Investigating the feasibility of vehicle telemetry data as a means of predicting driver workload

Phillip Taylor, Nathan Griffiths, Abhir Bhalerao,
Zhou Xu, Adam Gelencser, Thomas Popham

August 4, 2016

Abstract

Driving is a safety critical task that requires a high level of attention and workload from the driver. Despite this, people often also perform secondary tasks such as eating or using a mobile phone, which increase workload levels and divert cognitive and physical attention from the primary task of driving. If a vehicle is aware that the driver is currently under high workload, the vehicle functionality can be changed in order to minimize any further demand. Traditionally, workload measurements have been performed using intrusive means such as physiological sensors. Another approach may be to monitor workload online using readily available and robust sensors accessible via the vehicle's Controller Area Network (CAN). In this paper, we present details of the Warwick-JLR Driver Monitoring Dataset (DMD) collected for this purpose, and announce its publication for driver monitoring research. The collection protocol is briefly introduced, followed by statistical analysis of the dataset to describe its structure. Finally, we announce the public release of the dataset, for use in both driver monitoring and data mining research.

Keywords: Driver monitoring, Data collection, EDA, ECG, CAN-bus

1 Introduction

Drivers have limited attentional resources that must be divided between the various driving tasks, including perceiving the driving environment and controlling the vehicle speed and direction (Young & Regan, 2007). These resources are often also allocated to tasks unrelated to driving, such as holding a conversation in the vehicle, using a phone, or choosing a radio station (Stutts, Reinfurt, Staplin, & Rodgeman, 2001). The attention of a driver may also change with the time of day, fatigue, and tiredness. In some cases, the demand for attentional resources can become too high for the driver to handle, causing them to be inattentive, lowering driver performance, and increasing the risk of a crash (Stutts et al., 2001; Regan, 2005; Regan, Hallett, & Gordon, 2011). Understanding how drivers divide their attention to tasks, and when drivers are inattentive, is therefore important to ensure driving safety and to aid the driver in managing their attentional resources.

The attention of a driver can be managed by either increasing their attentional resources, or by decreasing their demand. For instance, the resources available are lowered for a driver that is tired or fatigued, which can be increased by opening a window, cooling the interior of the vehicle, or encouraging the driver to take a break (). The demands of a driver can be reduced, through both design and real time adaptation in the vehicle. Intuitive interfaces with lower complexities, for instance, are less demanding to use than those that are more complicated (Regan et al., 2011). Further, if the vehicle is able to determine the current inattention status of the driver, it may be possible to change the information provided to them or withhold certain event updates entirely ().

The allocation of attentional resources can be considered as driver workload, which describes the impact of tasks on drivers (Mehler, Reimer, & Coughlin, 2012). Workload is usually measured using subjective or physiological measures, or by analysing the performance or behaviour of the driver. Subjective measures require that drivers report their perceived workload demand, either while driving or afterwards. Reporting while driving can itself increase the workload levels (), and other biases can be introduced because of the different perceptions and limited memories of drivers (). To produce continuous measures of workload, physiological parameters and driver performances can be used. Physiological measures include Heart Rate (HR) and Skin Conductance (SC), which both increase during periods of increased workload (Mehler et al., 2012). To gain reliable measurements, however, the equipment is often intrusive and impractical for everyday driving. Other methods measure the driver's head position or eye parameters using a driver facing camera, but these are expensive and can be unreliable in poor light conditions (Reimer, Mehler, Wang, & Coughlin, 2012). A alternative approach, is to use telemetry data to estimate driver performance and behaviour (Mehler et al., 2012).

In this paper, driver workload is investigated using subjective, physiological, and performance based measures in the Warwick-JLR Driver Monitoring Dataset (DMD). Specifically, we induce increased workload in the form of cognitive distraction using the *N*-back task in a study with thirteen participants. Differences are observed in subjective responses, the HR, HR variance (HRV), Electrodermal Response frequency (EDR), SC, and several driving parameters taken from the vehicle telemetry data. We then build predictive models on the telemetry data to output the workload status of the driver, given by the physiological measures.

The remainder of the paper is structured as follows. Related work on driver inattention monitoring is discussed in Section 2. The study is outlined in Section 3, and statistical analysis of the data collected is presented in Section 4. A data mining methodology is proposed in Section 5 for building predictive models for the driver distraction problem. In Section 6 the results of applying this methodology to the data collected are presented. Finally, Section 7 concludes this paper.

2 Driver workload monitoring

Drivers who are inattentive are more at risk of being involved in a crash than those who are not (Stutts et al., 2001). A taxonomy of driver inattention provided by Regan et al. (2011), who divides it into *diverted* (performing tasks unrelated to driving), *restricted* (fatigued or unwell), *misprioritized* (prioritizing unimportant driving tasks above critical tasks), *neglected* (lack of due care because of familiarity to the road environment), and *cursory* (rushed or panicked driving). Dong, Hu, Uchimura, and Murayama (2011) uses describes activities where the required attentional resources of the driver are increased as *distractions*, and *fatigue* for when driver attention is reduced generally. In this paper, we consider distractions and in particular their effects on workload when attentional resources are diverted from the driving task.

Driver activities each require different amounts of attentional resources in varying combinations of *visual*, *auditory*, *physical* or *cognitive* workload (Dong et al., 2011). For example, selecting a radio station induces a combination of cognitive, to think of the station and its frequency, physical to select the station, and visual and auditory feedback to determine whether or not the correct station is selected. In researching driver inattention, secondary tasks are often used to increase the workload in these dimensions. The *N*-back task (Mehler et al., 2012; Reimer et al., 2012), for example, primarily increases the cognitive workload, but the task also requires auditory attention to listen to the stimuli. Other tasks, including the lane change task (McCall, Wipf, Trivedi, & Rao, 2007), interacting with a laptop or tablet (Ersal, Fuller, Tsimhoni, Stein, & Fathy, 2010), memory recall tasks (?, ?) have also been used to increase driver workload.

Isolating and measuring the different kinds of workload is non-trivial, and studies often include a mixture of subjective reports, physiological parameters, or driving performances (Cain, 2007). Subjective measures are derived from questionnaires such as the NASA Task Load Index (TLX) (Hart & Staveland, 1988), and ask drivers to report their estimated workload levels during periods of driving. The TLX asks participants to rate on a scale of 0–20 their performance, effort, frustration, mental, physical and temporal demands. It was developed originally for use in the aviation domain, but has since been applied in many studies including those of driver workload. An alternative to the TLX that was developed specifically for the automotive domain, is the Driver Activity Load Index (DALI) (Pauzie, 2008). This rates the driver experiences in dimensions of attentional effort, task interference, visual, auditory, temporal demands. Drivers are often required to recall their experiences after a study has ended, meaning that drivers may inaccurately report their experiences due to forgetfulness as well as subjectiveness.

Physiological measures include HR, SC, eye blink parameters and pupilometry change during periods of high workload (Mehler et al., 2012; Taylor, Griffiths, & Bhalerao, 2015; Li et al., 2014). They are typically captured from sensors attached to the driver, or from driver-facing cameras. Traditionally physiological sensors such as ECG electrodes are intrusive, and they are not suitable for monitoring driver workload on a daily basis. Recently, however, sports watches and image processing have provided less intrusive methods, although they may be inaccurate in some circumstances (?, ?, ?). Driver-facing video and infra-red cameras can be used to estimate the head position, gaze direction, blink rate and other eye parameters (Zhang, Owechko, & Zhang, 2008; Dong et al., 2011). These have been found useful in several studies measuring workload and fatigue, but are often unreliable in poor light conditions or when the driver wears glasses.

Many studies focus predominantly on physiological measures for workload, and consider relatively few performance measures (Healey & Picard, 2005; Rodrigues, Vieira, Vinhoza, Barros, & Cunha, 2010; Reimer et al., 2012; Mehler et al., 2012; Flores, Armingol, & de la Escalera, 2011; Jo, Lee, Park, Kim, & Kim, 2014). This is likely due to the higher responsiveness of physiological measures to workload, and performances measures extracted from vehicle telemetry data are often used only as secondary inputs (Wollmer et al., 2011). Vehicle telemetry data includes measurements from all devices in the vehicle, and can be recorded via the Controller Area Network (CAN)-bus. Common measures used for driver workload monitoring include features extracted from the steering wheel, vehicle speed, and pedal positions (Wollmer et al., 2011; Mehler et al., 2012). The mean or standard deviation (STD is often extracted from signals over whole distraction or normal driving periods, and are often minutes long. For example, Mehler et al. (2012), present a statistical analysis of mean values of the heart rate, skin conductance level and vehicle speed, and STD of steering wheel reversal rates and gaze dispersion. Although these results show that features of physiological and telemetric signals share a relationship with inattention, they are of little use in a real-time detection system as distraction states change in a matter of seconds. Other authors on the other hand, such as Torkkola, Massey, and Wood (2004); Tango and Botta (2009); Wollmer et al. (2011), present methods that process inputs signals in smaller windows and are more appropriate, but rely on image processing or physiological measures in addition to vehicle telemetry.

3 Collection protocol

The DMD was collected using a Range Rover Sport with automatic transmission on a test track located at Jaguar Land Rover's principal engineering facility at Gaydon, Warwickshire, UK (pictured in Figure 1). The driving environment is representative of a highway and is approximately 3.8 miles long with four lanes and two main straights with two major corners at the end of each. In comparison to public roads it is quiet, as it is used solely by automotive engineers for research and development purposes. During the trial each participant was instructed to drive as if it were a highway, at speeds of around 70mph and changing to an outer lane to overtake when necessary.

The protocol used for each trial is outlined in Table 1, and is similar to that used by Reimer et al. (2012) and Mehler et al. (2012). Upon arrival at the test facility, the driver was first briefed on the study and consent forms are signed. Three point ECG electrodes are then attached on the driver's chest and EDA electrodes to

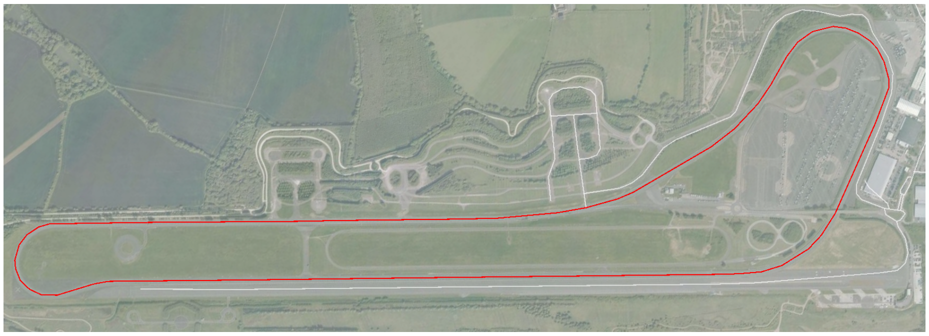


Figure 1: Map of the Gaydon emissions track used for the driver monitoring trials.

	Stage	Mean duration (s)	STD
1	Habituation	1302	269
2	Baseline	280	99
3	0-back (introduction)	10	2
4	0-back	82	9
5	0-back (recovery)	256	59
6	1-back (introduction)	10	2
7	1-back	100	12
8	1-back (recovery)	300	81
9	2-back (introduction)	11	6
10	2-back	113	15
11	2-back (recovery)	294	127

Table 1: The protocol for the WarwickDMD study, employing three *N*-back tests of different difficulties, presented in a random order to each participant.

the underside of the index and middle fingers of their non-dominant hand. The ECG electrodes are positioned closer together than may be usual, as this was found to reduce noise generated through shoulder movement while driving and produce a cleaner signal. Gel electrodes with adhesive pads were used for both the ECG and EDA sensors. The EDA electrodes were secured further using surgical tape, while still ensuring the driver was comfortable and had full movement of their fingers.

The driver was then seated and instructed to make themselves comfortable in the driving position, by adjusting the seat, steering wheel and mirrors. The electrodes were connected to a GTEC USB biosignal amplifier (USBamp)[1], which resides in the rear of the vehicle. The wires from the ECG electrodes came out of the top of the participant's shirt and over the seat, while the EDA wires were positioned to the side of the non-dominant hand. The USBamp was then connected to a laptop, and data recording was commenced at 256Hz using MathWorks Simulink[2].

As well as the physiological data, over 1000 signals were recorded from the vehicle's CAN-bus at a sample rate of 20Hz, using a data logging system located under the passenger seat. Many of these signals are

[1] http://www.gtec.at/Products/Hardware-and-Accessories/g.USBamp-Specs-Features

[2] http://uk.mathworks.com/products/simulink/

Stimulus	1	5	9	3	0	2	3	3	2	9	&	&
0-back	1	5	9	3	0	2	3	3	2	9		
1-back	-	1	5	9	3	0	2	3	3	2	9	
2-back	-	-	1	5	9	3	0	2	3	3	2	9

Table 2: Example of the *N*-back test with a block of 10 numbers. In place of "&" the word "and" is said by the experimenter, requiring the participant to provide a response. Where there is a "-" no response is required by the participant.

unlikely to be of relevance to driver workload, such as the window wiper speeds or air conditioning controls. The full set of relevant signals is not known, however, so we recorded the full set during the study and propose to use feature selection techniques to filter them afterwards (Kohavi & John, 1997; Guyon & Elisseeff, 2003; Hermana, Zhanga, Wanga, Yec, & Chena, 2013).

To induce increased workload, a secondary distraction task in the form of an *N*-back test was used. The *N*-back test requires that the participant repeats digits provided to them with delays of increasing difficulty. Delays of 0, 1 and 2 were used, referred to as the 0-, 1- and 2-back tests respectively, which have been shown to have increasing impacts on physiology and driving style (Mehler et al., 2012; Reimer et al., 2012). An example block of 10 digits is shown in Table 2, with expected responses for the 0-, 1- and 2-back tests. In the 0-back test, the participant is required to repeat digits back directly, whereas digits are repeated with delays in the 1- and 2-back tasks. Before starting the trial, the participant must show a minimum proficiency of 8 out of 10 correct responses for two consecutive blocks of each task.

When the driver is comfortable, data is being recorded successfully, and the driver has displayed the minimum proficiency in the secondary tasks, the vehicle is driven onto the track and the trial protocol listed in Table 1 is performed. A habituation period of driving under normal conditions is used (stage 1), to familiarise the participant with the vehicle and track environment. Once the habituation period is completed a baseline period of normal driving is recorded (stage 2). After this reference period, between stages 3–11, the protocol alternates between *N*-back tests and recovery periods of normal driving. Each participant undergoes each of the 0-, 1- and 2-back tests in a random order. In each of the *N*-back test stages, 4 blocks of 10 digits were presented to the participants, with a pause of 2 seconds between digits and 5 seconds between each block. Before each *N*-back task, a brief explanation and reminder of it is provided (stages 3, 6 and 9), taking around 10 seconds. After each *N*-back test there was a recovery period of normal driving, with no secondary task. The protocol ends with a recovery period, after which the vehicle is taken off the track and data recording is ended.

In each of the *N*-back tests, all of the digits were repeated regardless of the shift (in contrast to Mehler et al. (2012)). This meant that the 1-back task was in effect one digit longer and the 2-back task was two digits longer than the 0-back task, which is reflected in the mean durations shown in Table 1. Other variances in durations were due both to safety concerns, recording quality or human variations. Some events on the road such as low flying birds or overtaking vehicles, for example, led to a pause in the protocol or the extension of a stage due to safety concerns.

From both the ECG and EDA physiological data streams collected, two measures were extracted. The ECG signal provides both HR, from the frequency of *R*-peaks (highlighted by the red dots in Figure 2), and HRV from their variation computed using the Standard Deviation of Successive Difference (SDSD) method (Mehler, Reimer, & Wang, 2011). Both the HR and HRV are computed over 15 seconds of data in a sliding window. The SC is derived from the absolute value of the EDA signal, and the frequency of spikes in the EDA signal (highlighted by the red dots in Figure 3), provides the EDR measure. The EDA sensor readings for each participant are different, and the sensor requires individual configuration to ensure

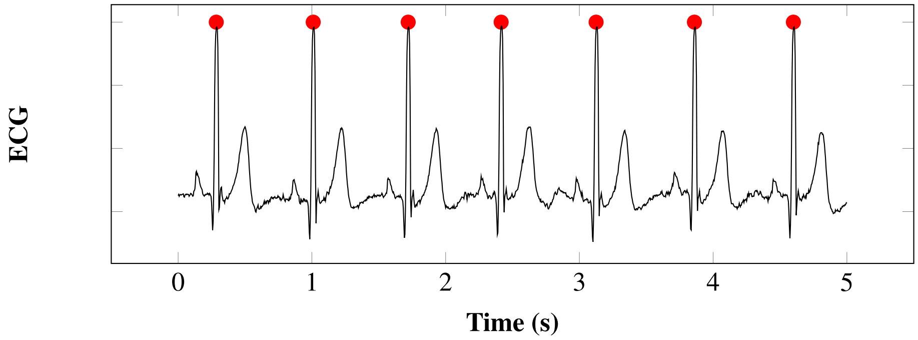


Figure 2: Five seconds of an ECG signal recorded during driving. The dots highlight the *R*-peaks, which can be used to compute the HR and HRV.

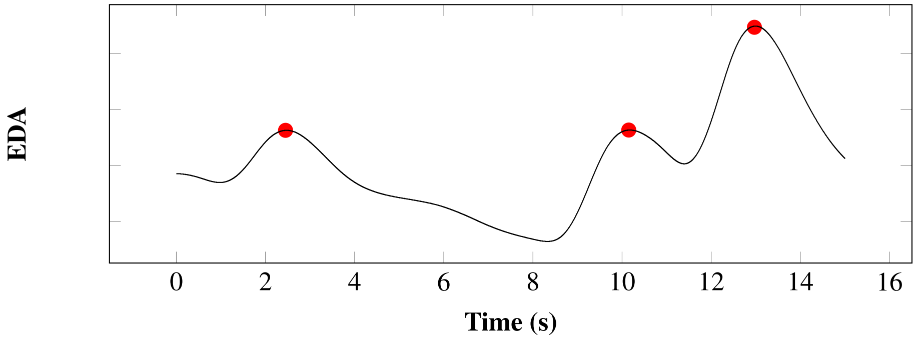


Figure 3: Fifteen seconds of an EDA signal recorded during driving. The dots highlight EDR, which increase in frequency under workload. The SC is given by the absolute value of the signal.

the signal is within a measurable range. The SC values derived from it cannot therefore be directly compared across the different drivers in the trial, and was normalized between 0 and 1 for each participant based on the measurements during the baseline period.

4 Statistical analysis

In this section, the dataset is analysed statistically for any significant findings. First, the task performance and subjective ratings are inspected, followed by statistical analysis of the data streams recorded.

4.1 Task performance and subjective ratings

The error rates for the digit recall tasks are shown in Figure 4. The number of incorrect responses overall was less than 2.45%, and the majority of participants made no errors. The number of errors increased in a linear relationship with the task difficulty, with participants making the most errors for the 2-back test. In some cases of the 2-back test the participant stopped responding for a block of numbers, and the remainder of the block was counted as incorrect responses. In some other cases, that were also counted as errors, the participant responded in the 2-back test as if it were the 1-back test.

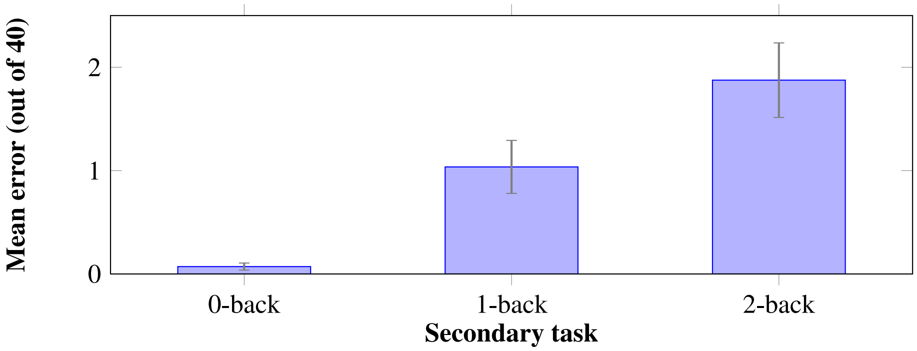


Figure 4: Mean error rates (out of 40 recalled digits) of participants for each of the secondary tasks. Error bars represent the standard error.

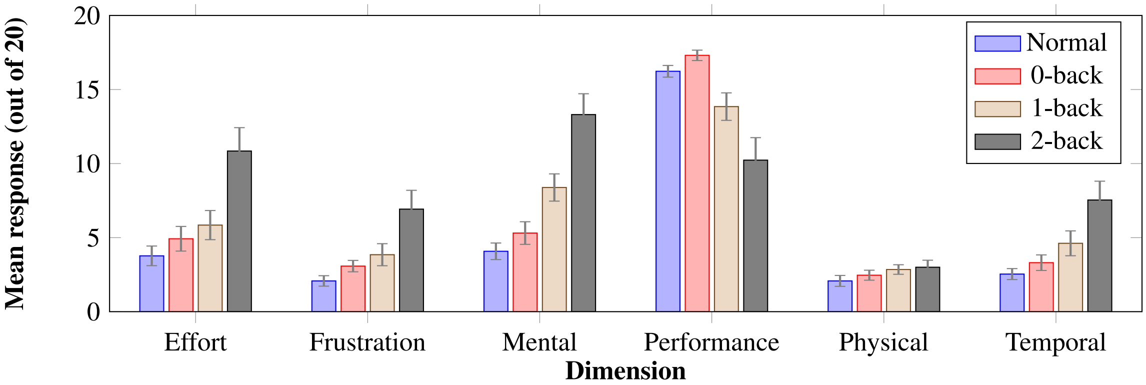


Figure 5: Mean responses to NASA TLX questions. Error bars represent the standard error.

Signal	p	**N × D**	**N × 0**	**N × 1**	**N × 2**	**0 × 1**	**0 × 2**	**1 × 2**
HR	**0.031**	**0.006**	1.000	0.422	**0.050**	1.000	1.000	1.000
HRV	0.554	0.283	1.000	1.000	0.996	1.000	1.000	1.000
SC	**0.000**	**0.000**	1.000	**0.003**	**0.005**	0.452	0.537	1.000
EDR	**0.034**	**0.004**	0.605	0.265	0.122	1.000	1.000	1.000

Table 3: p-values from two way t-test and ANOVA for physiological data streams. In the heading **N** represents periods of normal driving, **0**, **1**, and **2** represents periods of the 0-, 1- and 2-back tests respectfully, and **D** is periods where any of the N-back tasks were being performed.

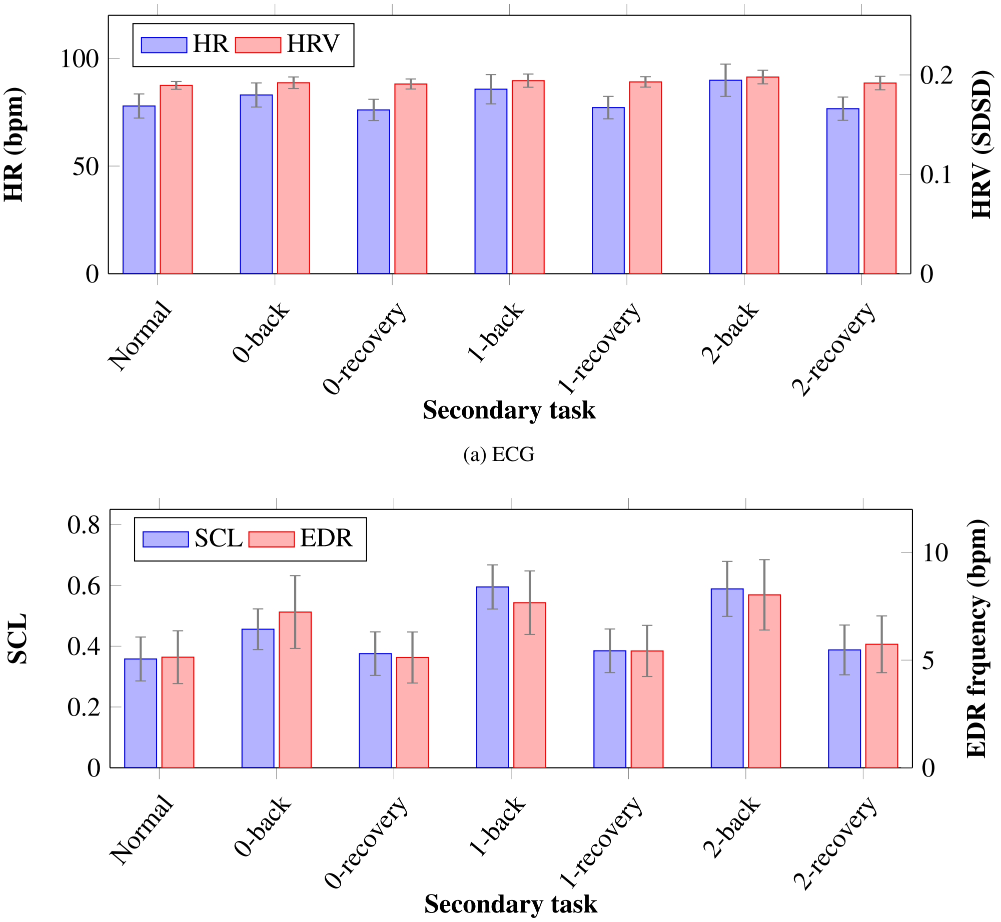
After the trial, when the vehicle was stationary, the participants were asked to fill in four NASA-Task Load Index (TLX) (Hart & Staveland, 1988) questionnaires – one for normal driving and for each of the N-back tests. The TLX asks participants to score their experiences out of 20 in 6 dimensions, namely: mental, physical, and temporal demand, performance, effort, and frustration. The mean responses to the TLX questionnaires are shown in Figure 5, which indicate that the perceived levels of workload increased with the task difficulty. The largest changes were reported for the mental demand and effort dimensions, which was to be expected because of the cognitive nature of the secondary tasks used. The estimated performances decreased with the 1- and 2-back tasks, while reported performance increased on average for the 0-back test over normal driving.

4.2 Analysis of data streams

Results of statistical analyses of the physiological measures are shown in Table 3, comparing normal and distracted conditions in two ways to detail properties of the dataset. First, a two-way t-test is used to compare the mean value of the measures over all subjects during normal driving periods (baseline or recovery) and distracted periods (with secondary tasks). Second, an Analysis of Variance (ANOVA) is performed to determine if there was a significant difference in means during individual secondary task periods. In follow-up to this a four way pairwise t-test, normalized by the Bonferroni correction, was performed. All results in this table produced p-values of less than 0.1 in at least one of the t-test and the ANOVA and any p-value smaller than 0.05 is highlighted in bold. The authors accept that conclusions made from this analysis are limited because it is a multiple comparisons procedure, but a two-way ANOVA, including all recorded signals is impractical due to their number.

The two way t-test showed a significant difference in the HR, SC, and EDR with $p < 0.01$, and the ANOVA produced a significant difference between at least one of the baseline or task periods ($p < 0.05$), shown in Table 3. The change in HR during the 2-back task from the normal driving periods was significant ($p < 0.05$), and the change in SC was significant for both the 1-back and 2-back tasks ($p < 0.01$). Figure 6 shows the mean values of the four physiological measures taken from the (a) ECG and (b) EDA signals, computed over all participants for the baseline, task, and recovery periods. The results reflect the statistical analysis and show that each of the HR, SC, and EDR frequency increased during the task periods, and decreased to the baseline levels during the recovery periods. The HRV did not change during the different trial periods, and had no significant difference in any of the statistical tests.

Table 4 shows the same t-tests and ANOVA for the representative signals of the vehicle telemetry data. As well as the raw signal values, the standard deviation (STD) was computed for each signal over a one second sliding window (See Section 5 for details). This produces a feature of the signals where sample values are equal to the STD of the twenty samples before and after the respective sample in the signal. Signals directly related to the driving controls, such as the pedals and steering wheel were expected to have a



(a) ECG

(b) EDA

Figure 6: Mean values of (a) HR and HRV and (b) SC and EDR frequency over all subjects for the different periods of the trial. Each recovery period is presented separately and error bars represent the standard error.

Signal	**F**	*p*	**N × D**	**N × 0**	**N × 1**	**N × 2**	**0 × 1**	**0 × 2**	**1 × 2**
ACCC	STD	0.232	0.056	1.000	0.419	1.000	1.000	1.000	1.000
Brake on	STD	0.239	0.057	1.000	0.436	1.000	1.000	1.000	1.000
Engine Speed	raw	0.237	0.063	0.414	1.000	1.000	1.000	1.000	1.000
Engine Torque	raw	0.053	**0.016**	0.067	1.000	0.672	1.000	1.000	1.000
Engine Coolant	STD	0.190	**0.036**	0.362	1.000	1.000	1.000	1.000	1.000
Gear Selected	raw	0.085	**0.012**	0.207	1.000	0.556	1.000	1.000	1.000
SWA Momentum	STD	**0.003**	0.066	1.000	0.087	0.055	**0.030**	**0.020**	1.000
SWA	STD	**0.024**	0.471	0.555	0.423	0.968	**0.039**	0.095	1.000
Suspension Height	STD	0.091	0.213	0.089	1.000	1.000	0.527	0.228	1.000
Throttle Position	raw	**0.044**	**0.010**	0.068	1.000	0.473	1.000	1.000	1.000
Yaw Rate	STD	**0.022**	0.532	0.422	0.679	0.715	**0.048**	0.051	1.000
...									

Table 4: *p*-values from two way *t*-test and ANOVA for selected signals of the vehicle telemetry. In the heading **N** represents periods of normal driving, **0**, **1**, and **2** represents periods of the 0-, 1- and 2-back tests respectfully, and **D** is periods where any of the *N*-back tasks were being performed.

strong relationship to driver workload. The analysis confirms this, with the throttle position and STD of SWA (Steering Wheel Angle) momentum changing significantly between the driving periods ($p < 0.05$ in both the two way *t*-test and ANOVA).

Signals with indirect relationships to the vehicle controls, such as suspension movements and yaw rate, were expected to have weaker relationships to the driving conditions. These had larger *p*-values in general than measures of the vehicle controls, but the engine speed and the gear selected by the automatic transmission, had relationships more similar to those of the vehicle controls. Other signals that have no obvious link to the driver were of course expected to have large *p*-values, and for the majority this was the case. A small number, including Adaptive Cruise Control Cancel (ACCC), engine coolant temperature, and others redacted from Table 3, had small *p*-values for the two way *t*-test and can only be explained by chance.

5 Data Mining of the DMD

Data mining aims to discover patterns and build models from data, and has been successfully applied in several disciplines from market research to weather and environment prediction (Witten, Frank, & Hall, 2011; Aggarwal, 2013). Vehicle telemetry mining in the automotive domain has been applied in various domains, including safety improvement, fault detection, and efficiency gains (Crossman, Guo, Murphey, & Cardillo, 2003; Murphey, Crossman, Chen, & Cardillo, 2003; Murphey et al., 2008; Kruse, Steinbrecher, & Moewes, 2010; X. Huang, Tan, & He, 2011). Here, we apply a data mining methodology to build predictive models for driver workload. The methodology is based on the general data mining process described by John (1997), and is outlined in Figure 7.

The data mining process commences with creating a database to describe a problem such a driver workload estimation. In this case, the problem definition is “to estimate driver workload from vehicle telemetry data”. The data mining methodology then aims to produce models that are capable of predicting driver workload using inputs of vehicle telemetry data from the CAN-bus. In Section 3 the protocol for data collection was outlined, by which data relating to driver workload estimation was collected. Next, the data is analysed to ensure it was collected correctly and that it describes the driver workload problem, as presented in Section 4.

To frame driver workload estimation as a binary classification task, a labelling from each of the physiological measures (HR, HRV, EDR and SC) was generated for each driver. The mean value of the measures

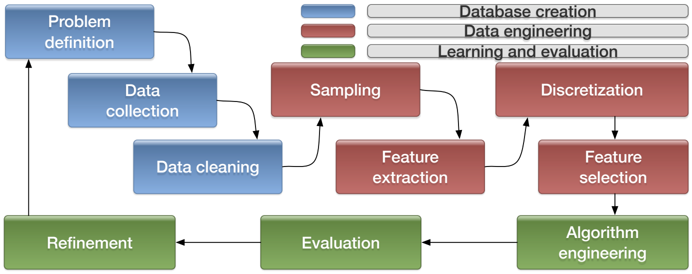
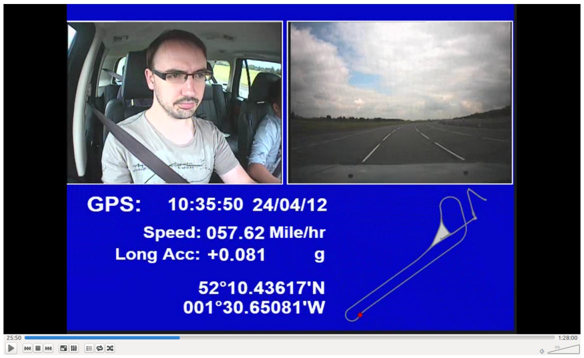


Figure 7: Diagram showing the data mining process.



GPS: 10:35:50 24/04/12
Speed: 057.62 Mile/hr
Long Acc: +0.081 g
52°10.43617'N
001°30.65081'W

Figure 8: Screen shot of the video recorded during the trials, with driver and forward facing cameras, as well as GPS details overlaid.

during the baseline period was taken to be representative of normal driving, and values close to this were labelled as 0. Values more than one standard deviation (STD) for HR, HRV and EDR and 0.25 STDs for SC (again computed over the baseline period) were labelled as 1, to signify periods where the physiology of driver changed. A final labelling was applied using the timings of tasks taken from video streams (illustrated in Figure 8), which was synchronized via the GPS times also present in the CAN-bus. When the driver was under a normal driving scenario with no secondary task the label was 0, and during tasks the label was 1.

5.1 Temporal feature extraction

In vehicle telemetry mining, it is advantageous to include trend information about signals (Antunes & Oliveira, 2001; Tango & Botta, 2009; Wollmer et al., 2011; Aggarwal, 2013). We refer to this process of incorporating historical information into the current sample as *temporal feature extraction*, although in some literature it is referred to as motif extraction (Aggarwal, 2013). A feature, $f(\cdot)$, such as the mean or STD is extracted from a signal, S, at time t, with a window length of l,

$$f(s_t, s_{t-1}, \dots, s_{t-l+1}) = f(S_{t,l}), \tag{1}$$

where $f(S_{t,l})$ is the temporal summary of S between times t and $t - l + 1$. If $t < l$, because it is at the beginning of the recorded signal, t samples are used in extracting the feature. This is performed for all values of t, ensuring that a signal with n samples produces a feature that contains n samples also to line up with the target variable, Y.

Type	Feature
Statistical	Min, Max, Mean, Standard deviation, Entropy, Fluctuation.
Structural	Raw value, First, Second and Third derivatives, First 5 and Max 5 DFT coefficient magnitudes, Max 5 DFT coefficient frequencies, Convexity, Gradient direction, Integral, and Absolute integral.

Table 5: List of features extracted from each signal from the DMD over sliding temporal windows of sizes 0.5s, 1s, 2.5s and 5s.

From each signal in the vehicle telemetry we extract the features listed in Table 5 over various temporal windows to produce a set of features to learn from. We believe that these features, extracted over these window lengths, are diverse enough to capture important historical information from the signals to be used in models. There are other features that could be extracted, or features can be extracted using automated and supervised methods (Guo, Jack, & Nandi, 2005; Mierswa & Morik, 2005; Hamel & Eck, 2010).

5.2 Feature selection

The DMD contains numerous features that are either irrelevant to the predictive task or redundant to others. Both irrelevance and redundancy in a feature set have a negative effect on the performance and complexity of models built on data (Kohavi & John, 1997; Guyon & Elisseeff, 2003; Hermana et al., 2013). The door lock status is unlikely provide any insight into the workload level of the driver, for example, and engine speed is highly redundant to the vehicle speed. For simplicity, the features in the DMD were reduced by hand to the 33 listed in Table **??**. Of course, this includes the features extracted from each of the signals and therefore there are a total of 3828 features that could potentially be used to build a model.

We apply supervised feature selection to choose the features from this full set, and in particular we use Symmetric Uncertainty (SU) (Witten et al., 2011) and minimal Redundancy Maximal Relevancy (mRMR) selection (Peng, Long, & Ding, 2005; Hermana et al., 2013; Taylor et al., 2014). SU is a variant of Mutual Information (MI), that is normalized by the mean entropy of the two variables to mitigate the bias MI has towards features of high dimensionalities. MI is defined as,

$$MI(X,Y) = \sum_{v_1 \in vals(X)} \sum_{v_2 \in vals(Y)} p(v_1, v_2) \log_2 \frac{p(v_1, v_2)}{p(v_1)p(v_2)}, \tag{2}$$

where $vals(Y)$ is the set of values of Y, $p(v_1, v_2)$ is the join probability distribution of X and Y, and $p(v)$ is the marginal probability distribution. Entropy of a variable, X, is

$$H(X) = \sum_{v \in vals(X)} p(v) \log_2 p(v), \tag{3}$$

and the SU between two variables is then,

$$SU(X,Y) = 2\frac{MI(X,Y)}{H(X) + H(Y)}, \tag{4}$$

Minimal Redundancy Maximal Relevancy (mRMR) is a selection algorithm that iteratively selects features that are least redundant to already selected features and most relevant to the labels in each step. The redundancy of a prospective feature is calculated as the mean SU with already selected features, and the relevancy as the SU with the labels. The feature that maximises the difference between the relevancy and redundancy is then chosen, and added to the selected feature set. This is repeated until a given number of features is chosen from the set. In this paper, as in Taylor et al. (2014), we first select one extracted feature from each signal, before combining them in a second stage of selection. In the classification evaluations first fifteen selected features are used.

To apply information based approaches such as MI or SU to numeric or continuous data the probability density functions of the variables must be estimated and integrated (Kwak & Choi, 2002; Sun, Wang, Zhang, & Zhao, 2010). In general this is non-trivial, so we discretize the features before selection so that the probabilities can be computed easily (Fayyad & Irani, 1993; Hermana et al., 2013). Possibly the most commonly used discretization method, and the one applied to data in this paper, is the Minimum Description Length (MDL) method (Fayyad & Irani, 1993). MDL recursively splits the variable domain into multiple discrete levels while maximizing the information gain of each cut point (Fayyad & Irani, 1993). Other methods can be used to estimate entropies of continuous variables, such as Parzen windows, but these require the selection of parameters that cannot be determined easily from the data (Hermana et al., 2013).

5.3 Evaluation

In an evaluation, a learning approach must be applied on training data to build a model that is then used to make predictions for testing samples, which we refer to as a *train-test cycle*. A train-test cycle should be performed several times with different training and testing samples to produce a more robust performance estimate (Hand, Mannila, & Smyth, 2001; Witten et al., 2011; Japkowicz & Shah, 2011). Here, we split the data into four sections by the tasks being performed by the drivers, one for the baseline period (of normal driving), and one for each of the task difficulties (0-, 1, and 2-back tests) and their associated recovery periods. Each train-test cycle is made up from a different combination of these sections, with data from the baseline period always being in the training data. For example, the training data in one train-test cycle contains the data from the baseline period, as well as the data from the 0-back test and 0-back recovery periods of the drivers. The testing data in this cycle is then made of the remainder of the data, from the 1- and 2-back tests and the recovery periods following them.

To estimate the performance of model, the predictions it makes for testing samples are compared to the labels (Japkowicz & Shah, 2011; Jensen & Cohen, 2000; Witten et al., 2011). The success rate, or accuracy, describes the proportion of correct predictions the model made, but is not a reliable measure in domains with class imbalance such as driver workload monitoring. The Area Under the Receiver Operator Characteristic Curve (AUC) accounts for this issue by considering class distributions, and has been adopted by many researchers in numerous domains (J. Huang & Ling, 2005; Japkowicz & Shah, 2011).

6 Classification results

In an initial set of classification simulations models were built using data from individual drivers to predict, each of the labellings (Task, HR, HRV, EDR, and SC). In each case, three train-test cycles were used with training data made of the baseline period and two tasks with their associated recovery periods. The testing data contained samples from the third task and recovery periods. All predictions from the testing phase were combined to produce overall AUC values for each evaluation, shown in Table **??**. The AUC performances in almost all cases were poor, and is possibly a result of over-fitting. Models over-fit when they learn patterns in the training data that do not describe the underlying concepts, that are often considered as noise. Having said this, for some drivers and some labellings the AUC performances were higher. This shows that, some measures of workload can be predicted from vehicle telemetry for some drivers.

These low performances may also as a result of the vehicle telemetry data not in fact being related to

	Task	HR	HRV	EDR	SC
1	0.523	0.552	0.573	0.548	0.518
2	0.563	0.380	0.579	0.534	0.437
3	0.596	0.501	0.558	0.182	0.556
4	0.562	0.580	0.510	0.650	0.457
5	0.410	0.320	0.348	0.484	0.565
6	0.509	0.317	0.000	0.436	0.422
7	0.512	0.433	0.466	0.736	0.560
8	0.559	0.595	0.439	0.327	0.565
9	0.416	0.351	0.650	0.465	0.613
10	0.481	0.461	0.431	0.441	0.503
11	0.466	0.586	0.132	0.669	0.561
12	0.507	0.395	0.528	0.459	0.589
13	0.429	0.278	0.649	0.852	0.569

Table 6: AUC performances of models built with data from baseline period and two tasks and tested on the other

	Task		HR		HRV		EDR		SC	
1	0.822	(0.007)	0.793	(0.012)	0.764	(0.064)	0.823	(0.021)	0.725	(0.015)
2	0.764	(0.003)	0.741	(0.005)	0.779	(0.006)	0.775	(0.007)	0.675	(0.004)
3	0.729	(0.001)	0.716	(0.002)	0.742	(0.002)	0.745	(0.003)	0.653	(0.001)
4	0.706	(0.001)	0.698	(0.001)	0.717	(0.001)	0.724	(0.002)	0.638	(0.001)
5	0.690	(0.001)	0.686	(0.001)	0.700	(0.001)	0.710	(0.001)	0.628	(0.000)
13	0.618	N/A	0.639	N/A	0.649	N/A	0.670	N/A	0.597	N/A

Table 7: AUC performances for models built and evaluated on all the data of different numbers of drivers.

driver workload. To investigate this, models were built using training data made of the baseline period and the three task and recovery preiods, from different subset combinations of the drivers. The models were then used to make predictions for samples in the training data, and AUC performances were computed. The mean AUC performances for models built with data from between one and five, and all thirteen, drivers are shown in Table 7. In general the AUC performances were much higher when the training and testing data were the same, showing that there is some correlation between the vehicle telemtry data and driver workload. For each of the five labellings the AUC performance decreased as data from more drivers was included. This again indicates that different drivers are affected differently by increased workload, and models for driver workload monitoring should be driver specific.

7 Conclusions

In this paper we have further presented and analysed the DMD. The collection protocol was described in detail and statistical analysis of the physiological measures and some vehicle telemetry signals was presented. The dataset produced, including the physiological and vehicle telemetry data is available via www.dcs.warwick.ac.uk/dmd/ in a comma separated variable (csv) format. The vehicle telemetry data is sampled at 20Hz, while the physiological data is sampled at 256Hz.

The statistical analyses showed that the physiological measures were affected significantly by increased workload. These results replicated findings by (Mehler et al., 2012) and (Reimer et al., 2012), who also found a linear increase in HR over the three *N*-back task difficulties. The changes in EDR and SC were less significant between the three task difficulties, but the change from a normal driving situation were more significant. HRV was not found to be a good indicator of workload for any of the task difficulties. Some signals in the vehicle telemetry data showed similar changes to those of the physiological measures. In particular, signals relating to the SWA and throttle changed significantly during periods with secondary tasks. This is to be expected, as these are the signals most closely related to the driver.

In classification evaluations, the majority of AUC performances were low when the training and testing datasets were taken from different task periods. This was likely as a result of model over-fitting, and models trained and tested on the full datasets had higher performances. This over-fitting is possibly caused due to the vehicle telemetry data not being a good indicator of driver workload. For instance, the data may not properly describe the underlying concepts relating to driver workload, and the models fit to noise unrelated to the problem.

Through this research we conclude that using vehicle telemetry may not be the most appropriate source of information for real-time driver monitoring. Instead, visual based approaches may be used as non-intrusive measures of workload and distraction. There are issues with these also, as gathering images in poor light conditions leads to poor performance in their analysis. Further, eye gaze analysis and pupilometry is non-trivial when drivers wear glasses.

Acknowledgements

We are grateful to EPSRC and Jaguar Land Rover (JLR) for funding this work, and to JLR for supporting the vehicle trials and data collection.

References

Aggarwal, C. (2013). *Managing and mining sensor data*. Boston, MA: Springer.

Antunes, C., & Oliveira, A. (2001, August). Temporal data mining: An overview. In *Kdd workshop on temporal data mining* (pp. 1–13). ACM New York, NY.

Cain, B. (2007, July). *A review of the mental workload literature* (Technical Report). Defence Research and Development Toronto (Canada).

Crossman, J., Guo, H., Murphey, Y., & Cardillo, J. (2003, July). Automotive signal fault diagnostics - part I: Signal fault analysis, signal segmentation, feature extraction and quasi-optimal feature selection. *IEEE Transactions on Vehicular Technology*, *52*(4), 1063–1075.

Dong, Y., Hu, Z., Uchimura, K., & Murayama, N. (2011, June). Driver inattention monitoring system for intelligent vehicles: A review. *IEEE Transactions on Intelligent Transportation Systems*, *12*(2), 596–614.

Ersal, T., Fuller, H., Tsimhoni, O., Stein, J., & Fathy, H. (2010, September). Model-based analysis and classification of driver distraction under secondary tasks. *IEEE Transactions on Intelligent Transportation Systems*, *11*(3), 692–701.

Fayyad, U., & Irani, K. (1993, September). Multi-interval discretization of continuous valued attributes for classification learning. In *International joint conference on articial intelligence* (Vol. 2, pp. 1022–1027).

Flores, M., Armingol, J., & de la Escalera, A. (2011, December). Driver drowsiness detection system under infrared illumination for an intelligent vehicle. *IET Intelligent Transport Systems*, *5*(4), 241–251.

Guo, H., Jack, L., & Nandi, A. (2005, February). Feature generation using genetic programming with application to fault classification. *IEEE Transactions on Systems, Man and Cybernetics, Part B: Cybernetics*, *35*(1), 89–99.

Guyon, I., & Elisseeff, A. (2003, March). An introduction to variable and feature selection. *Journal of Machine Learning Research*, *3*, 1157–1182.

Hamel, P., & Eck, D. (2010, August). Learning features from music audio with deep belief networks. In *The international society of music information retrieval* (pp. 339–344).

Hand, D., Mannila, H., & Smyth, P. (2001). *Principles of data mining* (1st ed., Vol. 1). Cambridge, MA: The MIT Press.

Hart, S., & Staveland, L. (1988). Development of NASA-TLX (task load index): Results of empirical and theoretical research. In P. Hancock & N. Meshkati (Eds.), *Human mental workload* (Vol. 52, pp. 139–183). North-Holland.

Healey, J., & Picard, R. (2005, June). Detecting stress during real-world driving tasks using physiological sensors. *IEEE Transactions on Intelligent Transportation Systems*, *6*(2), 156–166.

Hermana, G., Zhanga, B., Wanga, Y., Yec, G., & Chena, F. (2013, December). Mutual information-based method for selecting informative feature sets. *Pattern Recognition*, *46*(12), 3315–3327.

Huang, J., & Ling, C. (2005, March). Using AUC and accuracy in evaluating learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, *17*(3), 299–310.

Huang, X., Tan, Y., & He, X. (2011, June). An intelligent multifeature statistical approach for the discrimination of driving conditions of a hybrid electric vehicle. *IEEE Transactions on Intelligent Transportation Systems*, *12*(2), 453–465.

Japkowicz, N., & Shah, M. (2011). *Evaluating learning algorithms: A classification perspective*. New York, NY: Cambridge University Press.

Jensen, D., & Cohen, P. (2000, March). Multiple comparisons in induction algorithms. *Machine Learning*, *38*(3), 309–338.

Jo, J., Lee, S., Park, K., Kim, I.-J., & Kim, J. (2014, March). Detecting driver drowsiness using feature-level fusion and user-specific classification. *Expert Systems with Applications*, *41*(4), 1139–1152.

John, G. (1997). *Enhancements to the data mining process* (Unpublished doctoral dissertation). stanford university, Stanford, CA.

Kohavi, R., & John, G. (1997, December). Wrappers for feature subset selection. *Artificial Intelligence*, *97*(1-2), 273–324.

Kruse, R., Steinbrecher, M., & Moewes, C. (2010, March). Data mining applications in the automotive industry. In M. Beer, R. Muhanna, & R. Mullen (Eds.), *International workshop on reliable engineering computing* (pp. 23–40). Research Publishing Services.

Kwak, N., & Choi, C.-H. (2002, December). Input feature selection by mutual information based on Parzen window. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*(12), 1667–1671.

Li, L., Werber, K., Calvillo, C., Dinh, K., Guarde, A., & König, A. (2014). Multi-sensor soft-computing system for driver drowsiness detection. In V. Snášel, P. Krömer, M. Köppen, & G. Schaefer (Eds.), *Soft computing in industrial applications* (Vol. 223, pp. 129–140). Springer International Publishing.

McCall, J., Wipf, D., Trivedi, M., & Rao, B. (2007, September). Lane change intent analysis using robust operators and sparse bayesian learning. *IEEE Transactions on Intelligent Transportation Systems*, *8*(3), 431–440.

Mehler, B., Reimer, B., & Coughlin, J. (2012, April). 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.

Mehler, B., Reimer, B., & Wang, Y. (2011, June). A comparison of heart rate and heart rate variability indices in distinguishing single-task driving and driving under secondary cognitive workload. In *International* *driving symposium on human factors in driver assessment, training, and vehicle design* (pp. 590–597). University of Iowa.

Mierswa, I., & Morik, K. (2005, February). Automatic feature extraction for classifying audio data. *Machine Learning*, *58*(2-3), 127–149.

Murphey, Y., Chen, Z., Kiliaris, L., Park, J., Kuang, M., Masrur, A., & Phillips, A. (2008, June). Neural learning of driving environment prediction for vehicle power management. In *Ieee international joint conference on neural networks* (pp. 3755–3761). IEEE.

Murphey, Y., Crossman, J., Chen, Z., & Cardillo, J. (2003, July). Automotive fault diagnosis - part II: A distributed agent diagnostic system. *IEEE Transactions on Vehicular Technology*, *52*(4), 1076–1098.

Pauzie, A. (2008, December). A method to assess the driver mental workload: The driving activity load index (dali). *IET Intelligent Transport Systems*, *2*(4), 315–322.

Peng, H., Long, F., & Ding, C. (2005, august). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(8), 1226–1238.

Regan, M. (2005, December). Driver distraction: Reflections on the past, present and future. *Journal of the Australasian College of Road Safety*, *16*(2), 22–33.

Regan, M., Hallett, C., & Gordon, C. (2011, September). Driver distraction and driver inattention: Definition, relationship and taxonomy. *Accident Analysis & Prevention*, *43*(5), 1771–1781.

Reimer, B., Mehler, B., Wang, Y., & Coughlin, J. (2012, February). 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), 454–468.

Rodrigues, J., Vieira, F., Vinhoza, T., Barros, J., & Cunha, J. (2010, September). A non-intrusive multi-sensor system for characterizing driver behavior. In *Ieee conference on intelligent transportation systems* (pp. 1620–1624). IEEE.

Stutts, J., Reinfurt, D., Staplin, L., & Rodgeman, E. (2001). *The role of driver distraction in traffic crashes*. Washington, DC: AAA Foundation for Traffic Safety.

Sun, H., Wang, H., Zhang, B., & Zhao, F. (2010, August). PGFB: A hybrid feature selection method based on mutual information. In *International conference on fuzzy systems and knowledge discovery* (pp. 2862–2871). IEEE.

Tango, F., & Botta, M. (2009, July). Evaluation of distraction in a driver-vehicle-environment framework: An application of different data-mining techniques. In P. Perner (Ed.), *Proceedings of the industrial conference on advances in data mining: Applications and theoretical aspects* (Vol. 5633, pp. 176–190). Springer Berlin Heidelberg.

Taylor, P., Griffiths, N., & Bhalerao, A. (2015, July). Redundant feature selection using permutation methods. In *Automatic machine learning workshop* (pp. 1–8).

Taylor, P., Griffiths, N., Bhalerao, A., Popham, T., Xu, Z., & Dunoyer, A. (2014, May). Redundant feature selection for telemetry data. In L. Cao, Y. Zeng, A. Symeonidis, V. Gorodetsky, J. Müller, & P. Yu (Eds.), *Agents and data mining interaction* (Vol. 8316, pp. 53–65). Springer Berlin Heidelberg.

Torkkola, K., Massey, N., & Wood, C. (2004, December). Detecting driver inattention in the absence of driver monitoring sensors. In *International conference on machine learning and applications* (pp. 220–226). IEEE.

Witten, I., Frank, E., & Hall, M. (2011). *Data mining: Practical machine learning tools and techniques*. San Francisco, CA: Morgan Kaufmann Publishers Inc., San Francisco, CA.

Wollmer, M., Blaschke, C., Schindl, T., Schuller, B., Färber, B., Mayer, S., & Trefflich, B. (2011, June). Online driver distraction detection using long short-term memory. *IEEE Transactions on Intelligent Transportation Systems*, *12*(2), 273–324.

Young, K., & Regan, M. (2007). Driver distraction: A review of the literature. *Distracted driving. Sydney, NSW: Australasian College of Road Safety*, 379–405.

Zhang, Y., Owechko, Y., & Zhang, J. (2008). Learning-based driver workload estimation. In D. Prokhorov (Ed.), *Computational intelligence in automotive applications* (Vol. 132, pp. 1–24). Springer Berlin Heidelberg.

Phillip Taylor received the M.Eng degree (with honors) in computer science from the University of Warwick, Coventry, UK., in 2011, where he is currently working towards a PhD. His postgraduate research areas is driver monitoring using data mining of the vehicle telemetry data. His main areas of research are data mining, feature selection and driver monitoring. Mr. Taylor is a member of the British Computer Society and the Institute of Electrical and Electronics Engineers.

Nathan Griffiths is an associate professor in the Department of Computer Science, University of Warwick, U.K. and a Royal Society Industry Fellow. His research is focussed on multi-agent systems, trust and reputation, social network analysis, and machine learning. Prior to joining Warwick he was a director in a software solutions company, and he retains links with industry and is involved in projects applying research to industrial problems. He has authored over 90 referred publications, and co-edited a book on agent-based service-oriented computing.

Abhir Bhalerao (M'00) is an Associate Professor (Senior Lecturer) in Computer Science. He received his B.Sc. and Ph.D. degrees in 1986 and 1992 respectively. He joined as faculty at Warwick in 1998 having completed 5 years as a post-doctoral research scientist with the NHS and Kings Medical School, London and as a Research Fellow at Harvard Medical School. His current interests are in modelling chronic lung diseases from CT, multi-camera reconstruction, and forensic image analysis. He has published about 70 refereed articles in image analysis, medical imaging, graphics and computer vision. He was the general co-chair and local organizer of the British Machine Vision Conference, 2007. He is co-founder and Research Director of Warwick Warp Ltd., a company specializing in biometric technologies.

Zhou Xu received the BEng in Electronics and Information Engineering from East China University of Science & Technology, and the MSc and PhD degrees in Online Test of Micro&Nano Systems from Lancaster University, UK. Dr. Xu joined the research department of Jaguar Land Rover in 2011. His current work includes vehicle on-board sensor fusion, data mining and machine learning, and is specialized in the intelligent control of ADAS system.

Adam Gelencser completed a BSc in Traffic Engineering and an MSc in Network Computing at Coventry University before joining Transport Research Laboratory's Intelligent Transportation Group as a researcher working on car-to-car communication systems. He later joined Jaguar Land Rover Research Department developing advancements to lane departure system and investigating other driver assistance related topics. Currently he is part of a team looking at data fusion and machine learning techniques to further enhance vehicle safety and comfort.

Thomas Popham graduated in 2003 with a MEng degree in Electronic Engineering with European Studies from the University of Warwick, and recevied the Ph.D. degree in Computer Science from the University of Warwick in 2010. He is currently working as a research engineer at Jaguar Land Rover developing image and signal processing algorithms for the future generation of vehicles. His interests lie in the areas of motion estimation, stereo vision and obstacle/pedestrian detection techniques