

Classifying Driver Workload Using Physiological and Driving Performance Data: Two Field Studies

Erin T. Solovey^{1,2}, Marin Zec², Enrique Abdon Garcia Perez², Bryan Reimer², Bruce Mehler²

¹Drexel University
Philadelphia, PA

²MIT AgeLab
Cambridge, MA

erin@cs.drexel.edu, marin.zec@tum.de, engarpe@mit.edu, reimer@mit.edu, bmehler@mit.edu

ABSTRACT

Understanding the driver's cognitive load is important for evaluating in-vehicle user interfaces. This paper describes experiments to assess machine learning classification algorithms on their ability to automatically identify elevated cognitive workload levels in drivers, leading towards the development of robust tools for automobile user interface evaluation. We look at using both driver performance as well as physiological data. These measures can be collected in real-time and do not interfere with the primary task of driving the vehicle. We report classification accuracies of up to 90% for detecting elevated levels of cognitive load, and show that the inclusion of physiological data leads to higher classification accuracy than vehicle sensor data evaluated alone. Finally, we show results suggesting that models can be built to classify cognitive load across individuals, instead of building individual models for each person. By collecting data from drivers in two large field studies on the highway (20 drivers and 99 drivers), this work extends prior work and demonstrates feasibility and potential of such measures for HCI research in vehicles.

Author Keywords

Cognitive workload; driving; physiological computing; heart rate; skin conductance; machine learning.

ACM Classification Keywords

H.5.2. Information interfaces and presentation (e.g., HCI): User Interfaces.

INTRODUCTION

In 2011, the U.S. National Highway Traffic Safety Administration reported that 3,331 people were killed and an estimated 387,000 were injured on U.S. roadways in motor vehicle crashes involving distracted driving [26]. As a consequence, technology usage in the vehicle is a major safety

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.
CHI 2014, April 26 - May 01 2014, Toronto, ON, Canada
Copyright 2014 ACM 978-1-4503-2473-1/14/04...\$15.00.

<http://dx.doi.org/10.1145/2556288.2557068>



Figure 1. Recording of physiology during on-road driving. Electrocardiogram (top right) and skin conductance (bottom right) sensor placement are shown.

concern. A casual observation of drivers suggests that they increasingly attend to mobile devices and interact with new technology built in their vehicles, creating distracted driving scenarios.

Driving itself is a dynamic, complex activity involving visual, cognitive and manual tasks: the driver has to form strategic goals, monitor the roadway environment and the vehicle systems, process information and make tactical action plans as well as execute control level activities [23]. Thus, the driving task imposes varying levels of workload on the driver. Understanding the workload induced during driving is important for preventing accidents and hazards on the road, and human factors researchers have studied driver workload in depth. It has been shown that operators perform better at intermediate levels of workload compared to extreme levels (i.e. too low or too high workload) [8]. The roots of this are in the Yerkes Dodson law of arousal [41] (inverted U) that suggests that during periods of underload, added workload may improve performance, while during heightened demand, higher workload may reduce performance. As they vie for driver attention, the workload imposed on the driver by in-vehicle user interfaces is constantly changing.

Advances in technology have led to a shift in the demands in many working environments from the largely physical to more supervisory oversight of automation, more cognitive demand, and increased frequency of vocal command interaction [34,42]. This development can be observed in the field of driving as well [16,37]. Modern cars provide driver support systems such as power steering, assistive cruise

control and lane keeping assist, which decrease the physical demands of driving. The next generations of automated vehicles will likely partially, if not completely, relieve the driver of safety critical control. However, the increasing level of automation is expected to place more variable demands on driver attention and cognitive activities. The complexity of in-vehicle interfaces will continue to increase at a rapid rate. Vehicles will likely contain more infotainment options, and proposed automated driving systems will demand new interfaces. In addition, the increased autonomy of the car may provide the driver with a sense that they have the capacity to undertake additional tasks while driving. Understanding how the evolving vehicle interface affects the driver's cognitive load is critical for optimizing the user experience and safety during system design and through adaptive interfaces. This work aims to explore automatic detection of driver cognitive workload through physiology and vehicle data.

Although mental workload is not directly observable [10], several measurement approaches to assessing mental workload are employed. Empirical measures are generally divided into three categories: subjective measures, performance measures and psychophysiological measures. Each method has advantages as well as disadvantages [19,43]. Subjective measures are cost-effective and suitable in prototype testing; however they suffer from time delay, subjective biases and are highly intrusive which ultimately makes them unsuitable for continuous workload assessment. Performance and psychophysiological measures can be measured non-intrusively and continuously throughout the tasks [19].

Recently, with sensing capabilities improving and costs decreasing, there has been a growing interest among automotive vendors in enabling their products to monitor and exploit driver and vehicle sensor data. Driving measures of speed, acceleration, location and inter-vehicle distance are more readily available for inferring the current situation. In addition, cameras and other sensors are increasingly able to measure driver data such as heart rate, gaze direction and other physiological measures. These vehicle and physiological measures provide the potential to monitor the dynamic state of the driver while actually driving and provide inputs to make adjustments in the characteristics of the vehicle or interfaces to improve performance [6].

Starting initially in a driving simulator [19] and moving to field studies [18,28,29], previous work demonstrated that vehicle sensor and physiological measures can both be collected in real-time and do not interfere with the primary task, making them potentially valuable for evaluating workload associated with automotive user interfaces. This work was largely based on normative assessment of group level data. While a simple threshold based assessment showed promise for detecting changes in cognitive load at the individual level [18], it remained to be shown how effectively this can be achieved using more advanced modeling.

This paper builds from this prior work with the goal of improving methods available for evaluating automotive user interfaces that may be used during driving tasks. It brings

us closer to automatic cognitive workload classification and makes the following contributions. 1) We report machine learning classification results along with details of the techniques and parameters for recognizing elevated cognitive load, based on data from a moderately sized and a large on-road field study. 2) We compare the value of heart rate, skin conductance and driving data for classification. 3) We explore the differences between training classifiers within individuals and across individuals. By successfully training across individuals, we demonstrate the potential of building models that can generalize to work for new drivers.

RELATED WORK

The following sections provide background on peripheral physiological measures for workload detection in driving.

Cognitive Load and Heart Rate and Skin Conductance

Cardiovascular measures have been reported to be sensitive to mental workload changes (e.g. [5,12,25,38,39]). The heart is innervated by both parts of the autonomic nervous system: the sympathetic nervous system (SNS) and the parasympathetic nervous system (PNS). Heart rate (HR) and heart rate variability (HRV) are widely adopted mental workload measures because they are easy to use and provide fundamental information about the autonomic nervous system [3,25]. While the SNS activation promotes arousal (e.g. "flight-or-fight" response in emergency), the PNS is responsible for maintaining bodily functions and resting. Sympathetic activation increases heart rate while parasympathetic activation decreases heart rate.

Electrodermal activity (EDA) refers to the electrical activity from (eccrine) sweat glands and their associated dermal and epidermal tissues [4]. There are two types of sweat glands in the human skin: the eccrine and the apocrine. Though related, eccrine and apocrine sweat glands are distinct in size, structure, function, distribution and nervous control [31]. Eccrine sweat glands are under control of the SNS. Thus, physiological arousal affects the production of ionic sweat by eccrine sweat glands resulting in changes of electrical resistance and conductance at the surface of the skin. EDA has been used as a measure for psychophysiological arousal for more than a century (e.g. [28]). EDA measures such as skin conductance level (SCL) and skin conductance response (SCR) have been reported to be sensitive to arousal and mental workload in driving as well [14,36]. SCL represents the overall tonic conductance level and SCR captures momentary (phasic) changes in electrical conductivity. This paper explores HR and SCL.

Measuring Physiological Signals in the Car

Literature on measuring physiological signals in the car can be divided into simulator studies [13,18,19,24,36] and field studies [14,15,18,28,30,33,40]. The major advantage of simulator studies is their relative ease of controlling experimental variables and conditions. However, there are disadvantages compared to a field study as well, e.g. lower risk perception or simulator sickness. Since driver-vehicle interfaces eventually have to be effective and safe in real-world settings with dynamic environmental factors such as light-

ing or soundscape, this paper utilizes data from two field-studies to examine classification performance in a more realistic setting than a driving simulator. In addition, while most of these studies report significant normative differences in physiological signals during driving, this paper works to automatically detect these differences at an individual level using machine learning. Below, we discuss some related field studies.

De Waard et al. [40] studied 28 male participants driving an instrumented car on two different roads, each associated with a different demand level. The data suggests significant effects of road type on heart rate measures, but they did not look specifically at workload.

Healey et al. [14] used physiological data (EKG, EMG, EDA, and respiration) from nine drivers in three real-world settings: rest, highway and urban. Three drivers accomplished seven drives each. Another six drivers completed one single drive. The three driving conditions were low (rest), medium (highway) or high (city) stress, respectively. This was validated through questionnaires. Using a linear discriminant function, an overall accuracy of 97.4% was reported for stress level classification. However, the generalizability is limited since 21 of the 27 drive datasets are attributed to three drivers only (each one of these repeated the course on multiple days).

In a larger study of 49 professional drivers, Jahn et al. [15] found heart rate to be sensitive to workload manipulation in their driving study. However, they conclude that the heart rate changes they observed reflect emotional strain or physical workload from steering actions as well. According to the authors, heart rate proved to be a sensitive but not selective measure for workload. In contrast, [25] claims HRV to be a selective measure for mental load. (See [22] for discussion of HR vs. HRV in high workload detection.).

The European HASTE study [8] produced mixed results on the relative sensitivity of heart rate and skin conductance in response to the demands surrogate secondary tasks in a sample of 24 drivers in the field. In contrast, Collet et al. [5] collected heart rate and skin resistance (inverse of SCL) data in 10 drivers on a closed driving track and found that both measures increased during secondary tasks requiring their attention (a phone conversation, a radio broadcast with content they would be quizzed on, and a conversation).

Schneegaß, et al. [33] present a field study with ten participants in which they collected EKG, skin conductance, and skin temperature data while participants drove in road environments presumed to induce differing levels of demand. This preliminary report only considered the skin conductance and skin temperature data, and found skin conductance for the group varied significantly across the road types and to be the more sensitive measure.

The in-vehicle evaluation framework and two on-road field experiments described in this paper extend this previous work by collecting data from a significantly larger, gender-balanced group of drivers across different age groups and looks at on-road responses to secondary task demands.

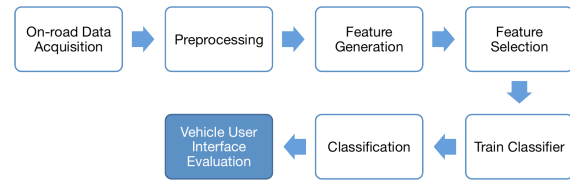


Figure 2. Steps required for on-road cognitive state classification for vehicle user interface evaluation.

CLASSIFYING DRIVER DATA FOR UI EVALUATION

The related work in this area has generated interest from automotive companies for utilizing physiological and vehicle sensor data in user interface evaluation. However, bringing this research into the in-vehicle user evaluation is still a challenge, as there are few standard methods in place for doing so. We detail our procedure and results as a step toward this. We employ machine learning techniques to learn a model for elevated workload based on physiological and vehicle data for automatic classification. This involves several steps as shown in Figure 2. We describe each of the steps in more detail below, along with the potential choices that need to be made to conduct on-road user evaluations using these methods. Research questions are highlighted.

First, there is a broad range of sensor technology available for *on-road data acquisition*, and more will be available in the future. In the studies described below, we investigate heart rate, skin conductance and vehicle telemetry because they have shown to be sensitive to changes in cognitive workload, and are relatively easy to measure during driving. We also assess their relative sensitivity to elevated cognitive demand to determine the most effective sensors. As additional sensors are shown to be indicative of cognitive workload in the future, they could be added to this framework. This leads to a research question addressed in this paper: *Can peripheral physiological and driver performance measures be used for accurate classification of elevated levels of driver workload for interface evaluation?*

Raw sensor data is often not suitable for direct analysis due to various forms of artifact and errors. Often, the first step in data *preprocessing* involves a manual or automated cleaning of the input signals. In addition, some useful data features cannot be directly measured but have to be computed first. The data pre-processing steps deal with such issues, but are often omitted in study descriptions. For each experiment below, we detail the pre-processing steps taken.

Feature generation and *feature selection* convert the pre-processed data into a set of values that could be used for learning a model for the various cognitive states. With sensor data, the recorded measurements form a time series that could be used directly as a set of features. Also, some features have to be extracted from raw data because they provide more information than the actual raw signal (e.g. heart rate is more informative than raw EKG). Other approaches aggregate the time series into summary functions such as the mean over a window. This reduces the dimensionality of the data. Previous work has shown that parameters such

as window size can have an impact on classification results in EEG data [11]. Understanding how these factors influence classification of other physiological and vehicle data will be addressed in the experiments below in order to answer the question: *How is classification accuracy affected by feature generation and selection parameters?*

With a set of features, we can employ machine learning techniques to *train a classifier* for detecting the cognitive workload of the driver, based on data previously acquired. Supervised learning techniques were employed using the empirical data from the two field studies with labeled data (*elevated* and *normal* workload levels). Supervised learning algorithms generally consist of two phases: a *training* phase and a *classification* phase. In the learning phase, the algorithm builds a model on the basis of training data and their labels. The resulting model is used in the classification phase to predict labels (i.e. class membership) for unlabeled data. The training period could happen just prior to classification, or the classification models could be built earlier, based on historical data. The quality of the training data will impact the performance of the classifier. We look at answering: *How does training data within and across individuals impact automatic cognitive workload recognition?*

Numerous classification algorithms have emerged and we report results with five different classifiers. We aim to investigate: *How does classification algorithm impact the ability estimate elevated workload with physiological data?*

The final goal with the real-time classifier is to *evaluate in-vehicle interfaces and technology* using the automatic cognitive workload detection in combination with more traditional measures. Using this framework, we can identify moments of interest during task performance where cognitive workload is elevated. This can be analyzed to identify interface, driving, and environmental conditions that may have induced elevated cognitive load. The rest of the paper focuses on practical issues for using this in realistic settings.

EXPERIMENT FRAMEWORK

We conducted two experiments based on data collected in field studies to explore the practical considerations for automatic classification of cognitive load during actual on-road driving. In both field studies, participants drove on an interstate highway while vehicle performance and physiology data were recorded. In addition to the primary driving task, participants had to complete a cognitive demand task described below as a proxy for secondary tasks that cause elevated workload.

We describe our approach to classification and the various parameters that could affect the classification results. These include: machine learning algorithm, window size, overlap of windows and features used. In the first experiment with 20 subjects, we look at classifying individual driver's workload levels by building unique models for each person. In the second experiment with 99 subjects, we explore building classification models across individuals, reducing the need for training on each individual. Common elements of the field studies are detailed below.

Stimulus	9	3	7	1	8	0	2	4	6	5
Response	.	.	9	3	7	1	8	0	2	4

Table 1. Example task block of auditory stimuli and the appropriate verbal responses in a 2-back secondary task performed during on-road driving.

Secondary Task Procedure

In the two field studies, an auditory presentation - verbal response delayed digit recall task was employed to impose additional mental workload while driving. This "n-back" task is documented in detail in [21]. The single digits 0-9 were presented one at a time at 2.25 second intervals in a randomly ordered sequence. As each new item was presented, participants were required to say out loud the digit two items back in the current sequence. An example set of a 2-back task is shown in Table 1. (The 0-back where a participant simply repeats each number as it is presented and 1-back where the number one item back in the sequence have also been used and elicit lower levels of demand.) This secondary task requires auditory perception and cognitive processing involving working memory. This has been shown in previous work to increase cognitive demand [20,29] and is being utilized as a calibration task that provides drivers with a consistent, and validated, dose of high cognitive demand that has been used in prior studies.

Vehicle Equipment and Physiological Sensors

In both experiments, a vehicle was instrumented for time-synchronized data collection from embedded vehicle sensors, a MEDAC System/3 monitoring system (NeuroDyne Medical Corp., Cambridge, MA). Vehicle performance data were logged at 10 Hz and physiological data at 250 Hz. The first data set was collected in a 2010 Lincoln MKS and the second data set was collected in a 2004 Volvo XC90.

Electrocardiogram (EKG) recordings employed a modified lead II configuration: the negative lead was placed just under the right clavicle (collar bone), the ground lead just under the left clavicle (Figure 1), and the positive lead on the left side over the lower rib. The skin was cleaned with isopropyl alcohol and standard pre-gelled silver/silver chloride disposable electrodes (Vermed A10005, 7% chloride wet gel) were applied. Skin conductance was measured utilizing a constant current configuration and non-polarizing, low impedance gold plated electrodes that allow electrodermal recording without the use of conductive gel. Sensors were placed on the underside of the outer segments of the middle fingers of the non-dominant hand and secured with medical grade paper tape. The thin surface, low profile design of the electrodermal sensors minimize interference with a natural grip of the steering wheel associated with the use of more traditional cup style electrodes. Figure 1 shows one of the two sensors. Measures of driving speed, steering wheel position, and acceleration data were recorded directly from the controller area network (CAN) bus of the vehicle.

A research associate was seated in the rear of the vehicle and was responsible for providing driving directions, ensuring safe vehicle operation, that participants understood and followed instructions, recording telemetry was working

properly and that the experiment proceeded according to a predefined script (Figure 1). The data acquisition system supported playing recorded audio and this ensured that primary instructions and tasks were presented consistently.

Data Preprocessing

A typical EKG waveform of a heartbeat consists of six components labeled P, Q, R, S, T and U. Each wave represents a specific stage in the underlying physiological process of a cardiac cycle. The most prominent pattern in the EKG waveform is the QRS segment since it usually contains a sharp spike in the signal (R-wave peak). However, EKG recordings often contain artifact (e.g. interference from skeletal muscle activity or electrical noise) and anomalies (e.g. as a result of heart conditions or equipment failures). We employed a QRS detection algorithm [2,7] to identify heart beats in the signal. The results of the heart beat detection were manually reviewed and edited. Heart rate (HR) and heart rate variability (HRV) have both been used successfully for assessing operator workload. However, prior work [22] found HR to be more robust than HRV during driving and a similar secondary task. This motivated our choice to use HR features instead of HRV features.

The skin conductance recordings were filtered using a wavelet transform to remove high frequency noise ([29]). Decomposition at level 4 using Coiflet wavelets with 5 vanishing moments showed best results during our exploratory analysis. Gross low frequency movement artifact was identified by manual inspection and removed.

Steering wheel reversal rates measure the frequency of steering wheel reversals exceeding a certain threshold angle (commonly referred to as gap). This reflects stability of control as distraction can cause quick large corrections. Steering wheel reversal rates were calculated using a 2nd order Butterworth filter as described in [27], and were provided as reversals per minute. Large reversals have gap size 3 and cut-off 0.6Hz whereas small reversals have gap size 0.1 and cut-off 2Hz. After preprocessing, signals were resampled to 10 Hz.

Feature Generation

In both experiments, similar feature generation methods were used for testing classification approaches. A labeled dataset was built from the synchronized sensor data. Data acquired during the cognitive demand task periods are labeled as *elevated* workload. A period is extracted from driving only periods and labeled *normal* workload.

To preserve information about the temporal dynamics, a sliding-window approach was used to aggregate attributes over specific time intervals (Figure 3). Each feature is computed using a fixed-length sliding window operator moving over the data. For each window, a set of features is computed. We take the mean, standard deviation, minimum, maximum and first derivative of the following measures: heart rate, skin conductance level (SCL), and vehicle velocity. In addition, we compute the number of small and large steering wheel reversals.

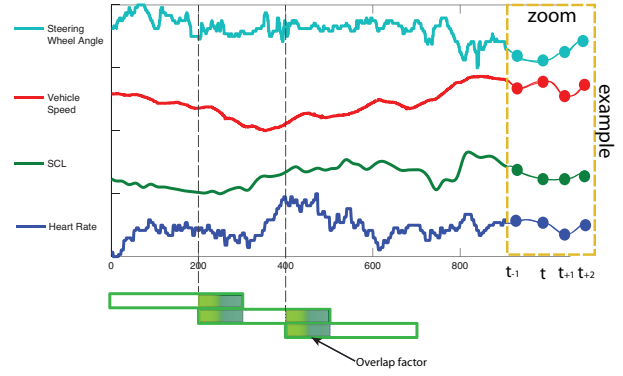


Figure 3. Sequential sensor data can be broken into fixed-length windows (green bars at bottom), which slide across the data. Within each window, we can calculate average, standard deviation, etc. The size of the window and the amount of overlap will affect the analysis.

Two parameters have to be specified to generate aggregate feature vectors: the window length and overlap factor. The window length determines the number of data points per signal to be considered for a single window. The overlap factor effectively determines the time offset between the first data points of two successive windows. While window length influences how much historical information is contained in a single window, the overlap factor influences how much historical information is shared among successive windows. In the results reported below, sliding windows of 10, 15, 20, 25, and 30 seconds are used with overlap percentages of 0%, 25%, 50%, and 75%.

Five feature based learning techniques are used in this analysis: decision trees, logistic regression, 1-nearest neighbor, multilayer perceptron, and naïve Bayes. These were chosen because they are simple learners, they generate explicit interpretable models and they can be implemented as incremental learners (i.e. they can adjust the existing model and do not have to relearn from scratch when confronted with new training data).

EXPERIMENT 1: AUTOMATIC CLASSIFICATION OF ELEVATED WORKLOAD IN INDIVIDUAL DRIVERS

In the first study, we collected data from 20 participants with the goal of examining the feasibility and practical considerations for automatic classification of elevated workload, based on an individual's vehicle and physiological patterns. We wanted to look at building individual models to account for individual differences between drivers.

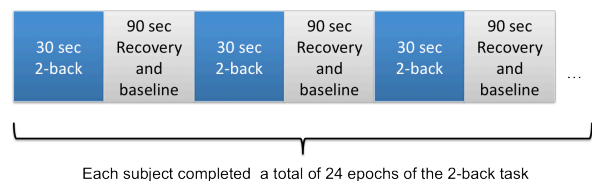


Figure 4. Experimental protocol for Experiment 1.

Participants

Twenty-six individuals driving more than three times a week and having a valid driver's license for at least three years were recruited. Participants had to report a driving record free of accidents for the past year. We collected reliable heart rate data from 20 participants (9 female, mean age of 23.9, standard deviation of 23.0). Due to recording issues, only 13 of the participants (7 female, mean age of 23.2, standard deviation of 2.6) had both reliable heart rate and skin conductance levels.

Procedure

Participants drove in urban traffic for approximately ten minutes before reaching an interstate highway. Subsequently, drivers were provided an additional twenty minutes of interstate driving to familiarize themselves with the vehicle and environment before a two minute single task driving reference period was established. Afterwards, subjects were presented 24 task periods consisting of the 2-back cognitive demand task described above, while they continued to drive on the highway. Each 30-second task period was followed by a ninety second recovery and baseline period (Figure 4). Throughout the study, heart rate, skin conductance level, speed, and steering wheel position were recorded.

Classification Approach

This section describes the signal processing, feature generation and classification approaches that we explored for automatic detection of elevated cognitive workload during the experiment.

We were interested in individual classification methods, and built separate datasets for each of the 13 participants and performed the classification within each set. For each dataset, we have 24 30-second examples of *elevated* cognitive load from the task periods, and 24 30-second examples of *normal* cognitive load extracted from the middle of the recovery and baseline periods when the participants were just driving. After the signal processing and feature generation steps described above, classification algorithms were run.

Ten-fold cross-validation was used for evaluating the approaches. To choose the window size and window overlap that yielded the highest classification accuracy, we performed an inner ten-fold cross-validation process within the training set. Our data was split first into a training and test set. Within that training set, the data was split into a training and validation set to choose the parameters that performed the best. The outer test set, thus, was not used in parameter selection and can provide an estimate of generalizability of the classification accuracy. In all iterations, when data was chosen for a training, validation or test set, the entire two-minute task period (including the n-back task and rest) was included. This is to ensure that data from a task period was not used for choosing parameters, or building the model as well as testing the accuracy of the method.

	All Features		Heart Rate	
	Mean	S.D.	Mean	S.D.
Decision Tree	75.0	10.8	72.8	12.8
Logistic Regression	75.5	10.9	73.9	11.3
Multilayer Perceptron	75.7	10.9	74.0	12.4
Naïve Bayes	75.0	12.5	74.1	11.8
Nearest Neighbor	69.4	11.6	71.5	10.3

Table 2. Mean and standard deviation for classification of elevated cognitive load from normal driving across 13 subjects using all features (20 subjects for heart rate only).

Classification Results

We report results as the average accuracy achieved from each of the datasets, using the nested cross-validation described above. The mean accuracy and standard deviation for each classifier are shown in Table 2. We looked at classification using *heart rate data* only for all 20 subjects and also ran the analysis using *all features* for the 13 subjects with complete data. For *all features*, a one-way analysis of variance shows that there is a significant difference in the accuracy of the five algorithms, ($p < 0.0001$). Tukey-Kramer post-hoc test showed that the nearest-neighbor classifier performed significantly worse than all four of the other algorithms. There were no additional significant results. For *heart rate only* (13 subjects), a one-way analysis of variance showed significant differences ($p < 0.0001$). Tukey-Kramer post-hoc test showed that the nearest-neighbor classifier performed significantly worse than logistic regression, multilayer perceptron and naïve Bayes, but not decision tree. There were no other significant results.

Discussion

The results of this experiment show that we could achieve reasonable classification accuracy, using simple features and classification methods. Even with only the heart rate data, the accuracy did not decrease by much, showing that this simple measure has promise for classifying cognitive workload for in-vehicle user interfaces. One thing to note is that the entire set of 24 trials translates to about 48 minutes of data. Thus, with 10-fold cross-validation, we were training on 90% of this data (or ~43 minutes). This makes sense for experiments and a proof-of-concept. However, this amount of training time for the classifier is not ideal for real-world evaluation. It is likely that future work would reduce this training time, and also improve the classification results. However, it still is not ideal to build individual models. Experiment 2 investigates classification across individuals that may reduce or eliminate this training time.

EXPERIMENT 2: ESTABLISHING METHODS ACROSS INDIVIDUALS

Experiment 2 moves toward having general classifiers that detect elevated cognitive workload without extensive training on individual drivers. We worked with data collected from 99 participants [20,29] with the goal of finding common features and algorithms that reliably can classify cognitive load automatically across individuals.

Age group (years)	Mean (SD)	Females	Males
20-29	24.75 (2.81)	17	18
40-49	44.74 (3.01)	16	16
60-69	63.97 (3.02)	16	16

Table 3. Age and gender of participants in Experiment 2.

Participants

Healthy individuals driving more than three times a week and having held a valid driver's license for at least three years were recruited. As in the previous study, participants had to report a driving record free of accidents for the past year. 146 individuals took active part in the driving portion of the study; however, 38 cases were excluded from the final dataset for reasons such as heavy traffic, poor weather conditions, or technical issues [20]. In addition, nine more subjects were removed from the dataset due to missing data or poor measurement quality in at least one signal domain. The dataset considered in this paper contains recordings from 99 subjects. Aiming for reliability in the real world, participants were balanced in age and gender (Table 3).

Procedure

Participants drove in urban traffic for approximately 10 minutes before reaching an interstate highway, on which they drove for 20 minutes to familiarize themselves with the vehicle and environment prior to a two minute single task driving period which established reference data. Afterwards, subjects were presented three task periods consisting of a series of four secondary task blocks each. In this study, there were periods of the 0-back, 1-back, and 2-back task described earlier. However, in this analysis, we focused on classifying the 2-back *elevated* periods from the single-task *normal* driving, as in Experiment 1. Future work will look at the other levels of demand. Each task period was followed by a two minute recovery period. The order of presentation of the three task difficulty levels was counterbalanced across participants. Figure 6 shows an overview of the experimental protocol.

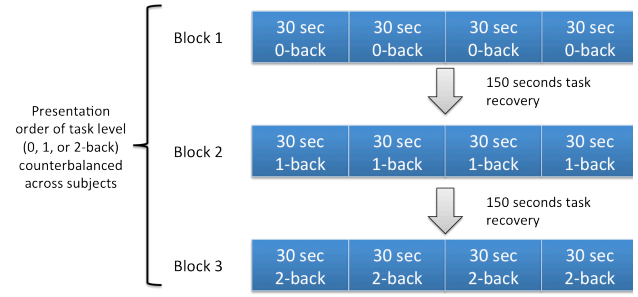


Figure 6. Procedure for Experiment 2.

Exploratory Analysis

The data shows that participants had a significant physiological response when presented and engaged in the secondary task. Figure 5 illustrates the value changes of heart rate during the experiment. Task periods are readily identifiable. Also, the counterbalanced presentation order can be recognized in each plot. The illustration support the hypothesis that cardiovascular measures are generally sensitive to changes in cognitive workload. We saw similar results looking at the skin conductance level and thus there is promise for classifiers that work across individuals.

Classification Approaches

For this experiment, we used similar processing and classification approaches as in the first experiment. However, the data set used included data from 99 participants, and classification was done across individuals. In addition, we have only the set of four consecutive 30-second *elevated* 2-back task periods per individual (24 trials in Experiment 1).

We were most interested in looking at high demand periods, and considered only the 2-back high demand periods and the single task driving periods. We looked at sliding windows of 10, 15, 20, 25, and 30 seconds and overlap factors of 0%, 25%, 50% and 75% to see the effects that these pre-processing parameters have on the classification accuracy. We were also interested in understanding the value of physiology features, driving features and their combination.

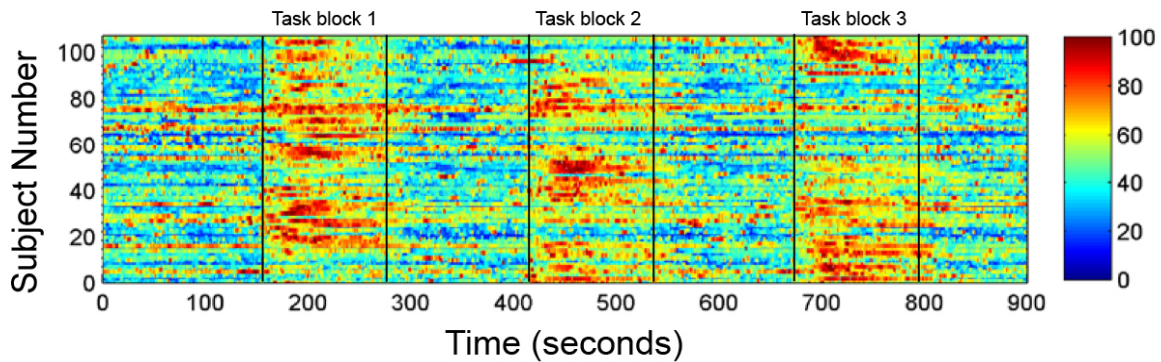


Figure 5. Heart rate change during experiment drive. Each row represents the change in heart rate of a single subject during experiment drive. Within three task block periods, subjects had to perform a secondary cognitive task (three levels of cognitive demand) in addition to the primary driving task. Red indicates maximum heart rate, blue color indicates minimum heart rate.

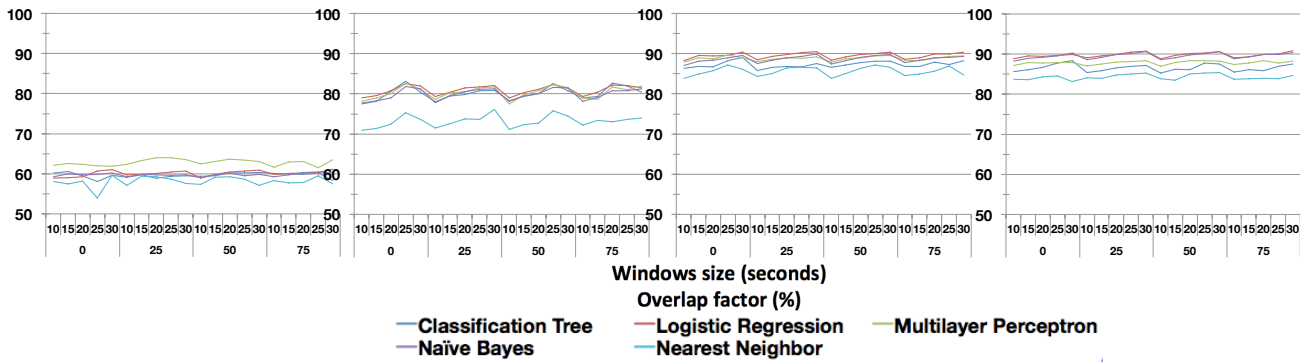


Figure 7. Classification accuracy using (from left to right) 1) driving features only 2) heart rate features only 3) all physiology features 4) physiology and driving features. Accuracy tends to increase with increased window size. The overlap factor had little influence. Nearest neighbor learning algorithm had the lowest classification accuracy, while the other algorithms had similar performance. The best performance was found in 3 and 4, indicating that HR was most sensitive to the cognitive load changes.

As in Experiment 1, ten-fold cross-validation was used for evaluating the approaches. However, we did not choose parameters using nested cross-validation. Instead, we report results from all choices of parameters as well as from five different classification algorithms for exploratory purposes. Folds were created based on individuals, so each fold held out 10% of the participants. Training and test sets never contained data from the same individual.

Classification Results

Our exploratory analysis resulted in a large amount of results (Figure 7). We highlight the main results that are significant for in-vehicle HCI research. Figure 7 shows the average accuracy for each combination of learning algorithm, sliding window size, overlap factor, and feature type.

Comparison of Physiological and Vehicle Features

The type of features had a clear effect on the classification results. When physiology features are ignored and classification algorithms are trained with features generated from *driving performance measures only*, we found the average classification accuracy to be 64% (1st plot in figure). We investigated using only *heart rate features* (mean, min, max, standard deviation, and first derivative), ignoring features extracted for electrodermal activity measures and driving data. With the exception of 1-Nearest Neighbor, all other learning techniques performed reasonably well, achieving an average of 80% accuracy (2nd plot). Using features generated from all available physiology data (heart rate and skin conductance level), logistic regression outperformed other classifiers achieving the highest performance with a 30 second sliding window. Multilayer perceptron and naïve Bayes had a significantly lower performance of 89% accuracy. The best classification performance using features from both physiology and driving performance data is similar to that using only physiology features. Logistic regression and naïve Bayes performed significantly better than all other classifiers.

Window size and window overlap

Figure 7 demonstrates the tradeoff that exists between window size and classification accuracy. Increasing the win-

dow size improved classification in all cases, except when the vehicle telemetry data was used by itself. This is important as the window sizes imply lag if the classification is operating in real time. As one might expect, the curves rise much more steeply as smaller window sizes increase to medium, and then seem to mostly plateau at the larger window sizes. It is interesting to note that the overlap factor doesn't have a significant impact on the classification accuracy.

Classification Algorithm

In the driving only analysis, the multilayer perceptron achieved the highest classification performance. In the analysis of the heart-rate data, the 1-nearest neighbor algorithm performed worse than the other algorithms. For most other analyses, the classifier choice did not make a large difference in the results, showing that feature generation and selection are key to accuracy in this domain.

DISCUSSION AND FUTURE WORK

Through the experiments described above, we examined several considerations for developing an automatic cognitive workload classifier for use in evaluating interfaces in real-world driving. We illustrate the impact that window size, classification algorithm, and training data set have on the robustness of the detection, and provide some guidance on the choice of these parameters. Similar to Grimes, et al. [11], we found a tradeoff between window size and classification accuracy. We did not find significant effects of window overlap. By using large datasets collected in real-world driving, we provide realistic estimates of the results of such systems. In experiment 1, we train models based on ~40 minutes of training data per individual. In experiment 2, we train models across individuals, using only 4 minutes of training data per person.

While our approach was successful in classifying cognitive load with high accuracy, there may be additional measures or algorithms that could also provide similar or improved results. In fact, it is likely that more sensitive driving performance measures, characterization of visual behavior and improved feature sets and algorithms will be developed in the future. In addition, there has been interest in classifying

workload from brain data during driving (e.g. [17]). However, these would all fit in with the framework in Figure 2.

From the results of Experiment 2, it is apparent that classification power in this dataset is coming from physiology features, specifically from heart rate derived features. On the other hand, for this experiment only a handful of driving performance measures were used (velocity and both large and short steering wheel reversals); other metrics may prove valuable for classification.

In Experiment 1, we take into account individual differences by training a classifier for each individual. The approach employed in Experiment 2 however, makes the assumption, that the response to workload is somewhat consistent across all types of drivers. In other words, the models built using the proposed approach do compensate for different physiological or behavioral responses among drivers, but they assume that the structural response pattern is the same for all drivers. In the future, we could move closer to a deployable system where the algorithm is trained on a large dataset, and then is able to classify new system interactions among a different set of drivers when they enter the vehicle. In a hybrid approach, automatic recognition algorithms could be trained on large datasets. Then, a new driver may spend a short time providing additional training data that is specific to the individual. This could fine-tune the model to be customized for the individual that is driving.

While our paper provides comparative results of five classification algorithms, there are additional algorithms and approaches that may achieve higher performance. Future efforts may wish to consider additional modeling approaches. In addition, this work focused on classifying high demand periods from single-task driving. We plan to integrate the 0-back and 1-back data that we collected into our models to get a comprehensive model of physiological changes as demand levels change. Future investigation could also look at classification results when the models trained on the n-back task data are used to classify workload in other, more realistic tasks. It may be useful to compare physiological sensing with cognitive modeling tools such as Distract-R [32] that also can be used for in-vehicle interface evaluation. Finally, while we used a somewhat invasive medical grade EKG, measures of heart rate could be integrated in future cars using steering wheel, seat back, or other sensor sites to provide physiological measures in less invasively.

CONCLUSION

In this paper, we have shown that machine learning techniques can be applied to vehicle sensor data as well as driver physiological sensor data to provide recognition of elevated cognitive load periods. In addition, we report on the results of experiments to investigate specific parameters and approaches. This work was motivated by the need for additional methods for evaluating novel in-vehicle user interfaces to provide effective and safe experiences on the road. Poorly designed in-vehicle user interfaces can lead to distracted and potentially unsafe driving. Thus, the goal of this work was to evaluate the feasibility and practical considerations of physiological workload detection during nat-

ural driving, with a large, balanced group of drivers. We consider this to be a foundation for concrete applications such as in-vehicle user interface evaluation. Most previous work was either done in a simulator or with a small number of participants. In addition, the methods we described would apply to classifying workload in other contexts, such as game user experience evaluation or passive, adaptive user interfaces (similar to passive BCI work [1,9,35,44]), and our sample size is larger than most papers in those areas. This has implications for broader applications for real-time cognitive load assessment and evaluating user interface technology in the wild, beyond driver user interfaces.

ACKNOWLEDGMENTS

Acknowledgment is extended to NSF (NSF Grant#1136996 awarded to Computing Research Association for the CI Fellows Project) and US DOT's Region I New England University Transportation Center at MIT for support. EAGP acknowledges financial support from CONACYT-DAAD grant 308755. MZ was supported by a scholarship from the Max Weber-Program Bayern.

REFERENCES

1. Afergan, D., et al. Dynamic Difficulty Using Brain Metrics of Workload. *Proc. CHI*, (2014).
2. Afonso, V.X., Tompkins, W.J., Nguyen, T.Q., and Luo, S. Multirate processing of the ECG using filter banks. *Computers in Cardiology*, (1996), 245–248.
3. Backs, R.W. et al. Cardiac measures of driver workload during simulated driving with and without visual occlusion. *Human Factors* 45, 4, (2003), 525–38.
4. Christie, M.J. Electrodermal activity in the 1980s: a review. *J Royal Soc of Medicine* 74, 8 (1981), 616–22.
5. Collet, C., et al. Physiological and behavioural changes associated to the management of secondary tasks while driving. *Applied Ergonomics* 40, 6 (2009), 1041–6.
6. Coughlin, J.F., Reimer, B., and Mehler, B. Monitoring, managing, and motivating driver safety and well-being. *IEEE Pervasive Computing* 10, 3 (2011), 14–21.
7. Eldar, Y.C., Oppenheim, A.V. Filter bank interpolation and reconstruction from generalized and recurrent nonuniform samples. *IEEE ICASSP* (2000) 2–5.
8. Engström, J., Johansson, E., Östlund, J. Effects of visual and cognitive load in real and simulated motorway driving. *Transportation Research F*, 2 (2005), 97–120.
9. Girouard, A, Solovey, E.T., Jacob, R.J.K.. Designing a passive brain computer interface using real time classification of functional near-infrared spectroscopy. *IJAACS* 6, no. 1 (2013): 26–44.
10. Gopher, D. and Donchin, E. Workload: An examination of the concept. In K.R. Boof, L. Kaufman and J.P. Thomas, eds., *Handbook of Perception and Human Performance*. (1986), 41–1–41–49.
11. Grimes, D., Tan, D.S., Hudson, S.E., Shenoy, P., and Rao, R.P.N. Feasibility and Pragmatics of Classifying Working Memory Load with an Electroencephalograph. *Proc. CHI*. (2008), 835–844.
12. Haigney, D., Taylor, R., Westerman, S. Concurrent mobile (cellular) phone use and driving performance:

- task demand characteristics and compensatory processes. *Transportation Research F* 3, 3 (2000), 113–121.
13. Hajek, W., et al. Workload-adaptive cruise control – A new generation of advanced driver assistance systems. *Transportation Research Part F* 20, 0 (2013), 108–120.
14. Healey, J.A. and Picard, R.W. Detecting Stress During Real-World Driving Tasks Using Physiological Sensors. *IEEE Transactions on Intelligent Transportation Systems* 6, 2 (2005), 156–166.
15. Jahn, G., et al. Peripheral detection as a workload measure in driving: Effects of traffic complexity and route guidance system use in a driving study. *Transportation Research Part F* 8, 3 (2005), 255–275.
16. Just, M.A., Keller, T.A., and Cynkar, J. A decrease in brain activation associated with driving when listening to someone speak. *Brain research* 1205, (2008), 70–80.
17. Kohlmorgen, Jens, et al. Improving human performance in a real operating environment through real-time mental workload detection. *Toward Brain-Computer Interfacing* (2007), 409–422.
18. Liang, Y.L.Y., Reyes, M.L., and Lee, J.D. Real-Time Detection of Driver Cognitive Distraction Using Support Vector Machines. *IEEE Transactions on Intelligent Transportation Systems* 8, (2007), 340–350.
19. Mehler, B., et al. Impact of Incremental Increases in Cognitive Workload on Physiological Arousal and Performance in Young Adult Drivers. *Transportation Research* 2138, 12 (2009), 6–12.
20. Mehler, B., Reimer, B., and Coughlin, J.F. 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 (2012), 396–412.
21. Mehler, B., Reimer, B., Dusek, J. MIT AgeLab Delayed Digit Recall Task (n-back). (2011) <http://agelab.mit.edu/mit-agelab-delayed-digit-recall-task-white-paper>.
22. Mehler, B., Reimer, B., Wang, Y. A comparison of heart rate and heart rate variability indices in distinguishing single task driving and driving under secondary cognitive workload. *Proc Driving Symposium on Human Factors in Driver Assessment, Training & Vehicle Design*, (2011), 590–597.
23. Michon, J.A. A Critical View of Driver Behavior Models: What Do We Know, What Should We Do? *Human behavior and traffic safety*, (1985), 485–520.
24. Miyaji, M., Danno, M., Kawanaka, H., and Oguri, K. Driver's cognitive distraction detection using AdaBoost on pattern recognition basis. *ICVES*, (2008), 51–56.
25. Mulder, L.J.M. Measurement and analysis methods of heart rate and respiration for use in applied environments. *Biological Psychology* 34, 2-3 (1992), 205–236.
26. National Highway Traffic Safety Admin. *Traffic Safety Facts: Distracted Driving*. Washington, DC, (2011).
27. Östlund, J., et al. *Adaptive Integrated Driver-Vehicle Interface (AIDE): Driving performance assessment - methods and metrics*. Gothenburg, Sweden, (2005).
28. Peterson, F., Jung, C.G. Psycho-Physical Investigations with the Galvanometer and Pneumograph in Normal and Insane Individuals. *Brain* 30, 2 (1907), 153–218.
29. Reimer, B., Mehler, B., Wang, Y., and Coughlin, J.F. A Field Study on the Impact of Variations in Short-Term Memory Demands on Drivers' Visual Attention and Driving Performance Across Three Age Groups. *Human Factors* 54, 3 (2012), 454–468.
30. Reimer, B. and Mehler, B. The impact of cognitive workload on physiological arousal in young adult drivers: a field study and simulation validation. *Ergonomics* 54, (2011), 932–42.
31. Saga, K. Histochemical and immunohistochemical markers for human eccrine and apocrine sweat glands: an aid for histopathologic differentiation of sweat gland tumors. *J of Investigative Dermatology* 6, 1 (2001).
32. Salvucci, D.D., Zuber, M., Beregovaya, E., and Markley, D. Distract-R: rapid prototyping and evaluation of in-vehicle interfaces. *Proc. CHI*. (2005), 581–589.
33. Schneegaß, S., et al. A Data Set of Real World Driving to Assess Driver Workload. *Proc of Auto UI*, (2013).
34. Singleton, W.T. *The Mind at Work*. Cambridge University Press, (1989).
35. Solovey, E.T., et al., Brainput: Enhancing interactive system with streaming fNIRS brain input. *Proc. CHI*. (2012).
36. Son, J., Reimer, B., Mehler, B., et al. Age and cross-cultural comparison of drivers' cognitive workload and performance in simulated urban driving. *Internl J of Automotive Technology* 11, 4 (2010), 533–539.
37. Stanton, N., & Young, M. Vehicle automation and driving performance. *Ergonomics*, 41(7) (1998)1014–1028.
38. Szabo, A., et al. Mental challenge elicits “additional” increases in heart rate during low and moderate intensity cycling. *Intl J Psychophysiology* 17, 3 (1994), 197–204.
39. Veltman, J.A. and Gaillard, A.W. Physiological workload reactions to increasing levels of task difficulty. *Ergonomics* 41, 5 (1998), 656–69.
40. de Waard, et al. Effect of road layout and road environment on driving performance, drivers' physiology and road appreciation. *Ergonomics* 38, 7 (1995), 1395–407.
41. Yerkes, R. M., Dodson, J.D. "The relation of strength of stimulus to rapidity of habit-formation," *J of Comparative Neurology and Psychology* 18, (1908).
42. Young, M.S. Stanton, N.A. Malleable attentional resources theory: a new explanation for the effects of mental underload on performance. *Human Factors* 44, (2002), 365–375.
43. Young, M.S. and Stanton, N.A. Mental Workload. In N.A. Stanton, A. Hedge, K. Brookhuis, E. Salas and H. Hendrick, eds., *Handbook of Human Factors and Ergonomics Methods*. CNC Press, (2004), 1–9.
44. Zander T. O., Kothe C. Towards passive brain–computer interfaces: applying brain–computer interface technology to human–machine systems in general. *J. Neural Eng.* 8, (2011).