes: no task, lower difficulty cognitive task, and higher difficulty cognitive task) with and without normalization based on participants' no task data.

Table 2. Confusion matrices for the six machine learning models that used normalization.

kNN

Target \ Predicted	no task	1-back	2-back
no task	81.3%	14.8%	3.9%
1-back	25.4%	67.2%	7.4%
2-back	7.0%	14.2%	78.8%

ANN

Target \ Predicted	no task	1-back	2-back
no task	69.7%	25.6%	4.7%
1-back	19.9%	74.7%	5.5%
2-back	0.9%	16.2%	82.9%

SVM

Target \ Predicted	no task	1-back	2-back
no task	80.1%	13.9%	6.0%
1-back	24.4%	71.0%	4.6%
2-back	2.0%	10.9%	87.1%

NB

Target \ Predicted	no task	1-back	2-back
no task	80.1%	14.3%	5.6%
1-back	34.5%	56.7%	8.8%
2-back	3.6%	44.6%	51.8%

Decision Tree

Target \ Predicted	no task	1-back	2-back
no task	53.1%	38.6%	8.3%
1-back	19.2%	69.6%	11.1%
2-back	3.4%	4.6%	92.0%

LDA

Target \ Predicted	no task	1-back	2-back
no task	76.3%	18.2%	5.5%
1-back	36.3%	42.4%	21.2%
2-back	40.2%	19.9%	39.9%

Discussion and Practical Implications The modeling results revealed that SVM provided the highest classification accuracy (79.4%), followed by ANN (75.8%) and kNN (75.8%). Earlier studies that also used physiological predictors to classify drivers' cognitive taskload reported classification accuracies of 85-96% [22, 23, 25]; however, these earlier studies only focused on two-class problems, i.e., they only tried to distinguish between whether the driver was conducting a secondary task or not. In contrast, we classified driver taskload into three classes: no task, lower difficulty cognitive task (1-back), and higher difficulty cognitive task (2-back). From our confusion matrices, it can be observed that most models reached high classification accuracies for the no-task category. It is not surprising that detecting the mere existence of a secondary task would be an easier classification problem than detecting the existence and also the level of taskload. Further, the cognitive tasks that were used in some of these earlier studies appear to be more difficult than our task levels, which also makes the classification problem easier. For example, Solovey et al. [25] classified no-task from the 2-back task (our higher difficulty task level) and Kohlmorgen et al. [23] had their participants count down from a three-digit number by steps of 27. Overall, for these reasons, the accuracy level reached with the SVM model is promising for our classification problem. Systems that intervene (e.g., blocking incoming phone calls) when high levels of taskload is detected should be able to not only detect the existence of a secondary task but to distinguish between low and high levels of taskload. Interventions

during low taskload can result in nuisance alarms, which would in turn influence the drivers' system acceptance and use [41, 42].

As stated earlier, we built our models both on normalized and non-normalized data. Differences among drivers have been shown to impact the accuracy of driver state classification based on physiological data [38] and normalization has been suggested as a strategy to mitigate the effects of these differences. Our models showed that normalization does significantly enhance model accuracy; there was an increase of 26.5% in the SVM classification accuracy with normalization. Each participants' data was normalized with respect to their no-task responses in the training dataset. We used only the training dataset as the reference given that in a real-world application, the training dataset would be available whereas the test dataset would represent future values not yet observed. However, this normalization strategy requires the system to have prior data from the driver, or learn from the driver over time. We had a limited sample size of 33 people; it is possible that a model built on a much larger sample would make the normalization and hence learning from an individual driver less important.

1.4 Conclusion

We presented a case study demonstrating the usefulness of physiological signals in classifying different cognitive taskload levels experienced by drivers. In this case study, we explored a number of commonly used machine learning models, but the reader should be informed that there are other models, such as Hidden Markov [43] and Semi-Markov Models [44] and Recurrent Neural Networks [45], that are specifically built to capture state transitions. These models may perform better than the ones employed in this case study. Further, although excessive levels of cognitive load experienced by drivers is an important concern for traffic safety, there are other types of demands (e.g., visual tasks such as texting on a cell phone) that are also of concern and the prediction of these different demand types is an area of active research (e.g., [24, 46-48]). Detecting the demands experienced by drivers in real-time can be used to inform the behavior of adaptive interfaces [e.g., 49].

1.5 Further Reading

The reader is referred to the following body of literature for further readings: [12, 50] for cognitive load; [51] for general augmented cognition applications in human-computer interaction; and [19, 52] for driver state measures.

1.6 Follow-up Questions

While the authors are unable to provide the data used in this case study, there are a number of public EEG datasets, some at: https://github.com/meagmohit/EEG-Datasets. The reader is encouraged to explore these EEG datasets and practice the machine learning models and EEG data processing methods presented in this chapter and elsewhere in this book. We also provide some questions/suggestions below to guide your exploration:

- As has been discussed, the power of the frequency band is one feature we can elicit from EEG data, what other EEG features can be used to assess human operator state?
- EEG data is sensitive to artifacts such as eye blinks. The data that we used in this chapter was pre-processed by the data collecting software, so we did not need to manually filter artifacts. Practice design filters to remove eye blinks from the EEG data.
- Eye blinks can be indicator of workload. Thus, eye blinks can be considered an artifact but also a potentially useful feature of EEG data. Can you design algorithms to identify and capture the rate of eye blinks from EEG data?
- The machine learning models we explored in this chapter do not capture the temporal nature of the data. Are there other types of models that can leverage the temporal information for driver state detection? How would you implement them to a dataset that is similar to the one we present?

Acknowledgements
The funding for this study was provided by the Natural Sciences and Engineering Research Council of Canada (NSERC). We gratefully acknowledge Konstantinos Plataniotis, Cheng Chen Liu, Amirhossein S. Aghaei, Huei-Yen Winnie Chen, and George Liu for their role in the design of simulator study, and Nicole Wongsoo for her help in data collection.

References

1. Joel C. McCall and Mohan M. Trivedi. 2004. Visual context capture and analysis for driver attention monitoring. In *Proceedings of the 7th International IEEE Conference on Intelligent Transportation Systems*, 332-337.

2. Alex Pentland and Andrew Liu. 1999. Modeling and prediction of human behavior. *Neural Computation* 11, 1: 229-242.

3. Guosheng Yang, Yingzi Lin, and Prabir Bhattacharya. 2010. A driver fatigue recognition model based on information fusion and dynamic Bayesian network. *Information Sciences* 180, 10: 1942-1954.

4. Lex Fridman, et al. 2017. What can be predicted from six seconds of driver glances? In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 2805-2813.

5. Ashesh Jain, et al. 2015. Car that knows before you do: Anticipating maneuvers via learning temporal driving models. In *Proceedings of the IEEE International Conference on Computer Vision*, 3182-3190.

6. Sujitha Martin, Sourabh Vora, Kevan Yuen, and Mohan Manubhai Trivedi. 2018. Dynamics of driver's gaze: Explorations in behavior modeling and maneuver prediction. *IEEE Transactions on Intelligent Vehicles* 3, 2: 141-150.

7. Mercedes-Benz. 2016. *Safety. Decades of preparation. Millions of proof.* [cited 2019 March 09]; Available from: https://cars.mercedes-benz.ca/the-best-or-nothing/safety/safety.html.

8. Thomas A. Dingus, et al. 2016. Driver crash risk factors and prevalence evaluation using naturalistic driving data. *Proceedings of the National Academy of Sciences* 113, 10: 2636-2641.

9. Michael A. Regan, John D. Lee, and Kristie Young. 2008. *Driver Distraction: Theory, Effects, and Mitigation*. Boca Raton: CRC Press.

10. Miguel A. Recarte and Luis M. Nunes. 2000. Effects of verbal and spatial-imagery tasks on eye fixations while driving. *Journal of Experimental Psychology: Applied* 6, 1: 31.

11. Joanne L. Harbluk, Y. Ian Noy, Patricia L. Trbovich, and Moshe Eizenman. 2007. An on-road assessment of cognitive distraction: Impacts on drivers' visual behavior and braking performance. *Accident Analysis & Prevention* 39, 2: 372-379.

12. Fred GWC Paas and Jeroen JG Van Merriënboer. 1994. Instructional control of cognitive load in the training of complex cognitive tasks. *Educational Psychology Review* 6, 4: 351-371.

13. Maartje GA Ament, Anna L Cox, Ann Blandford, and Duncan Brumby. 2010. Working memory load affects device-specific but not task-specific error rates. In *Proceedings of the Annual Meeting of the Cognitive Science Society*,

14. Duncan P Brumby, Dario D Salvucci, Walt Mankowski, and Andrew Howes. 2007. A cognitive constraint model of the effects of portable music-player use on driver performance. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 1531-1535.

15. Marika Hoedemaeker and Mark Neerincx. 2007. Attuning in-car user interfaces to the momentary cognitive load. In *International Conference on Foundations of Augmented Cognition*, 286-293.

16. Kristie Young and Michael Regan. 2007. Driver distraction: a review of the literature, In *Distracted Driving*, I.J. Faulks, et al., Editors. Australasian College of Road Safety: Sydney, NSW, Australia. 379-405.

17. Bryan Reimer, et al. 2013. Impact of age and cognitive demand on lane choice and changing under actual highway conditions. *Accident Analysis Prevention* 52: 125-132.

18. Walter Piechulla, Christoph Mayser, Helmar Gehrke, and Winfried König. 2003. Reducing drivers' mental workload by means of an adaptive man–machine interface. *Transportation Research Part F: Traffic Psychology and Behaviour* 6, 4: 233-248.

19. Amirhossein S. Aghaei, et al. 2016. Smart Driver Monitoring: When Signal Processing Meets Human Factors: In the driver's seat. *IEEE Signal Processing Magazine* 33, 6: 35-48.

20. Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Journal of Chemical Information and Modeling* 53: 1689-1699.

21. Sarah Miller. 2001. Literature Review Workload Measures. ID: N01-006. University of Iowa, Iowa City, IA, USA.

22. Yu Kai Wang, Tzyy Ping Jung, and Chin Teng Lin. 2015. EEG-based attention tracking during distracted driving. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 23, 6: 1085-1094.

23. Jens Kohlmorgen, et al. 2007. Improving human performance in a real operating environment through real-time mental workload detection, In *Toward Brain-Computer Interfacing*, Guido Dornhege, et al., Editors. MIT Press: Cambridge, Massachusetts. 409-422.

24. Yulan Liang, Michelle L. Reyes, and John D. Lee. 2007. Real-time detection of driver cognitive distraction using support vector machines. *IEEE Transactions on Intelligent Transportation Systems* 8: 340-350.

25. Erin T. Solovey, et al. 2014. Classifying driver workload using physiological and driving performance data. In *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems - CHI '14*, 4057-4066.

26. Gianluca Borghini, et al. 2014. Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness. *Neuroscience and Biobehavioral Reviews* 44: 58-75.

27. Lisa R. Fournier, Glenn F. Wilson, and Carolyne R. Swain. 1999. Electrophysiological, behavioral, and subjective indexes of workload when performing multiple tasks: manipulations of task difficulty and training. *International Journal of Psychophysiology* 31: 129-145.

28. P. Wagner and T. Wagner. 2013. Galvanic skin response and investigation into cheating. *Human Psychophysiology* 8: 1-25.

29. Kilseop Ryu and Rohae Myung. 2005. Evaluation of mental workload with a combined measure based on physiological indices during a dual task of tracking and mental arithmetic. *International Journal of Industrial Ergonomics* 35, 11: 991-1009.

30. Bruce Mehler, Bryan Reimer, Joseph F. Coughlin, and Jeffery A. Dusek. 2009. Impact of incremental increases in cognitive workload on physiological arousal and performance in young adult drivers. *Transportation Research Record* 2138: 6-12.

31. B. Mehler, B. Reimer, and J. F. Coughlin. 2012. Sensitivity of physiological measures for detecting systematic variations in cognitive demand from a working memory task: an on-road study across three age groups. *Human Factors* 54, 3: 396-412.

32. Karel A. Brookhuis, Gerbrand de Vries, and Dick de Waard. 1991. The effects of mobile telephoning on driving performance. *Accident Analysis & Prevention* 23, 4: 309-316.

33. Yulan Liang and John D. Lee. 2010. Combining cognitive and visual distraction: less than the sum of its parts. *Accident Analysis & Prevention* 42, 3: 881-890.

34. Adrian M. Owen, Kathryn M. McMillan, Angela R. Laird, and E. Bullmore. 2005. N-back working memory paradigm: a meta-analysis of normative functional neuroimaging studies. *Human Brain Mapping* 25, 1: 46-59.

35. Dengbo He, Cheng Chen Liu, Birsen Donmez, and Konstantinos N. Plataniotis. 2017. Assessing high cognitive load in drivers through Electroencephalography. In *Proceedings of Transportation Research Board 96th Annual Meeting*, 17-02615.

36. Claudio Imperatori, et al. 2013. Modifications of EEG power spectra in mesial temporal lobe during n-Back tasks of increasing difficulty. A sLORETA study. *Frontiers in Human Neuroscience* 7: 109.

37. Samuel Sutton, Margery Braren, Joseph Zubin, and E.R. John. 1965. Evoked-potential correlates of stimulus uncertainty. *Science* 150, 3700: 1187-1188.

38. Chin Teng Lin, et al. 2005. EEG-based Drowsiness Estimation for Safety Driving Using Independent Component Analysis. *IEEE Transactions on Circuits and Systems I: Regular Papers* 52: 2726-2738.

39. Leif E. Peterson. 2009. K-nearest neighbor. *Scholarpedia* 4, 2: 1883.

40. Chih Chung Chang and Chih Jen Lin. 2011. LIBSVM: a Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology* 2: 1-27.

41. Birsen Donmez, Linda Ng Boyle, John D. Lee, and Daniel V. McGehee. 2006. Drivers' attitudes toward imperfect distraction mitigation strategies. *Transportation Research Part F: Traffic Psychology and Behaviour* 9, 6: 387-398.

42. Raja Parasuraman and Victor Riley. 1997. Humans and automation: Use, misuse, disuse, abuse. *Human Factors* 39, 2: 230-253.

43. Sean R. Eddy. 1996. Hidden markov models. *Current Opinion in Structural Biology* 6, 3: 361-365.

44. Shun-Zheng Yu. 2010. Hidden semi-Markov models. *Artificial Intelligence* 174, 2: 215-243.

45. Ronald J. Williams and David Zipser. 1989. A learning algorithm for continually running fully recurrent neural networks. *Neural Computation* 1, 2: 270-280.

46. Zhaojian Li, Shan Bao, Ilya V. Kolmanovsky, and Xiang Yin. 2017. Visual-manual distraction detection using driving performance indicators with naturalistic driving data. *IEEE Transactions on Intelligent Transportation Systems* 19, 8: 2528 - 2535.

47. Mauricio Munoz, et al. 2016. Distinguishing patterns in drivers' visual attention allocation using Hidden Markov Models. *Transportation Research Part F: Traffic Psychology and Behaviour* 43: 90-103.

48. Martina Risteska, Joyita Chakraborty, and Birsen Donmez. 2018. Predicting environmental demand and secondary task engagement using vehicle kinematics from naturalistic driving data. In *Proceedings of the 10th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 66-73.

49. Angelos Amditis, Luisa Andreone, Aris Polychronopoulos, and Johan Engström. 2005. Design and development of an adaptive integrated driver-vehicle interface: overview of the AIDE project. *IFAC Proceedings Volumes* 38, 1: 103-108.

50. Fred Paas, Alexander Renkl, and John Sweller. 2003. Cognitive load theory and instructional design: Recent developments. *Educational psychologist* 38, 1: 1-4.

51. 2007. *Foundations of Augmented Cognition*. CRC Press.

52. Fang Chen, et al. 2012. Multimodal behavior and interaction as indicators of cognitive load. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 2, 4: 22.