Driver State Classification: Identifying High Cognitive Load and Drowsiness through Driver Performance and Physiology

Suzan Ayas, Dengbo He, and Birsen Donmez, *Senior Member, IEEE*

***Abstract*— This study investigates the effect of high cognitive load and drowsiness on driving performance (speed, lane position, steering wheel movement) and driver physiology (cardiac activity, skin conductance) and uses these measures in classifying high cognitive load, alert, and drowsy driver states. A within subject driving simulator experiment was conducted with twenty-seven participants (14 females, mean age: 36.7). High cognitive load was induced via the n-back task (1-back, 2-back), a commonly used auditory-verbal recall task. Drowsiness was induced by monotonous driving (i.e., extended periods of low cognitive load), and was rated by trained observers. Mixed linear models were used to analyze the differences between the driver states, while machine learning models were used for multi-class classification. Compared to alert driving with no additional cognitive load, high cognitive load was associated with greater physiological arousal and speed variation and reduced speed and standard deviation of lane position (SDLP). Drowsiness was associated with lower physiological arousal and increased speed, SDLP, and standard deviation of steering wheel angle. Tree-based ensemble models (i.e., random forest, XGBoost) performed the best in classification. With simple features such as the average and SD, high cognitive load, drowsiness, and alert states were classified with up to 76% average accuracy. These measures could differentiate high cognitive load states with around 85% AUC and drowsiness with around 79% AUC within one model. These findings can help in the selection of metrics for driver monitoring systems that can differentiate driver cognitive overload and underload and inform the design of real-time intervention systems.**

***Index Terms*— attention, driver distraction, driver fatigue, driver monitoring systems, galvanic skin response, heart rate, machine learning.**

I. INTRODUCTION

THE degree of activation of cognitive resources during cognitive processing is known as cognitive load [1]. In general, when cognitive load demanded from a human is too high or too low, task performance can degrade, creating an inverted U-shaped relationship [2]. Demanding environmental conditions (e.g., fog, curvy roads), internal factors (e.g., stress), and secondary tasks (e.g., talking on the phone) can increase cognitive load, whereas driving under prolonged periods of low cognitive load due to monotonous road environments can lead to drowsiness [3], [4]. Drivers can become drowsy within 1 to 2 hours and sometimes within 30 minutes of monotonous driving due to low arousal and boredom [5]. Both high cognitive load and extended periods of low load/arousal can reduce driver awareness and alertness towards hazards and increase cognitive processing and reaction times [6]–[8], and therefore, crash risk [9], [10].

The prevalence of these risky driver states is of concern. In a US survey, 68.7% of drivers reported to having talked on the phone while driving at least once in the past month, while this rate ranged from 20.5% (UK) to 59.4% (Portugal) across seven European countries [11]. As for drowsiness, 19.5% of respondents to a Finnish survey reported falling asleep at least once in their driving experience, with 15.9% of respondents reporting recent instances of drowsy driving [12]. Similar rates of falling asleep while driving were reported in a survey of 19 European countries (17%) [13].

To prevent negative outcomes, driver monitoring systems (DMS) are being proposed to detect risky driver states in real time using various driving and driver related data. Upon detection of high cognitive load or drowsiness, the vehicle can initiate interventions to help drivers regulate their state. For example, adaptive user interfaces can be used to lower driver cognitive load, while alarms can be used to increase driver alertness [14]. A large body of literature focuses on driver state classification using driving and driver physiological data [15]–[20]; however, no consensus on the measures, models, sensors, or performance metrics have yet been established for real-world DMS applications. More recently, the European Union passed a regulation that mandates new motor vehicles to be equipped with driver attention warning systems [21], warranting more research towards building functional DMS products.

One of the challenges for DMS design is the overlap between different driver states, including high cognitive load and drowsiness, in terms of how they affect driver physiology and driving metrics (see Table I). As such, misclassifications can occur, and interventions triggered erroneously can further worsen these suboptimal states. For example, if an already overloaded driver receives DMS alerts to “wake them up”, they might experience increased stress and workload. The negative effects of alerting an already overloaded operator has been shown in other domains, including aviation [22]. Similarly, if driving automation takes over vehicle control after mistakenly classifying a drowsy driver as an overloaded one, the resulting reduction in workload can further worsen drowsiness [3]. Thus, accurately separating these states not only from alertness but also from each other is critical for understanding and minimizing classification errors in DMS applications. To the best of our knowledge, no study has examined whether driving and driver-related data can be used successfully to classify high cognitive load, drowsiness, and alert states in a single model to minimize misclassifications among these three states.

To address this gap, we conducted a within subject driving simulator study to (1) investigate the effects of high cognitive load and drowsiness on driving performance and driver physiology, and (2) explore the function of these measures in classifying these three driver states in a single classification model. For objective 1, we compared the effects of six different driver states ranging from high cognitive load to drowsiness (2-back, 1-back, alert, slightly drowsy, moderately drowsy, very drowsy) on driving performance and driver physiology through statistical analyses. For objective 2, we evaluated how these measures performed in machine learning (ML) models (random forest [RF], XGBoost [XGB], support vector machines [SVM], and k-nearest neighbors [KNN], and multilayer perceptron [MLP]) on classifying three driver states within a single model: high cognitive load, alertness, drowsiness.

II. Background

General cognitive workload literature and driving research show that high cognitive load and prolonged periods of low load can significantly impact driving performance and driver physiology [15], [23]–[25]. Table I presents a summary of how different measures respond to high cognitive load and drowsiness including the results of the current study (CS) that are presented in later sections.

Driving measures most commonly used in DMS have been speed, lane position, and steering wheel movement [17], [26]. Driving studies report lower speed [27]–[32] under high cognitive workload which is typically induced by cognitive tasks performed while driving (e.g., n-back tasks; [27]–[29], [31], [32]). Both positive [33], [34] and negative [35] correlations between speed and drowsiness have been reported, as well as no significant correlation [34], [36]. Standard deviation (SD) of speed showed an increase [27], decrease [28] and no significant change [31] under high cognitive load. While, the same measure was reported to decrease with monotonous driving [34], [36] and to increase [33] with sleep deprivation. SD of lane position (SDLP), often used as a cognitive load measure, has been shown to decrease [30]–[32], [37], under high cognitive load, with some studies reporting no effect [31], [37] or an increase [38]. SDLP have been found to both increase [33], [39]–[41] and to decrease [34] with drowsiness. Steering wheel reversal rate (SRR) did not change [30], [37] or increased [28]–[30], [37] under high cognitive load. Further, the mean and SD of steering wheel movements were found to increase [42], and SRR was found to decrease with drowsiness [35].

Cognitive load can also impact physiological response: high cognitive load can activate the sympathethic nervous system (i.e., fight-or-flight response), and low cognitive load can engage the parasympathetic nervous system (i.e., rest-and-digest response). In the general cognitive load literature and in driving research, cardiac and eye-tracking measures are the most commonly used vital signs, and cardiac measures like heart rate (HR) and skin conductivity are known to be sensitive to workload changes [15], [24]. In driving, high cognitive load has been associated with increased HR [27]–[29], [31], [37], [43] and skin conductance [27], [28], [31], [43]. Although the general cognitive load literature indicates that HRV is one of the more consistent measures of workload and decreases with high load [44], a driving study found it to increase [38]. Drowsiness while driving has been linked to decreased HR [45]–[47] and skin conductance [34] and to increased HRV [34], [46], [47].

TABLE I
SUMMARY OF THE LITERATURE AND THE **CURRENT STUDY (CS)** FINDINGS ON DRIVING PERFORMANCE AND DRIVER PHYSIOLOGY CHANGES WITH INCREASED HIGH COGNITIVE WORKLOAD AND DROWSINESS; ARROWS INDICATE DIRECTION OF CHANGE

Measure	High Cognitive Load	Drowsiness
Average speed	No change [31], [37] ↓ [27]–[32], CS	↑ [33], [34] No change [34], [36], CS ↓ [35]
Standard deviation (SD) speed	↑ [27], CS No change [31] ↓ [28]	↑ [33], [41], CS No change [36] ↓ [34], [36]
SD lane position	↑ [38] No change [31], [37] ↓ [30]–[32], [37], CS	↑ [33], [39]–[41], CS No change [36] ↓ [34]
Steering wheel reversal rate	↑ [28]–[30], [37] No change [30], [37], CS	↑ [35], [42] No change CS ↓ [35]
Heart rate	↑ [27]–[29], [31], [37], [43] No change [37], [38], CS	No change [34], CS ↓ [45]–[47], CS
Heart rate variability	↑ [38] No change [31], [37], [38]	↑ [34], [46], [47], CS
Average skin conductivity	↑ [27], [28], [31], [43], CS No change [37]	↑ CS ↓ [34]
SD skin conductivity	No change [31], [37], CS	No change CS

Overall, high cognitive load and drowsiness may affect some measures in similar ways and have conflicting findings across different studies. Yet, existing research focuses on classifying only one suboptimal state from baseline driving, simplifying the problem [17]–[19], [48]. A recent study on train operators used electroencephalogram (EEG) to classify drowsiness, alert, and cognitive load [49] with data collected from 7 train operators. The highest accuracy was 79% based on time-series ensamble tree models (RF, XGB, LightGBM). To the best of our knowledge, no automobile driving study explored if a single model trained on driving and driver-related data can successfully classify cognitive load, alert, and drowsy states to minimize misclassifying drowsiness and high cognitive load.

Our analyses focused on speed, lane position, SRR, HR, HRV, and skin conductance level informed by the relevant literature. We also collected eye tracking metrics via a dashboard-mounted system, but this data was excluded due to the low reliability of eye tracking, especially during the drowsiness states, in addition to synchronization issues.

III. Methods

A driving simulator study was conducted between November 2021 and April 2022 (University of Toronto, Research Ethics Board #41080). The experiment was a within-subject design with all participants first experiencing a high cognitive load period induced by a secondary cognitive task (i.e., n-back), followed by a long stretch of monotonous driving to induce drowsiness. All experiments were scheduled for three hours starting at 2pm to minimize the impact of circadian rhythm.

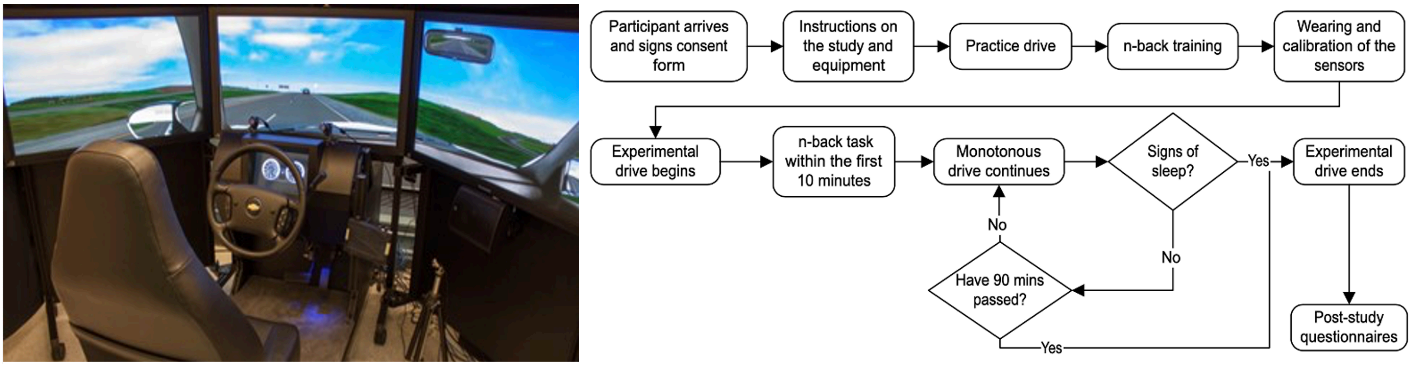

Participant arrives and signs consent form
Instructions on the study and equipment
Practice drive
n-back training
Wearing and calibration of the sensors
Experimental drive begins
n-back task within the first 10 minutes
Monotonous drive continues
Signs of sleep?
Yes
Experimental drive ends
No
No
Have 90 mins passed?
Yes
Post-study questionnaires

Fig. 1. NADS Minisim driving simulator (left) and experimental procedure (right)

A. *Participants*

Participants were recruited through university newsletters and listservs, online marketplaces (e.g., Kijiji.com), and posters placed in campus. Interested individuals filled out an online screening survey. To be eligible for the study, they had to have a valid driver's license for the last three years and to be a frequent driver. The latter criteria required them to drive almost every day or to have driven at least 5,000km in the last six months. These criteria for driving experience considered the COVID-19 lockdowns in place around the time of the study. The participants also had to not wear corrective eyeglasses (to minimize interference with the eye-tracker and camera recordings of eyelid closure; contact lenses were allowed), to have good written and spoken English, and to report having no sleep-related issues. Participants were asked not to consume any caffeine (tea, coffee, dark chocolate) and alcohol 5 and 24 hours before the experiment, respectively, to minimize any influence on HR and on drowsiness [50].

A total of 35 participants were recruited; however, due to equipment malfunction (n=4), recording errors (n=3), and participant withdrawal (n=1), data from 27 participants (14 females) could be analyzed. The average age was 36.7 (SD=14.4, range: 19 to 74) with no significant difference between genders (*t*(25)= -1.54, p=0.14). Participants reported driving almost every day (n=19), a few days a week (n=6), or a few times a month (n=2). As per our eligibility criteria, the participants who reported driving a few times a month also reported driving more than 5,000km in the last 6 months. Those who drove at least a few days a week reported driving 5,000-10,000km (n=15) or 10,001-20,000km (n=9) in the last 6 months; one did not know their mileage.

Participants were compensated at C$14/hour and received a C$8 bonus. They were told that they could receive up to C$8 based on their secondary task performance, but all participants were given the full amount regardless of performance.

B. *Apparatus*

A NADS Minisim driving simulator was used and collected driving data at 60Hz (Fig. 1-left). This simulator runs on a desktop computer and has a quarter cab, three 42" screens allowing for a 130° field of view, and two speakers for stereo sound. A low-frequency speaker under the driver seat creates vibrations to mimic the road surface.

Two webcams recorded the participant and the simulation environment at 30fps capturing the participant face and hand movement on the steering wheel, respectively. Driver physiological data was collected through Becker Meditec sensors at 256 Hz: three solid gel foam electrodes were placed on the participants' chest for electrocardiogram (ECG), and two solid gel foam electrodes were placed under their left foot for galvanic skin response (GSR) data.

C. *High Cognitive Load and Drowsiness Levels*

The experimental drive was designed as a large loop on a rural highway with a 96.6km/h (60mph) speed limit and low traffic density. Participants were instructed to maintain speed around the speed limit and to drive safely as they would in real life. They were asked to perform an auditory-verbal secondary task, starting around the 2-minute mark, and finishing around the 10-minute mark. Afterwards, they continued driving for 1.5hrs in total or until they started showing signs of falling asleep (e.g., long eye closures, see 'very drowsy' in Table II).

A modified n-back task with two difficulty levels (1-back and 2-back) was used to induce high cognitive load [27]. This task required participants to listen to a recording of a series of 10 letters (25s total), count the occurrence of specific patterns, and then respond verbally (during a 5s break in the recording). For the 1-back task, participants had to count how many times they heard two identical letters read to them back-to-back (e.g., 2 in CD**AA**LMX**BB**Z). For the 2-back task, they had to count how many times two identical letters were read to them with one letter in between (e.g., 1 in M**FXF**LDAATO). Each n-back level took about 2 mins, starting with a brief instruction (30s) followed by the participant completing the task three times (i.e., three series of 10 letters). There was around a 3-min period from the completion of the first n-back level to the start of the next. The order of the n-back levels was counterbalanced across participants: 16 participants started with the 1-back trials while 11 started with the 2-back trials. These two levels have been shown to induce differentiable performance decrements [27] and differentiable self-reported cognitive workload comparable to realistic tasks, like controlling the radio while driving [51].

Post-experiment, two independent raters reviewed webcam videos of the participant and assessed their level of drowsiness. For each minute of the video, the highest level of drowsiness was recorded. The raters received training on recognizing a set of predetermined indicators for five levels of drowsiness (Table II). These indicators were adapted from the sleepiness rating scale [52], which is commonly used for drowsiness ground truth ratings [53]. Similar to the guidelines in [53], to minimize fatigue, the raters performed assessments between 8am and 8pm and for a maximum of four hours a day, with a mandatory break after one hour of continuous assessment. An agreement

TABLE II
INDICATORS OF DROWSINESS LEVELS USED IN OBSERVER RATINGS (ADAPTED FROM KUNDINGER ET AL., 2020; WIERWILLE & ELLSWORTH, 1994).

DROWSINESS LEVEL	INDICATORS
Alert	Appearance of alertness present; normal facial tone; normal fast eye blinks; short ordinary glances; occasional body movements/gestures
Slightly drowsy	Still sufficiently alert; less sharp / alert looks; longer glances; slower eye blinks; first mannerisms as rubbing face/eyes, scratching, facial contortions, moving restlessly in the seat
Moderately drowsy	Eye-lid closures (1-2s); mannerisms; slower eye-lid closures; decreasing facial tone; glassy eyes; staring at fixed position
Very drowsy	Eyelid closures (2-3s); eyes rolling upward / sideways; no proper focused eyes; decreased facial tone; lack of apparent activity; large isolated or punctuating movements
Extremely drowsy	Eyelid closures (4s or more); falling asleep; longer periods of lack of activity; movements when transition in and out of dozing

of 80% was reached in ratings, and any disagreements were resolved by discussion to reach consensus. A third researcher, who prepared the training module, acted as a tiebreaker.

D. Procedure

On arrival, participants were asked to put their phones on silent mode and to store their phone and wristwatch outside of the simulator area for a distraction-free environment, and to confirm not having consumed caffeine and alcohol as instructed. Participants signed a consent form and were told that the study assessed general driving behavior. The specific purpose was not revealed until the end of the study.

First, participants completed a practice drive for around 10 mins to familiarize themselves with the simulator. Next, they received written and verbal training on the n-back task. The participants were then outfitted with the physiological sensors and completed the experimental drive. After the drive, participants were asked to consider the last 10 mins of the drive and fill out the Karolinska Sleepiness Scale (KSS) [54] and a risk perception scale [55]. KSS is a 10-level scale ranging from 1 (extremely alert) to 10 (extremely sleepy, can’t stay awake), with 5 being neither alert nor sleepy. The scale for risk perception ranged from 1 (driving on an easy road with no traffic, pedestrians, or animals while perfectly alert) to 10 (driving with my eyes closed; a crash is bound to occur every time I do this), with 6 being driving 20mph faster than traffic on an expressway. See Fig. 1-right for the procedure flowchart.

E. Data Preparation for Analysis

The raw ECG data was processed in MATLAB v.R2021b to remove noise using discrete wavelet transformation. Following this, average HR and average and SD of GSR, speed, lane position with respect to lane center, and steering wheel angle (SWA) were calculated over 30s periods with no overlap, starting from the n-back task periods (see Table III). 30s was chosen given that each n-back task took a total of 30s (25s listening + 5s responding). Averages and SDs were utilized similar to earlier studies for converting timeseries data to non-timeseries observations for training ML models [e.g., [27], [56]. HRV, more specifically root mean squared differences of successive R peaks (RMSSD), was calculated over 5-minute periods, as recommended by relevant guidelines [56]. The n-back task periods were excluded from the HRV analysis as each level only took two minutes in total. SRR was calculated per 30s as per SAE J2944 guidelines [57], [58]. A gap size of 1° was selected.

Even after de-noising, the ECG data of five participants were too noisy and had to be removed from HR analysis. Occasionally, the simulator generated a bug that required the participant to slow down, which might have temporarily alerted the participant. This happened once for five participants and twice for two participants, and all during the drowsiness periods. Approximately 6 mins around this bug were excluded from all analyses.

TABLE III
DESCRIPTION OF VARIABLES USED IN ANALYSIS.

DEPENDENT VARIABLES	DESCRIPTION
HR: Heart rate (beats per minute)	Average heart rate calculated every 30 s using inter beat interval
HRV: Heart rate variability	Root mean squared differences of successive R peaks (RMSSD) calculated every 5 min
GSR: Galvanic skin response (microSiemens)	Average and standard deviation (SD) calculated every 30 s
Speed (mph)	Average and SD calculated every 30 s
Lane deviation from lane center (feet)	SD calculated every 30 s
SWA: Steering wheel angle (degrees)	SD calculated every 30 s
SRR: Steering wheel reversal rate	Number of steering wheel movements >1° within 30 s

F. Statistical Modeling

Mixed linear models were built to analyze the variables listed in Table III, with driver state (2-back, 1-back, alert, slightly drowsy, moderately drowsy, very drowsy), n-back order (the order in which the n-back levels were presented), and their interaction as fixed factors and participant as random. Driver states were determined (and labeled) by the n-back task periods for high cognitive load and by the observer ratings for alert and drowsy states. The n-back order was included to control for potential carryover effects from the n-back tasks. Significant main effects of driver state were followed up with post-hoc tests. To correct for heteroskedasticity and non-normality, log transformations were applied. Jamovi 2.3, an R-based user interface, was used for analysis [59].

The 1-back and 2-back task performance (% correct trials) was compared using Wilcoxon signed-rank tests. For subjective data, participants were grouped into two based on the KSS score distribution, and t-tests were used to compare the perceived risk across the two groups.

G. Classification and Machine Learning Models

ML models were trained to distinguish three classes: high cognitive load (i.e., 1- and 2-back, labeled as per the secondary task periods) vs. alert vs. drowsy states (moderately and very drowsy, determined and labeled by observer ratings). Slightly drowsy states were excluded from the drowsy class because this state was deemed as a transition state as it often did not statistically significantly differ from alert or moderately drowsy states.

Following the process described in E. *Data Preparation for Analysis,* all variables listed in Table III were used except for

HRV, since HRV could not be calculated for the n-back task periods. Five participants with noisy ECG data were excluded from classification. High load, alert, and drowsy classes had 185, 369, and 1233 observations, respectively. Due to class imbalances, we used three resampling methods: down-, and up-sampling, and synthetic minority class oversampling technique (SMOTE [60]).

The data was randomly split into training (80%) and test datasets (20%) with proportionate representation from each driver state. For each participant, the mean and SD of each measure were calculated from their training data to standardize the training and test data, as in (1). This process ensured that individual differences were standardized, and no data leak occurred between training and test datasets. Five models were trained in R v.4.3.1 with 10-fold cross validation and tuned with grid search: RF, XGB, SVM, KNN, MLP. Area under the receiver operating characteristic curve (AUC) and average accuracy metrics were calculated.

$$x_{scaled} = \frac{x_{raw} - \bar{x}_{training}}{s_{training}} \qquad (1)$$

IV. Results

A. Descriptive Statistics

The experimental drive took on average 69 mins (SD=21). Participants were considered to be under high cognitive load during the n-back tasks, two minutes each for 1-back and 2-back levels. On the average, participants were rated to be alert for 10.0 mins (SD=7.3, n=25), slightly drowsy for 18.4 mins (SD=10.8, n=27), moderately drowsy for 29.4 mins (SD=19.2, n=27), and very drowsy for 6.1 mins (SD=6.9, n=9), see Fig. 2. Only nine participants reached the very drowsy state. The duration statistics for the "very drowsy" state is also affected by our stopping criterion – when the experimenter noticed that the participant was very drowsy, she stopped the experiment. Although sleepiness mostly steadily progressed (n=20), some participants briefly became more alert during drowsy states.

B. Secondary Task Performance and Subjective Measures

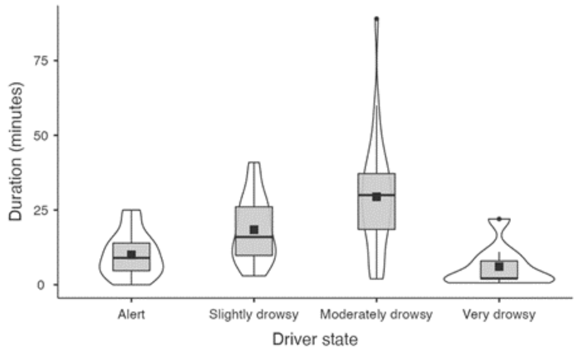


Fig. 2. Box and violin plots for duration of each drowsiness level as rated by observers. Black squares indicate the mean values, and violin plots illustrate the data distribution.

The average task performance was 72.9% (SD=13.2) for 1-back and 59.3% (SD=19.3) for 2-back, as shown in Fig. 3. As expected, the difference in 1-back and 2-back task performance was significantly different (W=52.5, p=0.009). The average KSS ratings for the last 10 mins of the experiment was 6.9 out of 10. As Fig. 3 illustrates, the majority reported their sleepiness level as either 6 or 8. Fourteen participants reported KSS≥7, indicating they felt sleepy with some difficulty to stay alert. Five of them were rated by the observers to be very drowsy at the end of the experimental drive, while the rest were rated to be moderately drowsy. Out of those who reported KSS<7, four participants were rated as very drowsy, and the rest were rated as moderately drowsy.

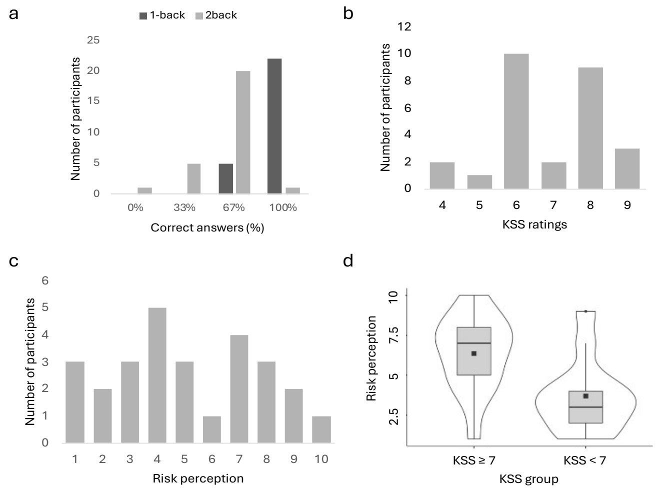


Fig. 3. a. 1-back vs. 2-back task performance (% correct answers out of 3 trials). b. Self-reported sleepiness levels (Karolinska Sleepiness Scale) in the last 10 mins of the experiment. c. Self-reported risk perception levels in the last 10 mins of the experiment. d. Box and violin plots of self-reported risk perception grouped by KSS ratings (KSS≥7 vs. KSS<7).

Average perceived risk was 5.07 (SD=2.66). Those who felt sleepy (KSS≥7) had higher risk perception (M=6.36, SD=2.37) than those who felt rather alert (KSS<7; M=3.69, SD=2.29, *t*(25)=2.967, p=0.007).

C. Statistical Analysis

Significant main effects were observed for driver state for some measures (p<.05) except for SRR (*F*(5, 104.5) =1.11, p= 0.36) and SD GSR (*F*(5, 104.4)=1.64, p=0.16); HRV had a marginally significant p-value (*F*(3, 264.7)=2.41, p=0.07). Significant F-test and post-hoc test results are reported in Table IV, and effect plots are shown in Fig. 4. Average speed and SDLP had a negative relationship with increasing cognitive load. SD speed was significantly higher during 2-back and very drowsy states compared to the alert state. SD SWA had larger values in moderate drowsiness than alert, 1-back, and 2-back states (the difference to 2-back was marginally significant).

Average HR showed significant increases with increasing cognitive load, but HR under alertness was not significantly different than that under the n-back and drowsy states. HRV showed increases as drowsiness levels increased. Overall, average GSR was not sensitive in differentiating high cognitive load vs. drowsiness; however, alert state was significantly different from the rest.

D. Classification Models

Confusion matrices from test predictions of all models for all resampling methods are given in Fig. 5. AUC for n-back vs. others and AUC for drowsy vs. others are given in Table V. Overall, the best performing models were RF and XGB with a range of AUC between 77-83% and 72-79% for n-back and

TABLE IV
SIGNIFICANT F-TEST RESULTS AND POST HOC TEST P-VALUES WITH SIGNIFICANT DIFFERENCES BOLDED

		Avg speed	**SD speed**	**SDLP**	**SD SWA**	**HR**	**HRV**	**Avg GSR (log transformed)**
F-Test for Driver State		F(5, 104.9)= 3.72, p=.004	F(5, 104.6)= 4.17, p=.002	F(5, 104.9)= 21.11 p< .001	F(5, 106.3)= 2.68, p=.03	F(5, 104.1)= 3.46, p= .006	F(3, 264.7)= 2.41, p=.07	F(5, 104.1)= 3.32, p= .008
1-back vs.	2-back	.26	**.002**	.42	.35	.98		.74
Alert vs.	2-back	**.01**	**< .001**	**.003**	.33	.11		**.01**
	1-back	.16	.53	**< .001**	.95	.11		**.004**
Slightly drowsy vs.	2-back	**.001**	**< .001**	**< .001**	.39	**.02**		**.02**
	1-back	**.03**	.80	**< .001**	.08	**.02**		**.01**
	Alert	.46	.70	.11	.07	.48	.82	.83
Moderately drowsy vs.	2-back	**< .001**	**.007**	**< .001**	**.06**	**.02**		.50
	1-back	**.02**	.64	**< .001**	**.005**	**.02**		.31
	Alert	.41	.28	**< .001**	**.005**	.51	.30	**.06**
	Slightly drowsy	.92	.47	.06	.28	.97	.31	.08
Very drowsy vs.	2-back	**.02**	.49	**< .001**	.27	**.003**		.75
	1-back	.12	.13	**< .001**	.08	**.003**		.93
	Alert	.59	**.05**	**< .001**	.08	**.05**	**.01**	**.04**
	Slightly drowsy	.97	.09	**.001**	.61	.14	**.04**	**.05**
	Moderately drowsy	.98	.23	**.05**	.81	.14	**.01**	.43

drowsiness classification, respectively. MLP generally performed well (AUC: n-back: 78-81% drowsiness: 74-78%). SVM performed well in differentiating n-back (AUC: 73-85%), while KNN consistently resulted in lower AUC values. For RF and XGB models, average accuracy ranged from 64-75% and 59-76%, respectively. However, as shown in the confusion matrices, under no resampling, models labeled majority of the observations drowsy due to class imbalance. When only the measures with significant differences in Table IV were utilized (i.e., excluding SRR and SD of GSR), the models performed slightly worse (average accuracy: 72%, average AUC: 72%).

V. Discussion

A within subject driving simulator study was conducted to examine the effects of two levels of high cognitive load and three levels of drowsiness on a variety of driving performance and driver physiological measures. Although many studies investigated driver states under high cognitive load and drowsiness separately, to the best of our knowledge, this is the first study to examine and compare both states. We found significant differences in speed, lane position, steering wheel movements, heart rate measures, and skin conductance levels across six levels of cognitive load states, which can inform future DMS designs. Further, we found that these measures can be used in classifying high cognitive load, alert, and drowsy states with around 80% AUC in a single ML model. This finding indicates that with more data and with more complex models, the three states can be separated, and misclassifications of high cognitive load and drowsiness can be minimized.

A. *Driving Measures*

We found significant reductions in speed with increased cognitive load, and higher speeds were observed when drivers were drowsy, similar to earlier studies (see Table I). SD speed was greater during n-back and drowsiness compared to alertness. Similar results have been reported in the literature, as well as conflicting findings (Table I). We observed no significant changes in the SD speed across drowsy states, but the data showed trends of increase under increased drowsiness. Overall, our findings imply some reduction in speed control under both high cognitive load and drowsiness.

Lane keeping performance improved (i.e., reduced SDLP) under high cognitive load compared to the alert state, consistent

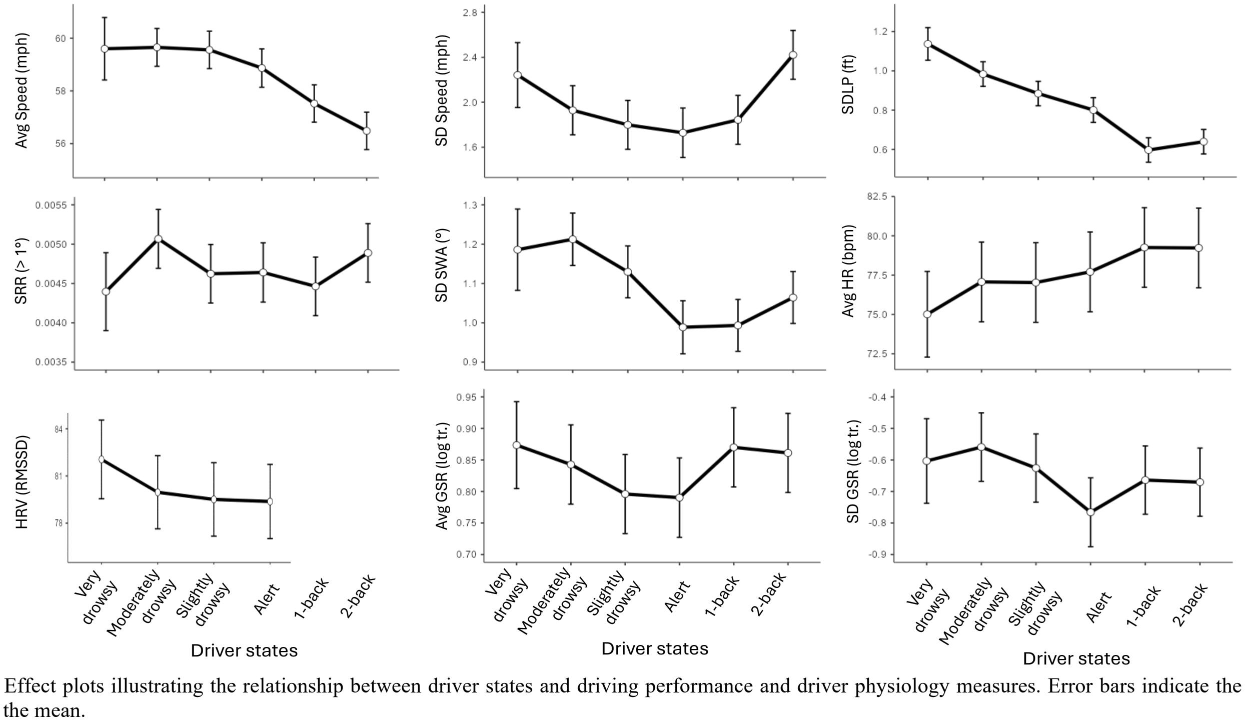


Fig. 4. Effect plots illustrating the relationship between driver states and driving performance and driver physiology measures. Error bars indicate the standard error of the mean.

with many other studies [32], [37], [61], [62]; however, conflicting findings also exist [31], [37], [38]. Improved SDLP is often coupled with increased visual concentration and decreased number of glances to the speedometer [63]–[65]. Although we could not use our eye-tracker data to support this, the drivers might not have noticed slowing down while completing the n-back tasks if they did not check the speedometer as often. Further, drivers might have had difficulties maintaining speed while trying to stay within the lane, as shown by the decreased speed and increased SD in speed during n-back tasks. Increased SDLP has been recorded as a consistent measure of drowsiness [33], [39]–[41]. Our results agree with this: as the drivers became drowsier, their ability to maintain lateral control decreased, which can support the use of SDLP for early drowsiness detection.

On-road studies have reported increased SRR during high cognitive load [28]–[30], [37]; however, we did not observe any significant differences between alert and n-back states, similar to other driving simulator studies [30], [37]. These differences might be due to limitations of the simulator setting, secondary task choice, or engagement levels in driving or the secondary task. We also did not see differences in SRR (1°) under drowsiness. Although decreases in SRR have been reported during drowsiness [35], gap size parameter might impact the findings (micro- vs. large-corrections). [57] investigated gap sizes for visual and cognitive secondary tasks; however, fine-tuning gap size for drowsiness need further research.

B. *Physiological Data*

Our analysis of cardiac measures exhibited expected trends: HR increased with increased cognitive load; however, the differences between adjacent driver states was not significant. We were unable to calculate HRV for high cognitive load due to short n-back task periods, but HRV is generally expected to decrease with increased cognitive load [15], [24]. In our study, a marginally statistically significant increase in HRV was observed under drowsiness. HRV might not be as sensitive as HR in differentiating high cognitive load levels [66], [16].

TABLE V
AUC VALUES FOR DETECTING HIGH COGNITIVE LOAD AND DROWSINESS

Model	No sampling	Up-sampling	Down-sampling	SMOTE	Average
Positive class: n-back					
RF	83%	82%	83%	80%	**82%**
XGB	79%	80%	77%	80%	79%
SVM	84%	76%	**85%**	73%	80%
KNN	75%	69%	75%	74%	73%
MLP	78%	79%	79%	81%	79%
Positive class: drowsy					
RF	76%	75%	77%	75%	76%
XGB	**79%**	78%	72%	78%	**77%**
SVM	75%	66%	73%	67%	70%
KNN	71%	67%	70%	68%	69%
MLP	76%	78%	74%	76%	76%
Avg	78%	75%	77%	75%	76%

RF: random forest. XGB: extreme gradient boost, SVM: support vector machine, KNN: k-nearest neighbor, MLP: multilayer perceptron

There were no effects on SD of GSR. But as expected, average GSR was higher during n-back compared to alert, slightly drowsy and moderately drowsy states. Parallel results have been shown for the mean [27], [28]. Our test could not distinguish GSR between 1-back and 2-back, similar to [43]. The increased GSR we found during drowsiness disagree with the expectations of lower skin conductivity [15]. Conflicting GSR patterns exist in the literature potentially due to the limitations of skin conductance in differentiating between high cognitive load levels [66]. It is also possible that sleep development while driving might stimulate a non-linear relationship in GSR level (e.g., [67]).

C. *Classification Models and Performance*

Our results showed that tree-based ensemble models (i.e., RF, XGB) performed the best in differentiating high cognitive load, alertness, and drowsy states. For detecting high cognitive load, the highest AUC was 85%, while this rate was 79% for

		No resampling			Up-sampling			Down-sampling			SMOTE-sampling		
		True values											
Models	Predicted values	nback	alert	drowsy	nback	alert	drowsy	nback	alert	drowsy	nback	alert	drowsy
RF	nback	26	5	10	30	7	14	46	16	56	36	14	30
	alert	9	35	17	4	50	24	8	74	96	10	61	41
	drowsy	33	88	387	34	71	376	14	38	262	22	53	343
XGB	nback	29	3	11	30	6	10	48	24	69	38	13	23
	alert	6	41	20	6	45	27	7	80	105	7	58	41
	drowsy	33	84	383	32	77	377	13	24	240	23	57	350
SVM	nback	28	8	10	6	1	0	47	22	59	6	3	0
	alert	1	14	7	0	33	14	12	89	143	0	32	12
	drowsy	39	106	397	62	94	400	9	17	212	62	93	402
KNN	nback	13	0	7	41	18	47	35	13	56	40	20	45
	alert	9	45	29	7	73	153	21	89	137	6	71	129
	drowsy	46	83	378	20	37	214	12	26	221	22	37	240
MLP	nback	22	8	15	36	17	56	45	32	91	47	24	77
	alert	11	51	32	22	82	97	6	63	75	11	68	71
	drowsy	35	69	367	10	29	261	17	33	248	10	36	266

Fig. 5. Confusion matrices for four machine learning models (RF: random forest. XGB: extreme gradient boost, SVM: support vector machine, KNN: k-nearest neighbor, MLP: Multilayer Perceptron). Columns indicate true values, and rows indicate predictions. Correct classifications are highlighted in gray.

detecting drowsiness. Our findings showed that with simple features such as the average and SD of driving and physiological measures, risky driver states can be classified with up to 76% average accuracy. For improved accuracy and real-time detection, additional measures and parameters can be explored, as well as time series approaches to consider temporal changes in data. For detecting high cognitive load, a simulator study achieved 85.6% average accuracy for classifying baseline, 0-back, 1-back, and 2-back using skin conductance level and SDLP with neural networks [56]. However, when HR and SRR data were added, the overall accuracy declined. The authors also showed that choice of window size (20s vs. 30s) also impacted accuracy. In a drowsiness classification study with cardiac measures of 5-min windows, accuracy rate of 80-85% was reported with RF for a binary classification, but the leave-one-out accuracy dropped to 58-64% when the RF was trained to classify three levels of drowsiness (alert, somewhat drowsy, drowsy) [68]. Overall, further research is needed on classification sensitivity with changing parameters, window sizes, models, and feature sets.

D. Secondary Task Performance and Subjective Measures

Our findings supported the validity of the modified n-back task in inducing high cognitive load while driving compared to no secondary task. In terms of accuracy, our study showed lower rates (1-back: 73%; 2-back: 59%) than in [27] who developed the modified task (1-back: 94%; 2-back 67%). The greater n-back performance in [27] might be due to differences in road environments, speed limits, and the additional training drive with n-back tasks that the authors implemented.

We also found that half of the drivers felt sleepy with some difficulty to stay alert (KSS≥7). However, some drivers might be unaware of their sleepiness levels, even if their eyes were shutting. For example, one participant was falling asleep after 40 mins of driving but rated their KSS as 4 (i.e., alert).

E. Limitations and Future Directions

Although the age range was not restricted and a gender balance was maintained among participants for greater generalizability, this driving simulator study was conducted with frequent drivers who restricted their caffeine and alcohol consumption prior to the experiment. More research is needed to understand whether the findings apply to a larger pool of driver demographics and how results may be affected by stimulants and medication. Other physiological measures (e.g., EEG) can be also explored [15], [20]. Physiological data can be subject to noise due to participant movement and requires detailed noise removal and pre-processing efforts, including manual data cleaning, even with research grade sensors. For practicality and real-time applications, however, more accurate wearable or non-contact sensors are needed.

This study took place in low-density traffic during daylight with minimal interaction with other vehicles, and the roads had a set speed limit. The complex real-life driving environments and individual differences might impact the driving measures in different ways. The use of driving measures would depend greatly on the driving context, sensor accuracy, and road conditions like existence of clear lane markers. Further, driving experience and familiarity with the simulator environment can also impact driving measures. We aimed to control for these factors through recruiting frequent and experienced drivers, as well as incorporating a practice drive prior to the experimental drive. As driving simulators can further limit the experience drivers have and reduce their perception of risk, on-road studies can shed more light to the use of these measures. We aimed to control for these factors by recruiting frequent and experienced drivers, as well as incorporating a practice drive prior to the experimental drive. As driving simulators can further limit the experience drivers have and reduce their perception of risk, on-road studies can shed more light to the use of these measures.

Drowsiness in this study was induced by monotonous driving for around an hour, and not through lack of sleep. A review on the impacts of sleep restriction [26] described similar patterns to our findings, indicating that our study could induce sleep-like responses through monotonous driving. Further, the observer ratings had high agreement for ground truth (80%) and findings overall matched participants' subjective sleepiness reports.

We used multiple resampling techniques to overcome the class imbalance problem, but these techniques can add bias. Additional data from n-back periods could improve the model performance. Lastly, we used data averaged over 30s periods, which removed the time series element of the sensor data. With more data, more advanced models like recurrent neural networks can be used to train on the temporal aspects for real-time detection models. There may be opportunities to improve detection performance by increasing sample size, engineering more relevant features from the measures collected, and training more complex deep learning models.

Another large gap is in understanding how DMS can be used in vehicles. Substantial research effort has gone into developing DMS (>700 studies) in the last decade, but only a limited number of studies (n=20) evaluated DMS-based drowsiness interventions [5]. Such interventions might impact how drivers interact with and trust DMS, especially under risky driver states. Interventions like emergency braking or automation takeovers due to false drowsiness detection might increase risk and can lead to annoyance, frustration, and stress. Warnings can lead to behavioral adaptations and over-reliance: drivers can get a false reassurance in their driving capability when DMS fails to detect drowsiness. As Fig. 5 shows, drowsiness and cognitive overload can be misclassified, in addition to being misclassified with alertness. Studies that focus on only one suboptimal state are unable to capture the risk of misclassification among different suboptimal states.

VI. Conclusion

In a driving simulator study, we found differences in driver performance and physiological data across a range of driver states. Overall, high cognitive load was associated with reduced speed, SDLP, steering wheel movement, and increased speed variation and high physiological arousal. On the other hand, drowsiness due to monotonous driving also led to reduced driving performance (e.g., greater speed, SDLP, SD SWA, SRR) and lower physiological arousal. Overlapping changes were also present in SD speed and average and SD GSR under both drowsiness and high cognitive load, which may lead to misclassifications. These measures could differentiate high cognitive load states with around 85% AUC and drowsiness with around 79% AUC within one model. These findings can be used in selecting appropriate measures to develop DMS to successfully differentiate cognitive overload and drowsiness both from alert states and from each other. Jointly training models for all three states can help avoid triggering unsuitable interventions in case of misclassifications, informing the design of real-time intervention systems.

Acknowledgment

We thank Yihan Wang, Mohamed Abdelwahab, Xing Tang, Davide Gentile, Mehdi Hoseinzadeh Nooshabadi, Dina Kanaan, Mattea Powell for their assistance in data collection.

References

[1] L. Longo, C. D. Wickens, G. Hancock, and P. A. Hancock, "Human Mental Workload: A Survey and a Novel Inclusive Definition," *Front. Psychol.*, vol. 13, 2022.

[2] J. F. Coughlin, B. Reimer, and B. Mehler, "Monitoring, managing, and motivating driver safety and well-being," *IEEE Pervasive Comput.*, vol. 10, no. 3, pp. 14–21, July 2011.

[3] S. D. Chong and C. L. Baldwin, "The origins of passive, active, and sleep-related fatigue," *Front. Neuroergonomics*, vol. 2, p. 765322, Dec. 2021.

[4] P. A. Desmond and P. A. Hancock, "Active and Passive Fatigue States," in *Stress, Workload, and Fatigue*, 0 ed., P. A. Hancock and P. A. Desmond, Eds. CRC Press, 2000, pp. 455–465.

[5] S. Ayas, B. Donmez, and X. Tang, "Drowsiness Mitigation Through Driver State Monitoring Systems: A Scoping Review," *Hum. Factors*, Nov. 2023.

[6] P. D'Addario and B. Donmez, "The effect of cognitive distraction on perception-response time to unexpected abrupt and gradually onset roadway hazards," *Accid. Anal. Prev.*, vol. 127, no. June 2016, pp. 177–185, 2019.

[7] S. S. Smith, M. S. Horswill, B. Chambers, and M. Wetton, "Hazard perception in novice and experienced drivers: The effects of sleepiness," *Accid. Anal. Prev.*, vol. 41, no. 4, pp. 729–733, 2009.

[8] Y. Ebadi, D. L. Fisher, and S. C. Roberts, "Impact of Cognitive Distractions on Drivers' Hazard Anticipation Behavior in Complex Scenarios," *Transp. Res. Rec.*, vol. 2673, no. 9, pp. 440–451, 2019.

[9] L. Wundersitz, "Driver distraction and inattention in fatal and injury crashes: Findings from in-depth road crash data," *Traffic Inj. Prev.*, vol. 20, no. 7, pp. 696–701, Oct. 2019.

[10] S. Brown, W. G. Vanlaar, and R. D. Robertson, "Fatigue-Related Fatal Collisions in Canada, 2000-2016," *Traffic Inj. Res. Found.*, 2020.

[11] R. B. Naumann and A. M. Dellinger, "Mobile Device Use While Driving — United States and Seven European Countries, 2011," *Morb. Mortal. Wkly. Rep.*, vol. 62, no. 10, pp. 177–182, Mar. 2013.

[12] I. Radun, J. Radun, M. Wahde, C. N. Watling, and G. Kecklund, "Self-reported circumstances and consequences of driving while sleepy," *Transp. Res. Part F Traffic Psychol. Behav.*, vol. 32, pp. 91–100, July 2015.

[13] M. Gonçalves, R. Amici, R. Lucas, T. Åkerstedt, F. Cirignotta, J. Horne, D. Léger, W. T. McNicholas, M. Partinen, J. Téran-Santos, P. Peigneux, L. Grote, and N. R. as S. Collaborators, "Sleepiness at the wheel across Europe: a survey of 19 countries," *J. Sleep Res.*, vol. 24, no. 3, pp. 242–253, 2015.

[14] C. A. DeGuzman, D. Kanaan, and B. Donmez, "Attentive User Interfaces: Adaptive Interfaces that Monitor and Manage Driver Attention," in *User Experience Design in the Era of Automated Driving*, A. Riener, M. Jeon, and I. Alvarez, Eds. Cham: Springer International Publishing, 2022, pp. 305–334.

[15] M. Lohani, B. R. Payne, and D. L. Strayer, "A review of psychophysiological measures to assess cognitive states in real-world driving," *Front. Hum. Neurosci.*, vol. 13, p. 57, Mar. 2019.

[16] K. Lu, A. Sjörs Dahlman, J. Karlsson, and S. Candefjord, "Detecting driver fatigue using heart rate variability: A systematic review," *Accid. Anal. Prev.*, vol. 178, p. 106830, Dec. 2022.

[17] A. Sahayadhas, K. Sundaraj, and M. Murugappan, "Detecting Driver Drowsiness Based on Sensors: A Review," *Sensors*, vol. 12, no. 12, pp. 16937–16953, Dec. 2012.

[18] G. Sikander and S. Anwar, "Driver Fatigue Detection Systems: A Review," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 6, pp. 2339–2352, June 2019.

[19] C. N. Watling, M. Mahmudul Hasan, and G. S. Larue, "Sensitivity and specificity of the driver sleepiness detection methods using physiological signals: A systematic review," *Accid. Anal. Prev.*, vol. 150, p. 105900, Feb. 2021.

[20] J. Ju and H. Li, "A Survey of EEG-Based Driver State and Behavior Detection for Intelligent Vehicles," *IEEE Trans. Biom. Behav. Identity Sci.*, vol. 6, no. 3, pp. 420–434, July 2024.

[21] "European Commission - Have your say," *European Commission - Have your say*, 21-Apr-2023. [Online]. Available: https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/13740-Road-safety-advanced-driver-distraction-warning-systems_en. [Accessed: 11-Mar-2024].

[22] Y. Zhao, K. Zhu, H. Xu, Z. Liu, P. Luo, and L. Wang, "Is red alert always optimal? An empirical study on the effects of red and blue feedback on performance under excessive stress," *Displays*, vol. 88, p. 103008, July 2025.

[23] G. Borghini, L. Astolfi, G. Vecchiato, D. Mattia, and F. Babiloni, "Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness," *Neurosci. Biobehav. Rev.*, vol. 44, pp. 58–75, July 2014.

[24] R. L. Charles and J. Nixon, "Measuring mental workload using physiological measures: A systematic review," *Appl. Ergon.*, vol. 74, pp. 221–232, Jan. 2019.

[25] E. Debie, R. Fernandez Rojas, J. Fidock, M. Barlow, K. Kasmarik, S. Anavatti, M. Garratt, and H. A. Abbass, "Multimodal Fusion for Objective Assessment of Cognitive Workload: A Review," *IEEE Trans. Cybern.*, vol. 51, no. 3, pp. 1542–1555, Mar. 2021.

[26] C. C. Liu, S. G. Hosking, and M. G. Lenné, “Predicting driver drowsiness using vehicle measures: Recent insights and future challenges,” *J. Safety Res.*, vol. 40, no. 4, pp. 239–245, Aug. 2009.

[27] D. He, B. Donmez, C. C. Liu, and K. N. Plataniotis, “High cognitive load assessment in drivers through wireless electroencephalography and the validation of a modified n-back task,” *IEEE Trans. Hum.-Mach. Syst.*, vol. 49, no. 4, pp. 362–371, 2019.

[28] B. Mehler, B. Reimer, and J. F. Coughlin, “Sensitivity of Physiological Measures for Detecting Systematic Variations in Cognitive Demand From a Working Memory Task: An On-Road Study Across Three Age Groups,” *Hum. Factors*, vol. 54, no. 3, pp. 396–412, June 2012.

[29] B. Reimer, B. Mehler, Y. Wang, and J. F. Coughlin, “A Field Study on the Impact of Variations in Short-Term Memory Demands on Drivers’ Visual Attention and Driving Performance Across Three Age Groups,” *Hum. Factors*, vol. 54, no. 3, pp. 454–468, June 2012.

[30] H. A. Jamson and N. Merat, “Surrogate in-vehicle information systems and driver behaviour: Effects of visual and cognitive load in simulated rural driving,” *Transp. Res. Part F Traffic Psychol. Behav.*, vol. 8, no. 2, pp. 79–96, Mar. 2005.

[31] R. Tarabay and M. Abou-Zeid, “Assessing the effects of auditory-vocal distraction on driving performance and physiological measures using a driving simulator,” *Transp. Res. Part F Traffic Psychol. Behav.*, vol. 58, pp. 351–364, Oct. 2018.

[32] B. Reimer, “Impact of Cognitive Task Complexity on Drivers’ Visual Tunneling,” *Transp. Res. Rec.*, vol. 2138, no. 1, pp. 13–19, Jan. 2009.

[33] J. T. Arnedt, M. A. C. Geddes, and A. W. MacLean, “Comparative sensitivity of a simulated driving task to self-report, physiological, and other performance measures during prolonged wakefulness,” *J. Psychosom. Res.*, vol. 58, no. 1, pp. 61–71, Jan. 2005.

[34] G. S. Larue, A. Rakotonirainy, and A. N. Pettitt, “Driving performance impairments due to hypovigilance on monotonous roads,” *Accid. Anal. Prev.*, vol. 43, no. 6, pp. 2037–2046, Nov. 2011.

[35] S. H. Fairclough and R. Graham, “Impairment of Driving Performance Caused by Sleep Deprivation or Alcohol: A Comparative Study,” *Hum. Factors*, vol. 41, no. 1, pp. 118–128, Mar. 1999.

[36] J. Ma, J. Gu, H. Jia, Z. Yao, and R. Chang, “The Relationship Between Drivers’ Cognitive Fatigue and Speed Variability During Monotonous Daytime Driving,” *Front. Psychol.*, vol. 9, p. 459, Apr. 2018.

[37] J. Engström, E. Johansson, and J. Östlund, “Effects of visual and cognitive load in real and simulated motorway driving,” *Transp. Res. Part F Traffic Psychol. Behav.*, vol. 8, no. 2, pp. 97–120, Mar. 2005.

[38] C. Gabaude, B. Baracat, C. Jallais, M. Bonniaud, and A. Fort, “Cognitive load measurement while driving,” in *Human Factors: a view from an integrative perspective*, 2012, pp. 67–80.

[39] M. Ingre, T. Akerstedt, B. Peters, A. Anund, and G. Kecklund, “Subjective sleepiness, simulated driving performance and blink duration: examining individual differences,” *J. Sleep Res.*, vol. 15, no. 1, pp. 47–53, Mar. 2006.

[40] S. Otmani, T. Pebayle, J. Roge, and A. Muzet, “Effect of driving duration and partial sleep deprivation on subsequent alertness and performance of car drivers,” *Physiol. Behav.*, vol. 84, no. 5, pp. 715–724, Apr. 2005.

[41] P.-H. Ting, J.-R. Hwang, J.-L. Doong, and M.-C. Jeng, “Driver fatigue and highway driving: A simulator study,” *Physiol. Behav.*, vol. 94, no. 3, pp. 448–453, June 2008.

[42] P. Thiffault and J. Bergeron, “Monotony of road environment and driver fatigue: a simulator study,” *Accid. Anal. Prev.*, vol. 35, no. 3, pp. 381–391, May 2003.

[43] B. Reimer, B. Mehler, J. F. Coughlin, K. M. Godfrey, and C. Tan, “An on-road assessment of the impact of cognitive workload on physiological arousal in young adult drivers,” in *Proceedings of the 1st International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, Essen Germany, 2009, pp. 115–118.

[44] D. Tao, H. Tan, H. Wang, X. Zhang, X. Qu, and T. Zhang, “A Systematic Review of Physiological Measures of Mental Workload,” *Int. J. Environ. Res. Public. Health*, vol. 16, no. 15, p. 2716, Jan. 2019.

[45] S. K. L. Lal and A. Craig, “Driver fatigue: Electroencephalography and psychological assessment,” *Psychophysiology*, vol. 39, no. 3, pp. 313–321, May 2002.

[46] F. N. Biondi, M. Lohani, R. Hopman, S. Mills, J. M. Cooper, and D. L. Strayer, “80 MPH and out-of-the-loop: Effects of real-world semi-automated driving on driver workload and arousal,” *Proc. Hum. Factors Ergon. Soc. Annu. Meet.*, vol. 62, no. 1, pp. 1878–1882, Sept. 2018.

[47] R. Buendia, F. Forcolin, J. Karlsson, B. Arne Sjöqvist, A. Anund, and S. Candefjord, “Deriving heart rate variability indices from cardiac monitoring—An indicator of driver sleepiness,” *Traffic Inj. Prev.*, vol. 20, no. 3, pp. 249–254, Mar. 2019.

[48] A. Kashevnik, R. Shchedrin, C. Kaiser, and A. Stocker, “Driver Distraction Detection Methods: A Literature Review and Framework,” *IEEE Access*, vol. 9, pp. 60063–60076, 2021.

[49] C. Fan, Y. Peng, S. Peng, H. Zhang, Y. Wu, and S. Kwong, “Detection of Train Driver Fatigue and Distraction Based on Forehead EEG: A Time-Series Ensemble Learning Method,” *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 8, pp. 13559–13569, Aug. 2022.

[50] T. L. Whitsett, C. V. Manion, and H. D. Christensen, “Cardiovascular effects of coffee and caffeine,” *Am. J. Cardiol.*, vol. 53, no. 7, pp. 918–922, Mar. 1984.

[51] B. Mehler, B. Reimer, J. Dobres, and J. F. Coughlin, “Assessing the Demands of Voice Based In-Vehicle Interfaces - Phase II Experiment 3 - 2015 Toyota Corolla,” 2015.

[52] W. W. Wierwille and L. A. Ellsworth, “Evaluation of driver drowsiness by trained raters,” *Accid. Anal. Prev.*, vol. 26, no. 5, pp. 571–581, Oct. 1994.

[53] T. Kundinger, C. Mayr, and A. Riener, “Towards a reliable ground truth for drowsiness: A complexity analysis on the example of driver fatigue,” *Proc. ACM Hum.-Comput. Interact.*, vol. 4, no. EICS, p. 78:1-78:18, June 2020.

[54] T. Åkerstedt and M. Gillberg, “Subjective and Objective Sleepiness in the Active Individual,” *Int. J. Neurosci.*, vol. 52, no. 1–2, pp. 29–37, Jan. 1990.

[55] O. Tsimhoni, D. Smith, and P. Green, “On-the-road assessment of driving workload and risk to support the development of an information manager,” *Ann Arbor MI*, 2003.

[56] J. Son and M. Park, “Estimating Cognitive Load Complexity Using Performance and Physiological Data in a Driving Simulator,” presented at the AutomotiveUI, Salzburg, Austria, 2011.

[57] G. Markkula and J. Engström, “A Steering Wheel Reversal Rate Metric for Assessing Effects of Visual and Cognitive Secondary Task Load,” in *Proceedings of the 13th ITS World Congress*, 2006.

[58] Society of Automotive Engineers, “Operational definitions of driving performance measures and statistics. Surface Vehicle Recommended Practice J2944.” 2015.

[59] “The jamovi project.” jamovi, 2022.

[60] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic Minority Over-sampling Technique,” *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, June 2002.

[61] J. He, J. S. McCarley, and A. F. Kramer, “Lane Keeping Under Cognitive Load: Performance Changes and Mechanisms,” *Hum. Factors*, vol. 56, no. 2, pp. 414–426, Mar. 2014.

[62] P. Li, G. Markkula, Y. Li, and N. Merat, “Is improved lane keeping during cognitive load caused by increased physical arousal or gaze concentration toward the road center?,” *Accid. Anal. Prev.*, vol. 117, pp. 65–74, Aug. 2018.

[63] J. Engström, G. Markkula, T. Victor, and N. Merat, “Effects of Cognitive Load on Driving Performance: The Cognitive Control Hypothesis,” *Hum. Factors*, vol. 59, no. 5, pp. 734–764, Aug. 2017.

[64] J. L. Harbluk, Y. I. Noy, P. L. Trbovich, and M. Eizenman, “An on-road assessment of cognitive distraction: Impacts on drivers’ visual behavior and braking performance,” *Accid. Anal. Prev.*, vol. 39, no. 2, pp. 372–379, Mar. 2007.

[65] M. A. Recarte and L. M. Nunes, “Mental workload while driving: effects on visual search, discrimination, and decision making,” *J. Exp. Psychol. Appl.*, vol. 9, no. 2, pp. 119–137, June 2003.

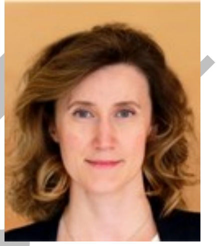
[66] P. Ayres, J. Y. Lee, F. Paas, and J. J. G. Van Merriënboer, “The Validity of Physiological Measures to Identify Differences in Intrinsic Cognitive Load,” *Front. Psychol.*, vol. 12, p. 702538, Sept. 2021.

[67] G. S. Larue, A. Rakotonirainy, and A. N. Pettitt, “Driving performance on monotonous roads,” 2010.

[68] A. Persson, H. Jonasson, I. Fredriksson, U. Wiklund, and C. Ahlström, “Heart Rate Variability for Classification of Alert Versus Sleep Deprived Drivers in Real Road Driving Conditions,” *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 6, pp. 3316–3325, June 2021.

Suzan Ayas received the B.S. degree in industrial engineering from Bogazici University, Istanbul, Turkey in 2017 and the M.A.Sc. degree in 2019 from the University of Toronto, Toronto, ON, Canada. She is currently a Ph.D. student at the Human Factors and Applied Statistics Lab at the University of Toronto, Toronto, ON, Canada.

Dengbo He received the bachelor’s degree in vehicle engineering from Hunan University, Changsha, China, in 2012, the M.S. degree in mechanical engineering from Shanghai Jiao Tong University, Shanghai, China, in 2016, and the Ph.D. degree from the University of Toronto, Toronto, ON, Canada, in 2020. He is currently an Assistant Professor with the Thrust of Intelligent Transpiration and Thrust of Robotics and Autonomous Systems, The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China.



Birsen Donmez (Senior Member, IEEE) received the B.S. degree in mechanical engineering from Bogazici University, Istanbul, Turkey, in 2001, and the M.S. and Ph.D. degrees in industrial engineering in 2004 and 2007, respectively, and the M.S. degree in statistics in 2007 from the University of Iowa, Iowa City, IA, USA.

She is a Professor with the Department of Mechanical and Industrial Engineering, University of Toronto, Toronto, ON, Canada, and the Canada Research Chair in Human Factors and Transportation. Her research interests include understanding and improving human operator behavior and performance in multitask and complex situations, using a wide range of analytical techniques.