Multi-Level Driver Workload Prediction Using Machine Learning and Off-The-Shelf Sensors

van Gent, Paul; Melman, T.; Farah, Haneen; van Nes, Nicole; van Arem, Bart

1 **Multi-Level Driver Workload Prediction Using Machine Learning and Off-The-Shelf**
2 **Sensors**
3
4

5 **Paul van Gent**
6 Delft University of Technology
7 Faculty of Civil Engineering and Geosciences
8 Stevinweg 1, 2628CN Delft, the Netherlands
9 Email: P.vanGent@tudelft.nl
10
11 **Timo Melman**
12 Delft University of Technology
13 Faculty of Mechanical, Maritime and Materials Engineering
14 Mekelweg 2, 2628CD Delft, the Netherlands
15 Email: T.Melman@tudelft.nl
16
17 **Haneen Farah**
18 Delft University of Technology
19 Faculty of Civil Engineering and Geosciences
20 Stevinweg 1, 2628CN Delft, the Netherlands
21 Email: H.Farah@tudelft.nl
22
23 **Nicole van Nes**
24 SWOV – Stichting Wetenschappelijk Onderzoek Verkeersveiligheid
25 Bezuidenhoutseweg 62, 2594AW The Hague, the Netherlands
26 Email: Nicole.van.Nes@swov.nl
27
28 **Bart van Arem**
29 Delft University of Technology
30 Faculty of Civil Engineering and Geosciences
31 Stevinweg 1, 2628CN Delft, the Netherlands
32 Email: B.vanArem@tudelft.nl
33
34
35 Word count: 6472 words text + 4 tables/figures x 250 words (each) = 7472 words
36
37
38
39
40
41 *Revised Paper*
42 Submission Date
43 **31-07-2017**

1 **ABSTRACT**
2 The present study aims to add to the literature on driver workload prediction using machine
3 learning methods. The main aim is to develop workload prediction on a multi-level basis, rather
4 than a binary high/low distinction as often found in litearature. The presented approach relies on
5 measures that can be obtained unobtrusively in the driving environment with off-the-shelf sensors,
6 and on machine learning methods that can be implemented on low-power embedded systems.
7   Two simulator studies were performed, one inducing workload using realistic driving
8 conditions, and one inducing workload with a relatively demanding lane-keeping task. Individual
9 and group-based machine learning models were trained on both datasets and evaluated. For the
10 group-based models the generalising capability, that is the performance when predicting data from
11 previously unseen individuals, was also assessed.
12   Results show that multi-level workload prediction on the individual and group level
13 works well, achieving high correct rates and accuracy scores. Generalising between individuals
14 proved difficult using realistic driving conditions, but worked well in the high demanding
15 lane-keeping task. Reasons for this discrepancy are discussed as well as future research directions.
16
17 *Keywords*: Driver workload, machine learning, workload prediction, random forest, support vector
18   machine, embedded workload prediction
19

1 **INTRODUCTION**
2 Research into driver workload has been conducted for at least three decades (*1, 2*). Recently,
3 research efforts have shifted to using powerful Machine Learning (ML) methods, giving
4 promising results (*5, 6*). ML methods have been used for other driver-related classification
5 problems, such as driver distraction (*7*), driver interruptibility (*8*) or driver identification (*9*). The
6 present study aims to fill the gaps in the existing research on predicting driver workload using ML
7 methods in several ways, as will be explained in the next paragraphs.
8   First, ML studies into predicting driver workload often focus on a binary classification
9 problem (high workload vs. low workload). A more fine-grained prediction of workload may be
10 desirable to enable adaptive interfaces for in-vehicle advice systems (IVIS), systems that may
11 simplify their content (*10*), or driver assistance systems that may incrementally increase their level
12 of support based on the level of driver workload. The experiments described in this paper attempt
13 to predict workload on 7- and 10-point workload scales.
14   Second, studies to date often use intrusive sensors or measure variables (i.e.
15 electroencephalogram, EEG) that are not practical in the driving environment (see for example (*5,*
16 *6*)). Additionally, it is unknown how well results obtained by the high-grade intrusive sensors used
17 in experiments translate to low-cost sensors. This work uses low-cost sensors that can be
18 integrated into the real-world driving environment, and uses measures that can be obtained
19 non-intrusively. This is important, since especially low-cost sensors are likely to be integrated into
20 the driving environment in real-world applications.
21   Lastly, the models generated in most studies are not generally publicly available for use
22 by the research community. The models developed in this study will be made available for
23 scientific use after publication of results (https://github.com/paulvangentcom).
24
25 **Research Objectives**
26 The previous section outlined the main research gaps and ways to add to the present literature. This
27 led to the formulation of three criteria for predicting driver workload in the present work: The main
28 goal is to develop a workload algorithm that (A) has usable accuracy when predicting multiple
29 workload levels, while generalising among individuals, (B) uses data that can be measured with
30 available low-cost, sensors that can be integrated into the driving environment, and (C) is
31 implementable on embedded hardware (for example in a smart steering wheel).
32   The first criterion (A), predicting workload at a higher resolution than the binary low/high
33 found in previous literature while generalising among individuals, is addressed in the experimental
34 design and data analysis presented in subsequent sections.
35   The second criterion (B) entails using sensor inputs from readily available, low-cost
36 sensors that are easy to implement in the driving environment. By using low-cost sensors, which
37 are likely to present more noise in the signal compared to high-end sensors, results will give a
38 better reflection of real-world performance compared to studies using high-end sensors. Apart
39 from having been used successfully in other workload prediction studies, selected variables should
40 be measurable non-intrusively in the driving environment. This led to the selection of heart rate,
41 skin response, blink rate and several performance measures (for an overview of the selection
42 process, see (*11*)). This criterion ensures any results are directly applicable to in-car settings at a
43 low cost, and that results obtained are likely to translate well to real-world applications.
44   Criterion C, ensuring the model is implementable on an embedded system, means it must
45 be efficient both in memory use as well as computational requirements. Two machine learning
46 algorithms were selected that can satisfy this criterion: 'Random Forest' and 'Support Vector
47 Machine' algorithms. Random Forests (*12*) are computationally efficient (*13*) but can have a large

1 memory footprint. Solutions have been proposed that allow embedded implementations while
2 maintaining performance (*14*), making it a suitable algorithm to use. Support Vector Machines
3 (*15*) implementations can suffer from computational complexity, as well as high memory footprint
4 for more complex models. Methods have been proposed, however, that achieve remarkable
5 efficiency increases without sacrificing performance (*16*, *17*), making SVM's also a suitable
6 candidate algorithm.
7   Two experiments were conducted to evaluate the feasibility of the previously defined
8 criteria. First, a simulator experiment was performed, where workload was induced using realistic
9 driving situations. Results of this experiment were explored futher using a dataset obtained from
10 another driving simulator experiment that induced workload with a demanding lane-keeping task.
11 Finally, results of both experiments are discussed and future steps are outlined.
12
13 **ESTIMATING WORKLOAD IN A REALISTIC DRIVING SCENARIO STUDY**
14 To assess the feasibility of predicting driver workload in realistic driving settings, a simulator
15 study was performed. The main goal was to evaluate the prediction of multi-level driver workload
16 in realistic driving conditions.
17
18 **Methods**
19 *Equipment*
20 The study was performed in a fixed-base, medium-fidelity driving simulator. A dashboard mockup
21 with three 4K-displays (resolution 4096*2160 px) provided roughly 180-degree vision. Actuators
22 consisted of a Fanatec steering wheel and pedals, and a custom blinker control. The simulation ran
23 in Unity3D. The simulated vehicle had an automatic gearbox and a top speed of 165 km/h.
24 FIGURE 1(A) illustrates the set-up.
25   Physiological data were recorded at 100Hz, using low-cost sensors powered by an Atmel
26 ATMega328p embedded processor board. Heart rate was recorded using a photoplethysmographic
27 (PPG) method (*18*) at the left index finger. Skin response was recorded at the middle and ring
28 finger of the same hand (see FIGURE 1(B)). Additionally, blink data were recorded using a GoPro
29 HERO+ camera on the dashboard, running at 1080p@30Hz. Simulator data were logged at 50Hz.
30
31 *Simulator Scenarios*
32 Two scenarios were created in Unity3D, one scenario with situations likely to induce high
33 workload ('high' workload' scenario) and one with situations that are not likely to induce high
34 workload ('low workload' scenario). Road geometry was based on a part of the Cooperative-ITS
35 (C-ITS) corridor in the Netherlands: the A67, a two-lane highway between Eindhoven and Venlo
36 with speed limit of 130km/h. Three weather conditions were designed for each scenario: clear
37 weather, and two degrees of fog with visibility of approx. 150 meters ('light fog') and below 25
38 meters ('heavy fog'). This gave a total of six scenarios.
39   To accurately design the road geometry, CAD drawings of the road segments were
40 secured from the open data program of the Dutch government (https://data.overheid.nl). Using
41 Autodesk 3DS Max, the data in the CAD files were converted to 3D models and textured. The
42 surrounding terrain was generated using height map data obtained from the Microsoft Bing Maps
43 API (https://www.bingmapsportal.com/). Canals and wooded areas were extracted automatically
44 from satellite imagery, and adjusted by hand where necessary. The location, shape, and content of
45 traffic signs was inferred from Google Streetview, designed in 3DS Max and manually placed at
46 the corresponding location in the scenario.
47

1
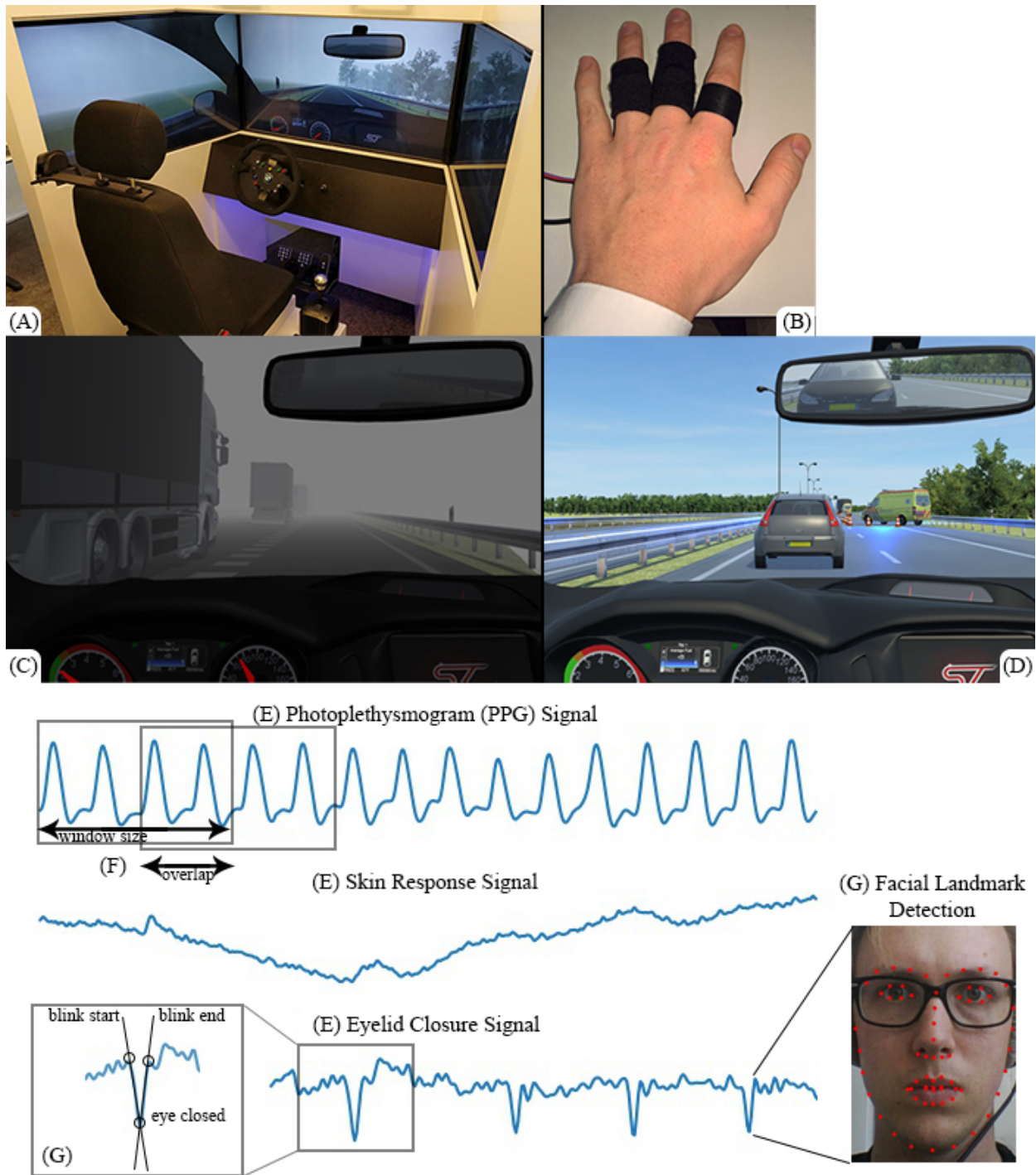
FIGURE 1 - Figure showing the simulator set-up (A), physiological sensors (B), the merging between a platoon of trucks in dense fog (C) and the accident site at the end of the 'high workload' scenario (D). Examples of the raw signal data are shown (E), the concepts of window size and overlap factor (F), an example of the facial landmark detection and the resulting process of analysing the blink rate signal (G).

The 'high workload' scenario was 15.9 km in length, and ran between Eindhoven and Someren. Participants would encounter several workload-inducing 'events' spread out across the scenario. After accelerating across an on-ramp, the first event was encountered: particpants had to

1   merge into a dense platoon of trucks (4-5 meters headway, FIGURE 1(C)), a manoeuvre shown to
2   increase workload on the driver (*19*). The second event was encountered two kilometres upstream,
3   and consisted of a segment of slow moving traffic on the right lane, designed to nudge the
4   participants to drive in the left lane. While passing the slow-moving traffic, an ambulance
5   approached from behind exhibiting auditory and visual signals, travelling at the legally allowed
6   max speed of 170km/h in the Netherlands (max. 40km/h difference with other traffic). This placed
7   the participant in the demanding situation of quickly having to find a gap in the much slower
8   moving lane to the right and perform a merging manoeuvre. The third event was a game of '20
9   questions' (*20*), intended to simulate an engaging (phone) conversation. By asking at most 20 polar
10   (yes/no) questions, participants had to guess which animal, object or person the experimenter had
11   in mind. The final event came near the end of the scenario. The right lane was closed off due to an
12   accident, with slow moving (< 15 km/h) traffic on the left lane (FIGURE 1 (D)). The 20 questions
13   game was played until the accident site was reached. If participants finished early, the game was
14   restarted with a different subject. After this, participants took the next exit and stopped the car.
15   The 'low workload' scenario consisted of self-paced driving in light traffic for 20.5km.
16   The simulated road was a replica of the A67 road between Someren and Venlo. There were no
17   events. Participants drove until reaching a designated exit, where they stopped the car.
18
19   *Experimental Procedure*
20   Approval for the study was obtained from the ethics committee at Delft University of Technology.
21   Participants drove the six scenarios spread out over three separate days, each day driving one
22   randomly assigned 'high workload' and one 'low workload' scenario. This approach was taken
23   because physiological measures can vary from day to day, as well as to avoid a fatigue effect from
24   occurring when asking participants to drive six 10-15-minute scenarios consecutively.
25   In the 'high workload' scenario, participants were asked to rate their experienced mental
26   effort and task difficulty on a 7-point scale after each event, leading to six workload data points per
27   run. In the 'low workload' scenario, the questions were asked at fixed positions in the scenario,
28   leading to four workload data points per run. The exact questions were '*How much mental effort*
29   *did the driving task take in the last few moments, on a scale of 1-7?*' and '*How difficult was the*
30   *driving task in the last few moments, on a scale of 1-7?*'. Scale labels ranged from very low/easy, to
31   very high/difficult, and were explained to participants before the experiment started. Note that we
32   did not use a standardised workload scale such as the NASA TLX or RSME, since we wanted to
33   keep interaction time with and demands on the driver to a minimum.
34   Participants that registered for the experiment received a copy of the informed consent. It
35   was signed and brought to the first session. After being seated in the simulator, a relaxation period
36   of three minutes was given to the participants. This was to allow the physiological measures of
37   each participant to return to its baseline. Sensors were attached, after which the signal quality was
38   checked. A physiological baseline was recorded first. After the baseline, it was briefly explained to
39   the participant that there would follow a drive on a segment of the A67 highway. Participants were
40   instructed to drive at their own pace, but not exceed the speed limit as indicated on road-side signs.
41   If a participant was unfamiliar with '20 questions', a test round was played to familiarise them
42   with the game.
43
44   **Data Analysis**
45   Participants were asked to rate their mental effort and driving task difficulty on a 7-point scale.
46   Since querying the driver might influence workload, the 'high workload' scenario was constructed
47   in such a way that at least one minute of driving was between each two events, to allow signals to

1   return to baseline. The data recorded between two events were not used in the analysis. In the case
2   of the 'low workload' scenario, one minute of data following each question were excluded from
3   the analysis.
4
5   *Preprocessing of Physiological Data*
6   An algorithm was developed to extract the most commonly used features from the measured heart
7   rate signal (*21*, *22*), using a sliding window approach (see FIGURE 1F). The output measures are
8   divided into time-domain (*23*) and frequency-domain measures (*24*). In the time-domain, the
9   measures included are BPM (beats per minute), IBI (inter-beat interval), MAD (median absolute
10 deviation of intervals between heart beats), SDNN (standard deviation of intervals between heart
11 beats), RMSSD (root mean square of successive differences between neighbouring heart beat
12 intervals), SDSD (standard deviation of successive differences between neighbouring heart beat
13 intervals), and the pNN50 and pNN20 (proportion of differences between successive heart beats
14 greater than 50ms and 20ms, resp.) In the frequency domain, included measures are LF (the low
15 frequency band: 0,04-0,15Hz), which is related to short-term blood pressure variation, and HF (the
16 high frequency band: 0,16-0,5Hz), which reflects breathing rate, and the LF/HF ratio, a measure of
17 sympathetic-parasympathetic balance (*24*, *25*).
18   Skin response consists of a tonic and phasic component (*26*). Tonic represents the
19 long-term, slow variation in the signal, indicative of general psycho-physiological arousal (*27*).
20 Phasic reflects relatively quick responses to discrete external stimuli, occurring generally between
21 1-3 seconds after stimulus onset (*27*). Power in the frequency spectrum of skin reponse between
22 0.03Hz-0.5Hz has been linked to short term workload changes (*28*). The mean, max-min
23 difference, MAD (median absolute difference), and 0.03-0.5Hz frequency spectrum were
24 extracted from the GSR signal, using the same window approach as for heart rate. Frequency
25 spectra were extracted using a trapezoidal integration of the area under corresponding frequency
26 bands in the power spectrum.
27   Blink data were detected offline from recorded video data. An algorithm was developed
28 to extract blink number, blink duration and inter-blink-interval. It functioned by detecting 68
29 'facial landmarks' (*29*), then calculating eyelid distance for each frame. Blinks were detected in the
30 resulting signal by finding large slopes, then finding the lowest point of reversal. The process is
31 displayed visually in FIGURE 1 (G).
32
33 *Driver Performance Data*
34 Performance measures reflect how the control the driver exerts over the vehicle varies across
35 conditions. We included steering wheel angle, steering wheel reversals, speed, variation in lateral
36 and longitudinal position, and headway and time to collision when available (for more
37 information, see (*11*)).
38
39 *Generating Machine Learning Sets*
40 Machine learning sets were generated from the raw data and labelled based on self-report data, by
41 varying window size and overlap factor. Window size refers to how much data is used for the
42 calculation of features, overlap factor refers to how much data any window W_i shares with the
43 previous window W_{i-1}. Both concepts are visualised in FIGURE 1 (F). Window sizes of 5, 10 and
44 30 seconds, and overlap factors of 0% and 50% were used, leading to a total of 6 sets.
45
46
47

1   *Model Development and Evaluation*
2   Two different machine learning algorithms were used: A Random Forest Regressor (RFR), and a
3   Support Vector Machines Regressor (SVR). The RFR creates an ensemble (forest) of regression
4   trees in which each tree is trained on a random subset of the features. They have been used in for
5   example (*30*). Support Vector Machines function by mapping the data to a higher dimensional
6   space, and solving an optimization problem to identify a set of hyperplanes that separate the
7   training data into classes. They have been used in for example (*7, 9*). With the SVR, the
8   Polynomial kernel (SVR(poly)), and the Radial Basis Function kernel (SVR(rbf)) were evaluated.
9   Algorithms that were used are taken from the SciKit-Learn repository (*31*).
10   The resulting models were evaluated using several metrics. Model error was evaluated
11   using mean absolute error (AE_μ) and median absolute error ($AE_{\mu1/2}$), both measures of the
12   accuracy of the predictions. The coefficient of determination (R^2) was also computed as a
13   goodness-of-fit measure. Performance for class-based predictions was also evaluated, expressed as
14   correct rate.
15
16   **Results**
17   *Participants*
18   19 participants took part in the experiment. Data from one participant were excluded because of a
19   failure to understand some tasks due to a language barrier. This left 18 participants, of which 12
20   were males and 6 were females. The average age was 34.56 years (SD 10.09). Of the 18
21   participants, 12 owned a car and reported using it three to four times a week on average, and
22   travelling between 2500 and 15000km annually. All participants held a valid driver's license. No
23   simulator sickness severe enough to terminate a driving session was reported. Reported mental
24   effort and perceived difficulty correlated with weather conditions and with scenario type
25   independently and in line with expectations, although no interaction effect was present (*11*).
26
27   *Individual Models*
28   The training and testing sets for the individual models were generated by dividing the dataset of
29   each driver into training and testing sets with an 80%/20% split ratio, respectively. This split ratio
30   was chosen to ensure sufficient training data, since individual datasets were relatively small.
31   The results indicated that the models functioned well, with the RFR outperforming the
32   SVR. For all individual models with a window size of 5s and overlap of 0%, the AE_μ was 0.343,
33   the $AE_{\mu1/2}$ was 0.129, R^2 was 0.679, Correct Rate (CR) was 76.30% when predicting discrete
34   classes, and 93.80% when miss-by-one errors were allowed (CR+/-1). This indicated that on
35   average, predictions were off by 0.343, and that half the predictions had an error less than 0.129,
36   from a total scale of 7 classes. See TABLE 1 for an overview of all results. Model performance
37   increased with a larger overlap factor. This was expected, since a larger overlap creates a larger
38   training set to fit the model to, and because a larger overlap factor indicates more shared variance
39   between adjacent samples. Interestingly, an inverse relationship between window size and model
40   performance was observed, contrary to what has been reported previously (*5*). Miss-by-one errors
41   indicate predictions that are 'almost correct', and still contain enough information about the true
42   workload states. For example, if workload is predicted as '6' while the true value is '7', the
43   information in the prediction is still useful: in either case workload is on the high end.
44
45   *Group Models*
46   The second step was to estimate the model performance within the entire group. The dataset
47   containing data from all drivers was split into training- and testing sets with a 60%/40% split ratio.

1   Since the size of the group dataset is much larger compared to individual dataset, a more stringent
2   split ratio could be chosen while maintaining a sufficiently large training set.
3   Results indicated group models performed well. The AE_μ for the model with window size
4   5s and 0% overlap was 0.605, the $AE_{\mu1/2}$ 0.406, R^2 0.661, CR 57.40%, and CR+/-1 90.60%.

TABLE 1 Performance Metrics RFR models

Window Size	5 sec		10 sec		30 sec	
Overlap Factor	0%	50%	0%	50%	0%	50%
Individual Model						
AE_μ	0.343	0.219	0.431	0.280	0.613	0.492
$AE_{\mu1/2}$	0.129	0.565	0.296	0.109	0.490	0.291
R^2	0.679	0.8716	0.590	0.794	0.071	0.306
CR	76.30%	85.21%	67.88%	80.77%	49.68%	60.82%
CR +/- 1	93.80%	97.61%	92.93%	96.13%	85.81%	89.55%
Group Model						
AE_μ	0.605	0.455	0744	0.553	0.898	0.801
$AE_{\mu1/2}$	0.406	0.250	0.565	0.344	0.628	0.652
R^2	0.661	0.774	0.564	0.709	0.372	0.504
CR	57.40%	69.57%	46.12%	62.48%	40.47%	43.82%
CR +/- 1	90.60%	93.81%	87.02%	91.42%	80.60%	84.56%
Generalising Model						
AE_μ	1.522	1.536	1.457	1.519	1.375	1.424
$AE_{\mu1/2}$	1.163	1.201	1.199	1.253	1.174	1.230
R^2	-0.538	-0.623	-0.460	-0.602	-0.299	-0.396
CR	20.07%	20.05%	19.81%	20.21%	20.21%	20.47%
CR +/- 1	55.18%	55.19%	55.46%	54.94%	57.21%	55.89%

5   **Performance metrics for the best performing (RFR) classifier. The table displays the mean (μ) and median**
6   **(μ1/2) absolute error metrics, the coefficient of determination (R2), the correct rate (CR) and the miss-by-one**
7   **correct rate (CR +/- 1).**

8   *Generalising Group Models*
9   The last step was to assess how models would perform in a realistic setting, e.g. a setting where
10   workload from an unknown driver is predicted based on data from a pool of other drivers. To
11   achieve this, data were sampled using a k-fold approach, with $k = N_{participants}$. For every k_i, the
12   training set consisted of all data except the held out participant k_i. Workload for participant k_i was
13   then predicted and model performance evaluated. This method simulated how the trained models
14   would perform when predicting data from previously unseen individuals. This obtained
15   performance measure reflects real-world settings, where it is impractical for models to be trained
16   on all possible drivers and generalising power is thus preferable.
17   Results showed that models did not perform well when generalizing to unknown drivers.
18   The AE_μ for all individual models with window size 5s and 0% overlap was 1.522, $AE_{\mu1/2}$ was
19   1.163, R^2 was -0.538, CR 20.07%, and CR+/-1 55.18%. The strongly negative coefficient of
20   determination suggests unsatisfactory performance (the mean of the data is a better predictor than
21   the trained model). The relatively low (though above chance level, not satisfactory) absolute error
22   rates given R^2 are explained by a class imbalance in the dataset, where two classes (workload level
23   1 and 2) dominate. To assess whether this was a possible cause for the poor performance of the
24   models, data were resampled using SMOTE (Synthetic Minority Over-Sampling Technique) (*32*).

1 This had little discernible effect on the model performance, and it was concluded that low
2 performance was not due to the class imbalance in the dataset. It was also observed that R^2
3 increases slightly with increasing window size, in accordance with earlier studies (5) and contrary
4 to the individual and group models in the present study.
5
6 **Conclusion**
7 The results of this study showed that predicting self-reported workload in a simulated realistic
8 environment was possible at the individual and group level, but proved difficult when generalising
9 to unknown drivers. Several causes can be identified. The simulated scenarios might not have
10 induced sufficient workload to be measurable with performance or physiological measures.
11 Indeed, most participants indicated that driving in the simulator felt very different from actual
12 driving, and was not that difficult at all. Since a self-report measure was used, which is a subjective
13 measure, it is possible that different participants had biased response tendencies. Lastly, it might
14 also be the case that different physiological response patterns to workload exist, in which case the
15 sample size of 18 could have been too small to account for all occurring patterns.
16   This raises the question whether workload prediction is at all possible on non-binary
17 scales, while generalising across drivers. To further explore this possibility, a dataset from a study
18 with a lane-keeping task was obtained. This study and the results are discussed in the next section.
19
20 **ESTIMATING WORKLOAD IN A FORCED-PACE SIMULATOR STUDY**
21 A dataset was re-used from a previously executed study by Melman et al. (in press, (33)) to further
22 assess multi-level workload prediction in drivers. The study featured a challenging lane-keeping
23 task, which had the potential to induce higher workload than the previous study. The same
24 physiological and performance measurements were used in as in the previously described
25 simulator study.
26
27 **Method**
28 *Equipment*
29 The study was performed in a fixed-base driving simulator at the faculty of Aerospace
30 Engineering, Delft University of Technology. The simulator consisted of a mockup dashboard with
31 three LCD projectors (BenQ W1080ST 1080p) that provided roughly 180-degree vision. The
32 simulated vehicle had an automatic gearbox and a top speed of 210 km/h.
33   Physiological data were logged using a biosignalsPlux wireless hub at 1000Hz. Heart rate
34 was recorded using three pre-gelled Ag/AgCl electrodes at the heart's v3-node. Skin response was
35 measured using the same pre-gelled electrodes, placed inside the palm and on the wrist of both
36 hands. Simulator data were logged at 100Hz.
37
38 *Scenarios*
39   The scenarios used to induce workload in drivers each consisted of a 25km long,
40 single-lane road. The road was divided into four 6km sections of different lane width (3.6m, 2.8m,
41 2.4m, 2.0m). Each section had seven curves, five with an inner radius of 750m and two with a
42 500m radius. Transitions between sections of different width always took place in a 750m radius
43 curve, and were preceded by a road sign indicating a narrowing road. The four sections were
44 identical, with the exception that the curves of segments 2 and 4 four were mirrored with respect to
45 section 1 and 3.
46   Cones were placed 8m apart on the road markings on both sides of the road. The main
47 task was to hit as few cones as possible. A cone hit was indicated to the participant visually by a red
48 dot on the side of the car where the cone was hit, and by a loud auditory beep. Extra difficulty in

1   lane-keeping was induced by a perturbation added to the vehicle's lateral motion. This
2   perturbation was an unpredictable multi-sine signal with five frequencies between 0.067Hz and
3   0.25Hz, with a maximum summed amplitude of 1,000N. Without the perturbation, lane keeping
4   (especially on straight segments) was not considered challenging enough. The width of the
5   simulated vehicle was 1.8m.
6   Three runs were driven with the aim of inducing different levels of workload: a self-paced
7   run and two forced-pace runs of 90km/h and 130km/h. In the self-paced run, participants had full
8   longitudinal control over the car and could drive at their own pace. In the forced-pace conditions,
9   however, the car's speed was automated and kept constant at 90km/h and 130km/h. This would
10 push participants into curves at high speeds, with the goal of raising their workload significantly.
11 The three runs were presented to the participants in randomised order.
12
13 *Procedure*
14 Participants read and signed an informed consent form, informing them of the purpose and
15 procedure of the study. Participants were instructed that the main task was to minimise the total
16 number of cone hits. Furthermore, participants were informed that during the experiment, a beep
17 would sound every 20 seconds. At the sounding of this beep, participants were asked to verbally
18 answer the question "*From 0 to 10, how much effort does the current driving task take you?*", with
19 0 being 'no effort', 5 being 'moderate effort' and 10 being 'a lot of effort'.
20   Before the experiment started, participants were familiarised with the simulator and the
21 procedure by driving two 3.7km trial runs. The first trial run was self-paced, the second was
22 forced-pace with speed at 110km/h. After the trial run, any question the participant had was
23 answered. The electrodes were attached, and a one-minute baseline was recorded.
24
25 **Analysis**
26 Participants rated their mental effort on a scale of 0-10, every 20 seconds. This rating was
27 annotated by the experimenter and added to the dataset. What data were logged, data
28 preprocessing, ML set generation, model development and evaluation are identical to what has
29 been described in the previous study.
30
31 **Results**
32 *Participants*
33 In total twenty-four participants took part in the experiment (17 male, 7 female). The average age
34 was 24.6 years (SD 2.4). Participants reported driving multiple times a week (11 participants), at
35 least once a month (7 participants) or less than one month (6 participants). All participants held a
36 valid driving license. Reported mental effort was sensitive to the lane width variations, although
37 regarding speed only to 130 km/h forced-pace condition (*33*).
38
39 *Individual Models*
40 As in the previous study, training and testing sets for the individual models were generated by
41 dividing the dataset into two stratified sets. More data per participant were collected than in the
42 previous experiment, so data were split with the more stringent 60%/40% split ratio.
43   Results were similar to the previous study, and indicated that the models performed well,
44 with RFR outperforming SVR. An inverse relationship between model performance and overlap
45 factor was observed, as well as increasing performance with increasing overlap factors, both as in
46 the previous experiment. For all individual models with a window size of 5s and overlap of 0%, the
47 AE_μ was 1.046, the $AE_{\mu1/2}$ 0.662, R^2 0.635, CR 40.74%, and CR+/-1 77.31%. The relatively larger

1 absolute errors, compared to individual models in the previous study, might have resulted from the
2 wider workload scale, the different nature of the driving task, or the more frequent reporting of
3 mental workload. More information is displayed in table 2.
4
5 *Group Models*
6 To evaluate performance at the group level, data were split with a 60%/40% split ratio. Results
7 indicated group models attained high performance. For the model with window size 5s and 0%
8 overlap, the AE_μ was 0.904, the $AE_{\mu 1/2}$ 0.638, R^2 0.774, CR 41.61%, and CR+/-1 82.30%. TABLE
9 2 displays the full results. Performance increased with larger overlap factors, and again an (weak)
10 inverse relationship between performance and window size was observed.

TABLE 2 Performance Metrics RFR Models

| *Window Size*	*5 sec*		*10 sec*		*30 sec*	
Overlap Factor	*0%*	*50%*	*0%*	*50%*	*0%*	*50%*
Individual Model						
AE_μ	1.046	0.823	1.213	0.853	1.127	0.870
$AE_{\mu 1/2}$	0.662	0.511	0.833	0.518	0.959	0.694
R^2	0.635	0.763	0.600	0.675	0.561	0.735
CR	40.74%	50.31%	33.93%	45.83%	20.83%	40.28%
CR +/- 1	77.31%	84.34%	70.83%	81.94%	65.83%	81.48%
Group Model						
AE_μ	0.904	0.730	0.984	0.808	1.084	0.876
$AE_{\mu 1/2}$	0.638	0.482	0.722	0.546	0.792	0.663
R^2	0.774	0.830	0.740	0.802	0.718	0.811
CR	41.61%	51.30%	35.12%	46.44%	34.22%	37.87%
CR +/- 1	82.30%	88.18%	80.32%	85.88%	73.21%	82.41%
Generalising Model						
AE_μ	1.878	1.988	1.988	1.989	1.809	1.717
$AE_{\mu 1/2}$	1.831	1.844	1.718	1.741	1.680	1.568
R^2	0.118	0.079	0.196	0.177	0.411	0.433
CR	14.09%	13.45%	12.62%	13.44%	15.72%	15.21%
CR +/- 1	41.92%	40.70%	44.15%	42.29%	47.16%	46.32%

11 **Performance metrics for the best performing (RFR) classifier. The table displays the mean (μ) and median**
12 **(μ1/2) absolute error metrics, the coefficient of determination (R2), the correct rate (CR) and the miss-by-one**
13 **correct rate (CR +/- 1).**

14 *Generalising Group Models*
15 Model performance when generalising to unknown individuals was then assessed, which did not
16 perform well in the first simulator experiment. Data sampling methods were identical to the
17 previous study.
18   Results indicated models performed moderately well. For the best performing model with
19 window size 30s and 50% overlap, the AE_μ was 1.717, the $AE_{\mu 1/2}$ 1.568, R^2 0.433, CR 15.21%,
20 CR+/-1 46.32%. Although model absolute error is relatively large, the coefficient of determination
21 indicated a moderate relationship between model and data. FIGURE 2 below displays the
22 predicted and true values for the first four participants. Individual model performance varied, with
23 workload being predicted well for some participants, while for others showed a correct trend but
24 with a constant offset error. These offset errors inflated the absolute error rates and deflated the

1   predictive accuracy despite good model performance. Generally, a decreased performance with
2   increased overlap factor was observed (except for the largest window size of 30s), as well as
3   increased performance with increased window size. The effect is similar to results for the model
4   generalisation step in the previous study, but more pronounced. The effect also corresponds with
5   what has been reported before (*5*).
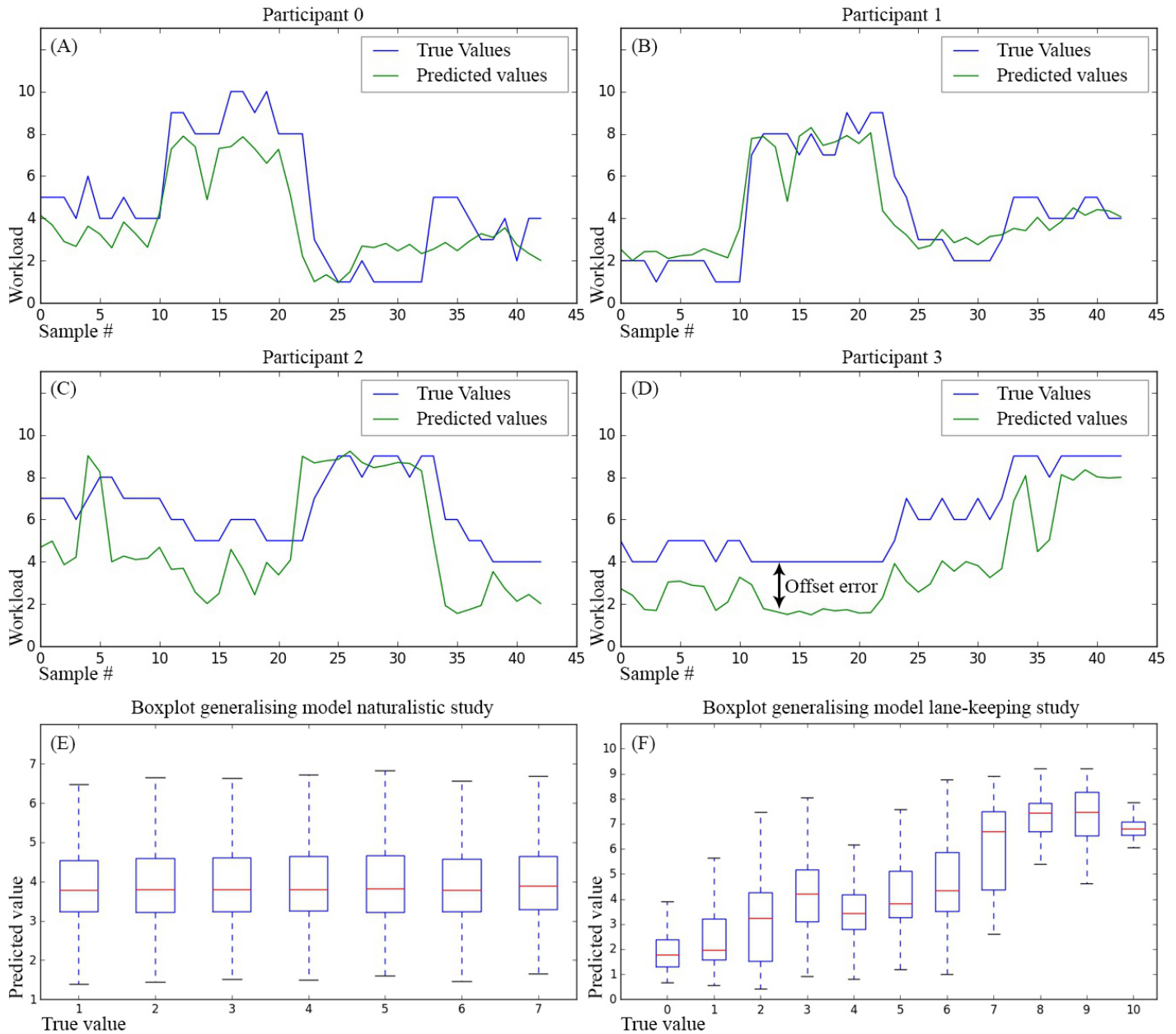
6

7   **FIGURE 2 – The top four windows (A-D) show plots from the first four participants indicate that the models**
8   **performed well, with the relatively large absolute errors likely resulting from individual scaling problems in**
9   **the predictions. These offset errors are indicated in (D): the general trend is predicted well but there is a**
10   **constant offset error. The last two windows (E-F) show box plots, further exploring the generalising models**
11   **from both studies.**

12

1   **Conclusion**
2   The results of this study show similarities with the previous study for individual and group-based
3   models. Additionally, this second experiment shows that, when predicting multi-level workload
4   (11 classes), generalising performance was satisfactory, although still with room for improvement.
5   This study seems to indicate that indeed non-binary workload prediction that generalises
6   to unknown individuals is possible using ML methods. Although models generalising between
7   individuals showed variations in performance based on which individual's workload was being
8   predicted, including constant offset errors in several participants, overall performance was
9   promising.
10
11 **OVERALL CONCLUSIONS AND DISCUSSION**
12 The present study tried to model driver workload using machine learning techniques that can run
13 on embedded systems, with data collected from low-cost-sensors. Results have shown that
14 individual models and within-group models functioned well in both a realistic driving setting as
15 well as an artificial lane-keeping task setting. When generalising to unknown drivers, only the
16 lane-keeping study produced usable results. As displayed in FIGURE 2 (E-F), in the first study the
17 generalised model learns to predict values around the mean to optimize accuracy, in the second
18 study the model learns to predict based on the reported workload.
19   Since the data we gathered in the study are time-series human physiological and
20 performance data, it likely exhibits strong autocorrelation from one sample to the next. This might
21 be a potential explanator for the higher performance in the individual and group models in both
22 studies. Since with random sampling, shared variance between samples from the training set and
23 the prediction set might bias the classifier towards a higher accuracy. To better assess performance,
24 training cases were included where the models had to generalize to unknown individuals. These
25 give a more accurate indication of performance, since with this approach there is no shared
26 variance between training set (all participants minus participant k) and the testing set (participant
27 k). As such, only the generalizing training case offers a reliable index of performance. This is an
28 important distinction, since it shows that although using machine learning to predict driver
29 workload can lead to promising results, care must be taken when interpreting the results. Without
30 care in selecting the sampling techniques used, model performance might be inflated.
31   Possible reasons for the discrepancy in generalizing performance between both studies
32 could include that the workload induced in the realistic settings was too low to be reflected in the
33 physiological or performance signals, that workload induced by artificial tasks is more easily
34 measurable than that induced by more realistic tasks, or that different physiological response
35 patterns to workload might exist and that the sample in the first study was either too small or
36 contained too much individual variation.
37   Possible limitations of the present study are that we employed a self-report measure as
38 ground truth of the experienced mental workload of the drivers. We did not employ standardised
39 workload scales such as NASA TLX, to keep interaction time and demand with the driver to a
40 minimum. However, this may have contributed to lower model performance through participant
41 response tendencies, and leaves some doubt as to what degree the data captures workload. In
42 addition to this, we did not look at compensatory behaviour drivers might employ to manage their
43 workload, such as reducing speed in complex or demanding situations.
44   Future directions are planned. These include feature space normalisation of the dataset to
45 attempt to reduce the offset errors observed in some individuals, as well as exploring additional
46 feature extraction methods. After this, on-road testing is planned to explore model performance in
47 real-world driving settings. Lastly, development of an embedded variant of the model is planned.
48

1
2   **REFERENCES**

3   1.   de Waard, D. *The Measurement of Drivers' Mental Workload*. Drukkerij Haasbeek, Alphen aan den
4   Rijn, 1996.
5   2.   Aasman, J., G. Mulder, and L. J. M. Mulder. Operator Effort and the Measurement of Heart-Rate
6   Variability. *Human Factors*, Vol. 29, No. 2, 1987, pp. 161–170.
7   3.   Matthews, G., L. E. Reinerman-Jones, D. J. Barber, and J. Abich. The Psychometrics of Mental
8   Workload: Multiple Measures Are Sensitive but Divergent. *Human Factors: The Journal of the
9   Human Factors and Ergonomics Society*, Vol. 57, No. 1, 2015, pp. 125–143.
10   https://doi.org/10.1177/0018720814539505.
11   4.   Mehler, B., B. Reimer, and J. F. Coughlin. Sensitivity of Physiological Measures for Detecting
12   Systematic Variations in Cognitive Demand From a Working Memory Task: An On-Road Study
13   Across Three Age Groups. *Human Factors: The Journal of the Human Factors and Ergonomics
14   Society*, Vol. 54, No. 3, 2012, pp. 396–412. https://doi.org/10.1177/0018720812442086.
15   5.   Solovey, E. T., M. Zec, E. A. Garcia Perez, B. Reimer, and B. Mehler. Classifying Driver Workload
16   Using Physiological and Driving Performance Data. *Proceedings of the 32nd annual ACM
17   conference on Human factors in computing systems - CHI '14*, 2014, pp. 4057–4066.
18   https://doi.org/10.1145/2556288.2557068.
19   6.   Jarvis, J., F. Putze, D. Heger, and T. Schultz. Multimodal Person Independent Recognition of
20   Workload Related Biosignal Patterns. *Proceedings of the 13th international conference on
21   multimodal interfaces - ICMI '11*, 2011, p. 205. https://doi.org/10.1145/2070481.2070516.
22   7.   Liang, Y., M. L. Reyes, and J. D. Lee. Real-Time Detection of Driver Cognitive Distraction Using
23   Support Vector Machines. *IEEE Transactions on Intelligent Transportation Systems*, Vol. 8, No. 2,
24   2007, pp. 340–350. https://doi.org/10.1109/TITS.2007.895298.
25   8.   Kim, S., J. Chun, and A. K. Dey. Sensors Know When to Interrupt You in the Car. *Proceedings of the
26   33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15*, 2015, pp. 487–
27   496. https://doi.org/10.1145/2702123.2702409.
28   9.   Moreira-matias, L., and H. Farah. On Developing a Driver Identification Methodology Using
29   In-Vehicle Data Recorders. *IEEE Transactions on Intelligent Transportation Systems*, Vol.
30   submitted, 2017. https://doi.org/10.1109/TITS.2016.2639361.
31   10.   Birrel, S., M. Young, N. Staton, and P. Jennings. Using Adaptive Interfaces to Encourage Smart
32   Driving and Their Effect on Driver Workload. Vol. 484, 2017, p. 764.
33   https://doi.org/10.1007/978-3-319-41682-3.
34   11.   Gent, P. Van, H. Farah, N. Van Nes, and B. Van Arem. Towards Real-Time, Nonintrusive Estimation
35   of Driver Workload: A Simulator Study. *Proceedings of the Road Safety and Simulation Conference
36   2017,* 2017.
37   12.   Breiman, L. Random Forests. *Machine learning*, Vol. 45, No. 1, 2001, pp. 5–32.
38   https://doi.org/10.1023/A:1010933404324.
39   13.   Sventnik, V., A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan, and B. P. Feuston. Random Forest: A
40   Tool for Classification and Regression in Compound Classification and QSAR Modeling. *Journal
41   of Chemical Information and Computer Science*, Vol. 43, No. 6, 2003, pp. 1947–1958.
42   https://doi.org/10.1016/j.rse.2008.02.011.
43   14.   Mishina, Y., R. Murata, Y. Yamauchi, T. Yamashita, and H. Fujiyoshi. Boosted Random Forest.
44   *IEICE Transactions on Information and Systems*, Vol. E98, No. D, 2015, pp. 1630–1636.
45   15.   Cortes, C., and V. Vapnik. Support-Vector Networks. *Machine Learning*, Vol. 20, No. 3, 1995, pp.
46   273–297. https://doi.org/10.1023/A:1022627411411.
47   16.   Theocharides, T., and S. Member. Embedded Hardware-Efficient Real-Time Vector Machines.
48   *IEEE transactions on neural networks and learning systems*, Vol. 27, No. 1, 2016, pp. 99–112.
49   https://doi.org/10.1109/TNNLS.2015.2428738.
50   17.   Bajaj, N., G. T. C. Chiu, and J. P. Allebach. Reduction of Memory Footprint and Computation Time
51   for Embedded Support Vector Machine (SVM) by Kernel Expansion and Consolidation. *IEEE
52   International Workshop on Machine Learning for Signal Processing, MLSP*, 2014.

https://doi.org/10.1109/MLSP.2014.6958875.

18. Jae Baek, H., H. Bit Lee, J. Soo Kim, J. Min Choi, K. Keun Kim, and K. Suk Park. Nonintrusive Biological Signal Monitoring in a Car to Evaluate a Driver's Stress and Health State. *Telemedicine and e-HEALTH*, Vol. 15, No. 2, 2009, pp. 182–189.

19. de Waard, D., A. Kruizinga, and K. A. Brookhuis. The Consequences of an Increase in Heavy Goods Vehicles for Passenger Car Drivers' Mental Workload and Behaviour: A Simulator Study. *Accident Analysis and Prevention*, Vol. 40, No. 2, 2008, pp. 818–828. https://doi.org/10.1016/j.aap.2007.09.029.

20. Kun, A. L., A. Shyrokov, and P. a. Heeman. Interactions between Human-Human Multi-Threaded Dialogues and Driving. *Personal and Ubiquitous Computing*, Vol. 17, No. 5, 2013, pp. 825–834. https://doi.org/10.1007/s00779-012-0518-1.

21. van Gent, P. Analyzing a Discrete Heart Rate Signal Using Python. http://www.paulvangent.com/2016/03/15/analyzing-a-discrete-heart-rate-signal-using-python-part-1/.

22. van Gent, P. Python Heart Rate Analysis Toolkit. *GitHub Repository*. https://github.com/paulvangentcom/heartrate_analysis_python.

23. Reimer, B., B. Donmez, M. Lavallière, B. Mehler, J. F. Coughlin, and N. Teasdale. Impact of Age and Cognitive Demand on Lane Choice and Changing under Actual Highway Conditions. *Accident Analysis and Prevention*, Vol. 52, 2013, pp. 125–132. https://doi.org/10.1016/j.aap.2012.12.008.

24. Montano, N., A. Porta, C. Cogliati, G. Costantino, E. Tobaldini, K. R. Casali, and F. Iellamo. Heart Rate Variability Explored in the Frequency Domain: A Tool to Investigate the Link between Heart and Behavior. *Neuroscience and Biobehavioral Reviews*, Vol. 33, No. 2, 2009, pp. 71–80. https://doi.org/10.1016/j.neubiorev.2008.07.006.

25. Billman, G. E. Heart Rate Variability - A Historical Perspective. *Frontiers in Physiology*, Vol. 2 NOV, No. November, 2011, pp. 1–13. https://doi.org/10.3389/fphys.2011.00086.

26. Lim, C. L., C. Rennie, R. J. Barry, H. Bahramali, I. Lazzaro, B. Manor, and E. Gordon. Decomposing Skin Conductance into Tonic and Phasic Components. *International Journal of Psychophysiology*, Vol. 25, No. 2, 1997, pp. 97–109. https://doi.org/10.1016/S0167-8760(96)00713-1.

27. Seitz, M., T. J. Daun, A. Zimmermann, and M. Lienkamp. Measurement of Electrodermal Activity to Evaluate the Impact of Environmental Complexity on Driver Workload. *Proceedings of the FISITA 2012 World Automotive Congress*, 2012, pp. 245–256. https://doi.org/10.1007/978-3-642-33741-3.

28. Shimomura, Y., T. Yoda, K. Sugiura, A. Horiguchi, K. Iwanaga, and T. Katsuura. Use of Frequency Domain Analysis of Skin Conductance for Evaluation of Mental Workload. *Journal of physiological anthropology*, Vol. 27, No. 4, 2008, pp. 173–177. https://doi.org/10.2114/jpa2.27.173.

29. Köstinger, M., P. Wohlhart, P. M. Roth, and H. Bischof. Annotated Facial Landmarks in the Wild: A Large-Scale, Real-World Database for Facial Landmark Localization. *Proceedings of the IEEE International Conference on Computer Vision*, 2011, pp. 2144–2151. https://doi.org/10.1109/ICCVW.2011.6130513.

30. Miyaji, M., M. Danno, H. Kawanaka, and K. Oguri. Driver's Cognitive Distraction Detection Using Adaboost on Pattern Recognition Basis. *Proceedings of the 2008 IEEE International Conference on Vehicular Electronics and Safety, ICVES 2008*, 2008, pp. 51–56. https://doi.org/10.1109/ICVES.2008.4640853.

31. Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay. Scikit-Learn: Machine Learning in Python. *Journal of Machine Learning Research*, Vol. 12, 2012, pp. 2825–2830. https://doi.org/10.1007/s13398-014-0173-7.2.

32. Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: Synthetic Minority over-Sampling Technique. *Journal of Artificial Intelligence Research*, Vol. 16, 2002, pp. 321–357. https://doi.org/10.1613/jair.953.

33. Melman, T., D. A. Abbink, M. M. van Paassen, E. R. de Boer, and J. C. F. Winter. What Determines

1   Drivers' Speed? An Empirical Investigation of Three Behavioural Adaptation Models. *Manuscript*
2   *submitted for publication*.
3
4
5