sensors

MDPI

Article

ADABase: A Multimodal Dataset for Cognitive Load Estimation

Maximilian P. Oppelt [1,2,*,†], Andreas Foltyn [3], Jessica Deuschel [3], Nadine R. Lang [1], Nina Holzer [3], Bjoern M. Eskofier [2] and Seung Hee Yang [4]

[1] Department Digital Health Systems, Fraunhofer IIS, Fraunhofer Institute for Integrated Circuits IIS, 91058 Erlangen, Germany
[2] Machine Learning and Data Analytics Lab (MaD Lab), Department Artificial Intelligence in Biomedical Engineering, Friedrich-Alexander-University Erlangen Nuremberg, 91052 Erlangen, Germany
[3] Department Sensory Perception and Analytics, Fraunhofer IIS, Fraunhofer Institute for Integrated Circuits IIS, 91058 Erlangen, Germany
[4] Artificial Intelligence in Biomedical Speech Processing Lab, Department Artificial Intelligence in Biomedical Engineering, Friedrich-Alexander-University Erlangen Nuremberg, 91052 Erlangen, Germany
* Correspondence: maximilian.oppelt@iis.fraunhofer.de
† Main contributing author.

Abstract: Driver monitoring systems play an important role in lower to mid-level autonomous vehicles. Our work focuses on the detection of cognitive load as a component of driver-state estimation to improve traffic safety. By inducing single and dual-task workloads of increasing intensity on 51 subjects, while continuously measuring signals from multiple modalities, based on *physiological* measurements such as ECG, EDA, EMG, PPG, respiration rate, skin temperature and eye tracker data, as well as *behavioral* measurements such as action units extracted from facial videos, *performance* metrics like reaction time and *subjective* feedback using questionnaires, we create *ADABase* (**A**utonomous **D**riving Cognitive Load **A**ssessment Data**base**) As a reference method to induce cognitive load onto subjects, we use the well-established *n*-back test, in addition to our novel simulator-based *k*-drive test, motivated by real-world semi-autonomously vehicles. We extract expert features of all measurements and find significant changes in multiple modalities. Ultimately we train and evaluate machine learning algorithms using single and multimodal inputs to distinguish cognitive load levels. We carefully evaluate model behavior and study feature importance. In summary, we introduce a novel cognitive load test, create a cognitive load database, validate changes using statistical tests, introduce novel classification and regression tasks for machine learning and train and evaluate machine learning models.

Keywords: cognitive load; affective computing; autonomous driving; machine learning; multimodal dataset

Citation: Oppelt, M.P.; Foltyn, A.; Deuschel, J.; Lang, N.R.; Holzer, N.; Eskofier, B.M.; Yang, S.H.; *ADABase*: A Multimodal Dataset for Cognitive Load Estimation. *Sensors* **2023**, *23*, 340. https://doi.org/10.3390/s23010340

Academic Editors: Soo-Hyung Kim and Gueesang Lee

Received: 29 November 2022
Revised: 21 December 2022
Accepted: 22 December 2022
Published: 28 December 2022

1. Introduction

The rapid development of novel sensor technology, powerful computing capabilities and methods using artificial intelligence has moved the prospect of autonomously driving vehicles into a potential candidate to transform the way people experience mobility. This development is a promising direction for traffic safety. Surveys [1–3] reporting the causes of traffic crashes in manual driving, have identified cognitive and emotional load among others as a major factor. However, until fully autonomous driving is available, the operator still needs to observe and, in critical situations, take control of the vehicle. To measure the required degree of manual interaction with the vehicle, the Society of Automotive Engineers (SAE) defines 6 levels of automation ranging from 0 (fully manual) to 5 (fully autonomous) [4]. While even lower level autonomous driving can reduce the complexity of traffic situations and therefore lead to task execution at levels with lower cognitive load [5], recent studies suggest, that inattention or distraction through additional tasks performed by the driver can lead to accidents for vehicles of level 1 (assisted driving), level 2 (partial

driving automation) and level 3 (conditional driving automation) [6,7]. When performing a secondary task, for example, talking to a car passenger while driving, inattention to the primary task, in this example, driving a car, can manifest itself on different levels: visual (not monitoring the road or the vehicle), manual (hands not on the steering wheel) or cognitive components [6], while visual and manual inattentions are important components for driver monitoring, we study the detection of phases with high cognitive states.

These states are of special interest, since engaging in dual- or multitask settings, a driver's attention is a finite resource and cognitive limitations are soon met [8]. The research community describes a subject's, cognitive state, with different terms like cognitive workload, mental workload, cognitive load or task load. We use Working Memory (WM), which is based on Baddeley's model [9], as storage of conscious information that is limited by the amount of information one can hold and process. Following this definition, we use Cognitive Load (CL) as a measurement to define the amount a subject puts on the WM during a task [10]. The definition of CL by [11] as a "... multidimensional construct representing the load that performing a particular task imposes ..." indicates that the prediction of when a person can attend to information (or process additional workload) is challenging.

One can study the concept of CL empirically, through meaningful measurements from four categories [12,13]. *Subjective measurements* can capture Perceived Mental Workload (PMWL) as shown in early studies, that used a mental-effort rating scale with 9-grades as symmetrical categories from very, very low mental effort (1) to very, very high mental effort (9) [14], while recent studies often use a multidimensional measure to evaluate perceived CL such as NASA-Task Load Index (TLX) [15–17]. Even though researchers have put a lot of work into developing these Likert-scale-styled subjective rating systems for different tasks, major issues remain: (a) subjective ratings rely on the participants' ability to introspect their current cognitive load [18], (b) the current task needs to be interrupted to answer the questionnaire and (c) subjective feedback can not be measured in real-time and continuously. *Performance measurements* are strongly related to the task imposed on the subject. Several metrics have been used in the past, such as the number of mistakes made during a complex learning task [19] or more general accuracy metrics such as hit-rate or reaction time, when solving standardized tasks such as *n*-back [20]. Others have used a dual-task load paradigm and evaluated the performance on the secondary task as an indicator of the cognitive load induced by the primary task [20]. In the area of driver monitoring, Engström et al. [6] reviewed the literature and identified inconsistent and counterintuitive findings, with an increasing, as well as a decreasing number of accidents, as a performance metric under increased cognitive load imposed by secondary tasks such as speaking while driving or listing to the radio. The measurement of CL using performance metrics is therefore limited by being strongly selective and task-dependent. A major advantage of performance metrics, such as response time of braking on brake light onset of leading cars, is, that these can be continuously measured during real-world driving situations [6]. *Behavioral Measurements* such as linguistic [21], speech [22–24], device inputs like pen [25,26] or computer-mouse movements [27] are commonly used in various applications, while metrics computed from behaviors such as lane keeping, steering wheel activity, mean vehicle speed or the number of initiated speed changes are more specific to driving [6]. A very prominent approach in the affective computing community is the extraction of action units from facial videos to detect emotions. These action units can also be useful to detect cognitive load while driving [28,29]. Another extensively researched modality is based on eye-tracking technology [30,31], that in addition to the behavioral dimension accounts to physiological measurements' with parameters such as pupil diameter. *Physiological Measurements*, include other properties of the eye, such as pupil diameter or pupil change responses [32]. Other physiological processes of the heart, muscles, lung, skin (temperature or conductance) and brain can be measured using biosignal acquisitions [33]. Both physiological and behavioral measurements have therefore the major advantage of being able to be recorded continuously. In addition to these biosignal recordings, biomarkers such as amylase or

cortisol, commonly measured in stress studies, may help to detect phases of high cognitive load [34].

Our attempt to measure cognitive load using multimodal data is influenced by other areas of research, that try to detect psychological states, especially contributions in the area of affective computing. Lisetti et al. have summarized early work in a review article, highlighting several contributions that show statistically significant changes in different elicited emotions [35]. Wang et al. attempted to recognize five emotional states of various drivers in an automotive simulator application setting [36]. With recent advances in machine learning over the last years, not only novel algorithms have been developed, but the requirement for high-quality data and therefore the collection and annotation process gained attention. Datasets such as DEAP [37], DECAF [38], cStress [39], ASCERTAIN [40], uulmMAC [41], AMIGOS [42] and WESAD [43] have been published in the affective sensing community.

In addition to the study of emotions and stress, other researchers already studied the concept of cognitive load while driving using multidimensional measurements. Haapalainen et al. studied the effect of various elementary cognitive tasks while recording the psycho-physiological activity using wearables [44]. Hussain et al. compared different modalities (videos of the face and electro-physiological measurements) for different cognitive load tasks and proposed a fusion technique of inputs [45]. Closely related to our contribution is the distracted driver dataset published by Taamneh et al., that records various driving conditions in a simulated environment [46]. CLAS [47] and MAUS [48] published datasets enabling the study of multidimensional data during different simulated cognitive load tasks. A noteworthy contribution, that shares a close relationship with our work on inducing cognitive load to drivers, is the eDREAM dataset [49,50]. A more practical approach for avoiding crashes through automatic measures is presented by Healey and Picard [51], that analyze cognitive load during real-world driving tasks using physiological signals. Another real-world driving study, conducted by Friedman et al. [52], tried to estimate cognitive load using only non-contact sensors during two experiments, a driving experiment and a version of an *n*-back task based on datasets, that have been introduced in [53,54].

Aside from the research gaps, described below and answered throughout this study, the analysis of related datasets makes it apparent, that issues with current data acquisitions setups and datasets remain:

- Some datasets are not publicly available to the research community. We release recordings of 30 subjects.
- Related work collected combinations of input modalities from subsets of relevant measures, while our work combines a diverse set of modalities form *physiological* and *behavioral* measurements, *subjective* questionnaires and *performance* metrics.
- In recent years, researchers introduced and studied various tasks, however, we as a research community failed to provide a database where the same subject participated in different tasks with different intensities of cognitive load in one recording session. Our setup fills this gap and therefore enables the analysis of distribution shifts and the evaluation of the robustness of predictors with representations that generalize across tasks and test the effects of subject-wise shifts during analysis.
- Missing information about metadata of the sample population like age, sex and personality traits, as well as the potential inclusion of subjects with undetected medical conditions or medication treatments, that might interfere with some measurements.

In this work, we focus on the estimation of cognitive load, collecting measurements of all four groups: physiological measurements, performance measurements, subjective measures and behavioral measures. For our subjective measurements, we employ the multidimensional NASA-TLX test. For performance measurements, we measure based on correct and incorrect hits recall and precision for our tests, as well as the reaction time. Our physiological measurements include Electrocardiography (ECG), Electrodermal Activity (EDA), Electromyography (EMG) Photoplethysmogram (PPG), respiration and

eye tracker data. For behavioral measurements, we use action units extracted from video recordings. We enrich these data records through a detailed analysis of our study cohort design. To explore the potential use of cortisol as a biomarker for CL, prominently used in stress studies, we collect salivary samples. Furthermore, our statistical analysis of various extracted statistical and medical-motivated expert features provides explainable measures of cognitive load. We train various machine learning models to predict CL of subjects while participating and observing a close-to real-world driving simulation of an autonomously driving car. This leads to our key major contributions:

- We implemented a simulation environment for autonomous driving software to induce different levels of cognitive load and a fully synchronous network of recording devices for multimodal measurements.
- We create a dataset that provides a wide range of physiological modalities, subjective ratings, performance metrics, behavioral data and biomarkers with precisely annotated phases of multiple levels of cognitive load.
- We conduct a robust statistical evaluation, and present various statistical and expert features with significant changes for multiple modalities.
- We identify several meaningful combinations of modalities to measure cognitive load using multimodal fusion techniques.
- We propose a novel continuous cognitive load value, combining subjective and performance measurements as a target for training.
- We release **A**utonomous **D**riving Cognitive Load **A**ssessment Data**base** (*ADABase*), containing 30 subjects to the public to enable the development of novel algorithms for multimodal machine learning.

In addition, research gaps in related works are clarified by providing a detailed analysis of 51 subjects with multimodal data recordings including modalities that are only briefly studied, such as action units and cortisol measurements at different levels of cognitive load. Furthermore, we combine multiple modalities that are until now, studied in subsets in related work, to a superset of modalities for cognitive load estimations, including remote measures such as eye tracking data and facial videos. We have created a new simulation with a close-to-real-world autonomous driving scenario, while also recording the extensively studied *n*-back test for easier alignment with related work.

2. Materials and Methods

To study the concept of cognitive load and develop machine learning algorithms, that utilize multimodal data for the detection of different cognitive load levels, this study focuses on the implementation of a fully integrated driving simulator. Our simulator includes the ability to record a wide range of physiological signals, face videos and eye tracker data, performance data, task-specific subjective feedback self-evaluations, personality traits and stress-related questionnaires while ensuring that all data points are synchronized across multiple acquisition platforms and modalities. In this study, we induce CL in two different ways. The first test is motivated by recent advances in the development of autonomously driving vehicles. We developed a test with subject/test-system interactions, that are similar to driver/vehicle interactions in lower-to-mid-level-autonomous vehicles. This test is introduced in Section 2.3 and contains different levels of cognitive load and the addition of a secondary task in the form of controlling an entertainment system while observing the vehicle. The second test is well-established in the research community: The *n*-back test introduced in Section 2.4 is conducted at three levels of difficulty and with single (visual) and dual-task workloads. However, due to the basic concept of our scenario and the conducted psychological assessment (*n*-back), commonly used in cognitive neuroscience to measure the working memory and its capacity, the results may generalize well to similar scenarios.

One of the many development contributions in this dataset is the usage of high-precision multimodal recordings and coordinated cognitive load simulation. The simulator for both tests is introduced in Section 2.2. In Section 2.1, we introduce statistics about our cohort design and relevant parameters such as driver experience. Our approach to measure the response of the hypothalamic-pituitary-adrenal (HPA) axis using salivary cortisol values is introduced in Section 2.5. Methods to evaluate subjective feedback are introduced in Section 2.6 and performance measures and metrics are described in Section 2.7. Section 2.8 describes our physiological measurements equipment and recording setup, including relevant references to current literature and a list of extracted expert and statistical features. The same information is provided for eye tracking data in Section 2.9 and behavioral data extracted from videos in Section 2.10. The handling of artifacts is described in Section 2.11. Before using the extracted expert features, motivated by medical practitioners and affective sensing literature, for the detection of phases with different levels of cognitive load we introduce our statistical evaluation protocol in Section 2.12. In Section 2.13, we introduce machine learning tasks using different sets of modalities for simple binary classification between low and high levels of cognitive load. Additionally, to this simple binary classification, we propose a three-class under-to-overload classification task and use subjective feedback and performance metrics to develop machine learning algorithms with a continuous cognitive load level as output. Our machine learning pipeline and algorithms are introduced in Section 2.14.

2.1. Participants and Cohort Description

The demographic data of all participants was acquired through self-reporting. Reported parameters were age, sex, weight, height, the highest obtained educational degree, state of employment, first language, state of driving license, as well as the duration of driving experience and handedness. Subjects with medical conditions or subjects under medications, that are known to have an impact on behavior, cognition, physiology or the HPA axis analyzed in this experiment, have been excluded upfront.

The acquired dataset consists of 51 (24 female, 26 male subject and 1 subject that did not want to state gender and age) subjects with an average age of 26.53 ± 5.93 years, where the youngest subject is 18 years and the oldest subject is 42 years old. The distribution of male and female participants is visualized in Figure 1. We measured body weights between 50.0 and 108.0 kilogram, with a mean weight of 71.2 ± 13.5 kg and a body height between 1.58 and 1.93 meters with a mean of 1.75 ± 0.09 m. We computed the Body-Mass-Index (BMI) yielding 22.9 ± 2.8 kg/m^2 in average, with a minimal value of 18.4 and a maximal value of 30.9 kg/m^2. Following the World Health Organization (WHO) expert consultation classification recommendations for adults, described in [55], we identify 1 subject as underweight (BMI < 18.5), 37 are in normal range ($18.5 \leq$ BMI < 25), 12 subjects are pre-obese ($25 \leq$ BMI < 30) and one participant was obese (BMI ≥ 30). We publish this meta information with our dataset to enable future work to develop algorithms that take the subject meta information into account. Eight of 24 female subjects were on contraceptive medications. Two subjects were using Levothyroxin to treat hypothyroidism. As our study involves various driving scenarios, we asked the number of years, since the driving license was acquired. Subjects without driving licenses were set to zero years of experience, yielding an average driving experience of 8 years of all participants. The detailed distribution is shown in Figure 1. Of 51 participants, 49 were right-handed and two left-handed. The experimentation setting was adjusted to match the handedness. The subjects were recruited from a diverse population of employees at the Fraunhofer Institute for Integrated Circuits and psychology students at Friedrich Alexander University Erlangen-Nuremberg. One complete session in our simulator, including preparation, measurements and debriefing took, depending on the subject's compliance, around three hours, with an average signal recording time of 155 ± 13 min per subject. The data was acquired on workdays from Monday to Friday at different times of the day. All subjects gave their informed consent for inclusion before they participated in the experiment. The study

was approved by the Ethics Committee of the Friedrich-Alexander-University Erlangen Nuremberg. (Ethics-Code: 129_21 B) The complete experimentation recording was 135 h for 51 subjects. We acquire psychological profiles as suggested by Gjoreski et al. when evaluating mental workload through self-reported questionnaires [56]. To analyze the impact of personality, we utilize the Big Five Inventory (BFI) as a model for the description of personality. To reduce testing time we use a short version, called Big Five Inventory-Kurz (BFI-K) introduced in [57] as an assessment questionnaire. We report the results for all five traits: extraversion vs. introversion, agreeableness vs. antagonism, conscientiousness vs. lack of direction, neuroticism vs. emotional stability and openness vs. closeness.

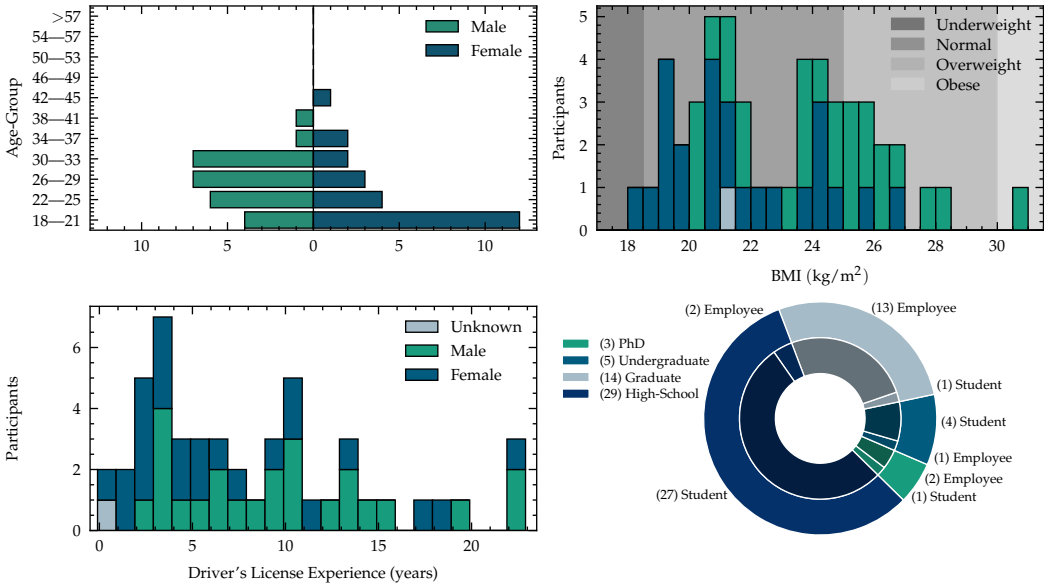

Figure 1. Demographic data was answered by subjects before participation. (**Upper-Left**): Pyramid with age groups over several subjects within that group (One subject did not answer the age and sex questionnaire). (**Upper-Right**): Histogram of BMI with highlighted classes according to WHO healthy lifestyle classification guidelines [55]. (**Lower-Left**): Histogram of driver experience in years. Subjects with no driver's license have been set to zero years. (**Lower-Right**): Pie-Chart with current occupation (student, employee) within different formal education (high-school, undergraduate, graduate, Ph.D.). Inner-Circle: Occupation, Outer-Circle: Highest Obtained Degree

2.2. Experiment Structure and Simulator Environment

The experimentation schedule consists of two main test phases, separated by a questionnaire phase to acquire self-reported subjective measures, targeting mental-well being, chronic stress and all demographic information reported in Section 2.1. The study setup was conducted in a controlled environment, to ensure, that no interruptions occur. An approximately constant room temperature was held during the completion of the experiments. To mitigate the effects of different body positions on parameters like Heart Rate Variability (HRV), the subject was in a sitting position in a car seat. The light conditions were the same for all subjects and no windows were present in the room (no sunlight). Issues with the circadian rhythm of the test subjects are taken into account, as all tests were conducted in fixed time slots, starting either at approximately 9 a.m. or 1 p.m. During this study we conducted two main tasks, in a randomized order:

- *n*-back Test: A established and commonly used continuous assessment of the Working Memory (WM), described in Section 2.4 [58].
- *k*-drive Test: An application task with multiple levels of cognitive load with a primary and secondary task,while observing an autonomously driving vehicle, described in Section 2.3.

During the complete experiment, the subjects were located in the center, sitting in an upright position in a simulator car seat (Playseat ®Evolution Pro Gran Turismo, Playseat

B.V. Innovatieweg 18-19, 7007 CD Doetinchem, The Netherlands). The user inputs were entered using the button controls of a steering wheel (Logitech G29 Driving Force, 7700 Gateway Blvd. Newark, CA 94560, USA). The different tasks were visualized using a multi-monitor setup (four 4K 55inch monitors) to enable an immersive experience and a simulation environment that is motivated by real-world car cockpits. All commands, instructions and custom tests were shown on the top monitor, while the three monitors below were used to show the driving simulation. The tablet was located on the side of the dominant hand. The test system was running depending on the active study phase, several custom applications, which were written in PsychoPy [59], exectued on a separate test system that is synchronized to the Biopac (BIOPAC Systems Inc., 42 Aero Camino, Goleta, CA 93117, USA) system, described in Section 2.8 using a Universal Serial Bus (USB) to TTL connection. A camera system, described in Section 2.10 and an eye tracking device, described in Section 2.9 were located above the steering wheel. The camera system is synchronized using analog triggers from the Biopac system and the eye tracker was connected to and synchronized with the PsychoPy test system. Images documenting the test environment are shown in Figure 2. An overview of the technical setup is shown in Figure 3.

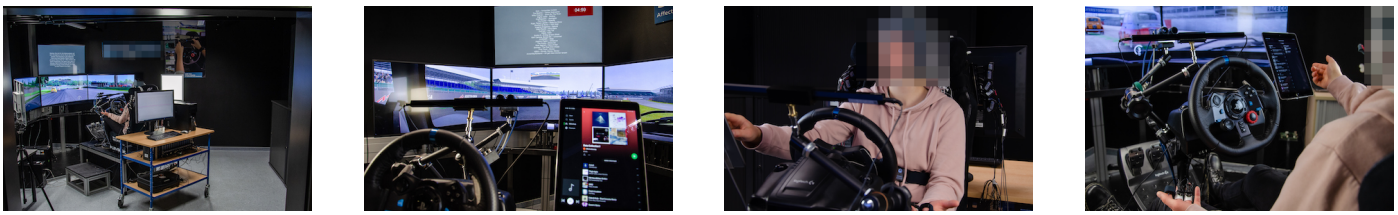

Figure 2. (**Left**)-to-(**Right**): Overview of the test environment with a depicted multi-monitor setup, subject in driving seat and acquisition setup; View from the subject's right shoulder perspective; View over the shoulder, with the steering wheel, camera, eye tracker and tablet cockpit; Frontal view of the subject fully connected to all bio-signals.

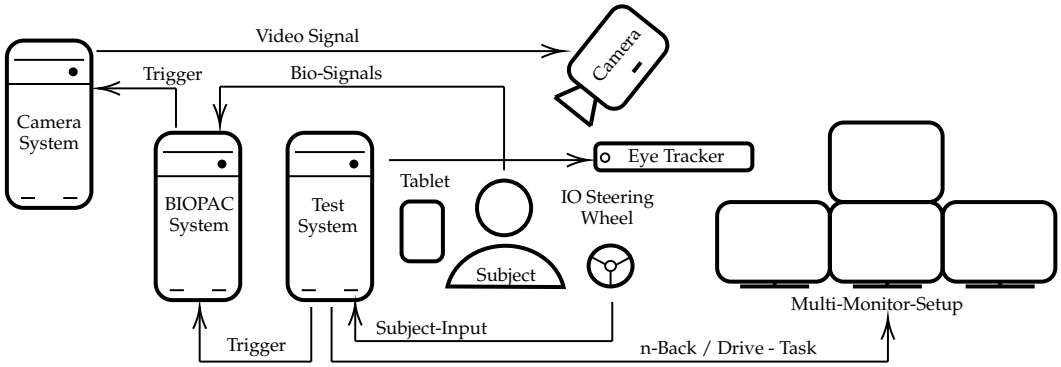

Figure 3. System components used during both experiments. Synchronization over multiple components mitigates effects such as clock drifts between different computer systems. The main test computer runs the experimentation software written in PsychoPy. The subject uses the buttons on a steering wheel as an input device during various experiments. The eye-tracker is synchronized with the main test computer. The camera system is triggered (on rising edge) with a constant low/high analog trigger signal having a frequency of 25 frames sent by the Biopac system. The camera is set to idle and automatically starts when the Biopac system is started at the beginning of the experiment and automatically stops as soon as the Biopac software stops. The Biopac system and the main test system are connected through a TTL line, sending the current state of the experiment at times of experiment phase transition as 255 bit-encoded signal, all interactions with buttons, as well as a continuous signal for synchronization.

The experimentation software is developed using PsychoPy version 2021.1.0. All experiments and user instructions are described in German language within the software. To mitigate any effects that can arise from personal interactions between the examiner

and the subject only a few interactions are required during the experimentation phase. At the beginning/end of the experiment, the subject is connected/disconnected to the Biopac system. During the questionnaire phase or before the driving test starts, the tablet is positioned by the examiner to ensure ergonomic accessibility. In addition to these interactions, the salivary samples are collected given the described schedule in Section 2.5. Deviations from the schedule are documented by the examiner on a predefined template.

2.3. Autonomous Driving Simulation

We are simulating a semi-autonomous driving experiment, where only little interaction of the subject with the vehicle is required and increasing levels of single- and dual-task load is put onto the subject. The steering, gas and brake control as well as gear shifting is autonomously controlled by the simulator. The subject's interaction is limited to three buttons, located on the steering wheel to detect three different maneuvers as a primary task. During level 1, the subject can use the ● button to indicate that the autonomously driving car is passing another car (overtaking). In level 2, the subject is additionally asked to indicate that the car is being overtaken, which can be detected by pushing the ■ button. For the last level (3), events of high acceleration and deceleration need to be detected by pushing the ✕ button. All driving sessions were simulated using the playback of a pre-recorded Assetto Corsa (Assetto Corso v.1.16, https://www.assettocorsa.it/en/, (accessed on 25 November 2022) KUNOS Simulazioni s.r.l., Via Degli Olmetti 39/B, 00060 Formello (RM), Italy) screen capture and were presented on the three lower monitors shown on the left side of Figure 4. During every session, the participant's car was driving on a racetrack (Assetto Corso Tracks: https://www.assettocorsa.it/tracks/ (accessed on 25 November 2022)), with twelve other cars that followed the road. The selected racetracks contained curvy and straight street segments without two-way traffic to simulate one-way highways. For easier reference, we are using the notation: k-drive, where k indicates the number of actions the user needs to react to and set $k \in \{1, 2, 3\}$ (Levels/Actions for reference: $k = 1$: ● $k = 2$: ● / ■ $k = 3$: ● / ■ / ✕).

Figure 4. Screenshot of experimentation setup during driving. The left image shows the split-screen visualized on the three monitors with driving instruments, such as position, back mirror, current gear and velocity. The right image shows the remaining time during the experiment, as well as a list of songs the subject is instructed to add to the playlist on the virtual tablet cockpit.

Similar to our dual-task n-back test (see Section 2.4), the subject is asked to solve a secondary task during level two and level three while observing the vehicle's actions. For our secondary driving task, we ask the subject to search and add songs to a playlist using a mobile music application (Spotify App v8.6.12, 03/2021, Spotify AB, Regeringsgatan 19, SE-111 53 Stockholm, Sweden), that is presented on an iPad (iPad version 2021 Apple Inc., 1 Infinite Loop, Cupertino, CA 95014) next to the steering wheel. The songs that need to be added are shown on the top screen in our multi-monitor setup shown on the right side of Figure 4. To ensure a correct understanding we provide a detailed description of every task to the subject upfront and propose solution strategies, such as using the current race position, the presence of rearview mirrors or a tachometer. In addition to these instructions presented in advance, the subject will have at least one training session of 50 s before the test level starts, which can be repeated if that is requested by the subject. Similar to the n-back training sessions the positive and negative hits are reported after every training session. Before starting the driving experiments, the subject will run through

three different baselines. During the first baseline, the subject will rest for five minutes and does not interact with the driving simulator, followed by a second baseline of two minutes, where the subject is observing the autonomously driving car in the test simulator without interaction or any imposed tasks. The third and last baseline phase (two minutes) is used to familiarize the subject with the music application, presented on the IPad. We run these baselines to ensure an ergonomic and familiar setting and record baselines for task-specific shifts, such as moving your eyes from the on-road screen to the iPad music application. We record positive (action was detected correctly) and negative hits (the wrong button was pressed, or no event occurred) as well as reaction time for our primary task. Additionally, we count the number of added songs that were added to the playlist for levels 2 and 3 for our dual-task experiment.

2.4. n-Back Test

In addition to our driving simulation, we conducted an *n*-back test [58] that is extensively used in the literature as a working memory paradigm. The experimentation scenario is inspired by [20]. In our single task schedule, we presented visuospatial stimuli as a blue square appearing at one of eight possible locations equally spaced and symmetrically around a white cross in the center on a black background. For dual-task tests, we additionally played recorded German consonants from the set {c, g, h, k, p, q, t, w} spoken by a male native speaker. Both versions are visualized in Figure 5. We repeated the test 6 times: 1-back to 3-back for visual-only stimuli and 1-back to 3-back for visual and audio-verbal stimuli, which were presented simultaneously for our dual-task workload paradigm. Prior to the complete *n*-back test run, the subject had different baseline phases. The schedule for the complete test run and the baseline phases is visualized in Figure 6. Each test was preceded by a short period of training that could be repeated on request by the subject, shown in Figure 7. Similar to Jaeggi et al. the stimuli were presented for 500 ms followed by an inter-stimulus-interval of 2500 ms [20]. The subjects were asked to react to repetitions using the buttons on the steering wheel. For a positive visual event, the subject was asked to press the and for an auditive event the , respectively.

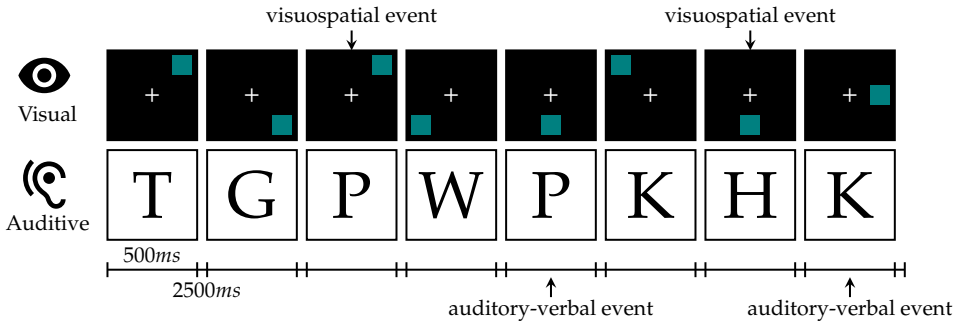

Figure 5. Example of visuospatial and auditory-verbal stimuli, presented during dual *n*-back test with level 2 (dual-task 2-back test) [20]. Stimuli are presented for 500 ms followed by an inter-stimulus-interval of 2500 ms. Example with two visuospatial and two auditory-verbal repetitions. The expected reactions are presented by arrows below and above the figure.

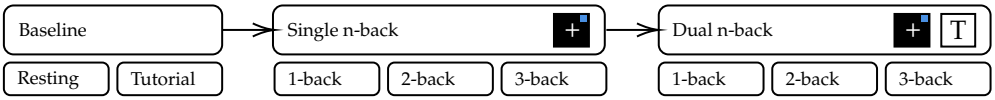

Figure 6. Visualization of the *n*-back test protocol. Baseline measurement consists of a resting phase of 5 min, followed by a 2 min tutorial, that shows stimuli and explains input controls to the subject. Two distinct phases of single and dual task *n*-back tests for $n \in \{1, 2, 3\}$ are presented to the subject.

We recorded both, positive hits (when the reaction to a stimulus was correct) and negative hits (no stimulus was played, or the wrong button was pressed). For positive hits, we recorded the reaction time from the beginning of the stimuli until the button was pressed. The evaluation protocol for positive/negative hits and reaction time is described in Section 2.7. The subject received performance feedback (number of correct and incorrect hits) automatically after each training session, visualized on the monitor, to ensure that the task was understood correctly, before starting the actual test. The system feedback was formulated neutrally, while the supervisor did not give any feedback during the complete study, to reduce the effects of perceived social stress.

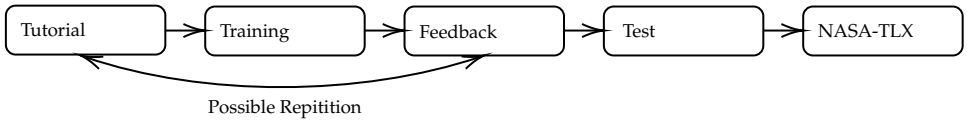

Figure 7. Protocol of a single *n*-back test sequence. The subject is able to repeat the tutorial and training phases. After training completion, the achieved score is reported to the participant (number of hits and false alarms are shown). Each test run is completed with a NASA-TLX assessment. The test phase takes 2 min to complete.

The appendix of this publication provides an overview of the experimentation sequence in Figure A2 of both conducted experiments.

2.5. Salivary Cortisol Responses

The measurement of glucocorticoids has proven to be a significant auxiliary value to detecting acute social stress responses using salivary samples [60,61]. Early studies identified an increase of cortisol, as a response to acute stress, transmitted by the HPA axis as a pathway to send signals to the organism. Cortisol is therefore a promising non-invasive biomarker for stress [62,63]. As our study does not impose any social evaluative thread but various levels of cognitive load, we are analyzing cortisol as a biomarker for our *k*-drive and *n*-back levels. Woody et al. have bisected the cognitive component and social evaluative component of stress and found that social evaluative thread in absence of cognitive load led to greater cortisol values while increased cognitive load had no main or additive effect [34] on cortisol. Contrary, Veltman and Gailard [64] have identified a high impact of different mental workload tasks and cortisol levels, while low task performance leads to higher cortisol levels. To capture the complete cortisol response [60,65], we measure four points in time for both tasks: Right before the test start (Pre-00), After finishing the test (Post-00), 10 min (Post-10) and 20 min (Post-20) after the test finishes.

2.6. Subjective Feedback

After every phase of the aforementioned *n*-back tests and drive scenarios, we ask the subject to self-evaluate and report the effort required to complete the given task. These subjective measures often vary. However, to a certain degree the self-reported ratings can reliably determine the CL put on a certain task [14,17,66]. In this work, we use the NASA-TLX, a well-established tool to analyze CL on a multidimensional, weighted scale [16]. We kept the dimensions as described by Hart [17]: mental, physical and temporal demand, performance, effort and frustration. The dimensions and associated questions were translated into German. Each dimension is also accompanied by a weighting factor through pairwise comparisons regarding their perceived importance. The final score is computed by a summation of the scores on each dimension times the corresponding weights (the number of times it was rated more important). The NASA-TLX score is then scaled to be between 0 and 100. In addition to the weighted version, we report the NASA-Raw Task Load Index (RTLX) without applying the weights. The overall score is computed using the average of all scores [15]. We collect three ratings for our single *n*-back test, three ratings for our dual workload *n*-back test and three ratings for our driving experiments. In addition, we evaluate perceived stress over the last month and therefore set a baseline for

our evaluations. We utilized the Perceived Stress Scale (PSS) by Cohen et al. [67] to enable respondents to indicate how stressful their life is. Analog to the work by Cohen et al. [67] the 5-scale ratings (from "Never" to "Very Often") are computed by summing scores of all items (including inversion for some questions). The respondents were able to express their positive and negative affect after the driving phase and the *n*-back test. Utilizing a 5-scale Positive and Negative Affect Schedule (PANAS) introduced in work by Watson et al. [68] with 10 positive feelings and 10 negative feelings, we report the affective state of the corresponding subject. The evaluation of affective states was conducted after both phases of this study. As PSS and PANAS questionnaires are used to measure a baseline stress level and affective states, but are not directly relate to specific phases of our experiment, we report the result in Appendix E.

2.7. Performance

Performance is commonly used to measure cognitive load as introduced in Section 1. Metrics to measure performance are highly task-dependent: We collect true positives (correct button pressed, TP), false positives (no event occurred, but the button is pressed, FP) and false negatives (the event occurred, but no input was recorded, FN) as well as reaction time for true positives. Both, for our *n*-back test, and our *k*-drive test, the maximal time between the beginning of an event and the time when the user could react to this event was set to 3 s. Similar to the nomenclature used in classification tasks we compute and report recall, precision and F_1-score. For this experiment we use:

$$\text{precision} = \frac{TP}{TP + FP} \tag{1}$$

$$\text{recall} = \frac{TP}{TP + FN} \tag{2}$$

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{3}$$

The same metrics are also computed for the secondary auditive *n*-back task. For the secondary tasks during levels 2 and 3 of *k*-drive we only collect the correctly added songs to the playlist (false positives are not recorded) and report only our recall-motivated performance metric.

2.8. Physiological Features

For data acquisition, we used a Biopac MP160 system and specialized Biopac modules ECG100C, EGG100C, EMG100C, EDA100C, PPG100C, RSP100C and SKT100C for the corresponding modalities. We used a sampling frequency of 2000 Hz for all modalities, as storage was not an issue in this simulator setting and the analog synchronization across multiple devices, modalities and time stamps is ensured to be consistent. For ECG recordings, we use the Einthoven, Lead I and Lead II with a limp sensor placed below the shoulder and lower torso using *Ag-AgCl* electrodes (torso limp positions: right arm, left arm, left leg and right leg as neutral). The skin was prepared with an isopropyl solution. The surface EMG was recorded to capture the activation of the trapezius muscle using three electrodes. The reference electrode is placed on the seventh cervical vertebra (C7), and the other two electrodes for measuring the activation are placed at the right lateral trapezius muscle. The EDA electrodes were placed on the palmar side of the index and middle fingers (digitus manus 2 and 3) at the medial phalanges. To compensate for changes in conductance at the beginning of the recording, the placement of the electrodes was done at least 10 min before the recording started. For skin preparation, only warm water was used. The PPG signal was measured on the tip of the middle finger (digitus manus 3). Skin Temperature (SKT) was measured using a sensor placed on the tip of the index finger (digitus manus 2). All, EDA, PPG and SKT were measured on the non-dominant hand, which was placed on the leg and could rest without movement as no interaction was

required. Respiration cycles were recorded using the BIOPAC MP3X respiratory effort transducer, measuring the change in thoracic circumference. During the early stages of our experimentation development, we recorded electroencephalographic recordings using a Mobita ConfiCab with 32 channels. We stopped the acquisition of these recordings during this experiment, as some subjects, reported headaches after the long duration of wearing the electrode cap.

We opted to extract features based on statistical signal properties and medical expert features for all biosignal modalities. Features extracted from electrocardiography recordings, reflecting the electrical activity of the human heart, are, except the heart rate itself, based on the concept of Heart Rate Variability (HRV). The HRV is impacted by modulations of biochemical processes, that correspond to the activation of the parasympathetic and sympathetic nervous systems. In addition to recordings of the electrical activity of the heart, we measure the mechanical activity using PPG sensors. Studies have extracted features based on the variability of the pulse waves, so call Pulse Rate Variability (PRV) features. In this publication, we are not interested in the estimation and computation of PRV features, because they have a strong correlation with HRV features, but are more sensitive to movement artifacts in our acquisition setup [69]. Recording PPG features as part of this database is still a valuable addition, as this modality enables the study of early fusion techniques reflecting measures like Pulse-Transit Time (PTT) [70] or detecting cognitive load using only wearable devices, sensing the activity of the heart using pulse waves in the extremities in the future. Based on Pham et al. [71], we solely use features reflecting the activity of the human heart on rhythmic scales and use the R-peaks for stable detection [72]. Our features are based on state-of-the-art techniques from the literature analyzing affective states, cognitive load, stress [33] and other biopsychosocial states. Our features are described in Table 1 and are among other techniques computed using time-based (RMSSD, SDNN, etc.) or frequency-based (HF, LF, etc.) methods. Another physiological response triggered by the Sympathetic Nervous System (SNS) is the activation of sweat glands. This electrodermal activity can be measured on the skin surface through conductance changes, that are caused by sweat and ionic permeability changes in and on the skin surface [73]. Current literature decomposes EDA signals into phasic and tonic components for the analysis of psycho-physiological responses [74]. We capture the tonic activation by computing the change in Skin Conductance Level (SCL) within a given interval and extract features of the faster-changing phasic component also interval-related by measuring Skin Conductance Response (SCR) and do not conduct any direct-stimuli related analysis of SCR events for our level-based experimentation setup. Both SCL and SCR based features are shown in Table 1. Past studies have also found a correlation between the activation of trapezius muscle activity using EMG recordings and cognitive load [75,76]. We have computed features based on the overall activity within a timeframe, e.g., by computing the root-mean-square of the preprocessed signal within a given interval, as well as features based on task-specific changes, such as the number of activations within that interval. To extract information from varying skin temperatures, that are associated with the cognitive load, we extract standard statistical features such as an increase or a decrease in temperature within a time window or mean and standard deviation and follow the methodology used in past studies, that found a correlation for both stress [33] and cognitive load [56]. We extract respiration features from our effort transducer values, by detecting lows (exhalation) and highs (inhalations) and use the amplitude difference between inhalation and exhalation, as well as the mean and standard deviation of respiration rates within a fixed window, as a measure for cognitive load [77]. All features used in this publication, are provided in Table 1. We compute all features with a rolling window with a predefined window size W.

Table 1. List of extracted expert and statistical features from the recorded biosignal modalities.

Modality	Features	Reference	Description
ECG	HR [1]	[33]	Mean heart rate
	$SDSD$	[33,71]	Standard deviation of successive NN intervals
	$SDNN$	[33,71]	Standard deviation of NN intervals
	$RMSDD$	[33,71]	Root mean square of successive differences
	LF, HF	[33,71]	Low and high-frequency component in the range of 0.05–0.15 Hz and 0.15–0.4 Hz of the Welch spectrum
	LF_n, HF_n	[33,71]	Normalized HF and LF spectrum
	LF/HF	[33,71]	Ratio of the LF and HF component
	$SD1/SD2$	[33,71]	Ratio of $SD1$ and $SD2$ as standard deviation along the identity lines of a Poincaré plot
	PSS	[78]	Percentage of short segments.
	PIP	[78]	Percentage of inflection points in NN intervals
PPG	PR[1]	[33]	Pulse rate, closely related to ECG HR (electrical vs. mechanical)
EDA	$\sum_t^W SCL'(t)$	[33,79,80]	Change of the Skin conductance level within a window
	$\#_{SCR}^{Peaks}/W$	[33,73,80,81]	Number of peaks as measure of the phasic component
	$\mu_{SCR}^{Amplitude}$	[33,73]	Mean amplitude values of SCR peaks
	μ_{SCR}^{Rise}	[33]	Mean rise time of SCR peaks
	$\mu_{SCR}^{Recovery}$	[33]	Mean recovery time to 50 percent of the maximal peak amplitude
EMG	$f_{RMS}(EMG)$	[82]	Root mean square of the EMG signal
	$\#_{Onsets}(EMG)/W$	[83,84]	Number of onsets per minute
	$\#_{Active}(EMG)/W$	[83,84]	Fraction of high activity (above threshold) per minute
	$max(EMG)$	[83]	Maximal amplitude of the EMG signal
SKT	μ_T, σ_T	-	Mean and standard deviation of the skin temperature
	min_T, max_T	-	Minimal and maximal of the temperature
	$\sum T'(t)/Time$	-	Increasing/Decreasing temperature per minute.
RSP	μ_{BR}, σ_{BR}	[77]	Mean and standard deviation of BR
	$\mu_{[E-I]}$	[77]	Mean of the ratio of inhalation and exhalation amplitude values

[1] We denote the mean μ and standard deviation σ for all modalities, except for ECG features, we use HR, $SDNN$ and PPG we use PR since these notations are more common in the affective computing domain.

2.9. Eye Tracking Features

Many studies have shown the relevance of eye-related characteristics for the prediction of a person's cognitive state [85]. For eye tracking, we utilize the Tobii Pro Fusion (Tobii Pro Fusion I5S, Tobii AB, Karlsrovägen 2D, S-182 53 Danderyd, Sweden) eye tracker, a stereoscopic system with two eye-tracking cameras for the left and right eye, respectively, that records the subject's eye movements and pupil size at a sampling rate of 250 Hz. We use the PyGaze library [86] to extract fixations (the activity of an eye looking at the same region for a certain duration), saccades (rapid eye movements between fixations), and blinks (defined by the absence of the signal for a certain time) from the acquired eye-tracking data. The fixation detection is based on a dispersion-duration method: if the dispersion of gaze data on the screen is below a certain threshold and has a duration of at least 100 ms, the data is considered a fixation. We select the dispersion threshold within an exploratory analysis as suggested by Salvucci et al. [87]. The duration threshold of 100 ms is a common choice [87–89]. Following Negi et al., we calculate fixations of three non-overlapping temporal ranges: 66–150 ms (short ambient), 300–500 ms (longer focal),

and above 1000 ms (very long). This distinction allows us to gain additional insights into an individual's conscious perception of objects [90]. Despite the strong dependence between pupil dilation and light reflexes, pupil size can provide information about the commitment of greater effort and the expectation of a difficult task [91,92]. High-frequency oscillations, also called hippus, can also indicate increased cognitive load [93].

In accordance with the literature, we extract multiple features based on fixations, saccades, blinks, and pupil dilation. An overview can be found in Table 2. For a window length W we extract the number and duration of fixations, saccades, and blinks as well as the mean saccade amplitude, thus the mean distance for a saccadic movement. As pupil-related features, we select the mean pupil size and two other features that take pupil change into account. To calculate the index of pupillary activity (IPA) we follow [93,94].

Table 2. List of extracted expert and statistical eye tracking features.

Concept	Features	Reference	Description
Fixations	$\#F_{>100}/W$	[85,94]	Fixation count per window for range $>$ 100 ms
	$\#F_{66-150}/W$	[90]	Fixation count per window for range 66–150 ms
	$\#F_{300-500}/W$	[90]	Fixation count per window for range 300–500 ms
	$\#F_{>1000}/W$	[90]	Fixation count per window for range $>$ 1000 ms
	μ_{FD}	[85,94]	Mean fixation duration
	med_{FD}	[90]	Median fixation duration
Saccades	$\#S/W$	[85]	Saccade count per window
	μ_{SA}	[85]	Mean saccade amplitude
	μ_{SD}	[85]	Mean saccade duration
	med_{SD}	-	Median saccade duration
Blinks	$\#B/W$	[85]	Blink count per window
	μ_{BD}	[94]	Mean blink duration
	med_{BD}	-	Median blink duration
Pupil	μ_{PS}	[94]	Mean pupil size
	IPA	[93,94]	Index of pupillary activity based on wavelet transformation

2.10. Videos

One of many indices in this study is the analysis of facial cues. These facial cues are sometimes closely related to messages, that express emotion, such as anger, fear, disgust, happiness, sadness or surprise. As we are highly interested in such behavioral expressions and their relationship to cognitive load, we record the subject's face during the complete experiment. A BASLER acA1920 155 µm (Basler AG, An der Strusbek 60–62, 22926 Ahrensburg, Germany) camera was used to record a video stream with a resolution of 1920 × 1080 at 25 frames per second, which was triggered using the BIOPAC analog output, to ensure a fully synchronized recording. Depending on the position of the subject, we extracted 1024 × 1024 pixels as a region of interest and resampled it to a 512 × 512 frame. For action unit extraction we used the SVM model developed by Cheong et al. [95], using Deng et al. [96] as face detection model, Albiero et al. Img2Pose [97] as pose correction model and Chen et al. MobileFaceNets [98] to extracted landmarks. To mitigate effects, such as many-to-one correspondences of certain behaviors and emotions, or varying interpretations across subjects, we opted to use a sign-judgment-based action unit system instead. The Facial Action Coding System (FACS) system, introduced by Ekman and Friesen [99] and improved by Ekman and Rosenberg [100] specifies different action units. Martinez et al. present the current state of research and its application [101], while predicting cognitive load using visual facial cues has been studied in Li and Busso [102], Yuce et al. [29] and Viegas et al. [28], on which we base our decision to extract the action units presented in Table 3. For our statistical analysis and machine learning experiments, we count the frames with active actions unit within the given window and denote the number of action units similar to our physiological features with # as count.

Table 3. List of extracted action units.

Action Unit Number	Reference Literature	Facial Action Coding System Name
AU1	[28,29,102,103]	Inner Brow Raiser
AU2	[28,29,102,103]	Outer Brow Raiser
AU4	[28,29,102,103]	Brow Lowerer
AU5	[28,102,103]	Upper Lid Raiser
AU6	[28,29,102,103]	Cheek Raiser
AU7	[28,29,102,103]	Lid Tightener
AU9	[28,102,103]	Nose Wrinkler
AU10	[28,29,102,103]	Upper Lip Raiser
AU12	[28,29,102,103]	Lip Corner Puller
AU14	[28,29,102,103]	Dimpler
AU15	[28,29,102,103]	Lip Corner Depressor
AU17	[28,29,102,103]	Chin Raiser
AU20	[28,102]	Lip Stretcher
AU23	[28,29,102,103]	Lip Tightener
AU24	[102]	Lip Pressor
AU25	[28,29,102,103]	Lip Part
AU26	[28,102,103]	Jaw Drop
AU28	[29,102]	Lip Suck
AU43	[28,29,102] [1]	Eyes Closed

[1] Used closely related AU45 - Eyes Open vs. Eyes Closed.

2.11. Preprocessing, Artifacts and Data Collection

The recording of biosignals is prone to various artifacts, while we designed our recording environment with great care, in some situations it is not evitable that artifacts degrade signal quality. Due to the complex nature of our environment, we were not able to prevent 50 Hz power line noise and removed this frequency component, if applicable, and all higher harmonics from the recorded signals using a notch filter. For our ECG feature extraction pipeline it is crucial to detect the R-peaks with high sensitivity and specificity. To ensure only sequences with sufficient quality are used during evaluation, we have computed Signal Quality Index (SQI)'s based on work by Zaho and Zhang of both leads and selected the lead with better SQI's or if both leads are corrupted, we exclude those intervals from the evaluation [104]. For video, we removed frames from evaluation if the subject's head moved outside the recording region. For eye tracker measurements we removed samples with a Tobii Fusion Pro system code indicating the detection of both eyes. For all other modalities, our artifact rejection was integrated by information from the examiner that noted events to the protocol during acquisition. These time intervals were manually removed. Otherwise, we computed features over a two-minute rolling window with a step size of five seconds of our continuous signal. We employed subject-wise z-score normalization (subtracted mean and scaled to standard deviation) using every feature computed within the complete phase of our *k*-drive and *n*-back test, including baselines, training and testing phases. For statistical evaluation and machine learning, we used the window that was in the center of the respective phase described in Sections 2.3 and 2.4 and removed outliers using the Median Absolute Deviation (MAD)-rule [105].

To summarize our feature-extraction and preprocessing protocol, for all records used in our statistical evaluation and machine learning pipeline, we start by rejecting artifacts as described above in this section and exclude corrupted parts of the signal. If the examiner reported recording issues during acquisition (e.g., ECG lead fall-off), we removed the corrupted parts from the sequence. Given the recorded, fully synchronized trigger signals, we define the windows of the baselines and levels. Following that, we computed all features for each modality using a rolling window over the complete experiments of *n*-back and *k*-drive and normalize the extracted trials using subject-wise z-score normalization. We used the MAD-rule to remove outliers of our features. When working with combinations of multiple features, we drop the complete instance (feature vector) that contains a missing value from our dataset and therefore out machine learning experiments. We do not employ

any imputation technique, for missing data, but remove the entire sample that contains a missing feature value. We report the percentages of available records after these steps in Appendix H for reference.

2.12. Statistical Evaluation

To analyze the effect of different experimentation levels (factors), on our extracted expert features (values) we conduct a one-way Analysis of Variance (ANOVA) for repeated measurements, observing if there are significant changes. We determine if the data is distributed normally, using a Shapiro–Wilk test. If the criterion of normality for ANOVA is not met, we conduct a non-parametric Friedman test as omnibus test instead. The assumption of sphericity was controlled by a Mauchly's test. Whenever necessary, we employ a Greenhouse-Geisser correction. As we conduct multiple comparisons we adjust our p-values using Holm–Bonferroni correction for our experiments (k-drive and n-back) separately. For features that changed significantly according to our repeated measures ANOVA, we conducted post hoc-tests. If the post hoc-test results show a p-value below our level of significance $\alpha = 0.05$ we report if the mean value is increasing or decreasing for that specific feature. If the conditions for normality are met, we use a two-sided paired t-test, otherwise, we use the non-parametric Wilcoxon test as a post hoc-test. These tests are executed for our features described in Sections 2.8–2.10.

2.13. Classification Tasks and Evaluation Protocol

Autonomous **D**riving Cognitive Load **A**ssessment Data**base** (*ADABase*) was recorded to study subjects under different levels of cognitive load while observing an autonomously driving vehicle, using multi-modal representations of physiological-, behavioral-, subjective- and performance-based measurements. We create various downstream tasks for classification and regression. As a first task, we separate our recording into two levels of cognitive load:

- A *Low Load Class*, without any task or undemanding tasks imposed on the subject.
- A *High Load Class* with tasks of high demand introducing high cognitive load.

Using this setup we evaluate three different experiments: For the visual-only n-back task we select the observation-only-baseline and the 1-back level as low load class and the 2-back and 3-back levels as high load class. The same phases are used for audiovisual classification. For complete n-back classification we used both baselines (resting and no-task observation), 1-back visual-only, 1-back audiovisual testing phase as low load class and 2-back and 3-back visual-only and audiovisual for high load. Using this classification scheme has several advantages of having a balanced class distribution, being able to evaluate single-only and dual-only vs. single-and-dual task loads, and capturing different levels of the load imposed by the given tasks. We propose to use this labeling scheme for k-drive, too. For our low load class, we use the baseline with an observation-only driving example as low load class and 1-drive as high load class. For our dual-task-load classification of k-drive, we use the baseline with observation-only driving and the baseline with secondary-task observation only as low load and 2 and 3-drive dual task as high load. We also train a classifier on the combination of n-back and k-drive schemes, leading to a single-task, dual-task and a combination of both binary classifications. It is important to note that these classification groups have been carefully selected to include similar tasks (e.g., only observation of n-back tests in baseline instead of simple rest periods) to mitigate the effects of distribution shifts, like not looking to the monitor or infotainment system and light patterns that influence the pupil dilation.

In addition to this very coarse-grained experiment of low vs. high load [48], we create an experiment with three classification levels, low-load, medium-load and high-load. We do not differentiate between a single and a dual classification task but study the combination of both tasks for n-back and k-drive. For k-drive, we use the observing-only-baseline and 1-drive for low load, 2-drive for medium load and 3-drive for high load. For n-back we use the observation-only baseline and visual-only 1-back as low load, 2-back visual only

and 1-back dual task for medium workload and 3-back visual-only and 2-back and 3-back audiovisual for high load. It is noteworthy that the records collected during the specific phases, using these splits are not equally distributed.

We propose several targets for our regression tasks. The first task is to use performance metrics only, averaging a score based on the recall of the primary task (scaled between 0 and 1, where 0 is the best score (highest recall) and 1 is the worst of the complete study population) and the reaction time of the primary task (scaled to be between 0 and 1, where 0 is an instant reaction and 1 is a reaction within 3 s as the maximal possible reaction time). As a second target, we use the NASA-RTLX score, scaled between 0 and 1, where 1 is the highest task load reported by the specific subject and 0 is the lowest reported task load of that subject. We drop all baselines from this dataset. For our third task, we compute the average of both targets and therefore use performance and subjective rating as a continuous target between 0 and 1 for our regression algorithms.

For evaluation, we conduct several experiments. We group our features modality-wise and train a classifier using the binary-task-classification to train and evaluate classifiers for all sub-tasks: single-*n*-back, dual-*n*-back, single-*k*-drive, dual-*k*-drive the combinations of both, single and dual-task load, and both experiments, *n*-back and *k*-drive. Using this experimentation schedule we can as an extension to our statistical evaluation in Section 2.12 determine the predictive power of unimodal and multimodal models.

2.14. Machine Learning Algorithms and Training

Past studies in the affective sensing multimodal human sensing community have shown promising results using eXtreme Gradient Boosting (XGBoost) as a machine learning technique for classification [106]. We follow their finding of using XGBoost and verify this approach, by training and evaluating a Support Vector Machine (SVM) with a linear and a radial kernel, as well as a k-Nearest-Neighbors (kNN) classifier using data from our experiment. To prevent data leakage in our machine learning pipeline, we use nested k-fold cross-validation with 10 inner folds as the validation set and 10 outer folds as the testing set [107], taking bias and variance of our models into account [108] and therefore provide reliable results. The data is split subject-wise. For our binary-classification task, we report Area Under the Receiver Operating Characteristic Curve (AUC) and F1-score. For our three-level classification task, we provide a confusion matrix. The predictive power of our regression task is expressed as R^2 and Mean Squared Error (MSE). For hyperparameter optimization, we run a tree-structured parzen estimator on the validation of every inner fold [109] and optimize for maximal AUC for our classification tasks and maximize R^2 for our regression task. We optimize for the number of estimators, depth, lambda and learning rate for XGBoost, gamma and C for radial basis SVM, C for linear SVM and neighbors and weight function for kNN. As input for our classification, we use the same windows used for our statistical tests, described in Sections 2.12 and 2.11. For training, validation and testing we use feature windows only recorded within the phases described in Section 2.13. In addition to our experiment, using different groups of unimodal and multimodal inputs, we report the feature importance of our XGBoost classifier to explain our model behavior. We use gain, weight and coverage for every feature and report the importance of the top-5 features for every modality.

3. Results

We structure our results similar to the sections in Section 2: After giving an overview of the recorded data in Section 3.1, we present the results of the statistical evaluation for subjective measurements in Section 3.2, for performance measurements in Section 3.3, for all physiological signals for *n*-back and *k*-drive in Sections 3.4–3.6. Behavioral statistical results are presented in Section 3.8. The personality traits of our cohort are presented in Section 3.8. We visualize our features representations using t-SNE in Section 3.9. The results of our machine learning experiments are presented in Section 3.10.

3.1. ADABase and Timescales

Each subject participated in both experiments described in Sections 2.3 and 2.4 in random order, resulting in 25 subjects with n-back as the first experiment and 26 subjects with k-drive as the first experiment. Due to the malfunction of the eye-tracking hardware, for six subjects the eye-tracker data is missing. A defect in the camera trigger line led to synchronization issues of video data with the rest of the test system and some subjects refused the recording of video signals with privacy concerns. We excluded the video data for 17 subjects, leading to missing action units for these subjects. For two subjects the skin temperature data saturated at maximum, caused by defective sensors connectors. For some subjects Eindhoven lead II of the ECG was noisy or the electrode fell off. As an alternative, we used the other lead for R-peak detection and feature computation. If temporary errors occurred during the recording (e.g., the subjects moved their face during the questionnaire phase outside the recording area), we drop only these specific intervals using our artifact correction mechanisms described in Section 2.11.

Figure 8 visualizes features computed for one session of a sample subject during both experiments. The color-coded background data highlights the phases on different timescales. Every test phase is proceeded by at least one training phase, for all tests of $n \in \{1, 2, 3\}$-back for visual-only and audiovisual task loads, 1-drive for single task load settings and $k \in \{2, 3\}$-drive dual task loads. The n-back baseline 1, and k-drive baseline 1 were resting phases of at least 5 min, where the subject was instructed to rest without any interference or additional tasks. During the second n-back baseline phase the subject observed an n-back sequence without any task imposed on the subject. In the second k-drive baseline an autonomous driving simulation was presented without any task imposed on the subject. The third k-drive baseline was introduced to the iPad infotainment application, used for dual-task load scenarios of k-drive. As introduced in Sections 2.3 and 2.4 these baselines enable the study of distribution shifts in future work. Salivary Cortisol, NASA-TLX and recall for the primary and secondary tasks are recorded after test phases or in the case of salivary biomarkers at time points when the test schedule required it. In addition to the fine-grained phases used in this publication, we would like to highlight the ability of our published *ADABase* to learn representations of cognitive load on different timescales. Starting with a continuous prediction, only limited by physiological processes in the test subject, followed by more medium-grained task-specific phases, that combine training and testing, to coarse-grained phases that differentiate between single and dual task loads tasks.

3.2. Subjective Evaluation

After every testing phase, we asked the subject to answer a NASA-TLX questionnaire. We conducted t-tests of raw NASA-TLX values between 1-back and 2-back, 2-back and 3-back and 1-back and 3-back and the same for all levels of k-drive. All Holm-adjusted p-values were with $p < 0.001$ below the level of significance $\alpha = 0.05$ to reject the null hypothesis (two-sided mean alternative) and show a significant change across the levels conducted in this study. Figure 9 shows the means of all six TLX dimensions during different phases of the experiments and Figure 10 shows boxplots of weighted and unweighted (raw) TLX values and the raw ratings for every dimension and level.

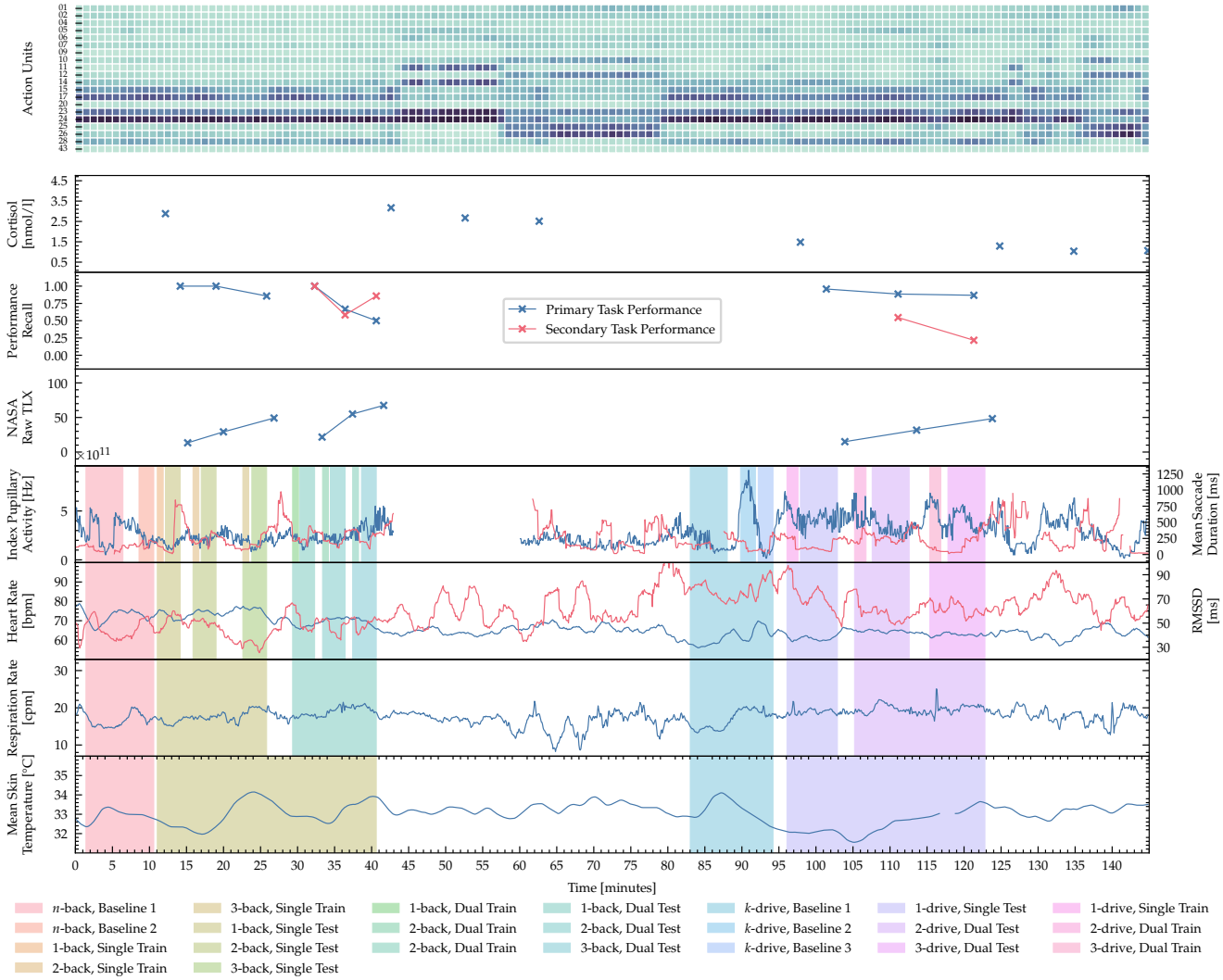

Figure 8. Visualization of the complete recording session for a single subject. Starting at the top-row: Continuously computed action units sampled with 25 fps, that are summed over 1-minute intervals and visualized as colors by occurrence (darker colors show higher activation). Followed by cortisol measurements, performance ratings and subjective feedback as NASA-RTLX scores. The last four rows visualize computed features with a rolling window of 2 min and step-size of 5 s, from eye-tracker, ECG (blue: *HR*, red: *RMSSD*), respiration and skin temperature measurements. Aligned with the overall time axis the different phases are presented as background colors. As described in Sections 2.2–2.4 the experiments (*n*-back and *k*-drive) are in randomized order. In this example, the *n*-back test was conducted first, followed by the questionnaire phase, and the driving phase. The *n*-back protocol was conducted as described in Figures 6 and 7: Single task *n*-back first in three levels, followed by dual-task *n*-back in three levels with training and testing phase. This subject did not repeat any training. If obvious measurement artifacts occurred during acquisition, the corresponding features were removed for downstream evaluations. This example shows corruption in the features of the eye tracker, as the subject's head moved outside the recording region of our eye tracker during the phase between *n*-back and *k*-drive.

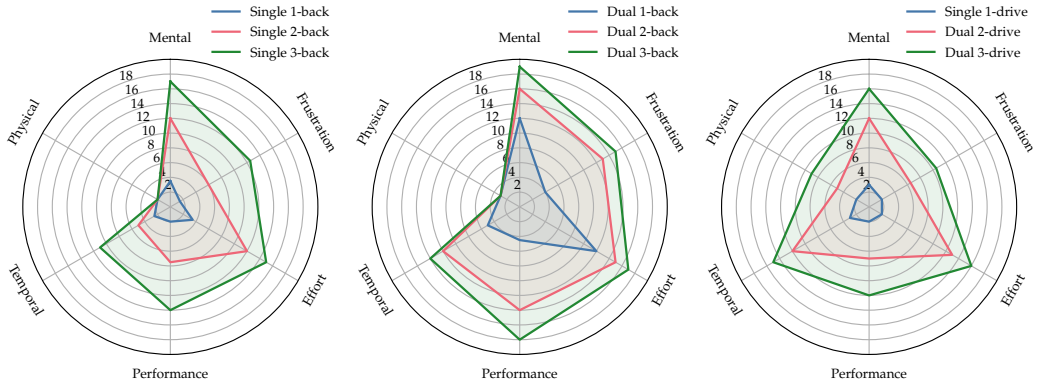

Figure 9. The three plots summarize the NASA-TLX answers of all six dimensions as the mean over all subjects. From left to right: Mean answers during single n-back, dual n-back and k-drive tasks. Reported numbers are not weighted and reflect the raw task load index described in Section 2.6. This radar visualization shows that mental, frustration, effort, performance and temporal dimensions increase with cognitive load tasks of higher intensities ($n \in \{1, 2, 3\}$-back) and dual vs. single, while the change in the subjectively perceived physical load is, compared to these dimensions, smaller. For k-drive of dual and single tasks, every NASA-RTLX dimension increases with a higher CL.

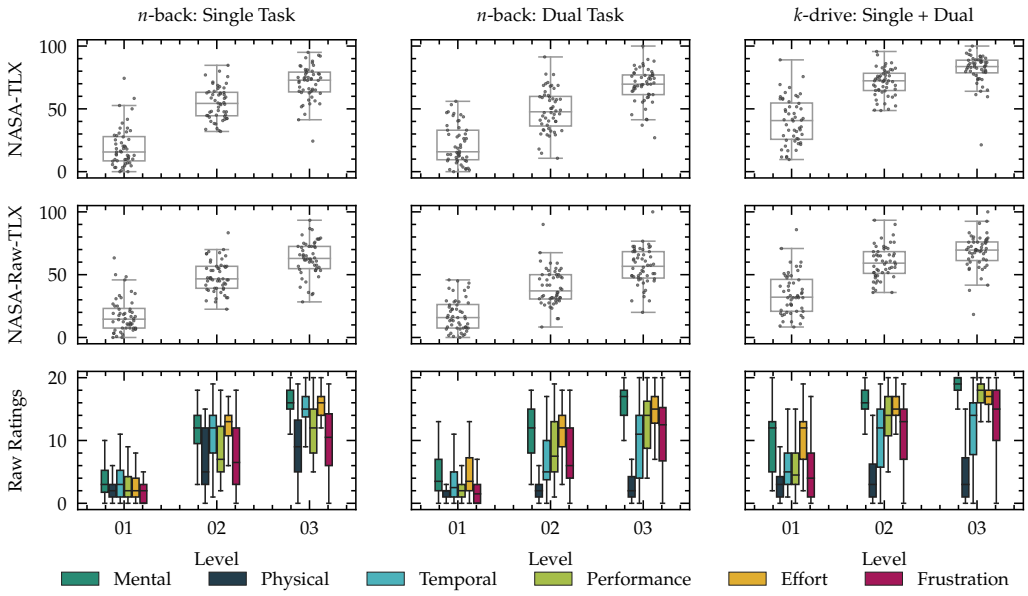

Figure 10. Results of subjective feedback for n-back and k-drive tests of increasing intensities of cognitive load. The first row shows the weighted NASA-TLX ratings during visual-only stimuli n-back, audiovisual n-back stimuli and single-task-only 1-drive and dual-task 2-drive and 3-drive tests. The second row shows the unweighted raw-TLX scores. The last row visualized the raw rating for every of the six NASA-TLX dimensions.

3.3. Performance

For every level in our single and dual task n-back tests and all our k-drive tests, we conducted t-tests for both recall and precision of the visual performance between the first level and the second level, the second level and the third level, as well the first and the third level. All Holm-adjusted p-values are smaller than 0.001. We report primary and secondary task performance and reaction time for all tests in Figure 11.

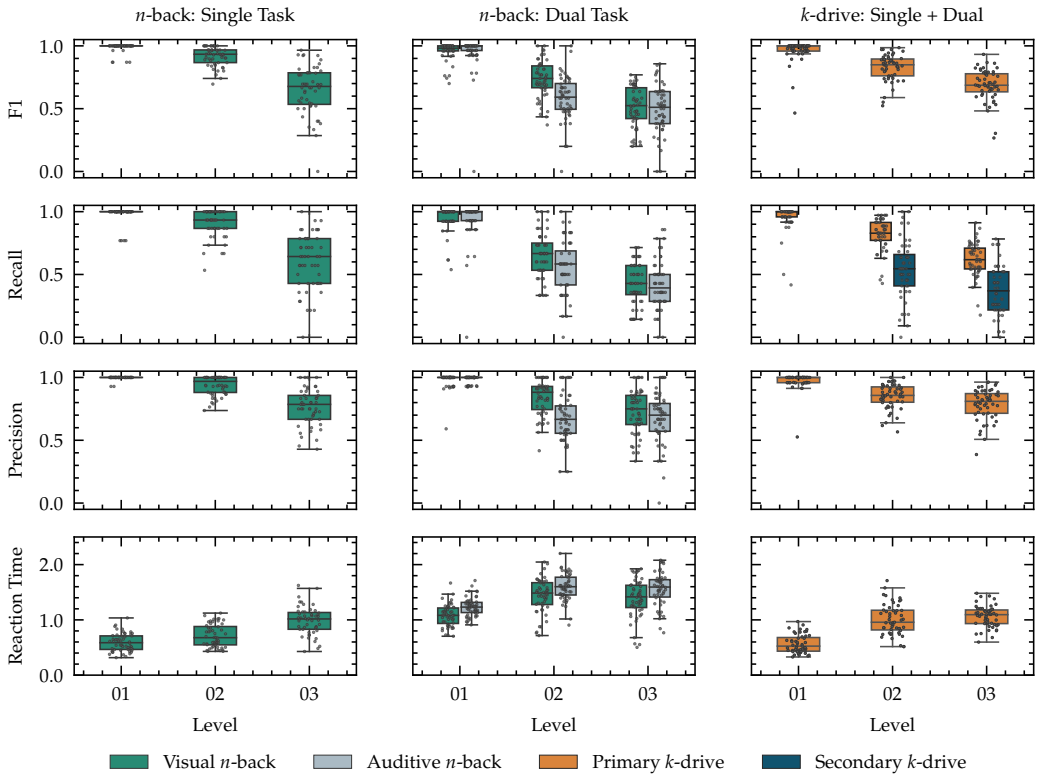

Figure 11. Recorded performance metrics during all conducted tests. For single-task *n*-back performance, we report *F*1-score, recall, precision and reaction time (**first column**). If a dual task was executed, we report these results as an additional box for the corresponding level (**second column**). For *k*-drive we report the same metrics for the single-task tests. As we only counted positive hits for our second task during dual-task *k*-drive, we report recall as a performance metric. (**last column**)

3.4. Biomarkers

We report the relative median values and quantiles of the raw non-normalized cortisol measurements in Figure 12. The null hypothesis of equal means before starting the experiment (Pre-00) and right after each levels of *n*-back (single and dual) and *k*-drive (Post-00) could not be rejected with a sufficient level of significance. This result is in line with other studies analyzing cortisol after phases of high cognitive load [34]. In addition to the reported relative values of *n*-back and *k*-drive tests, we report the cortisol values over the absolute daytime, showing the circadian variation of cortisol of the sample population.

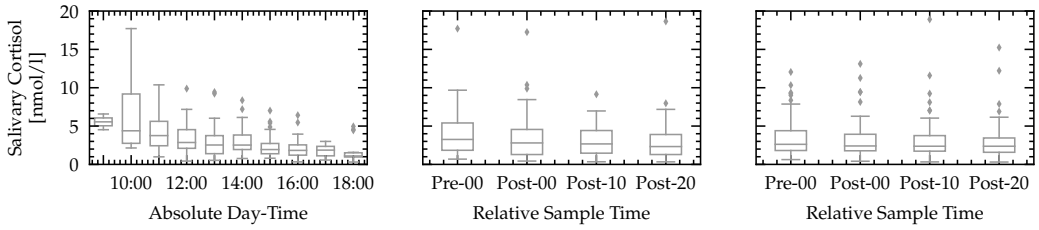

Figure 12. The start of each experiment was executed with a given daytime schedule described in Section 2.5. Some subjects required more time for task comprehension or were repeating the training phase multiple times. This leads to different absolute day-times of cortisol measurements while complying with the relative schedule described in Section 2.5. The first plot shows all measured cortisol values, throughout the day (accumulated one hour intervals). The second plot visualizes the cortisol values for the *n*-back experiment, while the last plot shows the results for *k*-drive.

3.5. Physiological Features—n-Back

For statistical evaluation of features measured during n-back (n-back: Section 2.4, features: Section 2.8) levels, we are using the evaluation protocol described in Section 2.12. We analyze our features extracted from the biosignal modalities and the eye tracker data in Table 4. The results of the omnibus test showed a significant change for features extracted from ECG, EMG, PPG, skin temperature, respiration and eye tracking data during resting baseline, observation only baseline and all n-back tests with single- and dual-task load. Comparing the features during the observation only baseline and the medium level 2-back and high level 3-back cognitive load task as post hoc experiment and evaluation of an increasing or decreasing mean for significant changes shows an increased heart rate, a decreased heart rate variability, a decreasing skin temperature, decreasing mean respiration rate and an increasing IPA. For the EDA activity we could not reject the null hypothesis for all repeated measures. We selected the baseline tutorial phase for our post hoc test because no cognitive load was introduced during this phase, while the subject was still observing an n-back sequence with similar lighting and pose conditions. All results are reported in Table 4.

3.6. Physiological Features—k-Drive

We evaluate our measurements of features during all levels of k-drive following our statistical evaluation protocol desribed in Section 2.12. Similar to our statistical evaluation in Section 3.5 we see a change for features extracted from ECG, EMG, skin temperature, respiration and eye tracker data using our ANOVA for repeated measure evaluation protocol described in Section 2.12. In contrast, the PPG showed no significance, while the EDA showed a significant change. For our omnibus tests, we used all three baselines described in Section 2.3 and all levels $k \in \{1, 2, 3\}$-drive. The post hoc test using the second baseline with driving simulation observations only and all three levels shows similar to the statistical evaluation of the n-back experiment an increasing heart rate, a decreasing heart rate variability, a decreasing skin temperature and an increasing IPA. All results are reported in Table 5.

3.7. Behavioral Features—n-Back and k-Drive

For both experiments n-back and k-drive, we extract action units from facial videos as described in Section 2.10. We evaluated the computed features using the same statistical setup as described in Sections 3.5 and 3.8 and report the results for our n-back test in Tables 6 and 7. The statistical evaluation of our behavioral features showed a significant change in the number of activations of outer brow raiser ($AU2$), chin raiser ($AU17$), lip tightener and pressor ($AU23$ and $AU24$) during n-back and k-drive for our omnibus test. Our n-back tests change the number of activations of brow lowerer ($AU4$), cheek raiser ($AU6$), lip corner depressor ($AU15$) and closed eyes ($AU43$). The k-drive tests showed a significant change in inner brow raiser ($AU1$), upper lid raiser ($AU5$), lip corner puller ($AU12$), dimpler ($AU14$), lip strecher ($AU20$) and jaw drops ($AU26$).

3.8. Personality Traits

When characterizing stress responses using multimodal physiological data, personality traits play an important role [61]. Figure 13 reports openness, conscientiousness, extraversion, agreeableness and neuroticism of our study population. As an additional test, we computed the spearman correlation between the mean overall NASA-TLX score of all levels for one subject and the reported personality and found $r = -0.401$, $CI = [-0.61, -0.14]$ for openness, $r = 0.053, CI = [-0.23, 0.32]$ for conscientiousness, $r = -0.082$, $CI = [-0.35, 0.2]$ for extraversion, $r = 0.164, CI = [-0.12, 0.42]$ for agreeableness and $r = -0.019$, $CI = [-0.29, 0.26]$ for neuroticism, with r as correlation coefficient and CI as 95% confidence intervals around r.

Table 4. Statistical evaluation of biosignal and eye-tracker based features during n-back experiment. The first column denotes the modalities, followed by the computed feature. The features are described in Section 2.8 for all bio-psychological features and all eye-tracker features are described in Section 2.9. The omnibus test and post hoc analysis are described in Section 2.12. All reported p-values are adjusted to compensate for multiple comparison problem using Bonferroni-Holm correction. For features that show a significant change (p-value below $\alpha = 0.05$) in the omnibus test, we conducted a two-sided t-test between a baseline phase and the 2/3-back for the visual-only and dual-task load. If the result of the post hoc test is significant, we report if the mean feature value is increasing (↑) or decreasing (↓).

Modality	Feature	Omnibus Test	Post Hoc Adjusted p-Values			
			Baseline Tutorial			
			2-back Single	3-back Single	2-back Dual	3-back Dual
ECG	HR	<0.001	0.001 (↑)	<0.001 (↑)	0.181	1.000
	$SDSD$	<0.001	1.000	0.008 (↓)	0.563	1.000
	$SDNN$	<0.001	0.010 (↓)	0.244	0.174	1.000
	$RMSSD$	<0.001	1.000	0.008 (↓)	0.563	1.000
	LF	<0.001	0.011 (↓)	0.003 (↓)	0.018 (↓)	1.000
	HF	0.001 †	1.000	1.000	1.000	1.000
	LF_n	<0.001 †	<0.001 (↓)	0.161	1.000	1.000
	HF_n	<0.001 †	<0.001 (↑)	0.161	1.000	1.000
	LF/HF	0.001 †	<0.001 (↓)	0.210	0.670	1.000
	$SD1/SD2$	<0.001	1.000	1.000	1.000	1.000
	PSS	<0.001	0.531	1.000	1.000	1.000
	PIP	<0.001	0.111	1.000	0.057	1.000
PPG	PR	0.029	1.000	1.000	1.000	1.000
EDA	$\mu_{SCR}^{Amplitude}$	0.186	-	-	-	-
	$\sum_t^W SCL'(t)$	0.130 †	-	-	-	-
	$\#_{SCR}^{Peaks}/W$	0.604	-	-	-	-
	μ_{SCR}^{Rise}	0.132	-	-	-	-
	$\mu_{SCR}^{Recovery}$	0.396 †	-	-	-	-
EMG	$\#_{Onsets}/W$	0.271	-	-	-	-
	$\#_{Active}/W$	0.132	-	-	-	-
	$max(EMG)$	0.029	1.000	1.000	1.000	1.000
	f_{RMS}	0.189	-	-	-	-
SKT	μ_T	0.008	0.013 (↓)	0.125	0.144	1.000
	σ_T	0.008	0.013 (↓)	0.125	0.144	1.000
	min_T	0.017	0.016 (↓)	0.080	0.059	1.000
	max_T	0.192	-	-	-	-
	$\sum T'(t)/T$	<0.001	0.987	0.563	0.031 (↑)	0.008 (↑)
RSP	μ_{BR}	<0.001	<0.001 (↑)	0.129	1.000	1.000
	σ_{BR}	0.060 †	-	-	-	-
	$\mu_{[E-I]}$	0.271	-	-	-	-
EYE	$\#F_{>100}/W$	<0.001	1.000	1.000	1.000	1.000
	$\#F_{66-150}/W$	0.153 †	-	-	-	-
	$\#F_{300-500}/W$	<0.001	1.000	1.000	1.000	1.000
	$\#F_{>1000}/W$	<0.001	1.000	1.000	0.016 (↓)	1.000
	$\mu_{FD_{>100}}$	0.035	1.000	1.000	0.245	0.085
	$med_{FD_{>100}}$	0.006	1.000	0.676	0.128	1.000
	$\#S/W$	<0.001	1.000	1.000	1.000	1.000
	μ_{SA}	<0.001	1.000	0.032 (↓)	1.000	1.000
	μ_{SD}	0.604	-	-	-	-
	med_{SD}	<0.001 †	1.000	1.000	1.000	1.000
	$\#B/W$	0.159	-	-	-	-
	μ_{BD}	<0.001	1.000	1.000	1.000	1.000
	med_{BD}	0.076	-	-	-	-
	μ_{PS}	0.076	-	-	-	-
	IPA	<0.001	<0.001 (↑)	<0.001 (↑)	<0.001 (↑)	<0.001 (↑)

† Non-parametric Friedman test.

Table 5. Tabular view of statistical results for features computed during k-drive, indicating significant changes, for multiple modalities. All statistical tests are conducted according to Section 2.12. p-values are adjusted using Bonferroni-Holm. Post Hoc tests are performed for $k \in \{1, 2, 3\}$-drive levels and driving tutorial baseline.

Modality	Feature	Post Hoc Adjusted p-Values			
		Omnibus Test	Baseline Tutorial		
			1-drive	2-drive	3-drive
ECG	HR	<0.001	0.076	<0.001 (↑)	<0.001 (↑)
	$SDSD$	0.003	0.008 (↓)	0.001 (↓)	<0.001 (↓)
	$SDNN$	<0.001	0.002 (↓)	<0.001 (↓)	<0.001 (↓)
	$RMSSD$	0.003	0.008 (↓)	0.001 (↓)	<0.001 (↓)
	LF	0.074	-	-	-
	HF	0.201	-	-	-
	LF_n	0.010	1.000	1.000	1.000
	HF_n	0.010	1.000	1.000	1.000
	LF/HF	0.243 †	-	-	-
	$SD1/SD2$	<0.001	1.000	1.000	1.000
	PSS	<0.001	1.000	0.438	1.000
	PIP	<0.001	1.000	1.000	1.000
PPG	PR	0.418	-	-	-
EDA	$\mu_{SCR}^{Amplitude}$	0.044	1.000	0.806	0.264
	$\sum_t^W SCL'(t)$	1.000	-	-	-
	$\#_{SCR}^{Peaks}/W$	<0.001	1.000	0.012 (↑)	0.070
	μ_{SCR}^{Rise}	0.022	1.000	0.110	1.000
	$\mu_{SCR}^{Recovery}$	0.013 †	0.102	1.000	1.000
EMG	$\#_{Onsets}/W$	0.001 †	1.000	1.000	0.176
	$\#_{Active}/W$	0.074	-	-	-
	$max(EMG)$	0.201	-	-	-
	f_{RMS}	<0.001 †	1.000	<0.001 (↑)	<0.001 (↑)
SKT	μ_T	<0.001	0.008 (↓)	<0.001 (↓)	0.004 (↓)
	σ_T	<0.001	0.008 (↓)	<0.001 (↓)	0.004 (↓)
	min_T	<0.001	0.004 (↓)	<0.001 (↓)	0.005 (↓)
	max_T	<0.001	<0.001 (↓)	<0.001 (↓)	<0.001 (↓)
	$\sum T'(t)/T$	0.002	<0.001 (↑)	0.657	0.233
RSP	μ_{BR}	0.001	1.000	1.000	1.000
	σ_{BR}	0.031 †	1.000	1.000	1.000
	$\mu_{[E-I]}$	1.000	-	-	-
EYE	$\#F_{>100}/W$	<0.001	1.000	<0.001 (↓)	<0.001 (↓)
	$\#F_{66-150}/W$	<0.001	0.772	<0.001 (↓)	<0.001 (↓)
	$\#F_{300-500}/W$	<0.001 †	0.672	1.000	1.000
	$\#F_{>1000}/W$	0.002 †	1.000	1.000	1.000
	$\mu_{FD_{>100}}$	<0.001	0.051	<0.001 (↑)	<0.001 (↑)
	$med_{FD_{>100}}$	<0.001	1.000	0.001 (↑)	0.001 (↑)
	$\#S/W$	<0.001 †	1.000	0.712	0.244
	μ_{SA}	<0.001	<0.001 (↓)	<0.001 (↑)	<0.001 (↑)
	μ_{SD}	0.418 †	-	-	-
	med_{SD}	0.020 †	0.404	1.000	1.000
	$\#B/W$	<0.001	0.184	0.089	0.048 (↓)
	μ_{BD}	0.025	0.020 (↑)	1.000	1.000
	med_{BD}	0.006	0.003 (↑)	1.000	1.000
	μ_{PS}	<0.001	0.173	0.449	1.000
	IPA	<0.001	1.000	<0.001 (↑)	<0.001 (↑)

† Non-parametric Friedman test.

Table 6. Behavioral changes in FACS action unit for 2/3-back tests for a single and dual-task load. Every action unit occurrence is counted over every frame within a given window.

Modality	Feature	Adjusted *p*-Values				
		Omnibus Test	Baseline Tutorial			
			2-back Single	3-back Single	2-back Dual	3-back Dual
AUS	#*AU*1	0.107	-	-	-	-
	#*AU*2	<0.001	1.000	1.000	1.000	1.000
	#*AU*4	<0.001	1.000	0.312	1.000	0.966
	#*AU*5	<0.001	1.000	1.000	0.312	0.741
	#*AU*6	0.509	-	-	-	-
	#*AU*7	0.722	-	-	-	-
	#*AU*9	0.509	-	-	-	-
	#*AU*10	0.509	-	-	-	-
	#*AU*11	0.509	-	-	-	-
	#*AU*12	0.987	-	-	-	-
	#*AU*14	0.987	-	-	-	-
	#*AU*15	0.004	1.000	0.015 (↓)	0.006 (↓)	0.002 (↓)
	#*AU*17	0.009	0.587	0.001 (↓)	0.025 (↓)	0.001 (↓)
	#*AU*20	0.294	-	-	-	-
	#*AU*23	0.001	1.000	1.000	1.000	1.000
	#*AU*24	0.009	1.000	1.000	1.000	1.000
	#*AU*25	0.509	-	-	-	-
	#*AU*26	0.091	-	-	-	-
	#*AU*28	0.722	-	-	-	-
	#*AU*43	0.002	1.000	1.000	0.011 (↑)	0.206

Table 7. Changes in action units during $k \in \{1, 2, 3\}$-drive test. Increase (↑) and decrease (↓) indicate data distribution shift between single (1-drive) and dual-task (2/3-drive) load.

Modality	Feature	Adjusted *p*-Values			
		Omnibus Test	Baseline Tutorial		
			1-drive	2-drive	3-drive
AUS	#*AU*1	<0.001	1.000	1.000	1.000
	#*AU*2	<0.001	0.606	0.023 (↑)	0.073
	#*AU*4	0.283	-	-	-
	#*AU*5	<0.001	0.051	<0.001 (↑)	<0.001 (↑)
	#*AU*6	0.268 †	-	-	-
	#*AU*7	0.554	-	-	-
	#*AU*9	0.164	-	-	-
	#*AU*10	0.078	-	-	-
	#*AU*11	0.867	-	-	-
	#*AU*12	0.003	1.000	0.003 (↓)	0.027 (↓)
	#*AU*14	0.043 †	1.000	0.049 (↓)	0.073
	#*AU*15	0.148	-	-	-
	#*AU*17	0.003	1.000	1.000	1.000
	#*AU*20	0.001	0.002 (↓)	0.075	0.314
	#*AU*23	0.011	0.888	0.350	0.252
	#*AU*24	0.043	1.000	0.008 (↑)	0.728
	#*AU*25	0.867	-	-	-
	#*AU*26	0.006	1.000	0.706	1.000
	#*AU*28	0.114	-	-	-
	#*AU*43	0.268	-	-	-

† Non-parametric Friedman test.

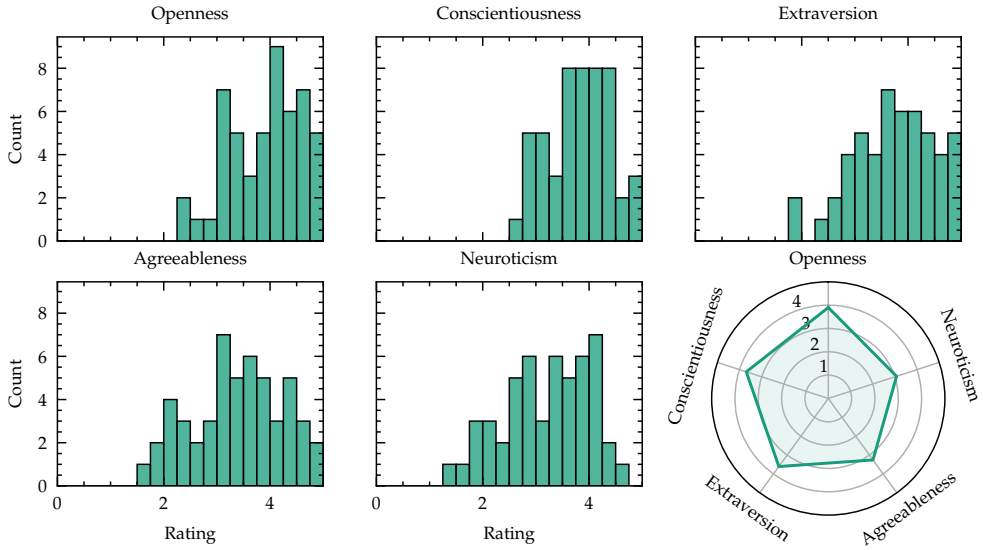

Figure 13. Big-Five personality traits: openness, conscientiousness, extraversion, agreeableness, neuroticism visualized as histogram and radar plot of mean values.

3.9. Representations

Using our extracted biosignal features, we have computed t-SNE representations with two components of different feature subsets. Representations are computed for all participating subjects. The colors indicate the coarse-grained binary classes, where blue contains the ground-truth labels n-back baselines, k-drive baselines, 1-back single task and level 1 drive task and red indicate phases with high cognitive load: 2/3-back single task, $n \in \{1, 2, 3\}$-back dual-task load and 2/3-drive tests. In other words: red indicates a high cognitive load, while blue indicates a low or no cognitive load. The last plot shows as a color map the regression target described in Section 2.13, with a darker blue for higher scores and a lighter green for lower scores. The separation of clusters visualized in Figure 14 as 2d t-SNE components of our high dimensional feature vector, described in Section 2.8 are a promising indication of useful representations for classification tasks in Section 3.10.

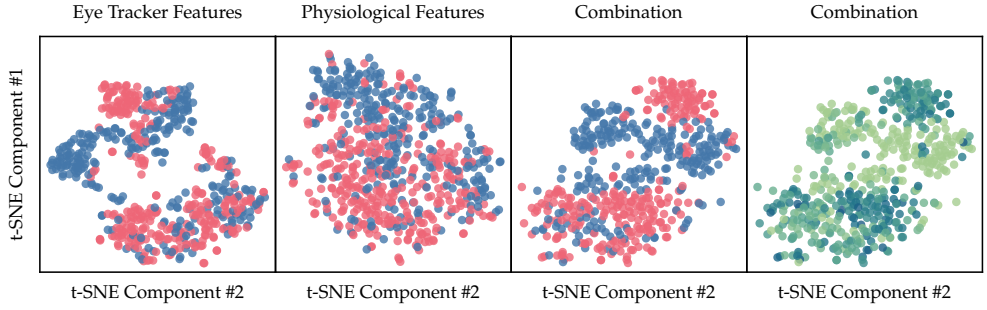

Figure 14. Visualization of 2-component t-SNE representations for computed biosignal- and eye-tracker-based features. The first plot shows t-SNE representation with eye-tracker features only, the second plot depicts only biosignal-based features and the third plot visualizes the combinations of both feature sets. The last plot presents the same representations from the third plot, but color coded as a linear combination of reaction time, recall and NASA-TLX score as color intensity.

3.10. Machine Learning

The automatic detection of cognitive load using computerized programs is a key ingredient for deployment in semi-autonomously driving cars. We conduct several experiments described in Section 2.13 using the machine learning pipeline described in Section 2.14 and *XGBoost* as model. Our first experiment determines the effect of using features from different and multiple modalities as input for binary task classification. The results are

reported in tabular form in Table 8. In addition, we train the models on various subsets of our collected data, containing single- and/or dual-task load *n*-back tests, single- and/or dual-task load *k*-drive tests and combinations of both. The AUC of models trained on the combination of all biosignal features is 0.81 ± 0.05, for the combinations of all *n*-back and *k*-drive tests. Models using only PPG as input achieve a AUC of 0.55 ± 0.06, using only respiration 0.65 ± 0.06, using only the trapezius activity 0.67 ± 0.06, using the electrical activity of the heart 0.75 ± 0.06 and models using the electrodermal activity achieve a AUC 0.68 ± 0.04. Using exclusively features based on action units data lead to a AUC of 0.69 ± 0.07. Models using solely features based on eye tracker data achieved a AUC of 0.84 ± 0.05. Models trained and evaluated on the combination of features from biosignals, action units and eye tracker data outperform unimodal features or other combinations with a AUC of 0.90 ± 0.04. This finding holds for single-and-dual-task *n*-back and *k*-drive tests. Another noteworthy finding is that eye tracker features perform equally well for all conducted tasks with a AUC of 0.86 ± 0.04 for single-and-dual-task *n*-back test and a AUC of 0.89 ± 0.06 for single-and-dual-task *k*-drive test, while other modalities such as EDA perform well for *k*-drive with 0.88 ± 0.03 but have a dropping performance for models trained and evaluated on single-and-dual task *n*-back tests with a AUC of 0.71 ± 0.08.

Table 8. Tabular data visualization of *F*1-Score and AUC scores for binary classification task. Rows contain various sets of features used for classification: Using only one modality (ECG, EDA, etc.) or combinations of various modalities: *Bio* containing all biosignal modalities, described in Section 2.8, *AU's* containing all action units extracted from video data and *Eye* containing all eye tracker features. The columns are separated into experiments described in Sections 2.3 and 2.4 and the combination of both experiments. The different groups for this two-level classification, detecting low and high cognitive load are described in Section 2.13. The second-level hierarchical columns differentiate between single and dual-task load (e.g., for the n-back task: visual vs. visual + auditive) and the combinations of similar cognitive load levels by intensities in *Comb.*, described in Section 2.13. All results were acquired using the evaluation pipeline described in Section 2.14 with nested cross-validation protocol and *XGBoost* as model for classification.

	***n*-back**						***k*-drive**						**Both**					
	Single		**Dual**		**Comb.**		**Single**		**Dual**		**Comb.**		**Single**		**Dual**		**Comb.**	
	F_1	*AUC*	F_1	*AUC*	F_1	*AUC*	F_1	*AUC*	F_1	*AUC*	F_1	*AUC*	F_1	*AUC*	F_1	*AUC*	F_1	*AUC*
PPG	0.61 ± 0.10	0.66 ± 0.09	0.48 ± 0.21	0.61 ± 0.09	0.53 ± 0.10	0.60 ± 0.09	0.58 ± 0.12	0.65 ± 0.12	0.59 ± 0.07	0.61 ± 0.12	0.53 ± 0.12	0.57 ± 0.06	0.52 ± 0.16	0.62 ± 0.09	0.54 ± 0.07	0.57 ± 0.04	0.52 ± 0.09	0.55 ± 0.06
Respiration	0.61 ± 0.14	0.71 ± 0.13	0.64 ± 0.12	0.69 ± 0.10	0.62 ± 0.11	0.65 ± 0.12	0.61 ± 0.11	0.68 ± 0.11	0.65 ± 0.11	0.72 ± 0.08	0.70 ± 0.10	0.76 ± 0.10	0.64 ± 0.05	0.65 ± 0.04	0.62 ± 0.09	0.67 ± 0.07	0.63 ± 0.05	0.65 ± 0.06
EMG	0.62 ± 0.09	0.66 ± 0.08	0.61 ± 0.09	0.60 ± 0.10	0.62 ± 0.08	0.62 ± 0.05	0.65 ± 0.11	0.69 ± 0.11	0.69 ± 0.09	0.75 ± 0.09	0.69 ± 0.08	0.73 ± 0.05	0.60 ± 0.18	0.66 ± 0.05	0.62 ± 0.10	0.66 ± 0.05	0.61 ± 0.10	0.67 ± 0.06
ECG	0.67 ± 0.11	0.73 ± 0.11	0.67 ± 0.08	0.73 ± 0.06	0.64 ± 0.14	0.71 ± 0.10	0.71 ± 0.15	0.87 ± 0.14	0.73 ± 0.07	0.88 ± 0.10	0.79 ± 0.09	0.87 ± 0.07	0.67 ± 0.11	0.74 ± 0.09	0.70 ± 0.08	0.78 ± 0.07	0.70 ± 0.07	0.75 ± 0.06
EDA	0.63 ± 0.07	0.69 ± 0.10	0.69 ± 0.06	0.76 ± 0.08	0.67 ± 0.10	0.71 ± 0.08	0.93 ± 0.11	0.96 ± 0.06	0.87 ± 0.12	0.97 ± 0.04	0.79 ± 0.04	0.88 ± 0.03	0.72 ± 0.07	0.75 ± 0.06	0.73 ± 0.08	0.81 ± 0.05	0.63 ± 0.05	0.68 ± 0.04
Biosignals	0.70 ± 0.13	0.79 ± 0.11	0.70 ± 0.09	0.78 ± 0.08	0.69 ± 0.07	0.77 ± 0.06	0.88 ± 0.08	0.97 ± 0.07	0.92 ± 0.04	1.00 ± 0.01	0.86 ± 0.06	0.96 ± 0.03	0.73 ± 0.07	0.81 ± 0.07	0.76 ± 0.08	0.87 ± 0.06	0.74 ± 0.05	0.81 ± 0.05
Eye Tracker	0.79 ± 0.08	0.92 ± 0.05	0.74 ± 0.08	0.82 ± 0.10	0.76 ± 0.06	0.86 ± 0.04	0.68 ± 0.12	0.82 ± 0.13	0.88 ± 0.07	0.96 ± 0.04	0.79 ± 0.05	0.89 ± 0.06	0.73 ± 0.11	0.86 ± 0.07	0.78 ± 0.06	0.90 ± 0.04	0.73 ± 0.07	0.84 ± 0.05
Action Units	0.60 ± 0.13	0.68 ± 0.06	0.70 ± 0.06	0.72 ± 0.10	0.54 ± 0.14	0.70 ± 0.06	0.64 ± 0.15	0.74 ± 0.09	0.68 ± 0.12	0.78 ± 0.11	0.71 ± 0.06	0.77 ± 0.08	0.52 ± 0.12	0.65 ± 0.09	0.59 ± 0.16	0.74 ± 0.08	0.64 ± 0.10	0.69 ± 0.07
Bio, AU's	0.70 ± 0.14	0.80 ± 0.11	0.71 ± 0.11	0.85 ± 0.06	0.70 ± 0.08	0.80 ± 0.06	0.89 ± 0.07	0.97 ± 0.05	0.91 ± 0.05	0.99 ± 0.03	0.85 ± 0.07	0.96 ± 0.03	0.73 ± 0.05	0.81 ± 0.05	0.80 ± 0.08	0.90 ± 0.06	0.73 ± 0.08	0.84 ± 0.06
Eye, AU's	0.80 ± 0.09	0.93 ± 0.06	0.78 ± 0.06	0.87 ± 0.07	0.77 ± 0.06	0.87 ± 0.05	0.68 ± 0.20	0.88 ± 0.11	0.90 ± 0.06	0.98 ± 0.03	0.80 ± 0.09	0.91 ± 0.07	0.75 ± 0.08	0.87 ± 0.05	0.81 ± 0.07	0.91 ± 0.04	0.77 ± 0.05	0.86 ± 0.05
Bio, Eye	0.82 ± 0.07	0.93 ± 0.06	0.75 ± 0.12	0.86 ± 0.09	0.78 ± 0.06	0.87 ± 0.05	0.91 ± 0.06	0.97 ± 0.05	0.92 ± 0.04	0.99 ± 0.02	0.87 ± 0.09	0.95 ± 0.04	0.81 ± 0.08	0.92 ± 0.06	0.83 ± 0.08	0.92 ± 0.06	0.81 ± 0.06	0.89 ± 0.04
Bio, Eye, AU's	0.82 ± 0.09	0.93 ± 0.06	0.78 ± 0.09	0.88 ± 0.09	0.77 ± 0.07	0.88 ± 0.05	0.88 ± 0.09	0.96 ± 0.06	0.94 ± 0.05	0.99 ± 0.01	0.89 ± 0.06	0.96 ± 0.03	0.84 ± 0.06	0.92 ± 0.06	0.83 ± 0.05	0.92 ± 0.04	0.82 ± 0.06	0.90 ± 0.04

To further study the impact of all acquired features on classification performance for our binary classification task, we analyze the models that used all tasks, levels and features from all modalities for training. As described in Section 2.14 we use *XGBoost* for classification and can therefore express the feature contribution to the final classification as *gain* metric, that implies the relative contribution of the corresponding features to the final prediction of the model. Figure 15 shows the gain importance of the top features for each modality used in our classification task, showing contributions form IPA and μ_{PS}, as high contributing features from the eye tracker modality, the change of skin conductance within a time interval as a feature with high gain in our *XGBoost* model and both heart rate and heart rate variability based features extracted from the electrical activity of the heart. In Figure A4, we present a supplementary analysis of the feature importance of our trained models.

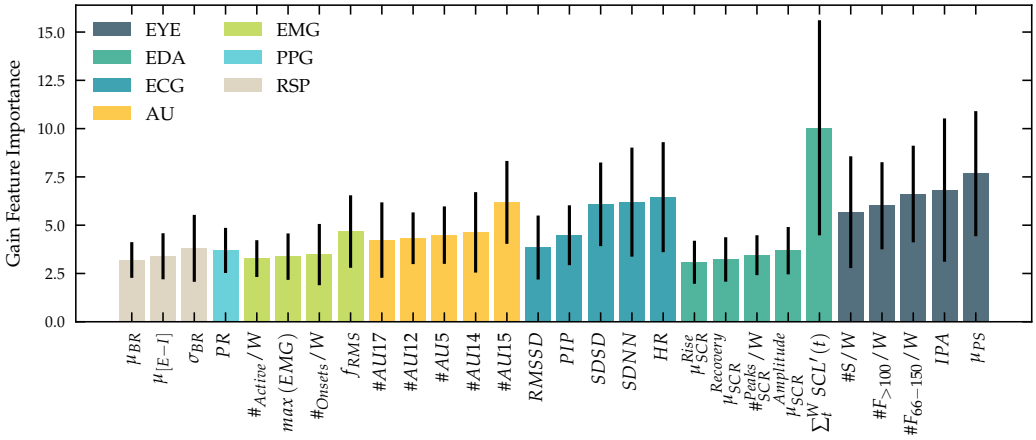

Figure 15. XGBoost provides among others *gain* as feature importance metric. We report the features with the five highest *gain* importance values, if available. *Gain* captures the feature importance by measuring the relative contribution of each feature in every tree to the final predictions, where higher values compared to other features indicate higher importance. Gain values are reported from models reported in Table 8 with bio-psychological features, action units and eye tracker-based features on combined (dual and single task load) with data from both experiments (10-fold-nested-cross-validation).

In addition to our low vs. high cognitive load task, we introduced a classification task for multiple intensities of cognitive load in Section 2.13. Table 9 visualizes confusion matrices across three different sets of features, using only biosignals, exclusively utilizing eye tracker-based features and a combination of eye tracker-based features, biosignals and action units. It becomes apparent that the detection performance drops significantly compared to simple binary classification only differentiating between low and high cognitive load, while low-load is still detected in 92 percent correctly for *k*-drive using all features, high-load is often confused with medium-load and only detects 69 percent correctly. This holds for different subsets of features and tasks. We can observe that the identification of multiple levels, degrades performance, leading to an $F1$-score of 0.72 ± 0.09 for using all features for the *k*-drive test.

Previous literature found that *XGBoost* models performed best for the detection of drivers' distraction from physiological and visual signals [106]. We confirm these results by training a SVM with linear and radial basis functions and a kNN classifier. All results are reported in Table 10.

In addition to our classification tasks, we train a regression model using a *XGBoost* regressor. The regression targets are described in Section 2.13. Using only performance-based targets, we were able to train a model with an R^2 score of 0.51 ± 0.07, using subjective ratings acquired using NASA-TLX scores, we get 0.48 ± 0.08, and for a linear combination (average of both), we achieve a R^2 score of 0.54 ± 0.14. All results are presented in Table 11.

Table 9. Results of three-level-intensity classification tasks, treating different levels as separate classes (columns). For three different combinations of features, based on biosignals alone, only eye-tracker data and the combination of features based on multiple modalities: biosignals, eye-tracker-based action units (rows). We report the confusion matrix for the conducted experiments, using only *n*-back levels, only Drive levels and a combination of both experiments. The grouping of different phases for all experiments is described in Section 2.13. The rows correspond to the true class.

		***n*-back**			***k*-drive**			**Both**		
		Low	**Med.**	**High**	**Low**	**Med.**	**High**	**Low**	**Med.**	**High**
Biosignals	Low	0.57 ± 0.23	0.17 ± 0.12	0.26 ± 0.14	0.82 ± 0.07	0.08 ± 0.07	0.09 ± 0.06	0.55 ± 0.10	0.21 ± 0.05	0.25 ± 0.08
	Medium	0.29 ± 0.15	0.49 ± 0.12	0.22 ± 0.18	0.09 ± 0.16	0.51 ± 0.19	0.40 ± 0.20	0.24 ± 0.13	0.42 ± 0.08	0.34 ± 0.07
	High	0.21 ± 0.10	0.24 ± 0.05	0.55 ± 0.09	0.07 ± 0.11	0.51 ± 0.26	0.42 ± 0.23	0.22 ± 0.08	0.27 ± 0.06	0.51 ± 0.07
Eye Tracker	Low	0.86 ± 0.12	0.07 ± 0.11	0.06 ± 0.08	0.84 ± 0.11	0.07 ± 0.06	0.09 ± 0.06	0.77 ± 0.16	0.09 ± 0.08	0.13 ± 0.09
	Medium	0.17 ± 0.22	0.49 ± 0.12	0.33 ± 0.19	0.00 ± 0.00	0.70 ± 0.19	0.30 ± 0.19	0.15 ± 0.16	0.55 ± 0.17	0.31 ± 0.14
	High	0.09 ± 0.07	0.32 ± 0.05	0.58 ± 0.05	0.02 ± 0.06	0.27 ± 0.22	0.71 ± 0.22	0.07 ± 0.07	0.33 ± 0.04	0.60 ± 0.08
Bio, Eye, AU's	Low	0.77 ± 0.17	0.10 ± 0.09	0.13 ± 0.11	0.92 ± 0.07	0.02 ± 0.04	0.06 ± 0.05	0.81 ± 0.12	0.08 ± 0.07	0.11 ± 0.06
	Medium	0.13 ± 0.12	0.67 ± 0.22	0.21 ± 0.17	0.03 ± 0.05	0.71 ± 0.18	0.26 ± 0.14	0.18 ± 0.15	0.50 ± 0.09	0.32 ± 0.10
	High	0.08 ± 0.05	0.27 ± 0.06	0.65 ± 0.07	0.03 ± 0.06	0.31 ± 0.17	0.66 ± 0.19	0.07 ± 0.06	0.32 ± 0.05	0.61 ± 0.05

Table 10. Comparison of different classifiers for all binary tasks. All classifiers have been trained using the complete set of features: bio-signals, eye-tracker data and action units from videos. The results in this table correspond to the last row in Table 8 for *XGBoost* based classification. *kNN* is a k-Nearest-Neighbor classifier, *SVM (rbf)* a support vector machine with a radial kernel and *SVM (lin)* is a linear support vector machine. All results have been acquired using nested-cross-validation.

	***n*-back**						***k*-drive**						**Both**					
	Single		**Dual**		**Comb.**		**Single**		**Dual**		**Comb.**		**Single**		**Dual**		**Comb.**	
	F_1	*AUC*	F_1	*AUC*	F_1	*AUC*	F_1	*AUC*	F_1	*AUC*	F_1	*AUC*	F_1	*AUC*	F_1	*AUC*	F_1	*AUC*
kNN	0.71 ± 0.09	0.78 ± 0.11	0.69 ± 0.11	0.72 ± 0.09	0.71 ± 0.09	0.76 ± 0.06	0.71 ± 0.10	0.86 ± 0.10	0.83 ± 0.07	0.94 ± 0.06	0.80 ± 0.05	0.90 ± 0.04	0.72 ± 0.05	0.79 ± 0.08	0.76 ± 0.07	0.84 ± 0.06	0.75 ± 0.03	0.81 ± 0.03
XGB	0.82 ± 0.09	0.93 ± 0.06	0.78 ± 0.09	0.88 ± 0.09	0.77 ± 0.07	0.88 ± 0.05	0.88 ± 0.09	0.96 ± 0.06	0.94 ± 0.05	0.99 ± 0.01	0.89 ± 0.06	0.96 ± 0.03	0.84 ± 0.06	0.92 ± 0.06	0.83 ± 0.05	0.92 ± 0.04	0.82 ± 0.06	0.90 ± 0.04
SVM (rbf)	0.79 ± 0.11	0.85 ± 0.10	0.69 ± 0.13	0.79 ± 0.06	0.73 ± 0.07	0.83 ± 0.07	0.80 ± 0.14	0.89 ± 0.09	0.88 ± 0.10	0.96 ± 0.06	0.86 ± 0.08	0.92 ± 0.07	0.72 ± 0.10	0.81 ± 0.10	0.80 ± 0.06	0.88 ± 0.04	0.77 ± 0.05	0.84 ± 0.05
SVM (lin)	0.75 ± 0.13	0.76 ± 0.25	0.69 ± 0.16	0.75 ± 0.22	0.73 ± 0.08	0.82 ± 0.09	0.63 ± 0.35	0.72 ± 0.34	0.89 ± 0.06	0.97 ± 0.05	0.84 ± 0.04	0.92 ± 0.05	0.73 ± 0.09	0.81 ± 0.09	0.80 ± 0.06	0.87 ± 0.06	0.76 ± 0.05	0.84 ± 0.07

Table 11. MSE and R^2 score of regression tasks using *performance* measures, *subjective* feedback and the linear combination of both as the target value for *n*-back, *k*-drive and the combination of both experiments.

	***n*-back**		***k*-drive**		**Both**	
	R^2	MSE	R^2	MSE	R^2	MSE
Performance	0.50 ±0.11	0.038 ±0.008	0.53 ±0.12	0.030 ±0.009	0.51 ±0.07	0.037 ±0.005
Subjective	0.43 ±0.13	0.043 ±0.013	0.59 ±0.06	0.030 ±0.006	0.48 ±0.08	0.041 ±0.011
Combination	0.53 ±0.10	0.041 ±0.010	0.58 ±0.08	0.033 ±0.005	0.54 ±0.08	0.041 ±0.009

4. Discussion

The presented results compiled in the previous chapter of this study provide a valuable addition to the research community assessing the effect of cognitive load using multimodal measurements.

With a carefully designed study population and a detailed cohort description, we conducted two experiments with various levels of cognitive load. As a reference experiment, that is comparable to experiments on cognitive load in related work, we asked the subjects to participate in a single- and dual-task load *n*-back test with visual-only stimuli and audiovisual stimuli inducing cognitive load. As an application-motivated test, we developed the *k*-drive test, which is inspired by the observations of semi-autonomously driving cars with only a few driver-car interactions required. This test imposes cognitive load by an increasing number of events the subject needs to respond to. In addition, a secondary task of interacting with a car infotainment system needs to be executed for dual-task load tests.

Our study compiled a recording setup for multiple modalities. Our *physiological* measurements include biosignals such as ECG, PPG, EDA, EMG, skin temperature and respiration recordings as well as eye tracker data. We found significant changes in established expert and statistical features during both, *n*-back and *k*-drive, which revealed an increasing heart rate, decreasing heart rate variability, an increasing number of peaks in the electrodermal activity, a decreasing mean skin temperature, a decreasing respiration rate and an increasing eye tracker IPA during cognitive load compared to non-load baselines, to name only a few. Our findings of a positive correlation between IPA and cognitive load align with previous research that has reported promising results using IPA in cognitive load measurement [93,110,111]. Additionally, our omnibus test has shown the statistical significance of fixation features which also prove to be relevant in cognitive load prediction. This finding complies with literature [30,85,112–114] In addition, we recorded *behavioral* measurements based on facial videos. Using simple features based on the number of activated action units within a certain time frame, we found several action units that changed significantly for increasing levels of task demand of *n*-back and *k*-drive levels. These observations are supported by the findings of Yuce et al. who analyzed the link between action units and cognitive load while driving [29]. We measured several *performance* metrics for all conducted *n*-back and *k*-drive tests and levels for both secondary and primary task loads, including precision and recall. In combination with the *subjective* feedback, acquired through NASA-TLX questionnaires following every test, we found properties of human responses, that are statistically significant across all analyzed dimensions: *physiological*, *behavioral*, *performance* and *subjective* measurements. Comparing our collected dataset with work analyzed by a recent review of related data collections by Seitz et al., confirms the validity of our multimodal approach for the computerized detection of cognitive load, differentiating our setup with a combination of multiple modalities, that all have been used in subsets, but not as a combination during the same protocol [115]. The review of Seitz et al. further helps to highlight our contributions, as we provide multiple potential targets for model training and evaluation: subjective ratings, performance-based measurements and task-specific level information, which all have been studied independently in related work and are presented as alternatives in this work [115].

Starting with a unimodal input we train several machine learning models and present the predictive power of each biosignal modality, eye tracker data and facial action units separating low and high levels of cognitive load. By combining features of various modalities, we present the predictive power of all biosignals, biosignals and eye tracker data and the combination of eye tracker data, biosignal and action units, leading to an overall AUC, using the complete dataset consisting of collected single- and dual-task load $n \in \{1, 2, 3\}$-back and $k \in \{1, 2, 3\}$-drive test of 0.91 ± 0.02. In general, we can observe that combining modalities leads to an improvement in classification accuracy. This is especially evident in the case of using data from both tasks and emphasizes the importance of a multimodal approach, given that other publications report similar results [116–118]. In addition, we report the gain feature importance values of our *XGBoost* classifier, showing important contributions of features like heart rate, heart rate variability, change of skin conductance within a given time interval, IPA, fixations and saccades of our eye tracker data as features with strong predictive power in our binary low-vs-high cognitive load

classification task. During *n*-back we were able to control the lighting conditions to be more stable, compared to the application-motivated *k*-drive test. This could be a reason, why IPA is the most important feature for the detection of cognitive load for one task while losing importance for the other task. Nonetheless, we found eye-tracking-based features to have a high contribution (see Appendix G) for the detection of cognitive load in this publication. This finding is in line with work by Ahmed et al. [119], who observed, that pupil-related features have the highest feature importance in comparison to ECG and respiration for multimodal classification. We evaluate several machine learning techniques using kNN, SVM with linear and radial kernel and find that XGBoost performs best using our complete set of features. We introduced a new task separating three levels of cognitive load: low, medium and high and found a high true positive rate for the detection of tasks that imposed only low loads onto the subject. The separation of levels with medium and high cognitive load leaves room for improvement in future work. It has also been observed in other publications that separation with more than 2 classes is challenging [56,117]. One possible reason for this could be a strong inter-subject variability, making fine-grained classification difficult.

In addition to our classification tasks, that separate the different levels of CL imposed onto the subject based on the task difficulty of each level, we introduced a regression task, that used *performance* metrics, *subjective* ratings and the linear combination of both measures as targets, resulting in a R^2 score of 0.54 ± 0.14 for all levels and tasks. To summarize our work, this work contributes:

- The introduction of *k*-drive, a novel and close-to-real-world autonomous driving task to study cognitive load.
- The collection of *physiological*, *behavioral*, *subjective* and *performance* measurement from 51 subjects while participating in levels of increasing task difficulty with single and dual workload scenarios.
- The extraction and evaluation of features from ECG, PPG, EMG, EDA, respiration belt, skin temperature, eye tracker data, facial video recordings using a detailed statistical evaluation protocol.
- The validation of the collected data using statistical methods before training and testing of machine learning models.
- The training and evaluation of machine learning models using unimodal and multimodal inputs for cognitive load estimation.
- The analysis of feature importance for models trained with multiple inputs.
- The introduction of novel machine learning tasks and training of baseline models as reference for future contributions by the affective sensing and machine learning community.

Accompanying this publication we release a subset of 30 subjects, containing the recorded and unprocessed sensor data, questionnaires, performance metrics and NASA-TLX scores, enabling the research community to develop and evaluate novel algorithms for cognitive load estimation. We encourage the research community to build novel algorithms and machine-learning methods using the released data.

Our future work will include the analysis of distribution shifts of machine learning models that are trained on one task and evaluated on another task. An important next step before deploying these algorithms in the wild is the analysis of the robustness and stability of fusion techniques for multimodal machine learning when one or more modalities are corrupted or perturbed. In direct line with this work will be the study of models trained in an end-to-end manner such as deep neural networks using the raw input signals for CL prediction. Another important factor is the inclusion of the reported personality traits as input to machine learning models to improve model performances using personalized information. In this publication, we have selected a constant window length of two minutes for all features and all inputs, based on results from literature introduced in Section 2.8. However, physiological responses manifest themselves on different timescales and the optimization of both latency and window length of all extracted features during or after phases with high CL may further improve the detection rate using computerized machine

learning models. Given the precise measurements of time with a resolution of only 0.5 ms for events and phases of different levels of CL and the results of this study this is a promising direction that will be explored and described in future work.

Author Contributions: Conceptualization, M.P.O., A.F., N.R.L.; methodology, M.P.O. and A.F., J.D.; software, M.P.O., A.F., J.D.; validation, M.P.O., A.F.; formal analysis, M.P.O.; investigation, M.P.O., A.F., J.D.; resources, M.P.O., A.F.; data curation, M.P.O., A.F.; writing—original draft preparation, M.P.O.; writing—1st review and editing, M.P.O., A.F., J.D.; writing—2nd review and revising, M.P.O., A.F., J.D., N.R.L., N.H., B.M.E., S.H.Y.; visualization, M.P.O.; supervision, N.R.L., N.H., B.M.E., S.H.Y.; project administration, N.R.L., N.H.; funding acquisition, N.R.L., N.H.; All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Bavarian Ministry of Economic Affairs, Regional Development and Energy through the Center for Analytics-Data-Applications (ADA-Center) within the framework of "BAYERN DIGITAL II" (20-3410-2-9-8). The hardware infrastructure is partly funded by the Federal Ministry of Education and Research under the project reference numbers 16FMD01K, 16FMD02 and 16FMD03.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Ethics Committee of Friedrich-Alexander-University Erlangen Nuremberg with protocol code 129_21 B on 21.04.2021.

Informed Consent Statement: Written informed consent was obtained from all subjects involved in the study. Subjects received a chocolate bar with delicious flavors as compensation after the experiment.

Data Availability Statement: Interested parties may use a subset (30 subjects) of the data presented here, after returning a signed End User License Agreement (EULA), to the Fraunhofer Institute of Integrated Circuits. The signed EULA should be returned in digital format by sending it to *adabase@iis.fraunhofer.de*. The usage of the dataset for any nonacademic purpose is prohibited. Nonacademic purposes include, but are not limited to: proving the efficiency of commercial systems, training or testing of commercial systems, selling data from the dataset, creating military applications and developing governmental systems used in public spaces.

Acknowledgments: We are extremely grateful to Dominik Weber for providing help setting up the BIOPAC systems. We would like to express our deepest appreciation for Elisabeth Städler's support when implementing the user interface of the test software and providing utilities to import the acquired data into our database. Furthermore, we would like to recognize the assistance of David Hartmann, who build a user interface for manual and visually assisted verification of data completeness and correctness. We acknowledge the contribution of Dominik Seuß, who helped during project initiation. We would like to thank Hannah Merz for valuable discussions about the marketing perspective of potential products. It was extremely motivating to see interest in our research for product development. We would like to acknowledge having discussions with Jens-Uwe Garbas at initiation and for the financial and administrative support during the whole project. We are exceptionally grateful for Hannah Böglers research about potential cognitive load tests and influences of different medications relevant to our project. Last but certainly not least, we would like to express our deepest thank you to Matthias Struck who helped clear the time to conduct this research, providing compute resources and most importantly has proven to be a strong reliant partner.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

ADABase	**A**utonomous **D**riving Cognitive Load **A**ssessment Data**base**
ANOVA	Analysis of Variance
AUC	Area Under the Receiver Operating Characteristic Curve
BFI	Big Five Inventory
BFI-K	Big Five Inventory-Kurz
BMI	Body-Mass-Index
BR	Breathing Rate
CL	Cognitive Load

ECG	Electrocardiography
EDA	Electrodermal Activity
EMG	Electromyography
EULA	End User License Agreement
FACS	Facial Action Coding System
HDF5	Hierarchical Data Format V5
HPA	Hypothalamic-Pituitary-Adrenal
HRV	Heart Rate Variability
IPA	Index of Pupillary Activity
kNN	k-Nearest-Neighbors
MAD	Median Absolute Deviation
MSE	Mean Squared Error
NAS	Negative Affection Sub-Schedule
PANAS	Positive and Negative Affect Schedule
PAS	Positive Affection Sub-Schedule
PMWL	Perceived Mental Workload
PPG	Photoplethysmogram
PRV	Pulse Rate Variability
PSS	Perceived Stress Scale
PTT	Pulse-Transit Time
RTLX	Raw Task Load Index
SCL	Skin Conductance Level
SCR	Skin Conductance Response
SKT	Skin Temperature
SNS	Sympathetic Nervous System
SQI	Signal Quality Index
SVM	Support Vector Machine
TLX	Task Load Index
USB	Universal Serial Bus
WHO	World Health Organization
WM	Working Memory
XGBoost	eXtreme Gradient Boosting

Appendix A

To ensure the highest interoperability between various research environments, we release the data in Hierarchical Data Format V5 (HDF5) [120] file format. Each subject is stored in a separate file and every file contains different hdf5-groups: The *META* group contains information such as sex, age and weight, *SUBJECTIVE* contains the feedback of that subject, like TLX answers of every dimension of the conducted tests, *PERFORMANCE* contains the metrics presented in Section 2.7 and *SIGNALS* contains all continuously collected data samples, like raw electrocardiograms or raw respiration measurements. *SIGNALS* also contains the extracted action units, as we do not release video data to protect the subjects' privacy. To save disk space and reduce transfer times, we resampled the signals to commonly used sampling frequencies. The data is stored in a table-like format and each sample has an associated timestamp in milliseconds starting at the beginning of the recording. As the files contain the continuously sampled signals, the user can identify the current experiment, using the *STUDY* field (*n*-back, *k*-drive, n/a), while the *PHASE* field (baseline, train, test, n/a) is more fine-grained containing information about the current task. The different levels (e.g., dual vs. single task *n*-back test) are contained in *LEVELS*. All three encoded states of the experiment: *STUDY*, *PHASE*, *LEVELS* have an associated timestamp in milliseconds to ensure fully synchronous access. As a reference python-like pseudocode interface one might implement code similar to this:

```
import pandas as pd

# Continuously recorded ECG during the nback test
data = pd.read_hdf(subject_file, "SIGNALS", mode='r')
data.loc[data["STUDY"] == 'nback', ["TS", "ECG_RAW"]].dropna()

# Reaction time of subject for level 1, n-back test.
data = pd.read_hdf(subject_file, "PERFORMANCE", mode='r')
data.loc[
(data["PHASE"] == 'nback') &
(data["LEVELS"] == 1), "VISUAL REACTION TIME"]

# Accessing available values for each table
pd.read_hdf(subject_file, "SIGNALS").columns
```

Appendix B

Depending on the task, frequent successive fixations or rapid blinking may indicate increased cognitive load [85]. Figure A1 shows the frequency of fixations over a two-minute cognitive load stimulus.

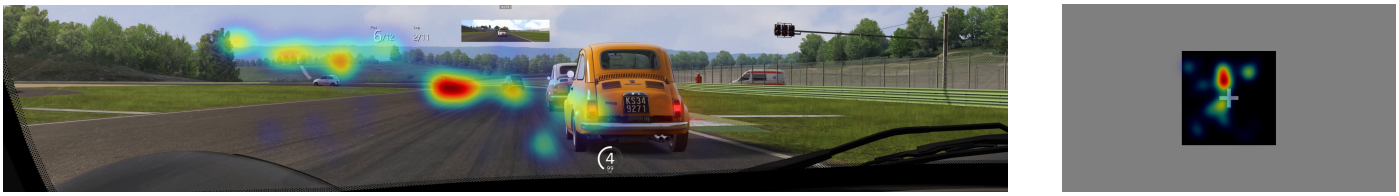

Figure A1. Fixation heatmaps for *k*-drive (**left**) and *n*-back (**right**) scenarios over a two-minute time interval. Darker (red) clusters show a high frequentness of fixations at this location.

Appendix C

Motivated by related work and our findings in Section 3.8, we have appended the ocean personality traits acquired using the BFI-K questionnaire to the feature vector for classification. The results are presented in Table A1. Using this simple method of simply appending the ocean scores to the feature vectors does not improve the predictive power outside the levels of confidence and therefore remains a task for future work.

Table A1. Results of the trained classifier using the same evaluation protocol described in Section 2.14 for a XGBoost classifier with and without attached ocean scores and all biosignal, eye tracker and action unit features.

	n-back						*k*-drive						Both					
	Single		Dual		Comb.		Single		Dual		Comb.		Single		Dual		Comb.	
	F_1	*AUC*	F_1	*AUC*	F_1	*AUC*	F_1	*AUC*	F_1	*AUC*	F_1	*AUC*	F_1	*AUC*	F_1	*AUC*	F_1	*AUC*
All, *OCEAN*	0.84 ± 0.04	0.93 ± 0.04	0.75 ± 0.09	0.85 ± 0.07	0.75 ± 0.09	0.87 ± 0.06	0.91 ± 0.13	0.97 ± 0.05	0.94 ± 0.04	1.00 ± 0.01	0.90 ± 0.05	0.97 ± 0.02	0.83 ± 0.06	0.91 ± 0.05	0.83 ± 0.11	0.92 ± 0.06	0.80 ± 0.06	0.90 ± 0.02
All	0.82 ± 0.09	0.93 ± 0.06	0.78 ± 0.09	0.88 ± 0.09	0.77 ± 0.07	0.88 ± 0.05	0.88 ± 0.09	0.96 ± 0.06	0.94 ± 0.05	0.99 ± 0.01	0.89 ± 0.06	0.96 ± 0.03	0.84 ± 0.06	0.92 ± 0.06	0.83 ± 0.05	0.92 ± 0.04	0.82 ± 0.06	0.90 ± 0.04

Appendix D

As an addition to our description in Sections 2.3 and 2.4, describing the detailed sequence of tests, Figure A2 shows a schematic sequence of different tests and experiments conducted in this publication. It is closely related to our visualization of extracted features during the complete experiment in Section 3.1.

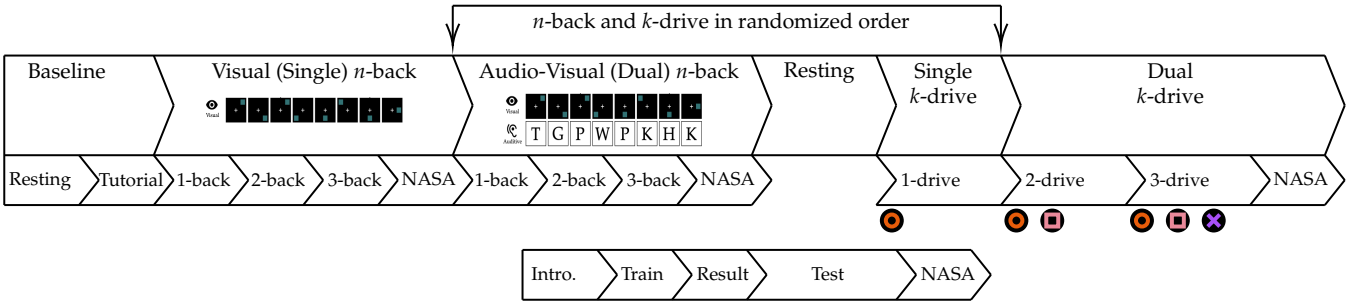

Figure A2. Visualization of a complete experimentation sequence with two experiments, *n*-back during single and dual-task load with three levels each, followed by a NASA-TLX questionnaire for each dimension. The *n*-back test is then concluded by a weighting of every NASA-TLX dimension and followed by a resting phase, where the subject answered the questionnaires described in Section 2.1. During *k*-drive the subject is conducting three levels $k \in \{1, 2, 3\}$ with tasks described in Section 2.3. 1-drive is conducted during a single-task load and 2/3-drive during dual-task load. The detailed sequence of one level for *n*-back and *k*-drive is depicted in the sequence diagram below the main diagram and holds for all *n*-back and *k*-drive tests.

Appendix E

The relationship between stress and cognition or performance has been extensively studied over the last few years [121–123]. Our line of work is not focused on the development of new models capturing the relationship between stress and cognition or performance. We are therefore not conducting any experiments in this direction. However, we asked the subjects to answer a PSS questionnaire to ensure that our sampled study population is not chronically stressed. The results are presented in the first column of Figure A3.

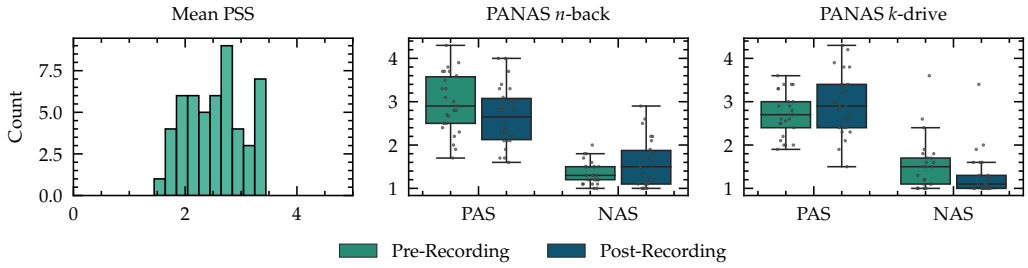

Figure A3. Results of PSS and PANAS questionnaires. The first column shows the distribution of mean values of Perceived Stress Scale (PSS) answered during the questionnaire phase, indicating the long-term stress values of the study population. The second (*n*-back) and third (*k*-drive) columns present the mean PANAS rating for positive affection sub-schedule (PAS) and negative affection sub-schedule (NAS) before and after the experiment.

Similar to interaction patterns between chronic or social stress and working memory performance, studies have found interactions between cognitive load and emotions [124]. The results are presented in Figure A3. After conducting a two-sided paired t-test using the PANAS scores, we find *p*-values of PAS for *n*-back of $p = 0.110$, NAS for *n*-back of $p = 0.043$ and PAS for *k*-drive of $p = 0.005$ and NAS for *k*-drive of $p = 0.027$, which indicates that *k*-drive test was a significantly more pleasing experience than *n*-back while eliciting similar levels of cognitive load, as shown in Section 3.

Appendix F

As an additional reference for researchers working with this database, we present our proposal for binary-two-level cognitive load classification, described in text form in Section 2.13 in Table A2.

Table A2. Tabular visualization of groups of experiments and levels used in our binary classification task. Described in greater detail in Section 2.4 we denote the n-back baseline 1 as resting phase and n-back baseline 2 as the tutorial phase. The baselines for k-drive are described in Section 2.3, where k-drive baseline 1 is resting, baseline 2 is tutorial-driving-observation and baseline 3 is tutorial-music-application.

		Low Cognitive Load	**High Cognitive Load**
n-back	Single Task	n-back baseline 2, visual: $n \in \{1\}$	visual: $n \in \{2,3\}$
	Dual Task	n-back baseline 2, audio-visual $n \in \{1\}$	audio-visual: $n \in \{2,3\}$
	Both Task	n-back baseline 1, 2, visual and audio-visual: $n \in \{1\}$	visual and audio-visual: $n \in \{2,3\}$
k-drive	Single Task	k-drive baseline 2	single: $k \in \{1\}$
	Dual Task	k-drive baseline 2, 3	dual: $k \in \{2,3\}$
	Both Task	k-drive baseline 1, 2, 3	$k \in \{1,2,3\}$
n-back, k-drive	Single Task	n-back baseline 2, visual $n \in \{1\}$, k-drive baseline 2	visual: $n \in \{2,3\}, k \in 1$
	Dual Task	n-back baseline 2, audio-visual: $n \in 1$ k-drive baseline 2, 3	audio-visual: $n \in 2,3$ and dual $k \in 2,3$
	Both Task	n-back baseline 1, 2, visual, audio-visual: $n \in 1$ k-drive baseline 1, 2, 3	visual audio-visual: $n \in 2,3$ and dual $k \in 1,2,3$

Appendix G

In addition to our reported *gain* feature importance in Figure 15 for binary classifiers trained on data from both experiments, we want to highlight the task-specific feature importance for classifiers trained on either n-back or k-drive tests. The XGBoost classifiers assigned features from eye-tracker and heart-rate-based measures a high contribution to the final classification result for our k-drive experiment shown in Figure A4. For our n-back test (see Figure A4) features based on the electrodermal activity and the eye-tracker had the highest contribution. It is noteworthy that the change of skin conductance levels within a window $(\sum_t SCL'(t))$ is the highest contributing feature (also with a very high variance) to the final classification for k-drive only classifications. We hypothesize this might be caused by movement artifacts (even though the left hand remained at rest, right arm movement might interfere with the electrodermal activity). However, the analysis of this effect remains for future work. The reported gain values in Section 3.10 reflect the relative contribution of a single feature to the final prediction. In addition to this metric, *coverage* remains an important tool to measure (count) the number of observations concerned by a single feature.

Comparing the feature importance of classifiers trained only on either k-drive or n-back indicates potential spurious correlations: Subjects in k-drive moved the dominant arm to the infotainment system for dual-task workloads (as described in Section 2.8 the EDA electrode is placed on the opposite side), which might increase skin conductance events per interval. During n-back the IPA feature extracted from eye tracker data showed increased importance. The detailed analysis of this behavior and more generally the identification of distribution shifts and spurious correlations is up to future work.

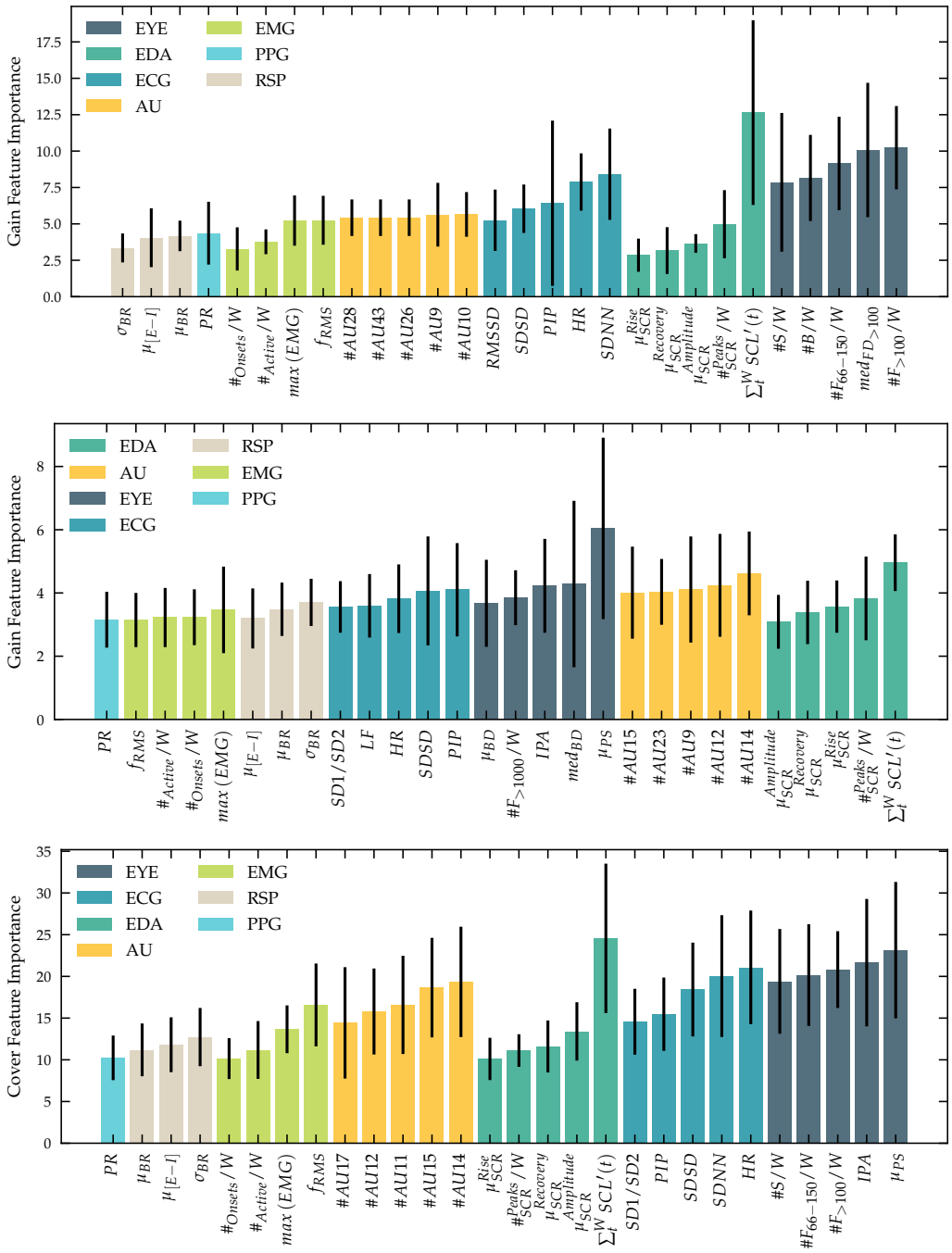

Figure A4. XGBoost *gain* feature importance of single and dual task *k*-drive (first row) and *n*-back (second row), using binary classification (high/low CL) and XGBoost *coverage* feature importance for single and dual-task *n*-back and *k*-drive in the last row. The feature-modality-groups (colors) are ordered by the importance of the complete group with the highest importance on the right and the lowest importance of the complete group on the left. We drop features with very low importance for visual clarity.

Appendix H

As described in Sections 2.3, 2.4 and 2.11, we designed our acquisition setup such that we reduce the number of instances with signal corruptions and artifacts. However, when recording human subjects some react to certain situations differently than others. These cases might lead to missing data, e.g., by ECG electrode fall-offs, respiration-belt misplacement (to tight vs. to wide) or by the movements of the face outside the recording

area of the video camera or the eye tracking region. The methods of handling these cases are described in greater detail in Section 2.11. For completeness, we present the number of instances used throughout our experiment for statistical evaluation and machine learning in Table A3.

Table A3. Number of records used for statistical evaluation, after removing corrupted records (see Section 2.11) and outliers in integer percent. The total number of subjects in this study was 51. Methodological details are presented in Section 2.11. The extracted features are described in Sections 2.8–2.10.

Modality	**Feature**	***n*-back**								***k*-drive**					
		Baseline		**Single**			**Dual**			**Baseline**			**Levels**		
		1	**2**	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**
ECG	HR	100	100	100	100	100	100	100	100	100	100	98	100	100	100
	$SDSD$	100	100	100	100	100	100	100	100	100	100	98	100	100	100
	$SDNN$	100	100	100	100	100	100	100	100	100	100	98	100	100	100
	$RMSSD$	100	100	100	100	100	100	100	100	100	100	98	100	100	100
	LF	100	100	100	100	100	100	100	100	100	100	98	100	100	100
	HF	100	100	100	100	100	100	100	100	100	100	98	100	100	100
	LF_n	100	100	100	100	100	100	100	100	100	100	98	100	100	100
	HF_n	100	100	100	100	100	100	100	100	100	100	98	100	100	100
	LF/HF	100	100	100	100	100	100	100	100	100	100	98	100	100	100
	$SD1/SD2$	100	100	100	100	100	100	100	100	100	100	98	100	100	100
	PSS	100	100	100	100	100	100	100	100	100	100	98	100	100	100
	PIP	100	100	100	100	100	100	100	100	100	100	98	100	100	100
PPG	PR	100	100	100	100	100	100	100	100	100	100	98	100	100	100
EDA	$\mu_{SCR}^{Amplitude}$	100	100	100	100	100	100	100	100	100	100	98	100	100	100
	$\sum_t^W SCL'(t)$	100	100	100	100	100	100	100	100	100	100	98	100	100	100
	$\#_{SCR}^{Peaks}/W$	100	100	100	100	100	100	100	100	100	100	98	100	100	100
	μ_{SCR}^{Rise}	100	100	100	100	100	100	100	100	100	100	98	100	100	100
	$\mu_{SCR}^{Recovery}$	100	100	100	100	100	100	100	100	100	100	98	100	100	100
EMG	$\#_{Onsets}/W$	98	98	98	98	98	98	98	98	98	98	96	98	98	98
	$\#_{Active}/W$	98	98	98	98	98	98	98	98	98	98	96	98	98	98
	$max(EMG)$	100	100	100	100	100	100	100	100	100	100	98	100	100	100
	f_{RMS}	100	100	100	100	100	100	100	100	100	100	98	100	100	100
SKT	μ_T	98	96	98	96	94	92	92	92	98	96	90	98	94	96
	σ_T	98	96	98	96	94	92	92	92	98	96	90	98	94	96
	min_T	98	96	98	96	94	92	92	92	98	96	90	98	94	96
	max_T	98	96	98	96	94	92	92	92	98	96	90	98	94	96
	$\sum T'(t)/T$	98	96	98	96	94	92	92	92	98	96	90	98	94	96
RSP	μ_{BR}	100	100	100	100	100	100	100	100	100	100	98	100	100	100
	σ_{BR}	100	100	100	100	100	100	100	100	100	100	98	100	100	100
	$\mu_{[E-I]}$	100	100	100	100	100	100	100	100	100	100	98	100	100	100
EYE	$\#F_{>100}/W$	86	86	86	86	86	86	86	86	86	86	86	86	86	86
	$\#F_{66-150}/W$	86	86	86	86	86	86	86	86	86	86	86	86	86	86
	$\#F_{300-500}/W$	86	86	86	86	86	86	86	86	86	86	86	86	86	86
	$\#F_{>1000}/W$	86	86	86	86	86	86	86	86	86	86	86	86	86	86
	$\mu_{FD_{>100}}$	86	86	86	86	86	86	86	86	86	86	82	86	86	86
	$med_{FD_{>100}}$	86	86	86	86	86	86	86	86	86	86	82	86	86	86
	$\#S/W$	86	86	86	86	86	86	86	86	86	86	86	86	86	86
	μ_{SA}	86	86	84	84	84	84	84	84	86	86	66	86	86	86
	μ_{SD}	86	86	84	84	84	84	84	84	86	86	68	86	86	86
	med_{SD}	86	86	84	84	84	84	84	84	86	86	68	86	86	86
	$\#B/W$	86	86	86	86	86	86	86	86	86	86	86	86	86	86
	μ_{BD}	86	86	84	86	84	86	86	86	86	86	64	86	86	86
	med_{BD}	86	86	84	86	84	86	86	86	86	86	64	86	86	86
	μ_{PS}	86	86	86	86	86	86	86	86	86	86	86	86	86	86
	IPA	86	86	86	86	86	86	86	86	86	86	86	86	86	86
AUS	$\#AU$	66	66	66	66	66	66	66	66	66	66	66	66	66	66

References

1. Stutts, J.C.; Reinfurt, D.W.; Staplin, L.; Rodgman, E.A. *The Role of Driver Distraction in Traffic Crashes: (363942004-001)*; Technical Report; AAA Foundation for Traffic Safety: Washington, DC, USA, 2001. [CrossRef]
2. McEvoy, S.P.; Stevenson, M.R.; Woodward, M. The prevalence of, and factors associated with, serious crashes involving a distracting activity. *Accid. Anal. Prev.* **2007**, *39*, 475–482. [CrossRef] [PubMed]
3. Engström, J.; Johansson, E.; Östlund, J. Effects of Visual and Cognitive load in real and simulated motorway driving. *Transp. Res. Part F Traffic Psychol. Behav.* **2005**, *8*, 97–120. [CrossRef]
4. On-Road Automated Driving (ORAD) committee. *Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles*; Technical Report; SAE International: Warrendale, PA, USA, 2021. [CrossRef]
5. Paxion, J.; Galy, E.; Berthelon, C. Mental workload and driving. *Front. Psychol.* **2014**, *5*, 1344. [CrossRef] [PubMed]
6. Engström, J.; Markkula, G.; Victor, T.; Merat, N. Effects of Cognitive Load on Driving Performance: The Cognitive Control Hypothesis. *Hum. Factors: J. Hum. Factors Ergon. Soc.* **2017**, *59*, 734–764. [CrossRef] [PubMed]
7. Banks, V.A.; Eriksson, A.; O'Donoghue, J.; Stanton, N.A. Is partially automated driving a bad idea? Observations from an on-road study. *Appl. Ergon.* **2018**, *68*, 138–145. [CrossRef] [PubMed]
8. Simon, H. Designing Organizations for a Information-Rich World. In *Computers, Communications, and the Public Interest*; Greenberger, M., Ed.; The Johns Hopkins Press: Balti-More, MD, USA, 1971. Available online: https://www.cs.purdue.edu/homes/ribeirob/pdf/HerbertSimon_waybackmachine.pdf (accessed on 25 November 2022).
9. Baddeley, A. Working Memory. *Science* **1992**, *255*, 556–559. [CrossRef]
10. Sweller, J.; van Merrienboer, J.J.G.; Paas, F.G.W.C. Cognitive Architecture and Instructional Design. *Educ. Psychol. Rev.* **1998**, *10*, 251–296. [CrossRef]
11. Paas, F.; Tuovinen, J.E.; Tabbers, H.; Van Gerven, P.W.M. Cognitive Load Measurement as a Means to Advance Cognitive Load Theory. *Educ. Psychol.* **2003**, *38*, 63–71. [CrossRef]
12. Chen, F.; Zhou, J.; Wang, Y.; Yu, K.; Arshad, S.Z.; Khawaji, A.; Conway, D. *Robust Multimodal Cognitive Load Measurement*; Human–Computer Interaction Series; Springer International Publishing: Cham, Switzerland, 2016. [CrossRef]
13. Nourbakhsh, N.; Chen, F.; Wang, Y.; Calvo, R.A. Detecting Users' Cognitive Load by Galvanic Skin Response with Affective Interference. *ACM Trans. Interact. Intell. Syst.* **2017**, *7*, 1–20. [CrossRef]
14. Paas, F.G.W.C. Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach. *J. Educ. Psychol.* **1992**, *84*, 429–434. [CrossRef]
15. Georgsson, D.M. NASA RTLX as a Novel Assessment Tool for Determining Cognitive Load and User Acceptance of Expert and User-based Usability Evaluation Methods. *Eur. J. Biomed. Inform.* **2020**, *16*, 8.
16. Hart, S.G. Nasa-Task Load Index (NASA-TLX); 20 Years Later. In Proceedings of the Human Factors and Ergonomics Society 50th Annual Meeting, San Francisco, CA, USA, 16–20 October 2006; p. 5.
17. Hart, S.G.; Staveland, L.E. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. *Adv. Psychol.* **1988**. *52*, 139–183. [CrossRef]
18. Zu, T.; Munsell, J.; Rebello, N.S. Subjective Measure of Cognitive Load Depends on Participants' Content Knowledge Level. *Front. Educ.* **2021**, *6*, 647097. [CrossRef]
19. Sweller, J.; Ayres, P.; Kalyuga, S. *Measuring Cognitive Load*; Springer: New York, NY, USA, 2011; pp. 71–85. [CrossRef]
20. Jaeggi, S.M.; Buschkuehl, M.; Perrig, W.J.; Meier, B. The concurrent validity of the N-Back task as a working memory measure. *Memory* **2010**, *18*, 394–412. [CrossRef] [PubMed]
21. Khawaja, M.A. Cognitive Load Measurement Using Speech and Linguistic Features. Ph.D. Thesis, UNSW, Sydney, Australia, 2010. [CrossRef]
22. Schuller, B.; Steidl, S.; Batliner, A.; Epps, J.; Eyben, F.; Ringeval, F.; Marchi, E.; Zhang, Y. The INTERSPEECH 2014 Computational Paralinguistics Challenge: Cognitive & Physical Load. In Proceedings of the INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association, Singapore, 14–18 September 2014; p. 5.
23. Huttunen, K.; Keränen, H.; Väyrynen, E.; Pääkkönen, R.; Leino, T. Effect of cognitive load on speech prosody in aviation: Evidence from military simulator flights. *Appl. Ergon.* **2011**, *42*, 348–357. [CrossRef]
24. Yin, B.; Chen, F.; Ruiz, N.; Ambikairajah, E. Speech-based cognitive load monitoring system. In Proceedings of the 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, Las Vegas, NV, USA, 31 March–4 April 2008; IEEE: Las Vegas, NV, USA, 2008; pp. 2041–2044, ISSN: 1520-6149. [CrossRef]
25. Ruiz, N.; Taib, R.; Shi, Y.D.; Choi, E.; Chen, F. Using pen input features as indices of cognitive load. In Proceedings of the Ninth International Conference on Multimodal Interfaces—ICMI '07, Aichi, Japan, 12–15 November 2007; ACM Press: Nagoya, Aichi, Japan, 2007; p. 315. [CrossRef]
26. Yu, K.; Epps, J.; Chen, F. Mental Workload Classification via Online Writing Features. In Proceedings of the 2013 12th International Conference on Document Analysis and Recognition, Washington, DC, USA, 25–28 August 2013; IEEE: Washington, DC, USA, 2013; pp. 1110–1114. [CrossRef]
27. Arshad, S.; Wang, Y.; Chen, F. Analysing mouse activity for cognitive load detection. In Proceedings of the 25th Australian Computer-Human Interaction Conference on Augmentation, Application, Innovation, Collaboration—OzCHI '13, Adelaide, SA, Australia, 25–29 November 2013; ACM Press: Adelaide, SA, Australia, 2013; pp. 115–118. [CrossRef]

28. Viegas, C.; Lau, S.H.; Maxion, R.; Hauptmann, A. Towards Independent Stress Detection: A Dependent Model Using Facial Action Units. In Proceedings of the 2018 International Conference on Content-Based Multimedia Indexing (CBMI), La Rochelle, France, 4–6 September 2018; IEEE: La Rochelle, France, 2018; pp. 1–6. [CrossRef]
29. Yuce, A.; Gao, H.; Cuendet, G.L.; Thiran, J.P. Action Units and Their Cross-Correlations for Prediction of Cognitive Load during Driving. *IEEE Trans. Affect. Comput.* **2017**, *8*, 161–175. [CrossRef]
30. Chen, S.; Epps, J.; Ruiz, N.; Chen, F. Eye activity as a measure of human mental effort in HCI. In Proceedings of the 15th International Conference on Intelligent User Interfaces—IUI '11, Palo Alto, CA, USA, 13–16 February 2011; ACM Press: Palo Alto, CA, USA, 2011; p. 315. [CrossRef]
31. Crundall, D.E.; Underwood, G. Effects of experience and processing demands on visual information acquisition in drivers. *Ergonomics* **1998**, *41*, 448–458. [CrossRef]
32. Xu, J.; Wang, Y.; Chen, F.; Choi, E. *Pupillary Response Based Cognitive Workload Measurement under Luminance Changes*; Series Title: Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2011; Volume 6947, pp. 178–185. [CrossRef]
33. Giannakakis, G.; Grigoriadis, D.; Giannakaki, K.; Simantiraki, O.; Roniotis, A.; Tsiknakis, M. Review on psychological stress detection using biosignals. *IEEE Trans. Affect. Comput.* **2019**, *13*, 440–460. [CrossRef]
34. Woody, A.; Hooker, E.D.; Zoccola, P.M.; Dickerson, S.S. Social-evaluative threat, cognitive load, and the cortisol and cardiovascular stress response. *Psychoneuroendocrinology* **2018**, *97*, 149–155. [CrossRef]
35. Lisetti, C.L.; Nasoz, F. Using Noninvasive Wearable Computers to Recognize Human Emotions from Physiological Signals. *EURASIP J. Adv. Signal Process.* **2004**, *2004*, 929414. [CrossRef]
36. Jinjun, W.; Yihong, G. Recognition of multiple drivers emotional state. In Proceedings of the 2008 19th International Conference on Pattern Recognition, Tampa, FL, USA, 8–11 December 2008; IEEE: Tampa, FL, USA, 2008; pp. 1–4, ISSN: 1051-4651. [CrossRef]
37. Koelstra, S.; Muhl, C.; Soleymani, M.; Lee, J.-S.; Yazdani, A.; Ebrahimi, T.; Pun, T.; Nijholt, A.; Patras, I. DEAP: A Database for Emotion Analysis; Using Physiological Signals. *IEEE Trans. Affect. Comput.* **2012**, *3*, 18–31. [CrossRef]
38. Abadi, M.K.; Subramanian, R.; Kia, S.M.; Avesani, P.; Patras, I.; Sebe, N. DECAF: MEG-Based Multimodal Database for Decoding Affective Physiological Responses. *IEEE Trans. Affect. Comput.* **2015**, *6*, 209–222. [CrossRef]
39. Hovsepian, K.; al'Absi, M.; Ertin, E.; Kamarck, T.; Nakajima, M.; Kumar, S. cStress: Towards a gold standard for continuous stress assessment in the mobile environment. In Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing—UbiComp '15, Osaka, Japan, 7–11 September 2015; ACM Press: Osaka, Japan, 2015; pp. 493–504. [CrossRef]
40. Subramanian, R.; Wache, J.; Abadi, M.K.; Vieriu, R.L.; Winkler, S.; Sebe, N. ASCERTAIN: Emotion and Personality Recognition Using Commercial Sensors. *IEEE Trans. Affect. Comput.* **2018**, *9*, 147–160. [CrossRef]
41. Hazer-Rau, D.; Meudt, S.; Daucher, A.; Spohrs, J.; Hoffmann, H.; Schwenker, F.; Traue, H.C. The uulmMAC Database—A Multimodal Affective Corpus for Affective Computing in Human-Computer Interaction. *Sensors* **2020**, *20*, 2308. [CrossRef]
42. Miranda-Correa, J.A.; Abadi, M.K.; Sebe, N.; Patras, I. AMIGOS: A Dataset for Affect, Personality and Mood Research on Individuals and Groups. *IEEE Trans. Affect. Comput.* **2021**, *12*, 479–493. [CrossRef]
43. Schmidt, P.; Reiss, A.; Duerichen, R.; Marberger, C.; Van Laerhoven, K. Introducing WESAD, a Multimodal Dataset for Wearable Stress and Affect Detection. In Proceedings of the 20th ACM International Conference on Multimodal Interaction, Virtual, 25–29 October 2018; ACM: Boulder, CO, USA, 2018; pp. 400–408. [CrossRef]
44. Haapalainen, E.; Kim, S.; Forlizzi, J.F.; Dey, A.K. Psycho-physiological measures for assessing cognitive load. In Proceedings of the 12th ACM International Conference on Ubiquitous Computing, Copenhagen, Denmark, 26–29 September 2010; ACM: Copenhagen, Denmark, 2010; pp. 301–310. [CrossRef]
45. Hussain, M.S.; Calvo, R.A.; Chen, F. Automatic Cognitive Load Detection from Face, Physiology, Task Performance and Fusion During Affective Interference. *Interact. Comput.* **2014**, *26*, 256–268. [CrossRef]
46. Taamneh, S.; Tsiamyrtzis, P.; Dcosta, M.; Buddharaju, P.; Khatri, A.; Manser, M.; Ferris, T.; Wunderlich, R.; Pavlidis, I. A multimodal dataset for various forms of distracted driving. *Sci. Data* **2017**, *4*, 170110. [CrossRef]
47. Markova, V.; Ganchev, T.; Kalinkov, K. CLAS: A Database for Cognitive Load, Affect and Stress Recognition. In Proceedings of the 2019 International Conference on Biomedical Innovations and Applications (BIA), Varna, Bulgaria, 8–9 November 2019; IEEE: Varna, Bulgaria, 2019; pp. 1–4. [CrossRef]
48. Beh, W.K.; Wu, Y.H.; Wu, A.-Y. MAUS: A Dataset for Mental Workload Assessmenton N-back Task Using Wearable Sensor. *arXiv* **2021**, arXiv:2111.02561.
49. He, D.; Donmez, B.; Liu, C.C.; Plataniotis, K.N. High Cognitive Load Assessment in Drivers Through Wireless Electroencephalography and the Validation of a Modified *N* -Back Task. *IEEE Trans. Hum.-Mach. Syst.* **2019**, *49*, 362–371. [CrossRef]
50. Liu, C.C. Towards Practical Driver Cognitive Load Detection Based on Visual Attention Information. Master's Thesis, University of Toronto, Toronto, ON, Canada, 2017. p. 126.
51. Healey, J.; Picard, R. Detecting Stress During Real-World Driving Tasks Using Physiological Sensors. *IEEE Trans. Intell. Transp. Syst.* **2005**, *6*, 156–166. [CrossRef]
52. Fridman, L.; Reimer, B.; Mehler, B.; Freeman, W.T. Cognitive Load Estimation in the Wild. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, Montreal, QC, Canada, 21–26 April 2018; ACM: Montreal, QC, Canada, 2018; pp. 1–9. [CrossRef]

53. Mehler, B.; Reimer, B.; Coughlin, J.F. Sensitivity of Physiological Measures for Detecting Systematic Variations in Cognitive Demand From a Working Memory Task: An On-Road Study Across Three Age Groups. *Hum. Factors J. Hum. Factors Ergon. Soc.* **2012**, *54*, 396–412. [CrossRef] [PubMed]
54. Reimer, B.; Mehler, B.; Dobres, J.; McAnulty, H.; Mehler, A.; Munger, D.; Rumpold, A. Effects of an 'Expert Mode' Voice Command System on Task Performance, Glance Behavior & Driver Physiology. In Proceedings of the 6th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, Seattle, WA, USA, 17–19 September 2014; ACM: Seattle, WA, USA, 2014; pp. 1–9. [CrossRef]
55. World Health Organization. *The SuRF report 2: Surveillance of Chronic Disease Risk Factors : Country-Level Data and Comparable Estimates*; World Health Organization: Geneva, Switzerland, 2005.
56. Gjoreski, M.; Kolenik, T.; Knez, T.; Luštrek, M.; Gams, M.; Gjoreski, H.; Pejović, V. Datasets for Cognitive Load Inference Using Wearable Sensors and Psychological Traits. *Appl. Sci.* **2020**, *10*, 3843. [CrossRef]
57. Rammstedt, B.; John, O.P. Kurzversion des Big Five Inventory (BFI-K). *Diagnostica* **2005**, *51*, 195–206. [CrossRef]
58. Kirchner, W.K. Age differences in short-term retention of rapidly changing information. *J. Exp. Psychol.* **1958**, *55*, 352–358. [CrossRef]
59. Peirce, J.W. PsychoPy-Psychophysics software in Python. *J. Neurosci. Methods* **2007**, *162*, 8–13. 2006.11.017. [CrossRef]
60. Mahesh, B.; Weber, D.; Garbas, J.; Foltyn, A.; Oppelt, M.P.; Becker, L.; Rohleder, N.; Lang, N. Setup for Multimodal Human Stress Dataset Collection. In Proceedings of the 12th International Conference on Methods and Techniques in Behavioral Research, and 6th Seminar on Behavioral Methods, Krakow, Poland, 18–20 May 2022; Volume 2. [CrossRef]
61. Saha, B.; Becker, L.; Garbas, J.U.; Oppelt, M.P.; Foltyn, A.; Hettenkofer, S.; Lang, N.; Struck, M.; Rohleder, N.; Mahesh, B. Investigation of Relation between Physiological Responses and Personality during Stress Recovery. In Proceedings of the 2021 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops), Kassel, Germany, 22–26 March 2021; IEEE: Kassel, Germany, 2021; pp. 57–62. [CrossRef]
62. Kirschbaum, C.; Hellhammer, D.H. Salivary Cortisol in Psychobiological Research: An Overview. *Neuropsychobiology* **1989**, *22*, 150–169. [CrossRef] [PubMed]
63. Smyth, N.; Hucklebridge, F.; Thorn, L.; Evans, P.; Clow, A. Salivary Cortisol as a Biomarker in Social Science Research: Salivary Cortisol in Social Science Research. *Soc. Personal. Psychol. Compass* **2013**, *7*, 605–625. [CrossRef]
64. Veltman, J.A.; Gaillard, A.W. Indices of mental workload in a complex task environment. *Neuropsychobiology* **1993**, *28*, 72–75. [CrossRef]
65. Abel, L.; Richer, R.; Küderle, A.; Gradl, S.; Eskofier, B.M.; Rohleder, N. Classification of Acute Stress-Induced Response Patterns. In Proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare, Trento, Italy, 20–23 May 2019; ACM: Trento Italy, 2019; pp. 366–370. [CrossRef]
66. Chen, S.; Epps, J.; Chen, F. A comparison of four methods for cognitive load measurement. In Proceedings of the 23rd Australian Computer-Human Interaction Conference on—OzCHI '11, Melbourne, VI, Australia, 30 November–2 December 2021; ACM Press: Canberra, Australia, 2011; pp. 76–79. [CrossRef]
67. Cohen, S.; Kamarck, T.; Mermelstein, R. A Global Measure of Perceived Stress. *J. Health Soc. Behav.* **1983**, *24*, 385. [CrossRef]
68. Watson, D.; Anna, L.; Tellegen, A. Development and Validation of Brief Measures of Positive and Negative Affect: The PANAS Scales. *J. Pers. Soc. Psychol.* **1988**, *54*, 1063–1070. [CrossRef]
69. Mejía-Mejía, E.; Budidha, K.; Abay, T.Y.; May, J.M.; Kyriacou, P.A. Heart Rate Variability (HRV) and Pulse Rate Variability (PRV) for the Assessment of Autonomic Responses. *Front. Physiol.* **2020**, *11*, 779. [CrossRef] [PubMed]
70. Hey, S.; Gharbi, A.; von Haaren, B.; Walter, K.; König, N.; Löffler, S. Continuous Noninvasive Pulse Transit Time Measurement for Psycho-physiological Stress Monitoring. In Proceedings of the 2009 International Conference on eHealth, Telemedicine, and Social Medicine, Cancun, Mexico, 1–7 February 2009; IEEE: Cancun, Mexico, 2009; pp. 113–116. [CrossRef]
71. Pham, T.; Lau, Z.J.; Chen, S.H.A.; Makowski, D. Heart Rate Variability in Psychology: A Review of HRV Indices and an Analysis Tutorial. *Sensors* **2021**, *21*, 3998. [CrossRef] [PubMed]
72. Kalidas, V.; Tamil, L. Real-time QRS detector using Stationary Wavelet Transform for Automated ECG Analysis. In Proceedings of the 2017 IEEE 17th International Conference on Bioinformatics and Bioengineering (BIBE), Washington, DC, USA, 23–25 October 2017; IEEE: Washington, DC, 2017; pp. 457–461. [CrossRef]
73. Li, P.; Li, Y.; Yao, Y.; Wu, C.; Nie, B.; Li, S.E. Sensitivity of Electrodermal Activity Features for Driver Arousal Measurement in Cognitive Load: The Application in Automated Driving Systems. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 14954–14967. [CrossRef]
74. Braithwaite, J.; Watson, D.; Jones, R.; Rowe, M. *Technical Report: A Guide for Analysing Electrodermal Activity (EDA) & Skin Conductance Responses (SCRs) for Psychological Experiments*; Technical report; Selective Attention & Awareness Laboratory (SAAL) Behavioral Brain Sciences Centre, University of Birmingham: Birmingham, UK, 2015. Available online: https://www.birmingham.ac.uk/documents/college-les/psych/saal/guide-electrodermal-activity.pdf (accessed on 25 November 2022).
75. Leyman, E.L.; Mirka, G.A.; Kaber, D.B.; Sommerich, C.M. Cervicobrachial muscle response to cognitive load in a dual-task scenario. *Ergonomics* **2004**, *47*, 625–645. [CrossRef] [PubMed]
76. Biondi, F.N.; Cacanindin, A.; Douglas, C.; Cort, J. Overloaded and at Work: Investigating the Effect of Cognitive Workload on Assembly Task Performance. *Hum. Factors J. Hum. Factors Ergon. Soc.* **2021**, *63*, 813–820. [CrossRef]

77. Grassmann, M.; Vlemincx, E.; von Leupoldt, A.; Mittelstädt, J.M.; Van den Bergh, O. Respiratory Changes in Response to Cognitive Load: A Systematic Review. *Neural Plast.* **2016**, *2016*, 1–16. [CrossRef]
78. Costa, M.D.; Davis, R.B.; Goldberger, A.L. Heart Rate Fragmentation: A New Approach to the Analysis of Cardiac Interbeat Interval Dynamics. *Front. Physiol.* **2017**, *8*, 255. [CrossRef]
79. Reimer, B.; Mehler, B. The impact of cognitive workload on physiological arousal in young adult drivers: A field study and simulation validation. *Ergonomics* **2011**, *54*, 932–942. [CrossRef]
80. Giakoumis, D.; Drosou, A.; Cipresso, P.; Tzovaras, D.; Hassapis, G.; Gaggioli, A.; Riva, G. Using Activity-Related Behavioural Features towards More Effective Automatic Stress Detection. *PLoS ONE* **2012**, *7*, e43571. [CrossRef]
81. 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. [CrossRef]
82. Smets, E.; De Raedt, W.; Van Hoof, C. Into the Wild: The Challenges of Physiological Stress Detection in Laboratory and Ambulatory Settings. *IEEE J. Biomed. Health Inform.* **2019**, *23*, 463–473. [CrossRef] [PubMed]
83. Schleifer, L.M.; Spalding, T.W.; Kerick, S.E.; Cram, J.R.; Ley, R.; Hatfield, B.D. Mental stress and trapezius muscle activation under psychomotor challenge: A focus on EMG gaps during computer work. *Psychophysiology* **2008**, *45*, 356–365. [CrossRef] [PubMed]
84. Der Bilt, V.; Der Glas, V. Detection of onset and termination of muscle activity in surface electromyograms. *J. Oral Rehabil.* **1998**, *25*, 365–369. [CrossRef]
85. Skaramagkas, V.; Giannakakis, G.; Ktistakis, E.; Manousos, D.; Karatzanis, I.; Tachos, N.; Tripoliti, E.E.; Marias, K.; Fotiadis, D.I.; Tsiknakis, M. Review of eye tracking metrics involved in emotional and cognitive processes. *IEEE Rev. Biomed. Eng.* **2021**, 1. [CrossRef]
86. Dalmaijer, E.S.; Mathôt, S.; Van der Stigchel, S. PyGaze: An Open-Source, Cross-Platform Toolbox for Minimal-Effort Programming of Eyetracking Experiments. *Behav. Res.* **2014**, *46*, 913–921. [CrossRef]
87. Salvucci, D.D.; Goldberg, J.H. Identifying Fixations and Saccades in Eye-Tracking Protocols. In Proceedings of the Symposium on Eye Tracking Research & Applications—ETRA '00, Palm Beach Gardens, FL, USA, 6–8 November 2000; ACM Press: Palm Beach Gardens, FL, USA, 2000; pp. 71–78. [CrossRef]
88. Ware, C. *Information Visualization: Perception for Design*, 3rd ed.; Interactive Technologies, Morgan Kaufmann; Elsevier: Amsterdam, The Netherlands, 2013.
89. Jacob, R.J.; Karn, K.S. *Eye Tracking in Human–Computer Interaction and Usability Research*; Elsevier: Amsterdam, The Netherlands, 2003; pp. 573–605. [CrossRef]
90. Negi, S.; Mitra, R. Fixation Duration and the Learning Process: An Eye Tracking Study with Subtitled Videos. *J. Eye Mov. Res.* **2020**. *13*. [CrossRef]
91. McCloy, D.R.; Lau, B.K.; Larson, E.; Pratt, K.A.I.; Lee, A.K.C. Pupillometry shows the effort of auditory attention switching. *J. Acoust. Soc. Am.* **2017**, *141*, 2440–2451. [CrossRef]
92. Winn, M.B.; Wendt, D.; Koelewijn, T.; Kuchinsky, S.E. Best Practices and Advice for Using Pupillometry to Measure Listening Effort: An Introduction for Those Who Want to Get Started. *Trends Hear.* **2018**, *22*, 233121651880086. [CrossRef]
93. Duchowski, A.T.; Krejtz, K.; Krejtz, I.; Biele, C.; Niedzielska, A.; Kiefer, P.; Raubal, M.; Giannopoulos, I. The Index of Pupillary Activity: Measuring Cognitive Load *vis-à-vis* Task Difficulty with Pupil Oscillation. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, Montreal, QC, Canada, 21–26 April 2018; ACM: Montreal, QC, Canada, 2018; pp. 1–13. [CrossRef]
94. Stoeve, M.; Wirth, M.; Farlock, R.; Antunovic, A.; Müller, V.; Eskofier, B.M. Eye Tracking-Based Stress Classification of Athletes in Virtual Reality. *Proc. Acm Comput. Graph. Interact. Tech.* **2022**, *5*, 1–17. [CrossRef]
95. Cheong, J.H.; Xie, T.; Byrne, S.; Chang, L.J. Py-Feat: Python Facial Expression Analysis Toolbox. 2021, p. 25, Framework. Available online: https://py-feat.org/pages/intro.html (accessed on 25 November 2022).
96. Deng, J.; Guo, J.; Ververas, E.; Kotsia, I.; Zafeiriou, S. RetinaFace: Single-Shot Multi-Level Face Localisation in the Wild. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; IEEE: Seattle, WA, USA, 2020; pp. 5202–5211. [CrossRef]
97. Albiero, V.; Chen, X.; Yin, X.; Pang, G.; Hassner, T. img2pose: Face Alignment and Detection via 6DoF, Face Pose Estimation. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; IEEE: Nashville, TN, USA, 2021; pp. 7613–7623. [CrossRef]
98. Chen, S.; Liu, Y.; Gao, X.; Han, Z. *MobileFaceNets: Efficient CNNs for Accurate Real-Time Face Verification on Mobile Devices*; Series Title: Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2018; Volume 10996, pp. 428–438. [CrossRef]
99. Ekman, P.; Friesen, W.V. *Facial Action Coding System: A Technique for the Measurement of Facial Movement*; Consulting Psychologists Press: Palo Alto, CA, USA, 1978.
100. Paul, E.; Rosenberg, L.E. *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System*; Oxford University Press: Oxford, UK, 2005.
101. Martinez, B.; Valstar, M.F.; Jiang, B.; Pantic, M. Automatic Analysis of Facial Actions: A Survey. *IEEE Trans. Affect. Comput.* **2019**, *10*, 325–347. [CrossRef]
102. Li, N.; Busso, C. Predicting Perceived Visual and Cognitive Distractions of Drivers With Multimodal Features. *IEEE Trans. Intell. Transp. Syst.* **2015**, *16*, 51–65. [CrossRef]

103. Giannakakis, G.; Koujan, M.R.; Roussos, A.; Marias, K. Automatic stress detection evaluating models of facial action units. In Proceedings of the 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), Buenos Aires, Argentina, 16–20 November 2020; IEEE: Buenos Aires, Argentina, 2020; pp. 728–733. [CrossRef]
104. Zhao, Z.; Zhang, Y. SQI Quality Evaluation Mechanism of Single-Lead ECG Signal Based on Simple Heuristic Fusion and Fuzzy Comprehensive Evaluation. *Front. Physiol.* **2018**, *9*, 727. [CrossRef] [PubMed]
105. Hall, P.; Welsh, A.H. Limit theorems for the median deviation. *Ann. Inst. Stat. Math.* **1985**, *37*, 27–36. [CrossRef]
106. Gjoreski, M.; Gams, M.Z.; Lustrek, M.; Genc, P.; Garbas, J.U.; Hassan, T. Machine Learning and End-to-End Deep Learning for Monitoring Driver Distractions From Physiological and Visual Signals. *IEEE Access* **2020**, *8*, 70590–70603. [CrossRef]
107. Stone, M. Cross-Validatory Choice and Assessment of Statistical Predictions. *J. R. Stat. Soc. Ser. B (Methodol.)* **1974**, *36*, 111–133. [CrossRef]
108. Cawley, G.C.; Talbot, N.L.C. On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation 2010; 29p. Available online: https://www.jmlr.org/papers/v11/cawley10a.html (accessed on 25 November 2022).
109. Bergstra, J.S.; Bardenet, R.; Bengio, Y.; Kégl, B. Algorithms for Hyper-Parameter Optimization. *Adv. Neural Inf. Process. Syst.* **2011**, *24*. Available online: https://proceedings.neurips.cc/paper/2011/file/86e8f7ab32cfd12577bc2619bc635690-Paper.pdf (accessed on 25 November 2022).
110. Weber, P.; Rupprecht, F.; Wiesen, S.; Hamann, B.; Ebert, A. *Assessing Cognitive Load via Pupillometry*; Springer: Cham, Switzerland, 2021; pp. 1087–1096. [CrossRef]
111. Fehringer, B. Optimizing the Usage of Pupillary Based Indicators for Cognitive Workload. *J. Eye Mov. Res.* **2021**, *14*. [CrossRef]
112. Wang, Q.; Yang, S.; Liu, M.; Cao, Z.; Ma, Q. An Eye-Tracking Study of Website Complexity from Cognitive Load Perspective. *Decis. Support Syst.* **2014**, *62*, 1–10. [CrossRef]
113. Broadbent, D.P.; D'Innocenzo, G.; Ellmers, T.J.; Parsler, J.; Szameitat, A.J.; Bishop, D.T. Cognitive Load, Working Memory Capacity and Driving Performance: A Preliminary fNIRS and Eye Tracking Study. *Transp. Res. Part F Traffic Psychol. Behav.* **2023**, *92*, 121–132. [CrossRef]
114. Korbach, A.; Brünken, R.; Park, B. Differentiating Different Types of Cognitive Load: A Comparison of Different Measures. *Educ. Psychol. Rev.* **2018**, *30*, 503–529. [CrossRef]
115. Seitz, J.; Maedche, A. Biosignal-Based Recognition of Cognitive Load: A Systematic Review of Public Datasets and Classifiers. *Inf. Syst. Neurosci.* **2022**, *58*, 35–52. [CrossRef]
116. Albuquerque, I.; Tiwari, A.; Parent, M.; Cassani, R.; Gagnon, J.F.; Lafond, D.; Tremblay, S.; Falk, T.H. WAUC: A Multi-Modal Database for Mental Workload Assessment Under Physical Activity. *Front. Neurosci.* **2020**, *14*, 549524. [CrossRef] [PubMed]
117. Wilson, J.C.; Nair, S.; Scielzo, S.; Larson, E.C. Objective Measures of Cognitive Load Using Deep Multi-Modal Learning: A Use-Case in Aviation. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **2021**, *5*, 1–35. [CrossRef]
118. He, D.; Wang, Z.; Khalil, E.B.; Donmez, B.; Qiao, G.; Kumar, S. Classification of Driver Cognitive Load: Exploring the Benefits of Fusing Eye-Tracking and Physiological Measures. *Transp. Res. Rec. J. Transp. Res. Board* **2022**, *2676*, 670–681. [CrossRef]
119. Ahmad, M.I.; Keller, I.; Robb, D.A.; Lohan, K.S. A Framework to Estimate Cognitive Load Using Physiological Data. *Pers. Ubiquit Comput.* **2020**. [CrossRef]
120. The HDF Group. *Hierarchical Data Format v5*; The HDF Group: Champaign, IL, USA, 2022.
121. Staal, M.A. Stress, Cognition, and HumanPerformance: A Literature Review and Conceptual Framework. 2004. p. 177. Technical Report NASA STI Program Office, Ames Research Center, Moffett Field, California. Available online: https://human-factors.arc.nasa.gov/flightcognition/Publications/IH_054_Staal.pdf (accessed on 25 November 2022).
122. Henderson, R.K.; Snyder, H.R.; Gupta, T.; Banich, M.T. When Does Stress Help or Harm? The Effects of Stress Controllability and Subjective Stress Response on Stroop Performance. *Front. Psychol.* **2012**, *3*. [CrossRef]
123. Sandi, C. Stress and cognition. *Wires Cogn. Sci.* **2013**, *4*, 245–261. [CrossRef]
124. Li, X.; Ouyang, Z.; Luo, Y.J. The effect of cognitive load on interaction pattern of emotion and working memory: An ERP study. In Proceedings of the 9th IEEE International Conference on Cognitive Informatics (ICCI'10), Beijing, China, 7–9 June 2010; IEEE: Beijing, China, 2010; pp. 61–67. [CrossRef]