

Article

Multimodal Assessment of Mental Workload During Automated Vehicle Remote Assistance: Modeling of Eye-Tracking-Related, Skin Conductance, and Cardiovascular Indicators

Fabian Walocha, Andreas Schrank, Hoai Phuong Nguyen and Klas Ihme *

German Aerospace Center, Institute of Transportation Systems, 38108 Braunschweig, Germany

* Correspondence: klas.ihme@dlr.de

Abstract: Remote assistance for highly automated vehicles (HAVs), i.e., third-party assistance from support staff outside the vehicle in times of the need for assistance, presents a solution to extend the capabilities of HAVs by integrating a third party for decision making in uncertain situations. Similar to other control center positions, we expect the remote assistance tasks to exert high mental demands on the human operators. Therefore, we assessed impact of elevated mental workload during HAV remote assistance in a controlled environment in a user study ($N = 37$) with the goal of identifying cues to differentiate workload levels based on eye-tracking-related, skin conductance, and cardiovascular indicators. The results provide evidence that (A) elevated workload induced via a secondary task depreciates performance, and (B) we can identify workload levels person-independently as differences in tonic skin conductance ($F(2,72) = 24.538$, $p < 0.001$, partial $\eta^2 = 0.405$) and pupil dilation ($F(2,72) = 13.872$, $p < 0.001$, partial $\eta^2 = 0.278$), resulting in a classification accuracy of 58% in a three-class classification task. The results provide evidence that we are able to differentiate operator workload during remote assistance in a time-resolved way with the ultimate goal to provide adaptations to counteract task deficiencies.

Keywords: remote operation; autonomous vehicles; remote assistance; user state monitoring; mental workload; physiology; eye tracking

Academic Editor: Andrea Sanna

Received: 15 December 2024

Revised: 13 January 2025

Accepted: 15 January 2025

Published: 17 January 2025

Citation: Walocha, F.; Schrank, A.; Nguyen, H.P.; Ihme, K. Multimodal Assessment of Mental Workload During Automated Vehicle Remote Assistance: Modeling of Eye-Tracking-Related, Skin Conductance, and Cardiovascular Indicators. *Information* **2025**, *16*, 64. <https://doi.org/10.3390/info16010064>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Despite advances in automated driving technologies, it will likely still take several years until fully autonomous vehicles (SAE level 5 [1]) will populate our roads in large quantities. Yet, technology is expected to soon mature to render SAE level 4 vehicles a reality. Such vehicles can drive autonomously in defined operational design domains but may reach their system's limit at certain times due to different external or internal events. In these cases, level 4 vehicles should be able to transition to a safe state for passengers and freight without human intervention (e.g., in case of obstacles, adverse weather conditions, or dirty sensors). However, in order to continue a ride after such safety maneuvers, input of a human may often be necessary. Recent legal changes in some countries (e.g., Germany [2]) enabled vehicle operation concepts in which this human fallback operator does not have to be in the vehicle itself but may act from a remote operation center.

In this way, one person may take care of several vehicles of a fleet, resulting in a lower number of operators needed, which may be beneficial given the current and expected future shortage of professional human drivers in public transport and logistics.

In principle, remote operation of vehicles can be realized in two different ways: as remote driving or remote assistance [1,3]. Remote drivers actually steer the vehicle from a distance via standard interfaces, such as the steering wheel, throttle, and brake, in real time, whereas remote assistants rather provide high-level guidance to deal with certain situations, such as setting waypoints or giving clearance [1,3]. At the moment, the current legislation in Germany only permits remote assistance as an implementation for remote operation of highly automated vehicles on public roads [2]. Since workplaces for remote assistance, unlike those for remote driving, will likely not at all resemble the vehicle cockpit, these have only been conceived recently [4,5], and little is known about the actual challenges of remote operation in practice. A recent overview paper [6] compiled a list of human-factor challenges that arise when implementing remote operation. One crucial human-factor issue is to design the human–machine interface, including the information management in the workplace, in a way that the remote operator’s mental workload is kept in an optimal range. Mental workload can be seen as the relationship between the mental demands placed on an operator and their capacity to deal with these (e.g., [7,8]). For optimal task completion, avoiding phases of too high or too low a workload is desired because these often come along with performance degradation (for a review, see [7]). Analyses of similar tasks in control rooms like air traffic control have supported this claim. For instance, a recent survey study with professional air traffic controllers investigated operator states with potential negative effects on task engagement [9]. The study’s result listed different states of degraded engagement that result from different levels of workload as potential threats to efficient and safe operation. These include inattention, deafness and blindness, overload, and perseveration (high workload) as well as task-related and task-unrelated mind wandering (low workload) [9]. The authors of [6] proposed that one approach to balance the remote operators’ workload is the design of workload-adaptive interfaces that adjust the task allocation and information presentation to the operator based on their momentary workload level. Integrating workload-adaptive human–machine interfaces in the remote assistance workstation could therefore help to avoid phases of low or high workload. This measure could make the workplace of remote assistants safer and less stressful and thus make an important contribution to increasing the efficiency and safety of level 4 vehicles in various scenarios. Still, in order to design such workload-adaptive user interfaces, a better understanding of remote operators’ mental workload is needed. Especially, research on potential indicators for the operators’ mental workload that can be recorded and interpreted during task execution is needed.

Up to the present day, to the best of the authors’ knowledge, there are no studies on indicators for the mental workload of operators in the context of remote assistance for highly automated vehicles (HAV). From a human factors perspective, when assessing mental workload in workplaces, it is desired that the methods for assessing the workload are as little task-invasive as possible, meaning that they should not hinder the remote assistant in the task completion. Hence, sensors are most suitable if these are either contactless or wearable without restricting the freedom of movement of the operator at his or her workplace. Theoretical considerations as well as earlier studies from similar task settings imply that eye-tracking-related, skin conductance, as well as cardiovascular indicators have the potential to be used for assessing mental workload (e.g., [10–14]). Generally, we can identify three related but separate constructs that might be indicative of overall experienced workload.

(A) Cognitive load describes the amount of effort extended to process information and perform tasks. When focusing on demanding tasks, we find that operators reduce

scanning behavior and narrow their focus on selected information sources (often referred to as cognitive tunneling) [15]. Pupil diameter, fixation duration, as well as fixation dispersion seem to be relevant indicators for cognitive load (e.g., [16,17]).

(B) Mental stress is a concept related to regulatory processes that occur when trying to cope with incoming stressors (such as during work). Stress responses are often associated with activity of the autonomic nervous system, such as cardiovascular responses and changes in skin conductance [18,19].

(C) Mental fatigue describes a state of mental exhaustion after prolonged task periods that require mental effort. Fatigue is related to hypo-vigilance and drowsiness and is likewise related to ocular activity, such as droopy eyelids or increased blink rate [20–22].

Assessing these indicators satisfies the abovementioned criteria for being not task-invasive when (cable-free) wearables and cameras integrated into the workstation are used. A combination of different indicators is advised for the assessment of mental workload [23], and hence, a combination of indicators as a potential feature space for mental workload assessment in remote assistance of automated vehicles should be used.

Despite the paucity of research in the field of remote assistance of HAVs, there are several studies that have explored mental workload assessment based on eye-tracking-related, skin conductance, and cardiovascular indicators in other contexts. For instance, recent research revealed that the pupils are more dilated with higher workload in an air traffic control task [24] or control room operation [25]. Wanyan and colleagues [26] reported changes in heart rate variability, pupil diameter, and eyelid opening with changing mental workload levels in a flight task. Furthermore, in a simulated naturalistic driving task, the skin conductance level was increased at a higher experienced workload [27]. Similarly, in a simulated driving task, Foy and Chapman [28] reported a higher skin conductance level as well as changing eye movement patterns with increasing workload during driving. Unni et al. [29] found that heart rate and heart rate variability, among others, changed with increasing working memory load in the context of simulated manual driving. These insights from studies in related domains further outline that a multimodal assessment of mental workload is necessary because the particular demands of the task (in terms of cognitive load, mental stress, and fatigue) influence which features are indicative.

Based on the aforementioned considerations, this work has two research objectives (ROs). RO1 is to assess whether the effects of increased mental workload during remote operation manifest as eye-tracking-related, skin conductance, and cardiovascular changes of the remote assistant and whether we can measure these changes effectively. In order to do this, we examined a set of candidate indicators from the set of sensors used during the study and performed a group comparison of them over the relevant task conditions. Specifically, we utilized pupil dilation, fixation duration, fixation dispersion, eyelid opening, blink rate, tonic skin conductance level, heart rate, and heart rate variability as candidate indicators. For addressing RO1, we formulated and tested the following hypotheses:

Hypotheses 1: *We can infer the workload level of the remote assistant by observing differences in eye tracking, skin conductance, as well as physiological indicators. We expect that our indicators are affected by different levels of mental workload (H1.1—multi-variate relationship). In particular, with increasing workload, we expect an increased pupil size (H1.2), a prolonged fixation duration (H1.3), a lower fixation dispersion (H1.4), a lower average eye opening (H1.5), a higher blink rate (H1.6), an elevated average tonic skin conductance level (H1.7), an increased heart rate (H1.8), as well as a decreased heart rate variability (H1.9).*

RO2 is to build a predictive model on the set of candidate indicators to predict the condition label that manipulated low, medium, or high workload as a multi-class classification task. The performance of the predictive model was evaluated in terms of

classification accuracy as well as precision and recall for the different classes (low, medium, and high workload).

In this study, we used a variant of the well-established n-back task [30] as secondary task to induce different levels of mental workload. In the n-back task, the participants have to retain n items in their working memory so that the mental workload increases with a higher number for n. The task has been used in various studies in different application domains to induce mental workload (e.g., [29]). This task has the advantage that work load increases robustly with the number of items to be retained in the working memory. Therefore, the n-back task is particularly suitable for inducing mental workload in a controlled fashion.

2. Materials and Methods

2.1. Design

The study was conceptualized as a dual-task study using a 4 (primary task variant: none plus three different scenarios) $\times$ 3 (secondary task difficulty: none, N = 1, N = 2) within-participants design. The primary task consisted of a set of three prototypical remote assistance problems that needed to be solved using the user interface of the remote operator workstation. The secondary task for this study was an auditory n-back task.

The data analysis focusing on eye-tracking, skin conductance, and cardiovascular indicators of the operator's workload presented here is part of a larger study. An evaluation of the design of the operator's workplace based on the same participant cohort has already been published (see [5]).

2.2. Participants

Of the N = 41 participants who took part in this study, four had to be excluded due to technical issues in the data collection process for any of the relevant data sources (only participants with complete datasets were included in the analysis). Hence, 37 participants (six female) with an age range from 22 to 31 years (M = 25.9, SD = 2.3) were included in the data analysis. Of these, 68% had experience in monitoring technical systems (e.g., airplanes, automated vehicles, wind channels, agricultural robots, pumps, and machines). All participants had normal or corrected-to-normal vision. The affinity for technology of the participants was high (Affinity for Technology Scale, ATI [31]: M = 4.88, SD = 0.5; scale poles 1: low to 6: high), and all of them possessed a valid driver's license for passenger vehicles. All participants had a university or state-certified technician degree in one of the following disciplines: mechanical, automotive, electrical, aerospace, and aviation engineering (according to the requirements posed to the Technical Supervisor, the German equivalent of the RO, as specified in the German Autonomous Driving Act [2]). Most participants (89%) stated that they drive a vehicle at least multiple times per month, 38% reported to drive several times a week or more. All participants had heard about HAVs in the past.

The participants provided written informed consent before starting the study and were allowed to stop at any time without consequences. They received EUR 25 as financial compensation for taking part in the study. The study procedure was in accordance with the Declaration of Helsinki and approved by the institute's ethics committee.

2.3. Operation Center Simulator

For the study, we used the remote operations center as described and evaluated in [5]. It consists of seven screens organized in a 2 (row) $\times$ 3 (columns) array (six regular 24" computer monitors) with a 24" touchscreen in front (see Figure 1). The screens in the upper row showed a live view from the supervised HAV (a pre-recorded simulation created

in Unreal Engine for the study). The lower row consisted of a screen displaying details on the current tasks, a notification screen, as well as a map screen. The operator could interact with the respective HAV via the touchscreen, e.g., by giving clearance, setting waypoints, or selecting alternative routes. A detailed description of the operator workplace can be found in [5]. For the study, the operator workplace was set up in the IDeE.Lab [32] of the German Aerospace Center in Braunschweig, Germany.



Figure 1. Remote operation center simulator together with the study set-up. The tablet on the bottom right was used for presentation of the secondary task.

2.4. Primary Task

Three different primary task scenarios were used in the study, which were selected based on a scenario catalog for remote assistants [3]. The scenarios were implemented with Unreal Engine and extracted as video clips. For an overview of the three scenarios, see Table 1, and for details, refer to [5].

Table 1. Remote assistance scenarios as primary task for remote assistant (RA).

Scenario	Name	Description
#1	Detected situation unclear	The supervised HAV detects obstacle (puddle) and informs the RA. RA has to assess the situation via the camera view and give clearance for HAV to continue driving.
#2	Blocked lane	A parking vehicle blocks the lane of the HAV. HAV stops and informs the RA. RA analyzes the situation with cameras and sets waypoints to calculate new trajectory using the lane for oncoming traffic. RA also has to provide clearance.
#3	Rerouting	The road on the designated route of the HAV is closed. RA has to choose a route from suggested alternatives on the touchscreen.

2.5. Secondary Task

We used a n-back task as secondary task to trigger different levels of mental workload for the remote assistant. The participants had to compare a presented item (a digit between 1 and 9) with the item presented n steps before. It is expected that a higher n induces higher mental workload because more items have to be maintained in the working memory. Digits were presented as played audio recordings with an interstimulus interval of 5 s on a tablet computer, which was not part of the operator workplace. The display of the tablet gave a visual feedback on the remaining length of the response window to indicate the remaining time until the next stimulus was played. The participants were instructed to listen to the sound only and they had to respond verbally by saying “correct” (in case the items matched) or “incorrect” otherwise. The responses were logged manually by the experimenter.

2.6. Sensor Set-Up

A stationary infrared camera-based eye tracking system with four cameras (SmartEyePro, SmartEye, Gothenburg, Sweden, 120 Hz sampling rate) was used to record pupil and gaze data. The different screens of the operator workstation were integrated into the eye tracker’s world model to reference participants’ gazes with the screens. Physiological data were recorded using a sensor for electrodermal activity (EDA) (EdaMove4, Movisens, Karlsruhe, Germany, 128 Hz sampling rate) on the palm of the left hand and an electrocardiogram (ECG) (EcgMove4, Movisens, Karlsruhe, Germany, 512 Hz sampling rate) placed on the participants’ chest. Three out of the forty-one participants were left handed. Before the experiment, we were assured by all participants that they were adept at using a mouse with their right hand, which was standardized for all participants.

2.7. Measures

Based on the literature on the association between physiological and behavioral parameters and mental workload, we extracted a set of indicators from the eye tracking, EDA, and ECG systems. For an overview, of the parameters, see Table 2.

Table 2. Description of collected candidate indicators.

	Description	Unit	Associated with
Eyelid opening	Average vertical distance of upper to lower eyelid between both eyes	millimeters	Mental fatigue [22]
Blink rate	Relative time that participant is blinking	percentage	Mental fatigue [21]
Pupil diameter	Average diameter of pupil in both eyes	millimeters	Cognitive load [16,17]
Fixation duration	Duration of fixation when detected	milliseconds	Cognitive load [16]
Fixation dispersion	Distance between gaze locations during fixation	degree	Cognitive load [33]
Tonic skin conductance level	Cleaned skin conductance level from raw EDA	millivolts	Mental stress [34]
IBI	Inter-Beat Interval between R-waves of successive heart beats.	milliseconds	Mental stress [19,34]
RMSSD	Root Mean Square of Successive Differences in IBI, calculated over 1 min intervals	milliseconds	Mental stress [19,34]

In the beginning, participants filled in a questionnaire on basic socio-demographic information, frequency of car usage, knowledge about HAVs, experiences with remote

operation, as well as the Affinity for Technology Scale (ATI, [31]). After each trial, participants filled in the NASA Task Load Index (NASA-TLX [35,36]). The NASA-TLX is a widely used self-report measure for mental workload and assesses the facets mental demand, physical demand, temporal demand, performance, effort, and frustration on scales from 1 (low) to 21 (high). It has to be mentioned that in the course of the study, participants also filled in other questionnaires that are, however, irrelevant for our research questions (for details on questionnaires and results, see [5]).

2.8. Procedure

Participants were welcomed to the DLR campus, and then, they were guided to the laboratory by the experimental instructor. In the lab, they were seated at the work station and had to fill out a form detailing data protection guidelines, participant rights, agreement on secrecy, and details for the financial reimbursement. After filling out these necessary forms, they were given a tablet with a link to the abovementioned questionnaires. Next, participants were introduced to the relevant information on the concept of remote operations for highly automated vehicles and were introduced to the remote operator workplace simulator. Then, the physiological sensors were attached. During a short training phase, participants could work through each primary and each secondary condition in order to make sure they understand their tasks.

After the training and eye tracker calibration, the recording of physiological and eye tracking data began. For a participant-specific baseline reading, participants were instructed first to relax and look at the screen in front of them for approximately 5 min. Next, participants worked in sequence through each primary task in isolation and then through each secondary task in isolation (random order each). After this, participants had to accomplish the primary and secondary tasks in combination in randomized order. Secondary task blocks always lasted 3 min, while primary task length was variable. On average, participants took 31.7 s for scenario #1, 50.9 s for scenario #2 and 49.6 s for scenario #3.

After the experiment, the recording was stopped, and we removed the physiological sensors, debriefed the participants, and thanked them for volunteering. In total, the procedure took roughly 2.5 h.

2.9. Manipulation Check

As indication of whether our experimental manipulation to induce mental workload with the secondary n-back task was successful, we analyzed the data of the NASA-TLX questionnaire as well as the performance data from the individual participants. In particular, we compared the NASA-TLX ratings after the three different secondary task conditions with each other. Moreover, primary task performance (in terms of task completion and task initiation time) was compared between the three secondary task conditions. Finally, we compared the secondary task performance between the N = 1 and N = 2 conditions.

2.10. Processing of Physiological and Behavioral Indicators

2.10.1. Preprocessing

In order to prepare the comparison of the eye tracking, skin conductance, and cardiovascular indicators, all indicators were collected and timestamped on a centralized server to ensure accurate synchronization. After verifying data quality, we used a light data pre-processing pipeline for removing missing data points and truncating the timeseries according to the respective condition blocks. Indicator averages were then collected from the remaining data in each condition block. Finally, we applied a feature-wise z-scoring to the truncated timeseries of each subject to normalize the data for the following analysis.

2.10.2. Feature Selection

To determine whether the eye tracking, skin conductance, and cardiovascular indicators varied significantly across the secondary task conditions, we conducted a MANOVA to determine the multivariate significance of our indicator set (independent variable: secondary task condition; dependent variables: candidate indicator set according to Table 2). Then, we ran post hoc univariate ANOVAs for each indicator and post hoc paired *t*-tests across conditions for all univariate significant indicators. The set of indicators differing significantly between the task conditions were then selected as input for the subsequent predictive modeling task.

2.10.3. Model Training

Finally, the set of selected features was used as the basis to train a machine learning model to predict the workload condition. For this, we employed gradient boosting trees using the XgBoost package [37], as tree-based models do not require a common scaling of the data, generally work well with tabular inputs, and provide an intuitive measure of feature importance in the form of leave-split frequencies. We trained this classifier in a multi-class setting using repeated LOSO-cross-validation, using a validation split and random search to find the optimal hyperparameter set.

3. Results

3.1. Manipulation Check

An exhaustive overview of the results of the questionnaire data and performance indicators was already presented in Schrank et al. [5]. The proposed manipulation check aimed to verify that the task manipulation of the secondary task indeed elicited a higher subjective experience of the mental load and that the secondary task condition impacted task performances. The results of the descriptive statistics of reported workload and performance indices can be found in Table 3. We calculated univariate repeated-measures ANOVAs to identify the impact of the condition on each index (*p*-values for workload, completion time, and initiation time were corrected for sphericity using the Greenhouse–Geisser correction). The results are found in Table 4. We found a significant effect of the secondary task condition on the reported task load as well as all performance indices. Therefore, we considered the manipulation check successful and continued using the secondary task block conditions *no secondary* vs. *N = 1* vs. *N = 2* as proxy for the workload states *low workload*, *medium workload*, and *high workload*, respectively.

Table 3. Descriptive statistics on task load and performance indices.

	No Secondary Task	N = 1	N = 2
Subjective mental workload (NASA-TLX)	6.62 ± 1.9	8.87 ± 2.3	10.49 ± 2.4
Task completion time (in s)	43.874 ± 16.48	40.757 ± 12.69	47.534 ± 18.81
Task initiation time (in s)	6.952 ± 3.71	6.115 ± 2.26	6.021 ± 2.17
Percentage of correct secondary task answers		99.92 ± 0.02	95.1 ± 0.06

Table 4. Results of repeated-measures ANOVAs of the secondary task conditions on task load and performance indices.

	DoF (Nom)	DoF (Denom.)	F	<i>p</i> -value	Partial η^2
Condition~Task load	2	72	86.246	<i>p</i> < 0.001	0.706
Condition~Completion time	2	220	12.306	<i>p</i> < 0.001	0.101
Condition~Initiation time	2	220	4.479	<i>p</i> = 0.02	0.039
Condition~Secondary performance	2	147	66.555	<i>p</i> < 0.001	0.312

3.2. Feature Selection

An overview on the descriptive statistics for the eye-tracking-related, skin conductance, and cardiovascular indicators can be found in Table 5. The MANOVA revealed a statistically significant effect of mental workload on the entire indicator set (Wilk's lambda = 0.5358, $F(8,323) = 34.9826$, $p < 0.001$). Next, we calculated post hoc univariate repeated-measures ANOVAs for each indicator, correcting for multiple comparisons using Bonferroni correction. The results are presented in Table 6. The analysis revealed statistical effects for the pupil diameter ($F(2,72) = 13.872521$, $p < 0.001$, Partial $\eta^2 = 0.27816$) and tonic skin conductance level ($F(2,72) = 24.538449$, $p < 0.001$, Partial $\eta^2 = 0.405337$) (see also Table 7 and Figure 2).

Table 5. Descriptive statistics on candidate indicators stratified by tasks.

	Raw Means			z-Scored Means		
	No Secondary	N = 1	N = 2	No Secondary	N = 1	N = 2
Eyelid opening	8.68 ± 1.4	8.774 ± 1.3	8.729 ± 1.2	0.117 ± 0.4	0.155 ± 0.4	0.137 ± 0.4
Blink rate	5.456 ± 3.4	5.045 ± 3.0	5.527 ± 3.7	−0.089 ± 0.1	−0.102 ± 0.1	−0.086 ± 0.1
Pupil diameter	3.09 ± 0.3	3.12 ± 0.4	3.201 ± 0.4	0.077 ± 0.3	0.123 ± 0.3	0.23 ± 0.3
Fixation duration	787.306 ± 413.68	723.952 ± 466.98	815.858 ± 499.5	−0.045 ± 0.4	−0.034 ± 0.3	−0.063 ± 0.3
Fixation dispersion	0.476 ± 0.1	0.482 ± 0.1	0.471 ± 0.1	0.017 ± 0.1	0.018 ± 0.2	−0.005 ± 0.2
Tonic SCL	26.321 ± 15.1	27.824 ± 14.1	28.18 ± 13.4	−0.165 ± 0.8	0.577 ± 0.6	0.727 ± 0.7
IBI	877.742 ± 103.7	874.51 ± 123.3	854.567 ± 116.8	0.243 ± 0.5	0.207 ± 0.5	−0.013 ± 0.6
RMSSD	45.53 ± 29.2	43.322 ± 20.9	45.361 ± 26.4	0.033 ± 1.0	−0.069 ± 0.8	−0.098 ± 0.9

Table 6. Univariate ANOVAs for each indicator.

	DoF (num.)	DoF (denom.)	F-Statistic	Uncorrected <i>p</i> -Value	Corrected <i>p</i> -Value	Partial η^2
Difficulty~Eyelid opening	2	72	0.793612	0.46	1.0	0.021569
Difficulty~Blink rate	2	72	0.932778	0.40	1.0	0.025256
Difficulty~Pupil diameter	2	72	13.872521	6 × 10^{−6}	6 × 10^{−5}	0.27816
Difficulty~Fixation duration	2	72	0.246608	0.78	1.0	0.006804
Difficulty~Fixation dispersion	2	72	0.735317	0.48	1.0	0.020017
Difficulty~Tonic SCL	2	72	24.538449	7 × 10^{−9}	6 × 10^{−8}	0.405337
Difficulty~IBI	2	72	4.351637	0.016	0.13	0.107843
Difficulty~RMSSD	2	72	0.500438	0.61	1.0	0.01371

Table 7. Pairwise two-sided *t*-tests on conditions of significant features with Bonferroni correction for multiple comparisons.

	Secondary Task Conditions	<i>t</i> -Statistic	Uncorrected <i>p</i> -Value	Corrected <i>p</i> -Value	Bayes Factor	Unbiased Cohen's <i>d</i>
Pupil diameter	No					
	Secondary — N = 1	−2.042008	0.048	0.14	1.106	−0.247022
	No					
	Secondary — N = 2	−5.291565	5 × 10 ^{−6}	1 × 10 ^{−5}	3833.512	−0.921550
	N = 1 — N = 2	−4.037813	2 × 10 ^{−4}	7 × 10 ^{−4}	108.054	−0.638012

Tonic SCL	No Secondary—N = 1	−5.562489	2×10^{-6}	6×10^{-6}	8540.83	−1.387125
	No Secondary—N = 2	−6.228975	2×10^{-7}	7×10^{-7}	6.241×10^4	−1.580610
	N = 1—N = 2	−1.554943	0.13	0.38	0.516	−0.342061

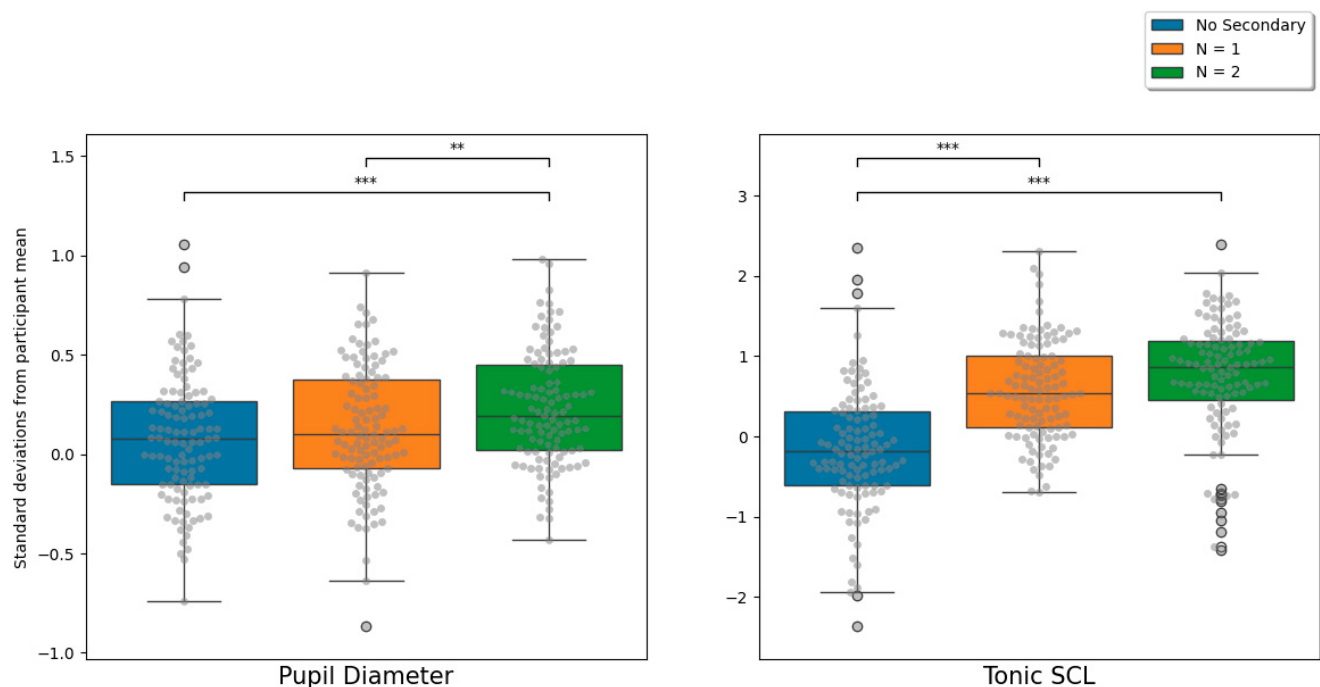


Figure 2. Boxplots showing the distribution of the univariate significant indicators pupil diameter and tonic SCL. ** $p \leq 0.01$, *** $p \leq 0.001$.

Using our set of significant univariate indicators, we calculated post hoc pairwise comparisons between workload conditions, correcting for multiple comparisons across conditions. The results are shown in Table 7 and visualized in Figure 2. We found significant differences between no secondary and N = 2 and between N = 1 and N = 2 for the pupil diameter. For the tonic skin conductance level, we found significant differences between no secondary and N = 1 and between no secondary and N = 2.

3.3. Model Training

Given our successful manipulation check, we can directly make inferences on the experimental condition as a proxy to the workload state of the participants. This results in a multi-class classification problem to be solved. In order to train a predictive model to make prediction on unseen data sources, we used a repeated nested leave-one-subject-out (LOSO) cross-validation with an inner-loop for hyper-parameter optimization to train the models. Then, we calculated confusion matrices and accuracy for each run, and the average of those yielded our estimate of the predictive power of our model. The parameters considered for hyper-parameter optimization can be found in Table 8. Across training loops, the optimal hyper-parameter configuration in the plurality of loops was {min_child_weight = 5, gamma = 1, colsample_bytree = 1.0, max_depth = 2} (see also Table 8). Figure 3 shows the results of the model training in the form of a confusion matrix. On average, we were able to differentiate between the three conditions with an accuracy of 57.66%. Given that the class labels are balanced, the baseline accuracy by predicting a random class was 33%. The precision values for the three classes were 0.66 for “no secondary”, 0.49 for “n = 1”, and 0.57 for “n = 2”. In contrast, the recall values were 0.65 for class

“no secondary”, 0.42 for “n = 1”, and 0.66 for “n = 2”. Using split frequency as a measure of feature importance, we found that the tonic skin conductance level was the most informative feature (71.34% split frequency), while the pupil diameter was less informative (28.66% split frequency). Collapsing the two classes “n = 1” and “n = 2” into a combined “workload” class to create a binary classification problem resulted in a classification accuracy of 77.17%. The precision values for the binary classification were 0.66 (no secondary) as well as 0.83 (workload), and the recall values were 0.65 (no secondary) and 0.83 (workload), respectively.

Table 8. Parameterizations considered for hyper-parameter optimization during training loops.

Feature Name	Feature Values
Min_child_weight	[1, 5, 10]
Gamma	[0.5, 1, 1.5, 2, 5]
Colsample_bytree	[0.6, 0.8, 1]
Max_depth	[1, 2, 3, 5]

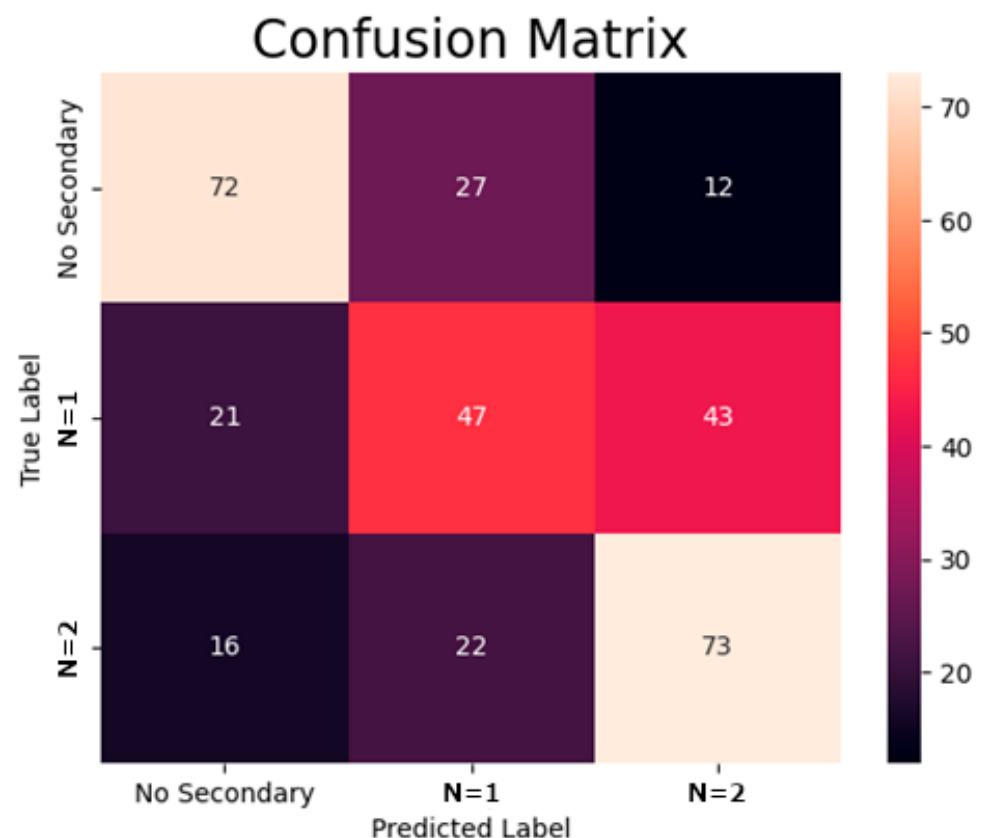


Figure 3. Total confusion matrix across all participants and training loops. The average accuracy during cross-validation achieved was 57.66% with a 33% baseline.

Lastly, given our trained model pipeline, we can use our trained model to make time-resolved predictions about the current workload state based on instantaneous predictions. To do this, we trained a boosting tree model using our set of best hyperparameters on the full dataset and predicted the time series on each point to annotate each time point with a predicted workload state. Finally, we used a sliding window approach to aggregate these predictions on 10 s intervals, yielding an assessment of continuous workload over a longer time period. The results of this approach are demonstrated on one exemplary participant in Figure 4. An exploratory visual analysis of this exemplary participant indicates that

workload tends to be highest at the beginning of task blocks. After some time, habituation seems to set in.

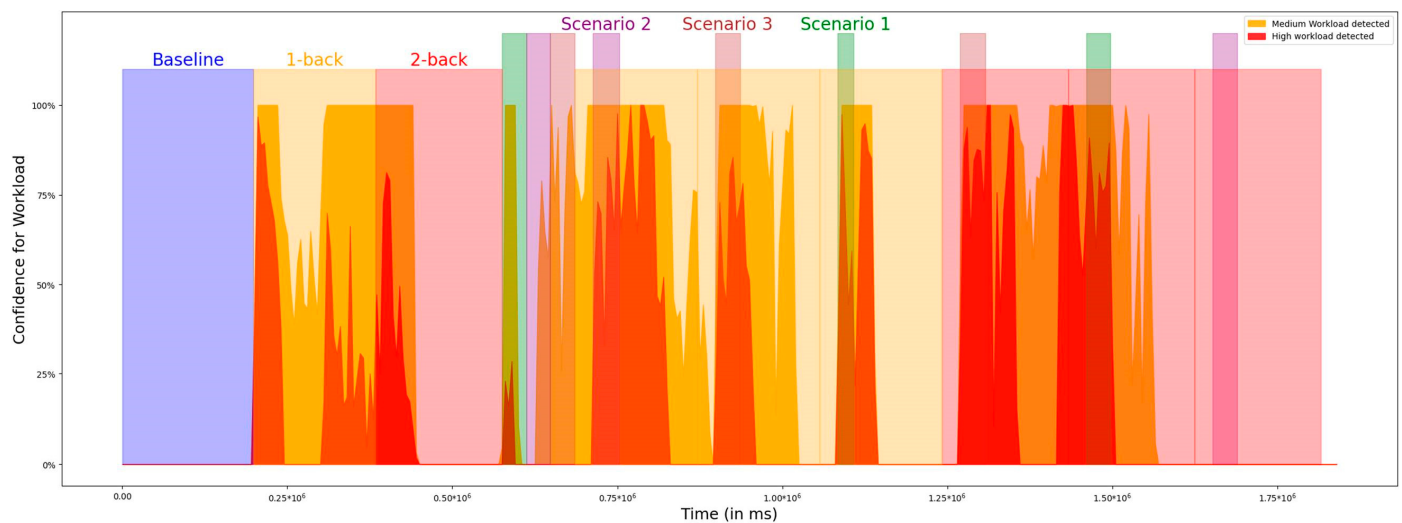


Figure 4. Workload prediction over 10 s intervals using a sliding-window approach for an exemplary participant.

4. Discussion

The aim of the presented study was to identify whether we can use changes in eye-tracking-related, skin conductance, as well as cardiovascular indicators to predict variations in mental workload via multivariate modeling. In order to do this, we collected self-assessments, task performance measures, and data from a set of predefined indicators. These collected data points were compared under various task conditions in a within-subject design across a variety of remote operator task blocks, during which a secondary workload-inducing task had to be solved. The results showed that, indeed, participants reported elevated workload in conditions with a more difficult secondary task, and their performance on both the secondary task and the remote assistance task decreased. Across these conditions, we also found that there is a significant multivariate effect on our indicator set, and we found significant differences for the tonic skin conductance level and the pupil dilation. Using a multi-class classification model, we were able to predict workload levels on unseen participants with an average accuracy of 58% with a 33% baseline.

By considering precision and recall for the three classes (low, medium, and high workload) of our classification problem, we can gain further insights for the application of the classifier in real workplaces. Practically speaking, we can interpret the results of the classification task in the following way: If our classifier returns “high workload”, we have a 57% certainty that the classification is accurate (as opposed to it actually being medium workload or low workload). If we are only interested in detecting elevated workload versus low workload (by combining medium and high workload predictions), the classification certainty when elevated workload is detected rises to 83% (corresponding to a binary accuracy of 77%). Depending on the required tolerance for a user-adaptive system using this classifier as a detection system, the behavior of the classifier can be further tweaked by adjusting the threshold on the classification probability for the active class, thereby achieving a higher precision at the expense of the total number of detections. This, however, has to be determined based on the exact purpose of the user interface adaptation and should therefore be subject for future studies.

As the focus of this work was to build a predictive model by using the most reliable indicators, our approach was to use a large set of candidate features and then use

hypothesis testing while accounting for multiple comparison to reduce this feature set. With this reduced set of indicators, classification accuracy is robustly above chance level. Furthermore, while no indicator associated with fatigue (blink rate and eyelid opening [21,22]) showed significant differences across secondary task conditions, this could likely be explained by the short duration of the experiment in comparison to total time an operator might be exposed to elevated task load throughout a work day. In order to provide a more accurate picture of fatigue effects on operator performance, future studies might focus on longer, repeated experimental task blocks to elicit these more naturally.

The induction of mental workload in this study was accomplished using a very controlled task, the well-established n-back task [30]. This has the advantage that our experimental set-up had an experimentally controlled character. However, this comes at the cost of ecological validity in task execution: Given the variety of tasks with different tasks requirements foreseen for the remote assistance of automated vehicles [3,4], it is likely that changes in the workload of the remote assistant will not solely result from changes in working memory load. Considering reports on indicative features for mental workload from other domains mentioned in the introduction [24–29], it appears necessary to validate the revealed findings with workload changes coming from more naturalistic changes in task demands. Therefore, future work needs to investigate how well the results from the controlled setting here transfer to more ecologically valid variations of mental workload, for instance, based on a stronger variety and complexity of tasks, decision points, and distractors. In addition, it may be worth studying how the effects found here transfer to other remote operation tasks, like remote driving (e.g., [38]).

For the current study, we employed a set of indicators that can be used with wearables or remote cameras (installed in the workstation) without task inference for the remote assistant. To improve the operator state assessment, it could be valuable to also include functional near-infrared spectroscopy (fNIRS) or electroencephalography (EEG) as further indicators for mental workload. Such technologies measure signals related to human cortical activity. With cortical activation patterns, the mental workload can be recorded directly where it arises, without having to determine it via physiology, which is of course also subject to the influences of movement, as is the case with the methods used here. Recent work has shown that EEG and fNIRS are promising for the assessment of mental workload in different domains (e.g., [29,39–41]). With the development of easier-to-use sensors for EEG and fNIRS [42–44], it will likely become more realistic that brain activity may be assessed without disturbing operators in their work.

5. Conclusions

We herein present a study on the effects and assessment of mental workload during remote assistance of HAVs. Our results indicated that pupil dilation as well as tonic skin conductance level may provide a relatively lightweight approach to assess mental workload during remote assistance. Based on such automated detection of mental workload, adaptive user interfaces for the operator workstation could be realized. These may then adjust the information management in the human–machine interface or the task allocation between different operators or the human operator and an automation according to the optimal operator workload. Future studies should investigate how the results transfer to changes in mental workload due to more naturalistic task demands.

Author Contributions: Conceptualization, F.W., A.S., and K.I.; data curation, F.W.; formal analysis, F.W.; investigation, F.W., A.S., and H.P.N.; methodology, F.W., A.S., and H.P.N.; project administration, F.W.; software, F.W. and H.P.N.; supervision, K.I.; validation, F.W.; visualization, F.W.; writing—original draft, F.W., A.S., and K.I.; writing—review and editing, F.W. and K.I. All authors have read and agreed to the published version of the manuscript.

Funding: The authors would like to thank all partners within the Hi-Drive project for their cooperation and valuable contribution. This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 101006664. The sole responsibility of this publication lies with the authors. Neither the European Commission nor CINEA—in its capacity of Granting Authority—can be made responsible for any use that may be made of the information this document contains.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by ethics committee of the German Aerospace Center (protocol code 05/22, date of approval: 14 June 2022).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The datasets presented in this article are not readily available because of the sensitive nature of the personal data collected for this work. Requests to access the datasets should be directed to the corresponding author.

Acknowledgments: The authors would like to thank the following people for their assistance during and after the study: Florian Rudolph and Nils Wendorff for helping to provide a technical framework for the study; Sarah Helweg and Thorben Brandt for their assistance during the data collection; as well as Chris-Leon Gorecki for their support in preparing the analysis.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. SAE International. Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles, J3016_202104, 2021 (J3016). Available online: https://www.sae.org/standards/content/j3016_202104 (accessed on 1 November 2024).
2. Deutscher Bundestag. Gesetz zur Änderung des Straßenverkehrsgesetzes und des Pflichtversicherungsgesetzes—Gesetz zum Autonomen Fahren, 2021. Bundesgesetzblatt (48). Available online: <https://www.gesetze-im-internet.de/stvg/BJNR004370909.html> (accessed on 25 October 2023).
3. Kettwich, C.; Schrank, A.; Avsar, H.; Oehl, M. A Helping Human Hand: Relevant Scenarios for the Remote Operation of Highly Automated Vehicles in Public Transport. *Appl. Sci.* **2022**, *12*, 4350. <https://doi.org/10.3390/app12094350>.
4. Kettwich, C.; Schrank, A.; Oehl, M. Teleoperation of Highly Automated Vehicles in Public Transport: User-Centered Design of a Human-Machine Interface for Remote-Operation and Its Expert Usability Evaluation. *MTI* **2021**, *5*, 26. <https://doi.org/10.3390/mti5050026>.
5. Schrank, A.; Walocha, F.; Brandenburg, S.; Oehl, M. Human-centered design and evaluation of a workplace for the remote assistance of highly automated vehicles. *Cogn. Tech. Work* **2024**, *26*, 183–206. <https://doi.org/10.1007/s10111-024-00753-x>.
6. Schrank, A.; Merat, N.; Oehl, M.; Wu, Y. Human Factors Considerations of Remote Operation Supporting Level 4 Automation. In *Road Vehicle Automation 11*; Meyer, G., Beiker, S., Eds.; Springer Nature Switzerland: Cham, Switzerland, 2024; pp. 111–125, ISBN 978-3-031-67465-5.
7. Young, M.S.; Brookhuis, K.A.; Wickens, C.D.; Hancock, P.A. State of science: Mental workload in ergonomics. *Ergonomics* **2015**, *58*, 1–17. <https://doi.org/10.1080/00140139.2014.956151>.
8. Wickens, C.D. Multiple resources and performance prediction. *Theor. Issues Ergon. Sci.* **2002**, *3*, 159–177. <https://doi.org/10.1080/14639220210123806>.
9. Migliorini, Y.; Imbert, J.-P.; Roy, R.N.; Lafont, A.; Dehais, F. Degraded States of Engagement in Air Traffic Control. *Safety* **2022**, *8*, 19. <https://doi.org/10.3390/safety8010019>.
10. Gramann, K.; Schandry, R. *Psychophysiologie: Körperliche Indikatoren psychischen Geschehens*; 4., vollst. überarb. Aufl.; Beltz PVU: Weinheim, Basel, 2009; ISBN 978-3-621-27674-0.
11. Gable, T.M.; Kun, A.L.; Walker, B.N.; Winton, R.J. Comparing heart rate and pupil size as objective measures of workload in the driving context. In *Adjunct Proceedings of the 7th International Conference on Automotive User Interfaces and Interactive Vehicular*

- Applications*, Nottingham, UK, 1–3 September 2015; Burnett, G., Gabbard, J., Green, P., Osswald, S., Pfleging, B., Kun, A., Eren, A., Antrobus, V., Eds.; ACM: New York, NY, USA, 2015; pp 20–25, ISBN 9781450338585.
12. Marquart, G.; Cabrall, C.; de Winter, J. Review of Eye-related Measures of Drivers' Mental Workload. *Procedia Manuf.* **2015**, *3*, 2854–2861. <https://doi.org/10.1016/j.promfg.2015.07.783>.
 13. Kosch, T.; Hassib, M.; Buschek, D.; Schmidt, A. Look into my Eyes. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, Montreal, QC, Canada, 21–26 April 2018; Mandryk, R., Hancock, M., Perry, M., Cox, A., Eds.; ACM: New York, NY, USA, 2018; pp 1–6, ISBN 9781450356213.
 14. Li, K.W.; Lu, Y.; Li, N. Subjective and objective assessments of mental workload for UAV operations. *Work* **2022**, *72*, 291–301. <https://doi.org/10.3233/WOR-205318>.
 15. Williams, L.J. Cognitive load and the functional field of view. *Hum. Factors* **1982**, *24*, 683–692. <https://doi.org/10.1177/001872088202400605>.
 16. van Gog, T.; Kester, L.; Nievelstein, F.; Giesbers, B.; Paas, F. Uncovering cognitive processes: Different techniques that can contribute to cognitive load research and instruction. *Comput. Hum. Behav.* **2009**, *25*, 325–331. <https://doi.org/10.1016/j.chb.2008.12.021>.
 17. Beatty, J. Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychol. Bull.* **1982**, *91*, 276–292. <https://doi.org/10.1037/0033-2909.91.2.276>.
 18. Ziegler, M.G. Psychological Stress and the Autonomic Nervous System. In *Primer on the Autonomic Nervous System*; Elsevier: Amsterdam, The Netherlands, 2012; pp. 291–293, ISBN 9780123865250.
 19. Taelman, J.; Vandeput, S.; Spaepen, A.; van Huffel, S. Influence of Mental Stress on Heart Rate and Heart Rate Variability. In *4th European Conference of the International Federation for Medical and Biological Engineering*; Magjarevic, R., Nagel, J.H., Vander Sloten, J., Verdonck, P., Nyssen, M., Hauelsen, J., Eds.; Springer Berlin Heidelberg: Berlin, Heidelberg, 2009; pp. 1366–1369, ISBN 978-3-540-89207-6.
 20. Dinges, D.F.; Mallis, M.M.; Maislin, G.; Powell, J.W. Evaluation of Techniques for Ocular Measurement as An Index of Fatigue and the Basis for Alertness Management, Washington, DC, USA, 1998. Available online: <https://trid.trb.org/View/647942> (accessed on 18 November 2024).
 21. Stern, J.A.; Boyer, D.; Schroeder, D. Blink rate: A possible measure of fatigue. *Hum. Factors* **1994**, *36*, 285–297. <https://doi.org/10.1177/001872089403600209>.
 22. Sundelin, T.; Lekander, M.; Kecklund, G.; van Someren, E.J.W.; Olsson, A.; Axelsson, J. Cues of fatigue: Effects of sleep deprivation on facial appearance. *Sleep* **2013**, *36*, 1355–1360. <https://doi.org/10.5665/sleep.2964>.
 23. Charles, R.L.; Nixon, J. Measuring mental workload using physiological measures: A systematic review. *Appl. Ergon.* **2019**, *74*, 221–232. <https://doi.org/10.1016/j.apergo.2018.08.028>.
 24. Rodríguez, S.; Sánchez, L.; López, P.; Cañas, J.J. Pupillometry to assess Air Traffic Controller workload through the Mental Workload Model. In *Proceedings of the 5th International Conference on Application and Theory of Automation in Command and Control Systems*, Toulouse, France, 30 September–2 October 2015; Feary, M., Feuerle, T., Rechea, C.G., Saez, F.J., Johnson, C., Martinie, C., Palanque, P., Pasquini, A., van Leeuwen, P., Winckler, M., Eds.; ACM: New York, NY, USA, 2015; pp. 95–104, ISBN 9781450335621.
 25. Bhavsar, P.; Srinivasan, B.; Srinivasan, R. Pupillometry Based Real-Time Monitoring of Operator's Cognitive Workload To Prevent Human Error during Abnormal Situations. *Ind. Eng. Chem. Res.* **2016**, *55*, 3372–3382. <https://doi.org/10.1021/acs.iecr.5b03685>.
 26. Wanyan, X.; Zhuang, D.; Zhang, H. Improving pilot mental workload evaluation with combined measures. *Biomed. Mater. Eng.* **2014**, *24*, 2283–2290. <https://doi.org/10.3233/BME-141041>.
 27. La Loeches De Fuente, H.; Berthelon, C.; Fort, A.; Etienne, V.; de Weser, M.; Ambeck, J.; Jallais, C. Electrophysiological and performance variations following driving events involving an increase in mental workload. *Eur. Transp. Res. Rev.* **2019**, *11*, 1–9. <https://doi.org/10.1186/s12544-019-0379-z>.
 28. Foy, H.J.; Chapman, P. Mental workload is reflected in driver behaviour, physiology, eye movements and prefrontal cortex activation. *Appl. Ergon.* **2018**, *73*, 90–99. <https://doi.org/10.1016/j.apergo.2018.06.006>.
 29. Unni, A.; Ihme, K.; Jipp, M.; Rieger, J.W. Assessing the Driver's Current Level of Working Memory Load with High Density Functional Near-infrared Spectroscopy: A Realistic Driving Simulator Study. *Front. Hum. Neurosci.* **2017**, *11*, 167. <https://doi.org/10.3389/fnhum.2017.00167>.
 30. KIRCHNER, W.K. Age differences in short-term retention of rapidly changing information. *J. Exp. Psychol.* **1958**, *55*, 352–358. <https://doi.org/10.1037/h0043688>.

31. Franke, T.; Attig, C.; Wessel, D. A Personal Resource for Technology Interaction: Development and Validation of the Affinity for Technology Interaction (ATI) Scale. *Int. J. Hum.-Comput. Interact.* **2019**, *35*, 456–467. <https://doi.org/10.1080/10447318.2018.1456150>.
32. Fischer, M.; Richter, A.; Plättner, J.; Temme, G.; Kelsch, J.; Assmann, D.; Köster, F. Modular and scalable driving simulator hardware and software for the development of future driver assistance and automation systems. In Proceedings of the Driving Simulation Conference 2024, Driving Simulation Conference 2024, Paris, France, 4–5 September 2014.
33. Sturman, D.; Wiggins, M.W. Drivers' Cue Utilization Predicts Cognitive Resource Consumption During a Simulated Driving Scenario. *Hum. Factors* **2021**, *63*, 402–414. <https://doi.org/10.1177/0018720819886765>.
34. Lazarus, R.S.; Speisman, J.C.; Mordkoff, A.M. The Relationship Between Autonomic Indicators of Psychological Stress: Heart Rate and Skin Conductance. *Psychosom. Med.* **1963**, *25*, 19–30.
35. Hart, S.G.; Staveland, L.E. Development of the NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Human Mental Workload*; Hancock, P.A., Meshkati, N., Eds.; Elsevier: New York, NY, USA, 1988.
36. Hart, S.G. Nasa-Task Load Index (NASA-TLX); 20 Years Later. *Proc. Hum. Factors Ergon. Soc. Annu. Meet.* **2006**, *50*, 904–908. <https://doi.org/10.1177/154193120605000909>.
37. Chen, T.; Guestrin, C. XGBoost. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016*; Krishnapuram, B., Shah, M., Smola, A., Aggarwal, C., Shen, D., Rastogi, R., Eds.; ACM: New York, NY, USA, 2016; pp 785–794, ISBN 9781450342322.
38. Meir, A.; Grimberg, E.; Musicant, O. The human-factors' challenges of (tele)drivers of Autonomous Vehicles. *Ergonomics* **2024**, 1–21. <https://doi.org/10.1080/00140139.2024.2346552>.
39. Shao, S.; Zhou, Q.; Liu, Z. Study of mental workload imposed by different tasks based on teleoperation. *Int. J. Occup. Saf. Ergon.* **2021**, *27*, 979–989. <https://doi.org/10.1080/10803548.2019.1675259>.
40. Tang, L.; Si, J.; Sun, L.; Mao, G.; Yu, S. Assessment of the mental workload of trainee pilots of remotely operated aircraft using functional near-infrared spectroscopy. *BMC Neurol.* **2022**, *22*, 160. <https://doi.org/10.1186/s12883-022-02683-5>.
41. Durantin, G.; Gagnon, J.-F.; Tremblay, S.; Dehais, F. Using near infrared spectroscopy and heart rate variability to detect mental overload. *Behav. Brain Res.* **2014**, *259*, 16–23. <https://doi.org/10.1016/j.bbr.2013.10.042>.
42. Uchitel, J.; Vidal-Rosas, E.E.; Cooper, R.J.; Zhao, H. Wearable, Integrated EEG-fNIRS Technologies: A Review. *Sensors* **2021**, *21*, 6106. <https://doi.org/10.3390/s21186106>.
43. Tsow, F.; Kumar, A.; Hosseini, S.H.; Bowden, A. A low-cost, wearable, do-it-yourself functional near-infrared spectroscopy (DIY-fNIRS) headband. *HardwareX* **2021**, *10*, e00204. <https://doi.org/10.1016/j.ohx.2021.e00204>.
44. Liao, L.-D.; Chen, C.-Y.; Wang, I.-J.; Chen, S.-F.; Li, S.-Y.; Chen, B.-W.; Chang, J.-Y.; Lin, C.-T. Gaming control using a wearable and wireless EEG-based brain-computer interface device with novel dry foam-based sensors. *J. Neuroeng. Rehabil.* **2012**, *9*, 5. <https://doi.org/10.1186/1743-0003-9-5>.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.