What’s on your mind? A Mental and Perceptual Load Estimation Framework towards Adaptive In-vehicle Interaction while Driving

Amr Gomaa  
German Research Center for Artificial Intelligence (DFKI)  
Saarbrücken, Germany  
amr.gomaa@dfki.de

Alexandra Alles  
German Research Center for Artificial Intelligence (DFKI)  
Saarbrücken, Germany  
alexandra_katrin.alles@dfki.de

Elena Meiser  
German Research Center for Artificial Intelligence (DFKI)  
Saarbrücken, Germany  
elena.meiser@dfki.de

Lydia Helene Rupp  
German Research Center for Artificial Intelligence (DFKI)  
Saarbrücken, Germany  
lydia_helene.rupp@dfki.de

Marco Molz  
German Research Center for Artificial Intelligence (DFKI)  
Saarbrücken, Germany  
marco.molz@dfki.de

Guillermo Reyes  
German Research Center for Artificial Intelligence (DFKI)  
Saarbrücken, Germany  
guillermo.reyes@dfki.de

ABSTRACT
Several researchers have focused on studying driver cognitive behavior and mental load for in-vehicle interaction while driving. Adaptive interfaces that vary with mental and perceptual load levels could help in reducing accidents and enhancing the driver experience. In this paper, we analyze the effects of mental workload and perceptual load on psychophysiological dimensions and provide a machine learning-based framework for mental and perceptual load estimation in a dual task scenario for in-vehicle interaction (https://github.com/amrgomaaelhady/MWL-PL-estimator). We use off-the-shelf non-intrusive sensors that can be easily integrated into the vehicle’s system. Our statistical analysis shows that while mental workload influences some psychophysiological dimensions, perceptual load shows little effect. Furthermore, we classify the mental and perceptual load levels through the fusion of these measurements, moving towards a real-time adaptive in-vehicle interface that is personalized to user behavior and driving conditions. We report up to 89% mental workload classification accuracy and provide a real-time minimally-intrusive solution.

KEYWORDS
Automotive Interfaces, User Experience, Adaptive Interfaces, Psychophysiological Measurements, Mental Workload, Perceptual Load, Machine Learning, Sensors Fusion, Dual Task Interaction

1 INTRODUCTION
Driving is an everyday task in which lapses in attention can have fatal consequences. The driving task is still one of the most complex tasks for the human brain despite large advancements in the automotive industry [35, 41]; drivers perform several visual and auditory mental sub-tasks to be able to drive adequately, such as perception, expectation, judgment, planning, and execution [29]. According to the WHO, approximately 1.3 million people die each year as a result of road traffic accidents1. Since almost half of all vehicular accidents involve driver inattention [95], driving safety could benefit from advances in attention research. Additionally, the number of available secondary tasks and therefore the number of distractions rises with the advancement of in-vehicle human-machine interfaces (HMI) [102, 105]. Thus, recent research advancements focus on designing interfaces with minimal interference on the driving task, while providing features aiding the driver’s comfort and safety [17, 37, 80]. One way of evaluating in-vehicle interfaces is measuring the mental effort they put on the driver, specifically, perceptual load (PL) and mental workload (MWL). MWL has been a topic of interest for the research community for the last 50 years, since Casali et al.’s work on defining and estimating mental effort [19]. PL has been discussed since the mid 1980s and solved a long-standing debate between early and late selection in visual

---

1https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries
attention processes [50]. Both MWL and PL play an important role in driver safety and the driving experience in terms of auditory and visual distractions. In this paper, we analyze different methods for objectively estimating the driver’s MWL and PL (as separate constructs) while performing the dual task of controlling a secondary system while driving. We provide an end-to-end machine learning (ML) based framework for MWL and PL estimation separately using non-intrusive easily integrable sensors that are in concurrence with current sensors in modern vehicles [2, 47, 98].

2 BACKGROUND AND RELATED WORK

2.1 Mental Workload (MWL)

Mental workload is an important factor in human-machine interaction and has often been studied in the context of driving [9, 15, 16, 45, 63, 65, 101]. Although no universal definition of MWL has been proposed to date, several authors agree that MWL is an interaction of task demands, other environmental factors, and human characteristics (e.g., attentional capacities or task experience) [65, 96]. Hart and Staveland [45] state that “workload is a hypothetical construct that represents the cost incurred by a human operator to achieve a particular level of performance”. This means that a driver is expending mental effort to maintain a subjectively safe driving behavior [9], while various mediating factors are continuously changing (e.g., different road conditions [24]).

There have been several attempts to explain MWL, however, the multiple resource theory by Wickens [104] is the most widely used approach. It assumes that several modality-specific attentional resources are used when performing a mentally demanding task. In dual task scenarios, several mental resources (e.g., visual or auditory) can process information simultaneously. This suggests that driving could be defined mainly as a mental, visual, and physical task [65]. One task commonly performed secondary to driving is talking and dialing on a phone or talking to a passenger. The latter task uses more auditory and verbal resources, while the former requires visual and manual resources [15]. Caird et al. [15] showed that dialing a number has highly detrimental effects on driving performance increasing accidents risk [45]. A performance loss may even occur when the secondary task uses different modalities than the driving task. Thus, hands-free and handheld phone conversations both had a negative influence on driving performance, compared to a single-task condition. This means that even using different modalities, MWL still rises with higher task demand [15].

Similar effects could also be detected when interacting with an in-vehicle interface while driving [89]. Traditionally, MWL can be estimated in three ways: using performance measures on the primary task, using self-reported subjective measures, or using psychophysiological measurements such as electroencephalography (EEG), electrocardiogram (ECG), electromyography (EMG), pupillometry, etc. [26, 63, 65, 73, 96]. However, subjective measures are hard to obtain in real time as the driver’s tasks must be interrupted to report his mental state. Additionally, most of the previous work studying MWL in the driving context use the n-back paradigm as the secondary task [4, 70, 71, 90, 92]. The 1,2,3-back versions of the task (the 0-back version is designed for vigilance and does not induce mental workload [56]) induce continuous task demand on the working memory, like monitoring, updating information, and rule-based task decisions [82]. As the performance of the primary task in dual task conditions is modulated by the modality of the secondary task [15, 104], we induce MWL by a secondary task with no attention interference on the driving (i.e., only MWL interference) through an auditory n-back task version.

2.2 Perceptual Load (PL)

Another factor affecting task demands is perceptual load. Lavie’s PL theory [58–60] states that perceptual capacity is a limited resource that is always involuntarily used. Thus, PL is the extent to which a task consumes available capacity. When perceptual capacity is exhausted due to a high PL, no additional task-irrelevant stimuli are processed, and an inevitable early selective process is triggered [93, 97]. When perceptual capacity is not exhausted, the remaining perceptual resources spill over to task-irrelevant stimuli. This leads to a late selective process that requires available cognitive resources [23]. When cognitive load is kept constant, a higher PL leads to less interference of irrelevant stimuli due to early processing, while a lower PL is related to a higher interference of stimuli. For the driving context, a higher PL would lead to a reduced awareness of unexpected stimuli (e.g., a deer crossing the street) or induce inattentional blindness as well as inattentional deafness [64, 91], possibly leading to a safety hazard in driving. Despite the load theory being relevant to different processes while driving, there is a lack of applied evidence for the model [76], specifically differentiating between PL and MWL. Murphy & Greene [75] manipulated PL separately from MWL via a visual search task and found significantly increased inattentional blindness and deafness in the high PL condition and higher reaction time to hazards, thus demonstrating that PL is an important factor in driving performance. Similarly, we implement a visual search task with variable difficulty. However, the visual search task is implemented on an in-vehicle HMI by varying the set size of stimuli to measure the effect of limiting available information, which is proven to be a factor influencing PL [59]. Furthermore, the visual search task mimics the task of navigating an in-vehicle HMI, providing ecological validity.

As far as we know, there is also a lack of studies aiming to connect PL to psychophysiological measurements. Chen & Epp [29] attempted to separate PL and MWL as two distinct sources of task difficulty and found a significantly higher variation in pupil dilation for high PL, while Oliva [81] focused on the selection mechanisms of the load theory and found larger peak pupil dilation for the high PL condition.

2.3 HMI Design in Automotive Domain

Car manufacturers are continuously adding features to modern cars moving towards smart autonomous vehicles and ease of driving [3]. However, these features add further complexity to the driving environment [7, 62, 92]. Thus, recent research advancement is oriented towards more user-centered design approaches for in-vehicle interfaces in order to alleviate the mental effort accompanying these added features [17, 37, 54, 77, 80]. Consequently, multiple designs emerged for seamless non-intrusive in-vehicle interfaces [1, 10, 12, 32, 36, 39, 40, 55, 57, 94, 99]. While these previous approaches and studies focus on enhancing user experience and reducing drivers’ MWL through offline
pre-design feedback (e.g., gathering users’ requirements and designing a universal semi-customizable interface), others focus on real-time (and semi-real-time) approaches to obtain user feedback and adapt the interface based on the user’s MWL and stress levels. Hence, a method for estimating MWL through psychophysiological measures [16, 34, 48, 63, 65, 101] and PL is needed [75]. Researchers [70, 92, 96] classified MWL in field studies using two types of psychophysiological dimensions: Skin Conductance Level (SCL) and Heart Rate (HR). However, while a field study provides better external validity, it imposes a safety restriction that limited these studies. They mostly used a simple driving task to avoid accidents, which meant that the primary driving task induced minimal mental load. On the other hand, Barua et al. [4] classify MWL using both psychophysiological methods and driving performance in a simulated environment. They were able to introduce a high mentally loading task in the simulation environment and assess the MWL based on the driving performance. Unlike these previous methods, this paper focuses on both MWL and PL estimation using psychophysiological dimensions. Our contributions can be summarized as follows: 1) We propose an MWL and PL estimation approach assessed on a dual task scenario involving a non-trivial driving task. This MWL and PL estimation approach can be used for adapting interfaces, triggering warnings, or managing the transfer of control of semi-autonomous vehicles. 2) We select psychophysiological dimensions that can be realistically obtained in a non-intrusive way for a real driving scenario (e.g., with a non-wearable eye-tracker and heart rate sensors that currently exist in modern vehicles [2, 47, 98]). 3) We provide a novel multimodal open-source framework with an anonymized dataset for MWL and PL classification in a dual task driving scenario.

2.4 Empirical Hypotheses

In this work, we propose an end-to-end predictive ML-based classification framework for MWL and PL estimation. Furthermore, we evaluate the sensitivity of psychophysiological measures in the detection of elevated levels of MWL and PL, paving the way towards adaptive in-vehicle HMIs reactive to momentary changes in MWL and PL. We formulated four hypotheses for this work as follows.

- **H1**: Correlations between measurements from a similar physiological channel (e.g., heart data) should be higher than correlations between different physiological channels (e.g., heart data and pupillometry) [69]. Previous studies show that convergent validity is higher than discriminant validity for selected MWL measurements [69]. However, since these measurements reflect the same latent construct, we should find a significant inter-correlation.

- **H2**: Psychophysiological measures can distinguish different n-back task levels while driving in a lane-change task, reflecting a difference in MWL levels [71]. This distinction between MWL levels should be reflected in both statistical analysis and the predictive ML classifier that would be deployed in the vehicle’s system.

- **H3**: Similar to [90], MWL levels are distinguishable by the driving performance during the lane-change task.

- **H4**: Similar to the previous hypotheses, PL induced by visual search task affects both psychophysiological measurements and driving performance in the lane-changing task [20, 81]. Similar to MWL, the distinction between PL levels should be reflected in both statistical analysis and the predictive ML classifier.

3 METHOD

3.1 Participants

In total, 49 participants were recruited for the study. Four participants were excluded for the following reasons: two due to recording errors, one due to a cardiac disease, and one due to incorrect task execution. The remaining 45 participants (42.2% female, 2.2% non-binary) with a mean age of 30.07 years ($SD = 10.38$) completed the entire driving route while doing a secondary task. The majority of participants were of German nationality (62.2%). 48.9% had a high school degree, 24.4% had a bachelor’s degree, and 26.7% had a master’s degree. Regarding handedness, 88.9% of participants were right-handed. As for visual acuity, 37.8% had corrected eyesight in the form of contact lenses only for compatibility with the eye tracker glasses. Concerning participants’ driving experience, participants had their driver’s license on average for 10.20 years ($SD = 8.74$). Most participants reported driving often (13.3% daily, 44.4% several times per week, 17.8% several times per month), and 71.1% reported previous experience with automatic transmission cars, while only 37.8% reported previous experience with a simulator.

3.2 Design

We designed a within-subject dual task counterbalanced driving experiment in a simulated study [67]. The participants completed a primary lane-changing task (LCT) on a straight three-lane road along each of the secondary tasks. For the secondary task, the participants performed both a visual search task (PL-correlated stimuli) and an n-back task (MWL-correlated stimuli) separately. Each of the two secondary tasks was conducted in three different difficulty levels (easy, medium, and hard), resulting in a 2x3 within-subject factor design. The experiment was piloted to mitigate design issues. For the primary driving task, our LCT was adapted from [68], a procedure commonly employed in driving simulator studies [13, 46, 87]. The target lane was marked with an upward arrow (↑), while

---

2The number of participants needed for the study was calculated with G’Power [30], using an effect size of $f = 0.229$ from [69]. It estimated $n = 45$. 

---
the other two lanes were marked with an (X) symbol. The lane-change task was divided into 6 road segments: one segment per difficulty level per secondary task, and it lasted between 120 to 160 seconds based on the driving speed with a 90-second pause between segments to allow the workload to return to baseline. There were no vehicular traffic or environmental changes to allow confounding factors.

For the visual secondary task, a visual search task [59] was presented on a tablet on the right side of the steering wheel. This task simulates the distraction of interacting with the in-vehicle infotainment systems. The task difficulty was manipulated by the number of items on the display (i.e., set size). The levels consisted of a 2x2, 3x3, and 4x4 matrix of items (see Figure 2). Each level had 40 stimulus arrays. Each stimulus array was presented for 2000ms with a 1000ms pause in between. During this stimulus, the participant had to identify the existence of an L-shaped target item among T-shaped distracting items and respond by pressing an “OK” button only (see Figure 2). All the items were randomly rotated between 0° and 360°. L-shaped targets were present in 50% of the trials. As for the auditory secondary task, an n-back task was utilized where participants had to listen to numbers and respond when the current number matched a previously presented number. This task simulates the distraction of talking on a phone or to a passenger. Numbers from zero to nine were used as stimuli and presented in German through automated speech. Each number was presented for the three difficulty levels consisting of a 1-back, 2-back, and 3-back task. The task was to listen to each number and to respond by saying “yes” in German (i.e., “Ja”) when the currently presented number was the same as the number presented one, two, or three positions before it, respectively. The numbers were presented in total 40 times, as in the visual search task; however, targets were present in 25% of the cases. Responses were coded manually and were verified with an inter-rater reliability analysis [86]. Random samples were assessed by a second rater from a video recording. An intra-class correlation (ICC) was calculated with SPSS using a two-way mixed model yielding a value of $r = .96$ ($F(17, 17) = 33.29, p < .001$) constituting excellent reliability.

### 3.3 Apparatus and Procedure

The driving simulator setup consisted of a driver’s seat surrounded by three 55-inch LCD screens, two were rotated at a 45-degree angle towards the driver to allow for a wider field of view, as in Figure 3. The simulator vehicle was on automatic transmission and controlled via a steering wheel, and gas and brake pedals. Heart rate data was collected through a commercial ECG-based sensor attached to the chest (Polar H10) and eye data was collected by a head-mounted Pupil Labs Core eye-tracker [52]. Participants were introduced to the sensors and completed a short training in the LCT and secondary tasks. Then, they completed the designed dual task procedure while instructed to prioritize the secondary tasks. Both the difficulty levels of each secondary task as well as the order of the tasks themselves were counterbalanced. Finally, they filled out a post-experiment questionnaire and received monetary compensation.

### 3.4 Data Preprocessing and Measurements

In this work, we rely on two non-intrusive psychophysiological (ECG and pupillometry) measurements, as well as primary and secondary task performance, to estimate the driver’s MWL and PL. These are described as follows.

**Psychophysiological Measurements.** An electrocardiogram (ECG) measures the electrical signals from a person’s heart to determine the heart rate (HR) in beats per minute. Heart rate variability (HRV) [74] is defined as the variation of the peak-to-peak time interval (RR interval) between two consecutive heartbeats. According to several previous studies, average HR and HRV can be used to estimate MWL levels [22] where an increase in MWL level would increase HR values while decreasing the HRV values. Multiple features can be extracted from both HR and HRV over the time interval, indicating the MWL and PL level. For HR, we used average HR, minimum, maximum, and standard deviation. As for HRV, there are several features divided into three categories: time domain, frequency domain, and non-linear features. However, we used only the time-domain-related root mean square of the successive differences between RR intervals (RMSSD) feature, which is the most used feature for mental load level estimation with HRV.

Pupillometry is defined as the study of pupil size (i.e., diameter) and reaction. According to [5, 6, 81, 100], it is possible to estimate MWL and PL levels using pupil diameter. However, pupil diameter is not lighting invariant, which makes it unsuitable for a generic light-independent setting [100]. Hence, researchers attempted different preprocessing steps that would result in a light-invariant pupil-dependent method for mental load estimation, such as the index of cognitive activity (ICA) [66], the index of pupillary activity (IPA) [28], and the low/high index of pupillary activity (LHIPA) [27]. While ICA is the most common method in previous studies, it is a patented approach. Therefore, Duchowski et al. developed the IPA and LHIPA as an open-source alternative for ICA. The IPA algorithm is constrained to a fixed-head setting only, which does not apply to a generic use-case such as driving, while LHIPA is an enhancement to the IPA algorithm with fewer constraints and more applicability. In our study, we use LHIPA as an input feature

![Image 2](image2.png)

**Figure 2:** The stimulus array for the visual search task per difficulty level. From left to right are the 2x2, 3x3, and 4x4 matrices.

![Image 3](image3.png)

**Figure 3:** Driving simulator setup overview showing the lane-changing task and the secondary visual search task.
to our algorithm. We calculate it for each eye separately based on their respective pupil diameter variations.

*Primary and Secondary Tasks Measurements.* For the driving performance, the participant’s position in the driving task was compared to a pre-calculated ideal path in terms of lateral deviation in meters with a 33 Hz sampling rate. A standardized procedure (ISO 26022:2010) commonly used to access the LCT driving performance [13, 14, 21, 46, 68, 85]. The results were sorted into their corresponding tasks and different metrics including average deviation, median, minimum and maximum, and standard deviation. The ideal path was defined to be the center of the lane where the driver should drive except when close to lane changes. Pilot runs showed that drivers needed approximately 36 meters to complete the lane change when driving undisturbed at 60 km/h, which was considered when defining the ideal path. Figure 1 highlights the structure of the ideal route and its comparison to the actual driving. Since the vehicle was on automatic transmission, further analysis of gear changing, clutch pressure, and other vehicle-related measurements is not applicable. In the visual search task, performance was measured as the reaction time between stimulus onset and the pressing of the “OK” button. As for the n-back task, performance was calculated by subtracting false positives from correct responses and dividing the result by the total number of possible correct responses.

### 3.5 Classification Framework

The classification task aims to differentiate between different MWL levels for the performing the n-back task while driving and different PL levels for performing the visual search task while driving. Figure 4 shows an overview of the MWL or PL level classification approach. It starts with the data collection and preprocessing stage (discussed earlier), followed by the input features extraction stage, and ends with a model training and prediction stage to determine the MWL or PL level (based on the load-inducing task). Features extraction and the ML models are described as follows.

#### 3.5.1 Input Features and Output Classes.

To classify the different MWL levels based on the acquired measurements, we utilize machine learning (ML) techniques to learn from the features generated from these measurements and create a prediction model for real-time inference during deployment. Table 1 shows the generated possible input features per output class. In total, our model has eight possible input features calculated per output class. The classification model uses these input features to classify between two pairs of three output classes (i.e., easy, medium, and hard) for each n-back task (inducing MWL changes) and visual search task (inducing PL changes).

#### 3.5.2 Dataset Split.

The entire dataset covers 45 participants and is split into training, validation, and test sets using an adaptation of the nested cross-validation [11] approach with 5-fold cross-validation in the outer loop and a one-third holdout test set in the inner loop. Note that all the splits are done on the participant level and not on the data point level. This way, no data from the same participant is used in training, validation, or testing for the same model. This is to ensure external validity and model generalization while avoiding data leakage [53].

#### 3.5.3 Machine Learning Models.

For our machine learning architecture, we utilize Fuerer et al.’s [31] auto-sklearn python library for classification (based on scikit-learn [84]) that automates finding the optimal machine learning model and its corresponding hyperparameters. This approach automatically searches for the optimum machine learning model using Bayesian optimization methods in a structured search space consisting of the combination of 15 classifiers, 14 feature preprocessing methods, and 4 data preprocessing methods. It also utilizes an ensemble approach that combines several optimum machine learning models to enhance prediction accuracy [18]. In this method, we compare an ensemble approach to one of the top 3 best-performing classifiers selected by the Bayesian optimization process. The best-performing classifiers for our dataset were the k-nearest neighbors (KNN) [38], the linear discriminant analysis (LDA) classifier [61], and AdaBoost [33].

### 4 RESULTS

Results are divided into two sections. Section 4.1 shows the statistical results, while section 4.2 shows the results of the multimodal machine learning based MWL and PL classification models.

#### 4.1 Statistical Results

For the main statistical analysis, statistical significance is determined using a p-value below 0.05 unless otherwise mentioned. Extreme outliers were excluded; thus, the number of participants varies between analyses.

##### 4.1.1 Descriptive Analysis.

The psychophysiological data was evaluated for each condition of each secondary task and averaged across all conditions. The subjects showed a mean HR of 79.58 bpm ($SD = 12.57$), a mean HRV-RMSSD of 35.29 milliseconds.
we performed paired t-tests per secondary task. For the n-back task, the performance rate (correct response - false positives / total number of possible correct responses) was used to indicate effects of induced MWL per level. Participants were significantly better at the easy condition ($M = 0.96; SD = 0.07$) compared to the medium level ($M = 0.85; SD = 0.15; t(42) = 4.72; p < .001$) and were significantly better at the medium level compared to the hard level ($M = 0.36; SD = 0.21; t(41) = 11.86; p < .001$). For the visual search task, we used the reaction time as a suitable indicator of task performance. This was the case because the performance rate was close to 100% for most participants in all three levels (easy: $M = 0.99; SD = 0.02$; medium: $M = 0.99; SD = 0.02$; hard: $M = 0.95; SD = 0.04$). Participants reacted significantly faster in the easy level ($M = 1.29; SD = 0.16$) compared to the medium level ($M = 1.44; SD = 0.18; t(44) = -10.28; p < .001$) and significantly faster in the medium level compared to the hard level ($M = 1.75; SD = 0.22; t(44) = -16.78; p < .001$). This indicates that the manipulation check was successful, and the levels were sufficiently different.

Additionally, we performed two separate MANOVAs to test the effect of MWL and PL on HR and HRV-RMSSD individually. To test $H2$ for differences in MWL levels across various physiological measures, we calculated a MANOVA with the n-back task levels for HR and HRV-RMSSD. The n-back task yielded a highly significant effect for HR and HRV-RMSSD ($F(1, 68.72.28) = 32.19; p < .001; \eta^2_p = .428$; Greenhouse-Geisser corrected) and likewise for HRV-RMSSD ($F(1, 55.66.48) = 19.50; p < .001; \eta^2_p = .312$; Greenhouse-Geisser corrected). Helmert contrasts revealed that for HR and HRV-RMSSD, this effect is derived from the difference between the easy level compared to the medium and hard ones combined (HR: $F(1, 43) = 47.46; p < .001; \eta^2_p = .525$; HRV-RMSSD: $F(1, 43) = 24.35; p < .001; \eta^2_p = .362$). No significant effect was found between the medium level and hard level, neither for HR ($F(1, 43) = 0.68; p = .413; \eta^2_p = .016$) nor for HRV-RMSSD ($F(1, 43) = 3.19; p = .081; \eta^2_p = .069$). This means that participants showed a lower heart rate in the easy level compared to the medium and hard levels combined, while the HR and HRV-RMSSD remained the same in the medium and hard load levels as seen in Table 2. A second MANOVA was performed to test $H4$ and analyze the influence of the different PL levels (set sizes in the visual search task) on the HR and HRV-RMSSD. There was a marginally non-significant effect for HR and HRV-RMSSD ($F(1, 74.76.67) = 3.27; p = .050; \eta^2_p = .069$; Greenhouse-Geisser corrected). Notably, the Helmert contrast for HR revealed that there was a significant effect for the easy level compared to the medium and hard ones in HR($F(1, 44) = 10.39; p = .002; \eta^2_p = .191$), while the medium level compared to the hard one was not significant ($F(1, 44) < 0.01; p = .984; \eta^2_p < .001$). There was no univariate effect in HRV-RMSSD for PL ($F(1, 77.77.96) = 1.03; p = .360; \eta^2_p = .023$).

### Table 2: Descriptive analysis (Mean±Std) of the psychophysiological measurements and driving performance per condition.

|                           | N-Back Difficulty | Visual Search Difficulty |
|---------------------------|-------------------|--------------------------|
|                           | Easy              | Medium                   | Hard         | Easy              | Medium                   | Hard         |
| Heart Rate                | 77.49 ± 12.60     | 82.54 ± 14.17            | 82.97 ± 14.78| 77.29 ± 11.81     | 78.58 ± 12.00            | 78.63 ± 12.7 |
| HRV-RMSSD                 | 37.79 ± 19.50     | 31.86 ± 15.99            | 30.34 ± 14.42| 38.37 ± 18.53     | 36.52 ± 17.20            | 37.70 ± 19.36|
| LHIPA Right Eye           | 2.38 ± 0.50       | 2.28 ± 0.31              | 2.29 ± 0.43  | 2.46 ± 0.58       | 2.39 ± 0.56              | 2.44 ± 0.58  |
| LHIPA Left Eye            | 2.37 ± 0.51       | 2.34 ± 0.34              | 2.29 ± 0.30  | 2.3 ± 0.22        | 2.3 ± 0.24               | 2.3 ± 0.24   |
| Driving Performance       | 0.15 ± 0.08       | 0.21 ± 0.13              | 0.23 ± 0.16  | 0.24 ± 0.11       | 0.24 ± 0.12              | 0.27 ± 0.12  |

$(SD = 16.72)$, and a mean LHIPA value of $2.37 (SD = 0.25)$ for the right eye and $2.33 (SD = 0.16)$ for the left eye. Mean driving performance was $0.22 (SD = 0.24)$. Further mean and standard deviation details are shown in Table 2.

#### 4.1.2 Correlations and Reliability

To investigate $H1$ (convergent and discriminant validity), we calculated Pearson’s correlation for the averaged psychophysiological measurements and additionally for all conditions. Only HRV-RMSSD and HR were significantly correlated ($r(44) = -.53; p < .001$). The higher the HR, the lower the HRV-RMSSD. The average values of the LHIPA did not reveal any significant relationship with any of the other dimensions (all $p$ values > .177). The same pattern was found when comparing each dimension for each secondary task per difficulty level. HR and HRV-RMSSD were without exception highly inter-correlated, but the values for LHIPA showed almost non-significant interactions, whether with each other or with any of the heart rate measurements. The detailed correlations for each secondary task per difficulty level and per measurement can be found in the supplementary materials. For reliability analysis, we calculated Cronbach’s alpha to analyze the internal consistency of the data [25]. We performed an analysis for each of the five dimensions for the six conditions. HR and HRV-RMSSD yielded significant internal consistencies (HR: $\alpha = .98$; HRV-RMSSD: $\alpha = .98$). As expected by the correlations, the reliability for LHIPA was unacceptable (right eye: $\alpha = .36$; left eye: $\alpha = .31$). Therefore the reliability criteria for LHIPA was not met for H2 (MWL distinction) and H4 (PL distinction) and had to be excluded from the analyses. The internal consistency of the driving performance showed a Cronbach’s alpha of $\alpha = .33$, which is considered as unacceptable. Consequently, we excluded driving performance for further analysis in the statistical testing [8, 78], since the reliability criteria for driving performance was not met for H3 and H4.

#### 4.1.3 Effects of Mental Workload and Perceptual Load

To test if the manipulation of the study for MWL and PL was successful, we performed paired t-tests per secondary task. For the n-back task, the performance rate (correct response - false positives / total number of possible correct responses) was used to indicate effects of induced MWL per level. Participants were significantly better at the easy condition ($M = 0.96; SD = 0.07$) compared to the medium level ($M = 0.85; SD = 0.15; t(42) = 4.72; p < .001$) and were significantly better at the medium level compared to the hard level ($M = 0.36; SD = 0.21; t(41) = 11.86; p < .001$). For the visual search task, we used the reaction time as a suitable indicator
4.2 Machine Learning Results

In this section, we show the results of the machine learning analysis using different learning models. As mentioned earlier, we compare an ensemble-based prediction approach vs. a single-model, and we test them using a 5-fold nested cross-validation approach. This approach was implemented for the two secondary tasks: the n-back task to estimate MWL level and the visual search task to estimate PL level. Table 3 highlights the classification results of the n-back task (i.e., MWL classification). It can be seen that using an ensemble model approach is superior to the use of a single model approach for all input feature types. This suggests high complexity in the input features that are difficult to learn using a single model approach for the multi-class classification use case. Moreover, it can be seen that the model combining eye features (i.e., LHIPA) and driving performance features alone (without the heart rate features) outputs random chance accuracy. This is in line with the statistical analysis which shows that both eye features and driving performance are unreliable (according to the Cronbach’s alpha analysis). This can also be seen when comparing the “all features” model with the “heart data only” model as both have almost identical accuracy for all the applied learning models. Moreover, the statistical analysis also showed that only the 1-back task (i.e., low load) is statistically significant with higher load levels. This is also seen in Table 3 as the accuracy values are on average only 49% with a maximum of 63%. Therefore, a further binary classification model was implemented to classify between the two classes low load and medium load. Table 4 highlights the results for this binary classification which also aligns with the statistical analysis. It shows that eye features and driving performance features are performing only at a random chance level, while the heart rate data has an average classification accuracy of 72.2%, up to 89% for a single fold. Additionally, it can be seen that using ensemble models is comparable to the use of a single model approach. This suggests that the difference in heart rate features is more pronounced for the binary classification approach compared to the multi-class one. As for the visual search task (i.e., PL classification), all the machine learning models resulted in random chance accuracy. This result also aligns with the statistical analysis that shows minor statistical significance, although the manipulation check was successful. This would suggest that while the visual search difficulty levels were comparable to each other, they had a small effect on the selected psychophysiological dimensions. Similar to the MWL classification, a binary classification approach was attempted. However, it did not improve the accuracy values compared to the multi-class approach.

5 DISCUSSION AND LIMITATIONS

5.1 User-Centered Design Implications

This study aims to evaluate the sensitivity of differing psychophysiological dimensions for MWL and PL as well as to build a predictive classification model that differentiates between different levels of MWL and PL. Our framework uses minimally invasive and widely used psychophysiological measurements that exist in modern cars [2, 47, 98] to classify MWL and PL levels towards a user-centered design approach that accommodates the driver’s mental capacities. We mimic the driver’s interaction with the vehicle’s interface while driving through the auditory n-back task and the visual search task, which are equivalent to a speech-based and a touch-based interaction respectively. While our work focuses on this general dual task approach of manipulating an interface while driving, it is also applicable to more specific applications such as transfer-of-control scenarios for automated and semi-automated vehicles [79]; simplified interfaces, or personalized warnings for mentally exhausting situations [88]; and increasing the user’s trust through system awareness and transparency [44]. Thus, an adaptive personalized interface observant of the driver’s mental capacities can be implemented. For example, high MWL and PL classification can be utilized to simplify the interaction approach with the in-vehicle HMI while driving [88], or to alert the user to stop driving entirely if needed until mental or perceptual “cool-down”. Finally, while the n-back task is widely used to control MWL levels in the driving context [4, 70–72, 90, 92], it is limited in regards to construct, concurrent, and ecological validity [49, 51]. Therefore, future studies could investigate other MWL-inducing forms such as traffic or road manipulation [29, 83].

5.2 Sensitivity of Psychophysiological Measurements

We found a significant correlation between heart rate and heart rate variability, indicating a convergent validity for heart features. The analysis of the performance measures of the two secondary tasks revealed that the manipulation of three load levels resulted in three difficulty levels each. For the n-back task, there was a significant effect of MWL on the psychophysiological dimensions. This was the case for HR and HRV between the low load and the higher ones. The visual search task showed a non-significant effect for PL for the same dimensions. However, in follow-up tests, there was a significant effect for HR only between the low level and above for PL. Similarly, our classification model was able to differentiate between the low MWL (i.e., 1-back task) and the higher ones (i.e., 2 and 3-back tasks), aligning with the statistical analyses. However,
were able to analyze this hypothesis for heart features only. Both
As for H4 (vergent vs. discriminant validity). Since there was only one negative
reasons, this hypothesis could not be analyzed for the driving
being momentary and thus masked in the averaging process.
being pronounced enough, or the effect of the mental distraction
due to the measured deviation from the reference ideal route not
affects driving performance in the LCT. This hypothesis could not
dimensions for higher load levels [103, 104].
such that its effect can no longer be seen in the psychophysiological
our study, the driver’s MWL reached a level of mental saturation
contrast to our study that utilized a 1-, 2-, and 3-back task due to our
similar setup), they used a 0-, 1-, and 2-back task as manipulation in
not be identified. While Mehler et al. [70] were able to identify
states that the psychophysiological dimensions
states that MWL
have shown significant correlations with heart rate dimension in previous studies [69]. LHIPA (a non-
patented alternative) fails to do so in our study. This could perhaps
be due to light variations as LHIPA is still not thoroughly tested
for light invariance and needs further investigation as mentioned
by the authors [27]. Moreover, while LHIPA was tested on several
controlled environments, it was applied to a much less controlled
one in our study (e.g., extreme light variance, head movements, eye
movements). H2 states that the psychophysiological dimensions
should show differences in MWL. Due to LHIPA’s unreliability, we
were able to analyze this hypothesis for heart features only. Both
HR and HRV showed significant effects for MWL and were able to
distinguish between the low load level compared to higher ones;
however, an effect between medium and high load levels could
not be identified. While Mehler et al. [70] were able to identify
differences for MWL on three load levels (for heart rate using a
similar setup), they used a 0-, 1-, and 2-back task as manipulation in
contrast to our study that utilized a 1-, 2-, and 3-back task due to our
interest in extremely high load levels. Thus, we hypothesize that
in our study, the driver’s MWL reached a level of mental saturation
such that its effect can no longer be seen in the psychophysiological
dimensions for higher load levels [103, 104]. H3 states that MWL
affects driving performance in the LCT. This hypothesis could not
be tested, because the reliability of the driving performance did
not meet the minimum acceptable requirements. This could be
due to the measured deviation from the reference ideal route not
being pronounced enough, or the effect of the mental distraction
being just momentary and thus masked in the averaging process.
As for H4, it relates to the PL effect. For the same previously men-
tioned reasons, this hypothesis could not be analyzed for the driving
performance. As for the psychophysiological measurements, the
statistical analysis revealed that there is a non-significant effect for
the heart features. However, this does not necessarily mean that
PL does not influence psychophysiological data. Previous studies
showed that other psychophysiological measures (e.g., eye features)
can be influenced by the PL of a task [20, 81]. Since we measured
the effects of HR and HRV in a dual task scenario, it could be that
having to perform a PL task as a secondary task compared to a
single task is already load-inducing for HR and HRV. Therefore,
these dimensions are more affected by the difference between driving
without a secondary task and driving while using an HMI [89].
Despite that, we found a highly significant effect for mean heart
rate when comparing low and higher PL levels (with a high ef-
fect size). This might suggest that mean heart rate is affected by
PL, especially in lower PL levels. However, since there is a lack of
studies analyzing PL effects on heart data, it should also be further
investigated in future studies. Finally, participants’ demographics
were limited in our study. Drivers were relatively young (30.07
mean age) and experienced (10.20 mean driving years). This might
have a different effect on their MWL, PL and driving performance
compared to older and less experienced drivers [16, 42, 83].

6 CONCLUSION AND FUTURE WORK
Driving remains a complex task that is performed by hundreds of
millions of people every day. With the increase in the number of
distractors in a car, more personalized and adaptive in-vehicle inter-
faces are needed. In this paper, we have shown possible approaches
to analyze mental and perceptual loads as well as classifying their
elevated levels. Elevated load estimation is quite important in antici-
patating a driver’s distraction level, fatigue, stress, and inatten-
tiveness. In our analysis, we were able to show a high correlation
between heart rate data and elevation in mental workload. We
propose a machine learning based framework for workload level
classification through non-intrusive real-time measurements that
is technologically ready for use in terms of deployment and ap-
licability. Moreover, our work highlights a potential for further
investigation of important factors in driver distraction and cognitive
resource management.

ACKNOWLEDGMENTS
This work is partially funded by the German Ministry of Education
and Research (BMBF) under project APX-HMI (Grant Number:
01IS17043) and project CAMELOT (Grant Number: 01IW20008).

REFERENCES
[1] Abdul Rafey Aftab, Michael von der Beeck, and Michael Feld. 2020. You Have
a Point There: Object Selection Inside an Automobile Using Gaze, Head Pose

Table 4: N-back binary-class average test accuracy and standard deviation (i.e., Mean%±Std). The random chance level is 50%.

|               | All Features | Eye & Drive | Heart & Eye | Heart & Drive | Heart Alone |
|---------------|--------------|-------------|-------------|---------------|-------------|
| Without       | LDA          | 66.7 ± 11.7 | 58.9 ± 10.3 | 70.4 ± 10.4   | 68.9 ± 9.0  | 66.7 ± 6.1  |
| Ensemble      | KN trivial    | 72.2 ± 7.9  | 54.4 ± 8.2  | 61.4 ± 9.3    | 70.0 ± 10.9 | 62.2 ± 4.2  |
| AdaBoost      | 73.3 ± 5.4   | 48.9 ± 8.9  | 67.9 ± 13.5 | 65.6 ± 5.4    | 72.2 ± 7.9  |
| With Ensemble | 70.0 ± 9.7   | 53.3 ± 10.9 | 69.0 ± 10.8 | 70.0 ± 10.3   | 70.0 ± 12.0 |
What’s on your mind? A Mental and Perceptual Load Estimation Framework

M Brucher, M Perrot, and E Duchesnay. 2011. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12 (2011), 2825–2830. https://doi.org/10.5555/1953048.2078195

Matthew J. Pitts, Lee Skrypchuk, Tom Wellings, Alex Attridge, and Mark A. Williams. 2012. Evaluating user response to in-car haptic feedback touchscreens using the lane change task. Advances in Human-Computer Interaction 2012 (2012), 1–13. https://doi.org/10.1155/2012/598739

Priya Ranganathan, CS Pramesh, and Rakesh Aggarwal. 2017. Common pitfalls in statistical analysis: Measures of agreement. Perspectives in clinical research 8, 4 (2017), 187. https://doi.org/10.4103/picr.PICR_125_17

Da Tao, Haibo Tan, Hailiang Wang, Xu Zhang, Xingda Qu, and Tingru Zhang. 2019. A Systematic Review of Physiological Measures of Mental Workload. International Journal of Environmental Research and Public Health 16, 15 (2019), 1–23. https://doi.org/10.3390/ijerph16152717

Veerle Ross, Ellen M.M. Jongen, Weixin Wang, Tom Brijs, Kris Brijs, Robert A.C. Ruiter, and Geert Wets. 2014. Investigating the influence of working memory capacity when driving behavior is combined with cognitive load: An LCT study of young novice drivers. Accident Analysis and Prevention 62 (2014), 377–387. https://doi.org/10.1016/j.aap.2013.06.032

Christopher D Wickens. 2008. Multiple Resources and Mental Workload. Human Factors 50, 3 (2008), 449–455. https://doi.org/10.1177/001872008X288394

Willem B. Verwey. 2000. On-line driver workload estimation: effects of road situation and age on secondary task measures. Ergonomics 43, 2 (2000), 187–209. https://doi.org/10.1080/001401300184558

Anil Iyengar, Fernando De La Torre, Wende Zhang, and Dan Levi. 2015. Driver Gaze Tracking and Eyes off the Road Detection System. IEEE Transactions on Intelligent Transportation Systems 16, 4 (2015), 2014–2027. https://doi.org/10.1109/TITS.2015.2398031

Christopher D. Wickens. 2007. How many resources and how to identify them? Commentary on Boles et al. and Vidulich and Tsang. Human Factors 49, 1 (2007), 53–56. https://doi.org/10.1177/001872007779589019

Helmut T Zwahlen, Charles C Adams, and David P DeBals. 1988. Safety aspects of CRT touch panel controls in automobiles. In Vision in Vehicles II. Second International Conference on Vision in Vehicles. 335–344. https://trid.trb.org/view/927092