IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 27 July 2023, accepted 11 August 2023, date of publication 15 August 2023, date of current version 22 August 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3305383

RESEARCH ARTICLE

Physiological Signals as Predictors of Cognitive Load Induced by the Type of Automotive Head-Up Display

GREGOR STRLE[1,2], ANDREJ KOŠIR[1], (Senior Member, IEEE), JAKA SODNIK[3], (Senior Member, IEEE), AND KRISTINA STOJMENOVA PEČEČNIK[3], (Member, IEEE)

[1] User-Adapted Communication and Ambient Intelligence Laboratory, Faculty of Electrical Engineering, University of Ljubljana, 1000 Ljubljana, Slovenia
[2] Research Centre of the Slovenian Academy of Sciences and Arts, 1000 Ljubljana, Slovenia
[3] Laboratory for Information Technologies, Faculty of Electrical Engineering, University of Ljubljana, 1000 Ljubljana, Slovenia

Corresponding author: Gregor Strle (gregor.strle@fe.uni-lj.si)

This work was supported in part by the Slovenian Research Agency within the Project Modeling Driver's Situational Awareness under Grant Z2-3204, in part by the Program Information Communication Technologies for Quality of Life (ICT4QL) under Grant P2-0246, and in part by the European Union's Horizon 2020 Research and Innovation Program for the Project HADRIAN under Grant 875597.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by Institutional Review Board on Ethics in Research under Application No. P2-0246, and performed in line with the Code of ethics for researchers and Guidelines for ethical conduct in research involving people issued by the University of Ljubljana.

ABSTRACT The visual information complexity of automotive head-up displays (HUDs) may affect cognitive load and reduce driver performance in critical situations. This study investigated whether physiological indicators of cognitive load can predict the type of HUD while driving. Physiological signals of heart rate variability (HRV), electrodermal activity (EDA), skin temperature, and pupil dilation were recorded from 28 participants using a motion-based driving simulator. Two types of HUD with different information complexities were compared: baseline HUD and augmented reality HUD. Heart rate and EDA were processed to create standardized biomedical features. Time-series analysis and basic statistics generated two sets of features for pupil dilation and skin temperature. The effect of signal combinations on classification performance was tested using signal fusion. Three gradient boosting classifiers (LGBM, HGBC, and XGB) were trained on physiological signals to predict HUD type. The fusion of HRV, EDA, and time-series features for skin temperature and pupil dilation yielded moderate performance, with average AUC ROC scores of XGB = 0.67, LGBM = 0.69, and HGBC = 0.70. Combining HRV, EDA, and basic statistical features for skin temperature and pupil dilation, the classifiers achieved an improved average AUC ROC score of 0.76. The best scores were 0.96 (LGBM and XGB) and 0.98 (HGBC). These results demonstrate the potential of physiological signals for modeling HUD-induced cognitive load and dynamically regulating its effects in real-time.

INDEX TERMS Advanced driver assistance system (ADAS), augmented reality, automotive head-up display, autonomous vehicles, classification, conditionally automated driving, human–machine interaction, machine learning, physiological signals.

I. INTRODUCTION

Automotive head-up display (HUD) allows the driver to view a variety of information as visual cues on a windshield in front of the driver. The HUD is an essential part of the advanced driver assistance system (ADAS), which provides enhanced visual and environmental monitoring of a vehicle, such as lane and distance keeping, dynamic cruise control, and vehicular communication, to reduce road accidents [1]. The main advantage of HUDs is that they provide a "eyes-on- road"

The associate editor coordinating the review of this manuscript and approving it for publication was Alberto Cano.

that keeps the driver's situational awareness on the road while conveying driving-critical information [2]. By reducing the time drivers take their eyes off the road, HUDs can increase safety, reduce driver distraction, and improve decision-making on the road [3]. At the same time, visual information in HUDs is becoming increasingly complex and includes advanced longitudinal and lateral visual guidance aids, in addition to standard information such as speed, vehicle diagnostics, and navigation.

Recent developments advocate the use of augmented reality in HUDs (AR-HUDs) to display complex driving and safety-related information [3], [4], [5]. The increased visual complexity of HUDs can negatively affect driver responses, which can be detrimental in safety-critical driving situations [5], [6], [7]. One of the key challenges is how and where to display information on HUD to provide an appropriate level of information complexity and to accommodate specific contexts, especially in safety-critical driving situations [3], [5], [6], [8]. Another key challenge is to provide this information to drivers without inducing excessive cognitive load [4], [9].

Much research has been conducted in driving simulators to investigate the safety issues of using HUDs in conditional automation, particularly in the event of a takeover (e.g. [4], [9], [10], [11]. However, there is limited work on the physiological effects of HUD on drivers. Existing studies have focused on physiological signals as indicators of cognitive load in takeover situations [12], classification of cognitive load in conditional automation [13], [14], or differences in cognitive load when switching between automated and manual driving [15].

Little is known about the effects of the cognitive load exerted on drivers by HUD. This study aimed to investigate the potential of physiological signals in assessing the cognitive load caused by HUDs while driving. The main research question was whether differences in drivers' physiological responses to cognitive load could predict HUD type. If such differences are measurable and predictable, physiological signals can be used to regulate the cognitive load induced by HUDs continuously and in real-time.

To this end, two types of HUD with different complexities of visual information were tested: baseline HUD and augmented reality HUD (AR-HUD). Physiological signals of heart rate, electrodermal activity (EDA), skin temperature, and pupil dilation were recorded in a motion-based driving simulator under conditional automation. Machine learning classifiers were trained using physiological signals, and signal fusion was performed to optimize the prediction of HUD type.

The present study makes a valuable contribution as it attempts to fill the gap in the current state-of-the-art in understanding physiological responses to cognitive load from different types of HUD. Its main contribution is the use of machine learning to uncover the potential of physiological signals as reliable predictors of the cognitive load. In addition, the study highlights the importance of signal fusion in improving the classification performance. The application of physiological signals as objective measures of workload has tremendous potential to shape the future of automated driving because there is an opportunity to dynamically regulate cognitive load to enhance driver performance and safety.

Related work on automotive HUDs and their effects, as well as human factors, is presented in Section II, along with research on physiological measures related to driver performance and cognitive load. The materials and methods used in this study are presented in Section III. The driving simulator and the two HUDs are presented along with the instruments and procedures used in the driving experiment and subsequent analyses. The results of the statistical analysis and machine learning are presented in Section IV. The article concludes with a discussion of the results of the research and potential for future work in Section V.

II. RELATED WORK

A. AUTOMOTIVE HEAD-UP DISPLAYS: DRIVER BEHAVIOR AND SAFETY

Several studies have reported the benefits of HUDs in improving driver behavior and safety compared to traditional head-down displays and, with recent technological advances, the further benefits of augmented reality HUDs (AR-HUDs). For example, [6] examined the interface designs of HUDs for conveying safety-related information from 13 major automotive manufacturers. Their results indicate that drivers' information-processing abilities should be considered when designing HUDs. They also reported that all commercial HUDs studied had a traditional design. The limited availability of commercial AR HUDs forces researchers to conduct experiments using driving simulators or virtual reality (for a review, see [3], [16]). Although both environments have several advantages over field studies, for example, they allow for controlled, reproducible, and safe experiments and data collection is easier to manage, validation in the real world is required [17].

The study by [18] showed that both visual attention and driving performance were improved with AR-HUD compared to traditional HDDs. The authors found differences in lateral and longitudinal vehicle control between the two displays, with drivers using HUD showing better driving performance and longer visual attention. Different gazes and driving performances were associated with each display, with the latter being influenced by the driving environment.

Similar results have been reported in [4], [11], and [19]. Blissing et al. compared driver behavior in mixed and virtual reality and found that driving behavior was different in each mode, with transverse and longitudinal driving behavior changing when transitioning between modes [19]. Park and Im investigated the effects of visual enhancements in AR-HUDs on driver performance and cognitive load, and found that visual enhancements both improve driver decision-making and reduce subjective cognitive load [4]. Jing et al. compared the effects of three different HUDs

(AR-HUD with arrow cues, AR-HUD with virtual shadows, and without AR-HUD) on driver behavior and acquisition efficiency [11]. Their results showed that both AR-HUDs performed better than AR-HUD: visual distraction was reduced and takeover efficiency was improved [11].

However, it is important to note that the effect of AR-HUDs visual enhancements depends on the type, complexity, and visual partitioning of the graphical elements in the interface. Poorly designed AR-HUD can negatively affect both driving performance and safety. Because of the increased information complexity of AR-HUDs, attentional allocation and switching challenges are common when using AR-HUDs (see also [3]). Several studies have shown that cognitive capture and inattentional blindness in AR-HUDs are symptoms of overused attentional resources, as is often the case in situations with complex and frequently changing information, which generates a higher cognitive load [8], [20], [21], [22], [23], [24].

To this end, [24] examined how different graphical layouts of AR-HUD affect driving performance. Three scenarios were compared: driving without AR-HUD, AR-HUD with a dispersed graphical layout, and AR-HUD with a dense graphical layout. While both AR-HUDs showed improved driving performance compared to driving without AR-HUD, the graphical layout in AR-HUDs also had an impact. The AR-HUD with the dispersed layout, which conformed to human-computer interaction principles and visual design rules and provided better distribution of visual information, showed better driving performance than the AR-HUD with a dense layout.

Kim and Gabbard evaluated the visual and cognitive distraction potential of AR-HUDs by comparing drivers' gaze behavior, situational awareness, confidence, and cognitive load with and without the use of AR-HUDs [8]. They report several important findings. First, the results show that AR-HUDs have an impact on drivers' visual attention allocation and that the perceptual forms of graphical elements on AR-HUD determine whether the interface is informative or distracting. The AR-HUD cuing of pedestrians and other critical objects with a bounding box and a virtual shadow were compared. Negative side effects, such as the cognitive capture effect and inattentional blindness, have been observed for AR-HUD using bounding boxes to identify pedestrians, resulting in reduced driver attention to pedestrians and other critical road elements [8]. Pedestrian perception was reduced by a larger number of bounding boxes, which drivers perceived as clutter, and obscured perceptual information that could predict pedestrian movement intentions. Consequently, the bounding box AR-HUD produces a higher cognitive load, leading to inattentional blindness. On the other hand, AR-HUD with virtual shadows had positive effects on visual attention and situational awareness, as more attentional resources were available for other critical objects and situations. Overall, situational awareness proved to be an important indicator of both visual and cognitive distractions to properly quantify the effects of using AR-HUD [8], [25].

Currano et al. examined the effects of using AR-HUD with different levels of visual complexity on drivers' situational awareness and perception [7]. The experiment was conducted by showing participants videos of driving situations. Two driving environments were tested with three variations of AR-HUD complexity: no HUD display, minimal information HUD (with cues to pedestrians and other critical road objects and signs), and complex information HUD (the most relevant information in the environment, including navigation and vehicle status). Although HUD complexity has a negative effect on situational awareness, the factors constituting the complexity of a scene may have a greater effect on situational awareness than the complexity of HUD design [7].

B. HUMAN FACTORS IN AUTOMOTIVE HUDs

A review of research on human factors in automated driving shows that most studies on HUDs focus on aspects related to safety and takeover performance rather than on user experience [3], [16], [26], [27], [28]. For a holistic understanding of AR-HUDs and their real-world applications, it is necessary to understand "how to design the automotive HUD system to best serve the driver" [2, p. 1936].

A user survey conducted by Beck et al. examined several issues related to user experience and user-perceived design improvement points for existing commercial HUDs [2]. Participants with extensive HUD experience participated in this survey. Eleven high-level HUD information items were assessed:1) speed control, 2) highway driving, 3) engine/transmission control, 4) wayfinding, 5) sign/warning recognition, 6) audio player control, 7) accuracy of HUD information, 8) individual and context-specific HUD information needs, 9) visibility of HUD images, 10) visual aesthetics of HUD interfaces, and 11) HUD location and layout issues. The results show that safety-related information (speed, speed limit, cruise control, and traffic signs) and navigation information displayed in the HUD are helpful for driving and complying with speed limits. However, the information needs of HUD users vary considerably depending on the driving environment. For example, some participants preferred a more realistic interface design, whereas others preferred a more appealing visual aesthetic. There are also different preferences regarding the amount of information and how/where it should be displayed [2].

Most of the studies discussed above were self-reports. Self-reports are susceptible to respondent bias (such as social desirability and agreeing responses), which may call into question the validity of these studies [29], [30], [31], [32], [33]. For example, [34] used physiological signals (cardiac, respiratory, and electrodermal signals) and facial features from 36 drivers as objective UX measures of driver emotions (measured as valence and arousal). The computerized estimates of drivers' emotional states were then compared with their self-reports on UX questions. The authors found "a discrepancy between the self-ratings and the algorithmic scores – drivers who answered the UX questions more positively experienced higher levels of stress, as evidenced by higher arousal scores and lower algorithmic scores for valence'' [34, p. 80].

One solution to respondent bias is to supplement self-reports with objective measures of driver physiology data [35].

C. PHYSIOLOGICAL MEASURES OF COGNITIVE LOAD

A major advantage of psychophysiological measures ``is the continuous availability of bodily data that allow strain to be measured at a high rate and with a high degree of sensitivity, even in situations in which overt behavior is relatively rare.'' [36, p. 270].

Therefore, physiological signals have been studied in detail as potential indicators of driver performance, particularly in relation to the cognitive load. Signals such as electroencephalography (EEG), cardiac (heart rate and heart rate variability), electrodermal (EDA indices of tonic and phasic skin conductance, skin temperature), respiratory activity, and eye-tracking measurements (e.g., gaze, pupil dilation, eye blinks) are all potential indicators of cognitive load [37], [38], [39], [40], [41].

For example, in a study by [39], several classification models were developed to predict cognitive load (low vs. high) based on several features extracted from the EEG, heart rate, HRV, EDA, respiration, pupil size, and eyeblinks of 14 participants. The EEG-based model performed the best (86% accuracy), followed by eye tracking measures (pupil size and blink rate) and EDA [39].

Several studies have shown that heart rate and skin conductance are robust indicators of cognitive load in several studies [37], [40], [42]. In addition, indices of heart rate variability (HRV), particularly the root mean square of successive differences in normal heartbeats (RMSSD), have proven to be reliable predictors of cognitive load [40], [42], [43], although not as robust as heart rate and skin conductance [42]. Another important indicator of cognitive load is pupil dilation, and several studies have found that pupil size increases with increasing cognitive load [39], [44], [45], [46]. A practical advantage of eye-tracking devices is that they are unobtrusive compared with devices that require a physical connection to the subject (e.g., EEG).

These physiological measurements are particularly important for understanding driver performance, particularly as indicators of the driver cognitive load. This topic has received increasing attention in studies of conditional automation [12], [14], [24], [42], [47], [48], [49], [50] because understanding driver cognitive load is critical to developing systems with efficient management of cognitive load that can reduce driver errors and thereby increase safety in critical situations. An overview of the selected studies is presented in Table 1, and the most relevant studies are discussed below.

To this end, [12] investigated how different driving situations affect driver cognitive load, using ECG and EDA as objective indicators of cognitive load along with self-reported cognitive load ratings. The study was conducted using a driving simulator, with 32 drivers divided into two equal groups. One group performed secondary tasks during automation (SAE Level 3), whereas the other group performed supervised driving (SAE Level 2). Their results showed that driver cognitive load was significantly higher during secondary tasks, with both physiological signals responding to variations in the cognitive load [12]. However, apart from a higher respiratory rate in the manual driving mode, likely owing to the sensitivity to the physical activity of vehicle control, no significant differences were found in the physiological measures and cognitive load between the manual and supervised driving modes. Other studies have also reported no significant differences in driver physical measures when switching driving modes [15], [51], [52].

A driving simulator study by [13] analyzed the cognitive load of 90 participants using physiological signals (ECG, EDA, and respiration) who drove in the conditionally automated mode for 25 min. The participants were divided into two groups: those who performed a secondary task and those who observed the environment of the vehicle. The study showed that the drivers' cognitive load was higher in the secondary task. The results also show that cognitive load can be accurately detected using machine learning classifiers (Random Forest Classifier, Support Vector Classifier, Multi-Layer Perceptron Classifier) based on physiological signals and their combinations. The lowest performance, with an accuracy between 6973% and 73%, was obtained for EDA alone, and the best accuracy of 92-94% was obtained by combining respiration and ECG, depending on the classifier [13].

Another application of machine learning to physiological data (ECG, EDA, and respiration) was conducted in a conditional automation study [14]. Participants ($n = 80$) were asked to perform a cognitive task unrelated to driving 15 times for 90 s while driving. The performance of the three machine learning models (Random Forest, Neural Network, and k-Nearest Neighbors) was evaluated by classifying the driver cognitive load as a function of task difficulty (no task vs. low vs. high) and task modality (visual cognitive task vs. auditory cognitive task). The authors reported the performance of the task difficulty classification models, with a weighted F1 score ranging from 0.51 to 0.71, depending on the model and feature combination. The best F1 score for task difficulty prediction was obtained using the EDA and respiration signals as inputs to a Random Forest classifier. The results also showed that the models had difficulty predicting the task modality (visual vs. auditory), with the best model achieving a weighted F1 score of 0.61 when ECG and RESP were used as signals. On average, the classifiers achieved weighted F1 scores with approximately 50% accuracy in predicting task modality. The authors suggested complementing physiological data with other data sources to support and improve the task modality prediction [14].

TABLE 1. Overview of selected studies investigating drivers' physiological responses to cognitive load.

References	Description	Measurements	Results
Mehler et al. [42]	A comparison of heart rate and heart rate variability indices in distinguishing single-task driving and driving under secondary cognitive workload (26 participants)	Heart rate, HRV, skin conductance.	A repeated-measures general linear model was used. Heart rate and skin conductance were robust indicators of driving with and without a task.
Yan et al. [47]	The driver mental workload prediction model was based on physiological indices, focusing on new drivers and high, medium and low levels of driving task complexities (26 participants).	Eye-tracking signals (pupil dilation, blink rate and duration, fixations), heart rate, HRV SDNN, nr. of errors, NASA TLX.	A regression model based on physiology was used to predict the workload represented by NASA TLX, with R2=0.745. No significant difference was found in pupil dilation, whereas mean HR and SDNN increased with task complexity.
McDonnell et al. [48]	Neural indices of driver workload and engagement during partial vehicle automation (71 participants).	The EEG was recorded while driving on the roadway, with partial vehicle automation engaged and disengaged.	The EEG showed no change in mental workload between manual driving and conditional automated driving.
Ma et al. [24]	A preliminary study on driving performance through a VR-simulated eye movement analysis of HDD and AR-HUDs and how different AR-HUD layouts affect driving performance (12 participants).	Obtained directly from the head-mounted device (HMD) were the pupil diameter, eye-opening size, and relative location of the eyes. Obtained outside the HMD were the windshield gaze point and gaze point distance, as well as pedaling information and steering wheel angle.	ANOVA statistics showed that the average blink frequency was significantly lower with the AR-HUD than without it, indicating that drivers were more relaxed with the assistance of the AR-HUD. The total gaze time was shorter with the AR-HUD than without it.
Radhakrishnan et al. [12]	Physiological indices of driver workload during car-following scenarios and takeovers in highly automated driving ($n = 32$).	HRV and EDA based physiological metrics were used as objective indicators of workload, along with self-reported workload ratings on a scale of 1–10 (lowest to highest).	Repeated measures ANOVA showed that the ECG and EDA signals were sensitive to variations in the workload.
Gomaa et al. [49]	Physiological indices of the mental workload and perceptual load from a dual task scenario for in-vehicle interaction, using machine learning (45 participants).	Mean, minimum, maximum, standard deviation of heart rate, HRV-RMSSD, pupillary activity index, and average deviation during lane switching.	Mental workload influenced some psychophysiological dimensions, whereas perceptual load has little effect. The best-performing classifiers were k-Nearest Neighbors (KNN), Linear Discriminant Analysis (LDA), and AdaBoost, using 5-fold nested cross-validation. The results for the binary classification (low load vs. medium load) showed eye features performed at a random chance level, whereas the heart rate data had an average classification accuracy of 72.2\%, up to 89\% for a single fold.
Meteier et al. [14]	Physiological indicators for assessing workload (induced by a non-driving cognitive task (N-back)) in conditionally automated driving in a simulator (80 participants).	The EDA, ECG, RESP signals, and subjective workload (on a 0–20 scale).	Three-class classification of workload using sensor fusion. The fusion of the signals yielded the best results. The best F1-score=0.713, using skin conductance and respiration signals as inputs of a random forest classifier.

D. PHYSIOLOGICAL MEASUREMENTS OF HUD AND ITS EFFECTS ON A DRIVER

Available literature on the physiological effects of HUDs on drivers is limited. A driving simulator study by [53] examined the effects of different display configurations (HDD vs. HUD) on a driver ($n = 19$). Driving performance, gaze behavior, physiological measures (heart rate, EDA, and temperature), and task completion times were measured repeatedly for both display configurations in two driving situations (driving only and driving with a task) on rural roads and in the city. The physiological measures showed no significant differences between the two display configurations or driving tasks [53].

[54] examined the effect of different cueing strategies on the cognitive load to provide guidelines for designing an optimal AR-HUD interface [54]. Three cueing strategies were

used: none (vehicle speed and direction), partial (in addition, relevant traffic signs, traffic lights, and road users relevant to the driver's path were highlighted), and all (in addition to partial cueing, all traffic signs, traffic lights, and road users were highlighted using different colors for different objects) [54]. The participants ($n = 36$) were tested for situational awareness and cognitive load using a driving simulator (six trips). Physiological signals (gaze and electrodermal activity (EDA)) were recorded along with subjective measures of cognitive load (NASA TLX) and trust in automation (Trust Score). The authors reported only the results of subjective measures, which showed that cueing strategies affect driver cognitive load, with partial cueing being the most helpful [54].

A VR-based driving simulator study was conducted by [24] to evaluate the utility of AR-HUD and how different AR-HUD layouts affect the driving performance and safety. The drivers ($n = 12$) were tested under three scenarios: without an AR-HUD, with a distributed layout, and with a dense layout. Each driver completed the three scenarios in an urban environment. The driver's eye movement data (pupil size, blink rate, eye aperture size, relative eye position, and gaze point), speed, and brake use were recorded to determine driving performance in the three driving scenarios. A significant difference in driving performance with and without AR-HUD was found for the average accelerator pedal amplitude, blink frequency, horizontal gaze angle, and vehicle speed. The results also show that the difference in AR-HUD interface displays affected the driver's allocation of cognitive resources, which was altered in AR-HUDs compared to normal driving [24].

To the best of our knowledge, none of the current studies has used machine learning to develop physiological models of the cognitive load induced by HUDs with varying information complexity.

III. MATERIALS AND METHODS

A. PARTICIPANTS

28 (14 males and 14 females) participated in the study. The ages of the drivers ranged from 21 to 57 years (M = 30.17 years, SD = 10.60 years) and they held a valid driver's license for an average of 11.77 years (SD = 10.12 years). 20% of the drivers had no experience with vehicles with automated features (any advanced driver assistance system (ADAS)), whereas 6.66% had driven a vehicle with multiple ADAS systems once, 13.3% a few times, and 60% several times.

The experiment was designed and conducted in accordance with the Code of Ethics for Researchers and the Guidelines for Ethical Conduct in Research with Human Subjects at the University of Ljubljana. Informed consent was obtained from all the participants. Participation in the study was voluntary. The participants were informed that they could stop the experiment at any time without providing a reason. Each participant received a gift voucher of 10 euros to participate in the study.

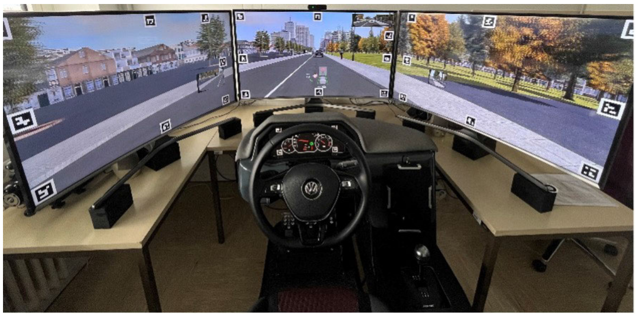


FIGURE 1. Nervtech motion-based driving simulator with a physical dashboard and 145Â° field of view of the driving environment.

B. DRIVING SIMULATOR

The study was conducted in a simulated driving environment consisting of a motion-based driving simulator [55] with real car parts (seat, steering wheel, and pedals) and a physical dashboard (see Figure 1). The dashboard was not designed as part of this study but mimicked the dashboard of a typical manually operated personal vehicle with an HDD. It displays the vehicle speed, engine rpm, fuel level, and status of the indicators and lights. The visuals were displayed on three 49-inch curved TVs that provided a 145Â° field of view of the driving environment.

The driving simulation scenario was 13 km long and lasted approximately 16 min (given speed limits). The scenario took place in an urban environment during the day with a low to moderate traffic density. During the driving scenario, several intersections with crosswalks and other road users formed the driving environment to create an object-rich test environment. During the driving scenario, the driver received four prompts to turn on the automated driving system to start automated driving and four prompts to take over control of the vehicle to continue driving manually.

C. AUTOMOTIVE HUDs

The baseline trial included a simple HUD along with a physical dashboard that was part of the driving simulator (the dashboard is shown in Figure 1). In the basic HUD, only two types of information are displayed: navigation instructions and takeover prompts. The navigation cues were simple bird's-eye view replicas of the cross-sections with arrows indicating the route. The takeover prompt was displayed as a visual notification consisting of the text "Take over!" and as a numeric countdown notification indicating the time remaining before automation was turned off. The takeover notification was displayed along with an audible notification in the form of a pure 4000 Hz tone at 65 dB.

The AR-HUD trial included multiple information and visual elements presented in two dimensions (2D) and using augmented reality. A preview of AR-HUD is shown in Figure 2.

The information displayed on the AR-HUD includes the vehicle speed, current speed limit, available ADAS features, and bounding boxes to indicate relevant objects

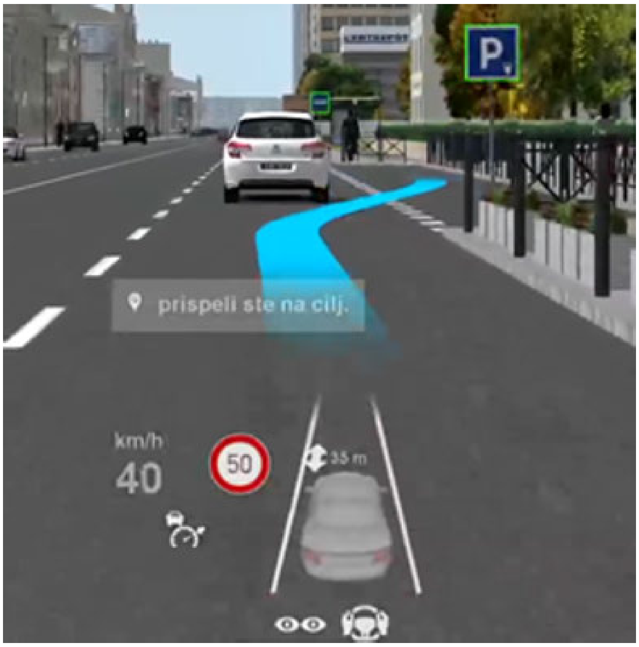
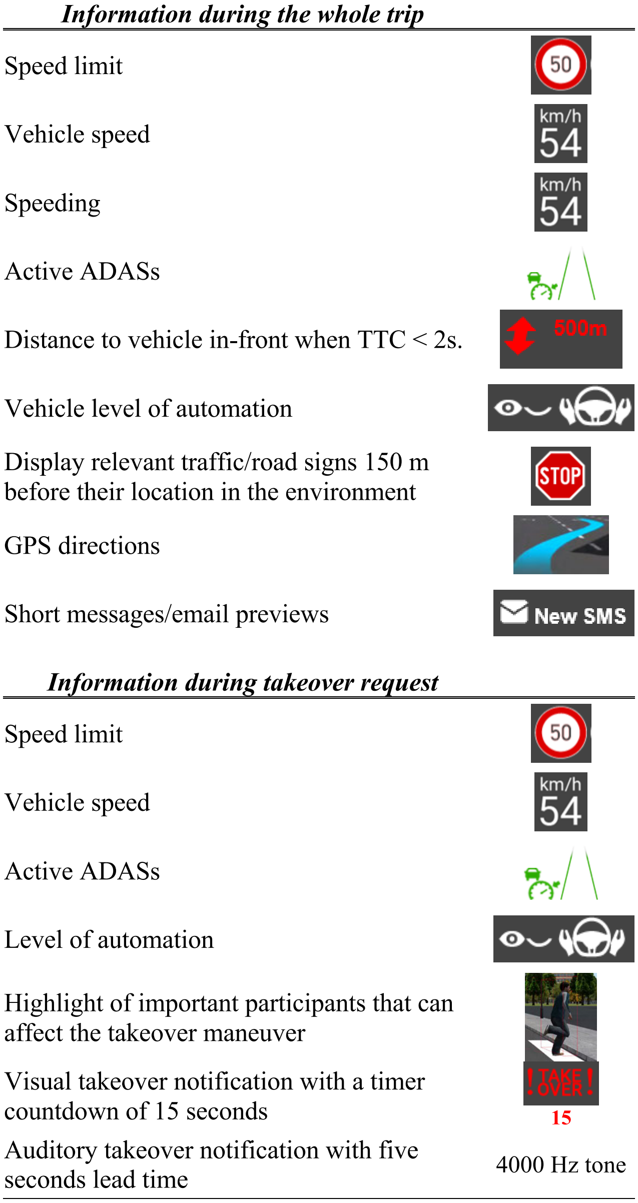


FIGURE 2. AR-HUD: visual elements presented during the drive.



Information during the whole trip	
Speed limit	50
Vehicle speed	km/h 54
Speeding	km/h 54
Active ADASs	
Distance to vehicle in-front when TTC < 2s.	500m
Vehicle level of automation	
Display relevant traffic/road signs 150 m before their location in the environment	STOP
GPS directions	
Short messages/email previews	New SMS
Information during takeover request	
Speed limit	50
Vehicle speed	km/h 54
Active ADASs	
Level of automation	
Highlight of important participants that can affect the takeover maneuver	
Visual takeover notification with a timer countdown of 15 seconds	! TAKE ! ! OVER ! 15
Auditory takeover notification with five seconds lead time	4000 Hz tone

FIGURE 3. Information shown in AR-HUD during driving and takeover.

(see Figure 3 for details). During the takeover request, AR-HUD displayed the same takeover notification as that used in the baseline HUD, along with the vehicle speed, current speed limit, and AR, highlighting the relevant road users that could affect the takeover maneuver. The information features of AR-HUD, which were displayed throughout the trip, and those that were displayed only during the takeover are shown in Figure 3.

D. INSTRUMENTS AND MEASURES

A Tobii Pro Glasses 2 eye tracker with a sampling frequency of 50 Hz was used to record pupillometry data [56]. Electrodermal activity (EDA), skin temperature, and heart rate variability (HRV) were recorded using the Empatica E4 wristband [57], which records data on galvanic skin response (EDA at 4 Hz), skin temperature (4 Hz), blood volume pulse (BVP, at 64 Hz), and interbeat interval (IBI, obtained by processing the BVP signal).

The physiological signals used in the analysis were the HRV, EDA, skin temperature, and pupil size. The physiological characteristics of these indicators are presented in Section IV. Physiological signals were used as inputs to machine learning classifiers to predict cognitive load as a function of HUD type: baseline HUD vs. AR-HUD.

The dependent variables for the subjective measures were driver user experience, system usability, and the acceptance of advanced traffic telematics.

User experience was assessed using the User Experience Questionnaire (UEQ), which consists of 26 questions designed to evaluate six aspects of perceived user experience: Attractiveness, Understandability, Efficiency, Reliability, Stimulation, and Novelty. The UEQ scores were calculated using the UEQ Data Analysis Tool available on the UEQ website. The UEQ rating scale ranges from -3 (terribly bad) to 3 (extremely good). The perceived usability of the system was assessed using the System Usability Scale (SUS), which consists of 10 usability questions. Responses were then used to calculate a score on a scale of 0-100, with a score of 68 set as a discriminatory cutoff: a score below 68 indicates below-average perceived ease of use, while a score above 68 indicates above-average perceived ease of use. Finally, system acceptance was assessed using the Acceptance of Advanced Transportation Telematics (AATT) questionnaire, which consists of nine questions. Two aspects were calculated based on the results of all AATT questions, representing the perceived usefulness and user satisfaction with the evaluated solution. The statistical analysis of these values is presented in Section IV-A.

E. PROCEDURE

The study had a 2 (driving mode: manual vs. auto) × 2 (display configuration: baseline HUD vs. AR-HUD) factorial design, with repeated measures for the first factor. All participants completed two trials: 1) a baseline trial in which test participants drove without the HUD and 2) a trial with the HUD, in which the HUD was displayed in addition to the dashboard, which was also used in the baseline trial. The order of the trials was randomized.

The study protocol was as follows.

Pre-trial:

1) Introduction: Instructions were provided in written form to ensure that each participant received the same amount of information.
2) Informed consent: The participants were asked to sign an informed consent form before the start of the study.
3) Demographic Questionnaire: The participants were asked to provide information on their age, sex, and driving experience. No personal data were collected for the study.
4) Test drive: The participants completed a test drive to familiarize themselves with the simulator and the study tasks. Biometric sensors were attached and calibrated before the test drive began.

Trial:

1) Drive 1: Complete the Baseline or AR-HUD trial (depending on the randomized order).
2) Self-Assessment: Completion of the UEQ, SUS, and AATT questionnaires regarding the first drive.
3) Drive 2: Complete the baseline or HUD trial (depending on the randomized order).
4) Self-Assessment: Complete the UEQ, SUS, and AATT questionnaires regarding the second drive.

F. STATISTICS AND MACHINE LEARNING

Data preprocessing and analyses were performed in Python v.3.10 [58] using pinguoun v.0.5.3 [59] for statistical analysis and mlxtend [60] and scikit-learn [61] libraries for machine learning.

1) STATISTICAL ANALYSIS

The nonparametric Mann-Whitney U and Kruskal-Wallis tests were used when the data were not normally distributed, and Welch's t-test was used when the data were normally distributed but had unequal variances. The significance level was set at $\alpha = 0.05$. Cronbach's alpha was used to test the reliability of self-reports from the UEQ, SUS, and AATT.

2) MACHINE LEARNING

Raw physiological data were preprocessed and normalized (z-score normalization) before further steps were performed to determine the features. All data were normalized and features were calculated per driver, driving interval, driving mode (manual vs. auto), and HUD type (baseline vs. AR-HUD).

Basic statistics and pycatch22 [62] were used to generate features from the skin temperature and pupil dilation signals. The basic statistical features include mean, standard deviation, minimum, maximum, skewness, and kurtosis. Pycatch22 is a widely used time-series characterization library with a collection of 22 time-series specific features describing symbolic, temporal, and frequency ranges, including the distribution shape, timing of extreme events, linear and nonlinear autocorrelation, incremental differences, and self-affine scaling (for an overview and feature definitions, see [62]). The effects of the two feature-generation approaches were compared later when evaluating classifier performance. NeuroKit2 v.0.2.1 [63], a Python library for biomedical signal processing, was used to process the data, generate features for heart rate variability (HRV), and extract tonic and phasic features from EDA signals. The mean and standard deviation of tonic and phasic EDA were used as features.

The physiological signal features were first tested for multicollinearity and variance, and all features with collinearity $> 95\%$ and/or zero variance were removed. Further analysis and selection were performed by recursive feature elimination (RFE) with 5-fold cross-validation and the LightGBM (LGBM) classifier. The most important features of each signal were retained.

The effects of signal fusion on classifier performance were analyzed using an exhaustive feature selector [60] by selecting and evaluating all possible signal combinations ($n = 15$) using 5-fold cross-validation and LGBM.

The gradient-boosting machine classifiers LGBM, HistGradientBoostingClassifier (HGBC), and XGBoost (XGB) from scikit-learn were used as machine learning models [61]. The advantage of these ensemble models is that they can handle missing data and are insensitive to scale differences in data. Repeated stratified k-fold cross-validation (n_splits = 10, n_repeats = 5) was used to evaluate classifier performance, with the ROC AUC serving as a measure of model performance. The optimization configurations for all classifiers were left at the default values.

SHapley Additive exPlanations (SHAP) analysis [64] was conducted to improve the interpretability of the best-performing model and to show how the cognitive load associated with the HUD type affects physiological responses.

IV. RESULTS

A. SELF-REPORT MEASURES

The self-reported results of 28 drivers for the UEQ, SUS, and AATT were analyzed. Good internal consistency was found for all three questionnaires, with Cronbach's alphas for the SUS ($\alpha = 0.79$), AATT ($\alpha = 0.81$), and UEQ ($\alpha = 0.71$).

TABLE 2. Definition of HRV and EDA features along with basic statistical features for pupil dilation and skin temperature.

Selected features	Description
HRV_RMSSD	The square root of the mean of the squared successive differences between adjacent RR intervals.
HRV_SDSD	The standard deviation of the successive differences between RR intervals.
HRV_MeanNN	The mean of the RR intervals.
HRV_SDNN	The standard deviation of the RR intervals.
temp_mean	average skin temperature
temp_std	standard deviation of skin temperature
temp_min	minimum skin temperature
temp_max	maximum skin temperature
pupil_mean	average pupil size
pupil_min	minimum pupil size
pupil_max	maximum pupil size
pupil_skew	skewness of pupil dilation.
EDA_Tonic_mean	average tonic component of the signal, or the Tonic Skin Conductance Level (SCL).
EDA_Tonic_std	standard deviation of the tonic component of the signal.
EDA_Phasic_mean	average phasic component of the signal, or the Phasic Skin Conductance Response (SCR).
EDA_Phasic_std	standard deviation of the phasic component of the signal.

TABLE 3. Definition of catch22 time-series specific features for pupil dilation and skin temperature.

Selected features	Description
temp_SC_FluctAnal_2_rsrangefit_50_1_logi_prop_r1	Rescaled range fluctuation analysis
temp_SC_FluctAnal_2_dfa_50_1_2_logi_prop_r1	Rescaled range fluctuation analysis
temp_CO_trev_1_num	Nonlinear autocorrelation
temp_DN_OutlierInclude_n_001_mdrmd	Negative outlier timing
pupil_CO_HistogramAMI_even_2_5	5-bin histogram mode
pupil_CO_Embed2_Dist_tau_d_expfit_meandiff	Embedding distance distribution
pupil_FC_LocalSimple_mean1_tauresrat	Change in autocorrelation
pupil_CO_f1ecac	Nonlinear autocorrelation
pupil_CO_trev_1_num	Nonlinear autocorrelation
pupil_DN_OutlierInclude_n_001_mdrmd	Negative outlier timing

The Mann-Whitney U test revealed no significant difference in the SUS scores between the two HUD types. However, the mean SUS scores were significantly higher for AR-HUD: HUD (M = 83.50, SD = 11.47) vs. AR-HUD (M = 87.47, SD = 9.87). No significant differences were found in AATT scores (Usefulness and Satisfaction) between the two HUD types. The average AATT scores for both HUDs differed only slightly in terms of Usefulness (baseline HUD (M = 1.20, SD = 0.70) vs. AR-HUD (M = 1.29, SD = 0.51)) and Satisfaction (baseline HUD (M = 0.95, SD = 0.54) vs. AR-HUD (M = 1.03, SD = 0.50)).

For the UEQ, a Welch's t-test showed a significant effect for Novelty, t(27) = -2.14, p = .004, with scores for AR-HUD (M = 1.23, SD = 0.80) significantly higher than for baseline HUD (M = 0.63, SD = 1.26). No significant effects were found for the other five dimensions of the UEQ (Attractiveness, Clarity, Efficiency, Reliability, and Stimulation).

B. PHYSIOLOGICAL SIGNAL ANALYSIS AND FEATURE SELECTION

This section investigates the physiological responses to the cognitive load induced by the type of HUD (baseline vs. AR-HUD). Kruskal-Wallis analysis of variance revealed no significant differences in the physiological responses to the two HUD types or between manual and conditionally automated driving.

Machine learning was employed to investigate physiological responses to the cognitive load induced by the HUD type. Two different feature generation approaches were compared, as mentioned in Subsection III-F2. Features generated using catch22 time-series analysis were compared with the basic statistical features of skin temperature and pupil dilation, with both sets having the same HRV and EDA features. This was performed to investigate whether the basic statistical approach can provide comparable performance to that of computationally demanding time-series analysis. If both sets would show comparable performance, then a set with a lower computational load would be preferable for a system working continuously and in a real-world setting.

First, feature selection using RFE with the LGBM classifier was performed to select the best predictors of HUD-induced cognitive load for each signal. 5-fold cross-validation was used to reduce model bias. The selected features for HRV and the tonic and phasic features of EDA were identical in both final feature sets. The basic statistical feature set is listed in Table 2.

The catch22 set included the same features for HRV and EDA (presented in Table 2), and the time-series specific catch22 features generated for pupil dilation and skin temperature (shown in Table 3).

The feature importance values calculated for the two sets of selected features are shown in Figures 4 and 5.

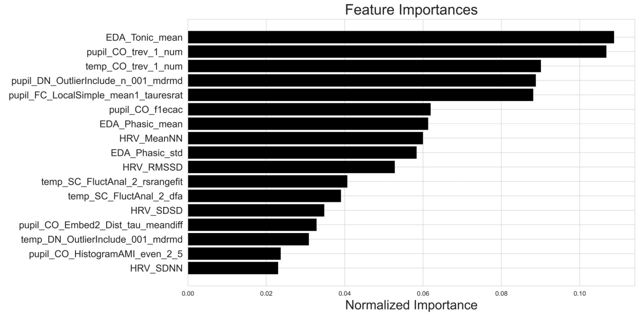


FIGURE 4. Time-series features: feature importance generated with LGBM classifier based on HRV, EDA, and the time-series (catch22) features for skin temperature and pupil dilation. Note that some catch22 feature names were shortened for presentation.

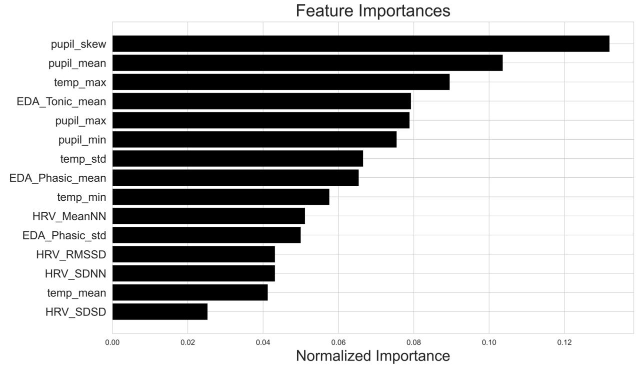


FIGURE 5. Basic statistical features: feature importance generated with LGBM classifier based on HRV, EDA, and the basic statistical features for skin temperature and pupil dilation.

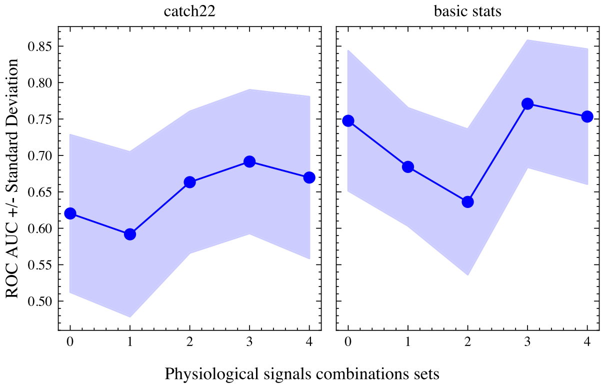


FIGURE 6. Comparison of physiological signal fusion for both feature sets, catch22 (left) and basic statistics (right). The ROC AUC scores are shown with the bounds for standard deviation of the scores.

1) EFFECTS OF SIGNAL FUSION ON CLASSIFICATION PERFORMANCE

Next, the two sets of selected features were used in the signal fusion to determine the optimal combination of signals for predicting the cognitive load induced by the two HUDs. The effects of signal fusion on classification performance were analyzed using the exhaustive feature selector [60] by sampling and evaluating all possible signal combinations ($n = 15$) with 5-fold cross-validation and LGBM.

Figure 6 shows a performance comparison of the two feature sets (based on the catch22 and the basic statistics) trained using the LGBM. The numbers on the x-axis in Figure 6 denote the five best-performing signal fusion sets listed in Tables 4 and 5.

TABLE 4. Catch22 time-series feature set. The signal fusion of HRV, EDA, and catch22 features for the skin temperature and pupil dilation. The effects of physiological signal combinations on the LGBM classifier performance.

Set	Signal fusion	AUC	CI_bound	STD	std_err
3	HRV, skin, pupil	0.69	0.03	0.09	0.01
4	HRV, skin, EDA	0.67	0.03	0.10	0.01
2	HRV, pupil, EDA	0.67	0.03	0.10	0.02
0	skin, pupil, EDA	0.63	0.03	0.11	0.02
1	HRV, skin, pupil, EDA	0.58	0.03	0.10	0.01

TABLE 5. Basic statistics feature set. Signal fusion of HRV, EDA, and basic statistical features for skin temperature and pupil dilation. The effects of physiological signal combinations on the LGBM classifier performance.

Set	Signal fusion	AUC	CI_bound	STD	std_err
3	HRV, skin, pupil	0.77	0.02	0.09	0.01
4	HRV, skin, EDA	0.76	0.03	0.09	0.01
0	HRV, pupil, EDA	0.75	0.02	0.09	0.01
1	skin, pupil, EDA	0.69	0.02	0.08	0.01
2	HRV, skin, pupil, EDA	0.63	0.03	0.10	0.01

Table 4 shows the five best-performing signal fusion sets based on the catch22 time-series feature set, whereas Table 5 shows the same for the basic statistical feature set.

Along with each fusion set, the tables provide the ROC AUC performance scores (AUC), confidence interval bounds of the cross-validation score average (CI_bound), standard deviations (STD), and standard errors (std_err.). Tables 2 and 3 present the features that represent each signal.

C. CLASSIFICATION OF COGNITIVE LOAD: HUD VS. AR-HUD

The three machine-learning classifiers, LGBM, XGB, and HGBC, were further tested on the top-ranking feature sets presented in Tables 4 and 5, respectively. The results of the classifiers trained on the signal fusion sets of HRV, skin temperature, and pupil size are shown in Figure 7. The classifiers built on the basic statistical feature set for skin temperature and pupil dilation showed better overall performance than those built on the time-series (catch22) features of the two signals. The average AUC ROC scores and the standard deviations of the scores for the classifiers based on the catch22 set were: XGB = 0.67 (0.09), LGBM = 0.69 (0.09), and HGBC = 0.70 (0.09). The average AUC ROC scores and standard deviations of the scores for the classifiers based on the basic set were better overall compared to the catch22 set: XGB = 0.76 (0.07), LGBM = 0.76 (0.09), and HGBC = 0.76 (0.07). The best performance for each classifier was also obtained with the basic statistical feature set, with ROC AUC scores of LGBM = 0.96, XGB = 0.96, and HGBC = 0.98.

D. MODEL INTERPRETABILITY

To further investigate the physiological responses to cognitive load associated with each HUD type, SHAP analysis

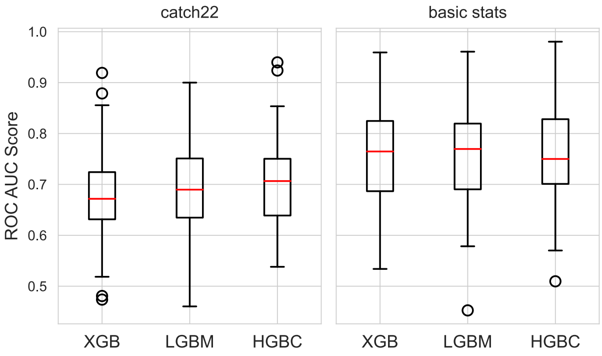


FIGURE 7. Gradient boosting classifier performances on the best performing fusion sets with features for skin temperature and pupil dilation generated with time-series (catch22) and basic statistics.

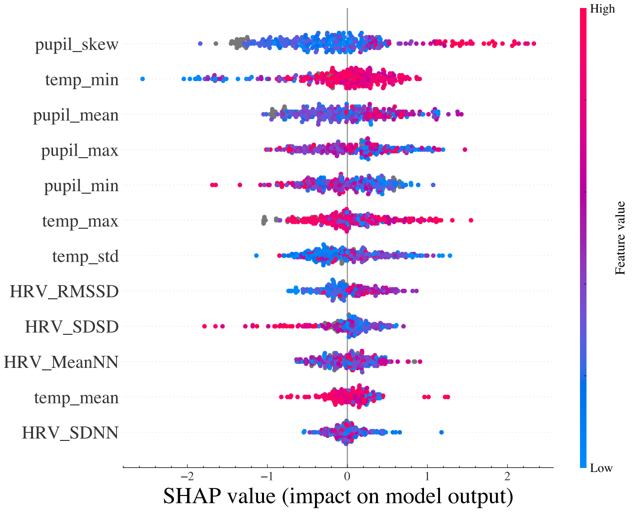


FIGURE 8. Physiological responses to cognitive load associated with AR-HUD. The impact of individual features on predicting the type of HUD is shown through positive and negative relationships between features and cognitive load. Features are ranked in descending order of importance.

was conducted on the features from the best-performing signal fusion set of HRV, pupil dilation, and skin temperature (Tables 5 and 2). To reduce model bias and substantiate its validity, stratified 10-fold cross-validation was used to train the LGBM classifier and calculate the SHAP values.

Figure 8 shows a SHAP summary plot of the impact of individual physiological features on predicting the cognitive load associated with AR-HUD. Note that for binary classification, the SHAP values for the baseline HUD and AR-HUD classes are symmetrical, as the contribution of a feature towards one class reduces the probability of its contribution to the other class by the same amount. The features are ranked in the descending order of importance. The dots represent the instances of each feature, and the horizontal position of the dot is determined by the SHAP value of that feature. The horizontal location of each dot shows the effect of its value on the prediction (cognitive load associated with HUD type). The density of each feature is observed from the swarm plot. The color indicates the original value of a feature, with high values denoted in red and low values denoted in blue [64].

Figure 8 shows that a high level of skew in pupil dilation has a high positive correlation with the cognitive load associated with AR-HUD. The increase in mean pupil size is also associated with AR-HUD, but to a lesser extent, and this effect is not as distinctive. The increase in mean pupil size may indicate an increase in cognitive load while driving with AR-HUD, as increases in load typically lead to increases in pupil dilation [46]. However, it might also show the effects of external conditions on pupil dilation, such as variations in luminance conditions associated with HUD type and/or road conditions [65]. A higher density of lower pupil skew is associated with the baseline HUD, whereas a higher skew in pupil dilation is associated with AR-HUD. This may indicate that AR-HUD elements had higher luminance and/or that the participants were attentive to different screen areas when using AR-HUD, which would cause a higher deviation from the mean pupil size and thus higher levels of skew in pupil dilation while driving with AR-HUD. The latter is in line with the eye-tracking study of AR-HUDs by [24], who reported that AR-HUD makes "the visual gaze more dispersed, [and] the AR-HUD-assisted driving allocates more driving resources to places other than the central driving area" [24, p. 129963].

Another interesting finding is the SHAP values for the HRV features, specifically the impact of RMSSD and SDSD on the predictive performance of the model. As shown in Figure 8, increased RMSSD correlates with the cognitive load associated with AR-HUD, whereas increased SDSD is negatively correlated with AR-HUD. The positive correlation between high RMSSD and AR-HUD might indicate that a lower cognitive load is associated with the AR-HUD as compared to that induced by the baseline HUD. This is in line with the results reported by [43], who investigated heart rate variability in resting, anticipatory, stressful, and recovery periods and its association with cognitive performance measured by a verbal learning task. Their results showed that RMSSD correlated negatively with the cognitive load induced during stressful periods but increased in the restful, anticipatory, and recovery periods. Their results also showed that the SDNN, which reflects the overall variability of the beat-to-beat RR intervals, increased during anticipatory and stressful periods. However, the SHAP values for the SDNN presented in Figure 8 are inconclusive, with a slight correlation between low SDNN values and the cognitive load associated with AR-HUD. This might be because SDNN represents phasic heart rate variability changes over longer periods of time and thus cannot capture shorter time intervals of the changing conditions (baseline vs. AR-HUD). In contrast, higher short-term (beat-by-beat) variability represented by SDSD is negatively correlated with the AR-HUD and positively correlated with the baseline HUD, as shown in Figure 8. Increased SDSD may indicate cognitive stress peaks during increased cognitive load. Similar results were obtained by [42], who studied physiological responses to cognitive load induced by driving with and without a secondary task. They reported that RMSSD and SDSD were the most robust measures, whereas SDNN failed to differentiate between the conditions. Moreover, in the present study, the lower minimum skin temperature values are being highly negatively correlated with AR-HUD, whereas the other SHAP results for skin temperature features are mostly inconclusive.

V. CONCLUSION AND FUTURE WORK

This study investigated whether differences in the physiological responses of drivers to cognitive load could predict the type of HUD used while driving. The present study found no statistically significant differences in the physiological responses to the two HUD types. This finding is consistent with the results of previous studies [15], [51], [52]. However, the results showed that physiological signals are reliable predictors of the cognitive load associated with HUD type. The impact of individual physiological features on predicting cognitive load was examined using the SHAP analysis on the best-performing feature set. The results showed that the most robust predictors of cognitive load associated with the two HUDs were pupil dilation and HRV signals, namely pupil skew, mean pupil size, HRV RMSSD, and HRV SDSD.

In terms of related work, no directly comparable studies have investigated drivers' physiological responses to the cognitive load induced by HUDs, as discussed in Section II. Most studies have investigated cognitive load associated with secondary tasks.

Studies by [12], and [13] showed that a higher driver cognitive load is associated with a secondary task. The results of [13], and [14] also showed that cognitive load can be accurately detected using machine learning classifiers based on physiological signals and their combinations. Their method is similar to that presented here, but they focus on the cognitive load induced by secondary tasks and takeover situations and report classification performance in terms of accuracy (e.g., accuracy between 69 and 73% for EDA and the best accuracy of 92-94% achieved by combining respiration and ECG), which is not directly comparable to our ROC AUC scores. In another classification study [14], driver cognitive load was assessed as a function of task difficulty (no task vs. low vs. high) and task modality (visual cognitive task vs. auditory cognitive task). The reported F1 metrics for task difficulty ranged from 0.51 to 0.71, depending on the model and feature combination. Their results also showed that the models had difficulty predicting the task modality (visual vs. auditory), with the best model using ECG and RESP as signals achieving a weighted F1 value of 0.61 and an average weighted F1 accuracy of approximately 50% [14].

The results presented show promising potential for physiological signals as indicators of cognitive load. The ability to predict the cognitive load induced by HUDs based on driver physiology means that dynamic regulation of cognitive load and its effects in real time is possible. This is an important step towards reducing excessive cognitive load and improving driver performance and safety.

This study has several limitations, most of which are characteristic of research conducted in simulated driving environments. One limitation of this study was the relatively small sample size of 28 participants, which may not be representative of the larger driving population. This should be addressed in future studies by recruiting a larger and more diverse driver sample. Another limitation is the use of a simulator, as it may not accurately capture the cognitive load in real-world driving situations given its controlled environment and lack of real-world unpredictability. Therefore, the results of this study should be investigated under actual driving situations.

Future work will consider a broader range of physiological indicators of cognitive load. Additionally, the type of HUD could be expanded to include more than just the baseline and AR-HUD, allowing for a more robust examination of the effects of different visual information complexities. It will examine HUDs in terms of visual information complexity, information cluster saturation, and placement in the driver's environment. Testing across additional forms of cognitive stimuli in the HUD should be considered. Further research could refine the application of machine-learning approaches within real-time systems to ensure practical feasibility, accuracy, and consistency. The study should also be conducted in actual driving situations, as opposed to a simulator, to improve the external validity.

ACKNOWLEDGMENT

This document reflects only the authors view, the Innovation and Networks Executive Agency (INEA) is not responsible for any use that may be made of the information it contains.

REFERENCES

[1] L. Masello, G. Castignani, B. Sheehan, F. Murphy, and K. McDonnell, "On the road safety benefits of advanced driver assistance systems in different driving contexts," *Transp. Res. Interdiscipl. Perspect.*, vol. 15, Sep. 2022, Art. no. 100670. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2590198222001300

[2] D. Beck, J. Jung, J. Park, and W. Park, "A study on user experience of automotive HUD systems: Contexts of information use and user-perceived design improvement points," *Int. J. Human–Computer Interact.*, vol. 35, no. 20, pp. 1936–1946, Dec. 2019.

[3] A. Riegler, A. Riener, and C. Holzmann, "Augmented reality for future mobility: Insights from a literature review and HCI workshop," *i-com*, vol. 20, no. 3, pp. 295–318, Dec. 2021.

[4] K. Park and Y. Im, "Ergonomic guidelines of head-up display user interface during semi-automated driving," *Electronics*, vol. 9, no. 4, p. 611, Apr. 2020, doi: 10.3390/electronics9040611.

[5] K. Chang and T. Seder, "Automotive augmented reality: User experience and enabling technology," *Inf. Display*, vol. 38, no. 1, pp. 12–18, Jan. 2022. [Online]. Available: https://sid.onlinelibrary.wiley.com/doi/abs/10.1002/msid.1272

[6] J. Park and W. Park, "A review on the interface design of automotive head-up displays for communicating safety-related information," in *Proc. Human Factors Ergonom. Soc. Annu. Meeting*, Nov. 2019, vol. 63, no. 1, pp. 2016–2017.

[7] R. Currano, S. Y. Park, D. J. Moore, K. Lyons, and D. Sirkin, "Little road driving HUD: Heads-up display complexity influences drivers' perceptions of automated vehicles," in *Proc. CHI Conf. Human Factors Comput. Syst.*, May 2021, pp. 1–15, doi: 10.1145/3411764.3445575.

[8] H. Kim and J. L. Gabbard, "Assessing distraction potential of augmented reality head-up displays for vehicle drivers," *Human Factors, J. Human Factors Ergonom. Soc.*, vol. 64, no. 5, pp. 852–865, Aug. 2022, doi: 10.1177/0018720819844845.

[9] F. Naujoks, K. Wiedemann, N. Schömig, S. Hergeth, and A. Keinath, "Towards guidelines and verification methods for automated vehicle HMIs," *Transp. Res. Part F: Traffic Psychol. Behaviour*, vol. 60, pp. 121–136, Jan. 2019.

[10] H. Detjen, S. Faltaous, B. Pfleging, S. Geisler, and S. Schneegass, "How to increase automated vehicles' acceptance through in-vehicle interaction design: A review," *Int. J. Human–Computer Interact.*, vol. 37, no. 4, pp. 308–330, Feb. 2021.

[11] C. Jing, C. Shang, D. Yu, Y. Chen, and J. Zhi, "The impact of different AR-HUD virtual warning interfaces on the takeover performance and visual characteristics of autonomous vehicles," *Traffic Injury Prevention*, vol. 23, no. 5, pp. 277–282, Jul. 2022.

[12] V. Radhakrishnan, N. Merat, T. Louw, R. C. Gonçalves, G. Torrao, W. Lyu, P. P. Guillen, and M. G. Lenné, "Physiological indicators of driver workload during car-following scenarios and takeovers in highly automated driving," *Transp. Res. F, Traffic Psychol. Behaviour*, vol. 87, pp. 149–163, May 2022.

[13] Q. Meteier, M. Capallera, S. Ruffieux, L. Angelini, O. A. Khaled, E. Mugellini, M. Widmer, and A. Sonderegger, "Classification of drivers' workload using physiological signals in conditional automation," *Frontiers Psychol.*, vol. 12, Feb. 2021, Art. no. 596038.

[14] Q. Meteier, E. De Salis, M. Capallera, M. Widmer, L. Angelini, O. A. Khaled, A. Sonderegger, and E. Mugellini, "Relevant physiological indicators for assessing workload in conditionally automated driving, through three-class classification and regression," *Frontiers Comput. Sci.*, vol. 3, Jan. 2022, Art. no. 775282.

[15] M. Lohani, J. M. Cooper, G. G. Erickson, T. G. Simmons, A. S. McDonnell, A. E. Carriero, K. W. Crabtree, and D. L. Strayer, "No difference in arousal or cognitive demands between manual and partially automated driving: A multi-method on-road study," *Frontiers Neurosci.*, vol. 15, Jun. 2021, Art. no. 577418.

[16] A. Riegler, A. Riener, and C. Holzmann, "A systematic review of virtual reality applications for automated driving: 2009–2020," *Frontiers Human Dyn.*, vol. 3, p. 48, Aug. 2021, doi: 10.3389/fhumd.2021.689856.

[17] J. Winter, P. Leeuwen, and R. Happee, "Advantages and disadvantages of driving simulators: A discussion," in *Proc. Measuring Behav. Conf.*, 2012, pp. 47–50.

[18] M. Smith, K. Bagalkotkar, J. Gabbard, D. Large, and G. Burnett, "Isolating the effect of off-road glance duration on driving performance: An exemplar study comparing HDD and HUD in different driving scenarios," *J. Hum. Factors Ergonom. Soc.*, vol. 65, no. 5, 2021, Art. no. 187208211031416, doi: 10.1177/00187208211031416.

[19] B. Blissing, F. Bruzelius, and O. Eriksson, "Driver behavior in mixed and virtual reality—A comparative study," *Transp. Res. F, Traffic Psychol. Behaviour*, vol. 61, pp. 229–237, Feb. 2019.

[20] J. Wolffsohn, N. McBrien, G. Edgar, and T. Stout, "The influence of cognition and age on accommodation, detection rate and response times when using a car head-up display (HUD)," *Ophthalmic Physiological Opt.*, vol. 18, no. 3, pp. 243–253, 1998, doi: 10.1016/S0275-5408(97)00094-X.

[21] K. W. Gish, L. Staplin, J. Stewart, and M. Perel, "Sensory and cognitive factors affecting automotive head-up display effectiveness," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 1694, no. 1, pp. 10–19, Jan. 1999.

[22] A. Pauzie, "Head up display in automotive: A new reality for the driver," in *Design, User Experience, and Usability: Interactive Experience Design* (Lecture Notes in Computer Science), vol. 9188, A. Marcus, Ed. Cham, Switzerland: Springer, 2015, pp. 505–516, doi: 10.1007/978-3-319-20889-3_47.

[23] Y. Wang, Y. Wu, C. Chen, B. Wu, S. Ma, D. Wang, H. Li, and Z. Yang, "Inattentional blindness in augmented reality head-up display-assisted driving," *Int. J. Human–Computer Interact.*, vol. 38, no. 9, pp. 837–850, May 2022.

[24] X. Ma, M. Jia, Z. Hong, A. P. K. Kwok, and M. Yan, "Does augmented-reality head-up display help? A preliminary study on driving performance through a VR-simulated eye movement analysis," *IEEE Access*, vol. 9, pp. 129951–129964, 2021.

[25] S. Langlois and B. Soualmi, "Augmented reality versus classical HUD to take over from automated driving: An aid to smooth reactions and to anticipate maneuvers," in *Proc. IEEE 19th Int. Conf. Intell. Transp. Syst. (ITSC)*, Nov. 2016, pp. 1571–1578. [Online]. Available: http://ieeexplore.ieee.org/document/7795767/

[26] J. L. Gabbard, G. M. Fitch, and H. Kim, "Behind the glass: Driver challenges and opportunities for AR automotive applications," *Proc. IEEE*, vol. 102, no. 2, pp. 124–136, Feb. 2014.

[27] A. Frison, Y. Forster, P. Wintersberger, V. Geisel, and A. Riener, "Where we come from and where we are going: A systematic review of human factors research in driving automation," *Appl. Sci.*, vol. 10, no. 24, pp. 1–36, 2020, doi: 10.3390/app10248914.

[28] J. L. Campbell, Z. R. Doerzaph, C. M. Richard, and L. P. Bacon, "Human factors design principles for the driver-vehicle interface (DVI)," in *Proc. Adjunct 6th Int. Conf. Automot. User Interfaces Interact. Veh. Appl.*, Sep. 2014, pp. 1–6, doi: 10.1145/2667239.2667305.

[29] A. Furnham, "Response bias, social desirability and dissimulation," *Personality Individual Differences*, vol. 7, no. 3, pp. 385–400, Jan. 1986.

[30] P. M. Podsakoff, S. B. MacKenzie, J.-Y. Lee, and N. P. Podsakoff, "Common method biases in behavioral research: A critical review of the literature and recommended remedies," *J. Appl. Psychol.*, vol. 88, no. 5, pp. 879–903, 2003.

[31] S. Bauhoff, "Systematic self-report bias in health data: Impact on estimating cross-sectional and treatment effects," *Health Services Outcomes Res. Methodology*, vol. 11, nos. 1–2, pp. 44–53, Jul. 2011.

[32] R. S. Kreitchmann, F. J. Abad, V. Ponsoda, M. D. Nieto, and D. Morillo, "Controlling for response biases in self-report scales: Forced-choice vs. psychometric modeling of Likert items," *Frontiers Psychol.*, vol. 10, p. 2309, Oct. 2019.

[33] A. Scott and A. T. Balthrop, "The consequences of self-reporting biases: Evidence from the crash preventability program," *J. Oper. Manage.*, vol. 67, no. 5, pp. 588–609, Jul. 2021.

[34] C. Spencer, I. A. Koc, C. Suga, A. Lee, A. M. Dhareshwar, E. Franzén, M. Iozzo, G. Morrison, and G. McKeown, "Assessing the use of physiological signals and facial behaviour to gauge drivers' emotions as a UX metric in automotive user studies," in *Proc. 12th Int. Conf. Automot. User Interfaces Interact. Veh. Appl.* New York, NY, USA: Association for Computing Machinery, Sep. 2020, pp. 78–81, doi: 10.1145/3409251.3411728.

[35] P. Ayres, J. Y. Lee, F. Paas, and J. J. G. van Merriënboer, "The validity of physiological measures to identify differences in intrinsic cognitive load," *Frontiers Psychol.*, vol. 12, Sep. 2021, Art. no. 702538.

[36] E. Galy, M. Cariou, and C. Mélan, "What is the relationship between mental workload factors and cognitive load types?" *Int. J. Psychophysiology*, vol. 83, no. 3, pp. 269–275, Mar. 2012.

[37] M. Malik, "Heart rate variability: Standards of measurement, physiological interpretation and clinical use. Task force of the European society of cardiology and the North American society of pacing and electrophysiology," *Circulation*, vol. 93, no. 5, pp. 1043–1065, 1996.

[38] L. R. Fournier, G. F. Wilson, and C. R. Swain, "Electrophysiological, behavioral, and subjective indexes of workload when performing multiple tasks: Manipulations of task difficulty and training," *Int. J. Psychophysiology*, vol. 31, no. 2, pp. 129–145, Jan. 1999.

[39] M. A. Hogervorst, A.-M. Brouwer, and J. B. F. van Erp, "Combining and comparing EEG, peripheral physiology and eye-related measures for the assessment of mental workload," *Frontiers Neurosci.*, vol. 8, p. 322, Oct. 2014.

[40] T. Heine, G. Lenis, P. Reichensperger, T. Beran, O. Doessel, and B. Deml, "Electrocardiographic features for the measurement of drivers' mental workload," *Appl. Ergonom.*, vol. 61, pp. 31–43, May 2017.

[41] F. Walker, J. Wang, M. H. Martens, and W. B. Verwey, "Gaze behaviour and electrodermal activity: Objective measures of drivers' trust in automated vehicles," *Transp. Res. F, Traffic Psychol. Behaviour*, vol. 64, pp. 401–412, Jul. 2019.

[42] B. Mehler, B. Reimer, and Y. Wang, "A comparison of heart rate and heart rate variability indices in distinguishing single-task driving and driving under secondary cognitive workload," in *Proc. 6th Int. Driving Symp. Human Factors Driver Assessment, Training, Vehicle Design Driving Assessment*. Iowa City, IA, USA: University of Iowa, 2011, pp. 590–597.

[43] K. Hilgarter, K. Schmid-Zalaudek, R. Csanády-Leitner, M. Mörtl, A. Rössler, and H. K. Lackner, "Phasic heart rate variability and the association with cognitive performance: A cross-sectional study in a healthy population setting," *PLoS ONE*, vol. 16, no. 3, Mar. 2021, Art. no. e0246968.

[44] S. T. Iqbal, X. S. Zheng, and B. P. Bailey, "Task-evoked pupillary response to mental workload in human-computer interaction," in *Proc. CHI Extended Abstr. Human Factors Comput. Syst.*, 2004, pp. 1477–1480.

[45] M. Pomplun and S. Sunkara, "Pupil dilation as an indicator of cognitive workload in human-computer interaction," in *Human-Centered Computing*. Boca Raton, FL, USA: CRC Press, 2019, pp. 542–546.

[46] P. van der Wel and H. van Steenbergen, "Pupil dilation as an index of effort in cognitive control tasks: A review," *Psychonomic Bull. Rev.*, vol. 25, no. 6, pp. 2005–2015, Dec. 2018.

[47] S. Yan, C. C. Tran, Y. Wei, and J. L. Habiyaremye, "Driver's mental workload prediction model based on physiological indices," *Int. J. Occupational Saf. Ergonom.*, vol. 25, no. 3, pp. 476–484, Jul. 2019.

[48] A. S. McDonnell, T. G. Simmons, G. G. Erickson, M. Lohani, J. M. Cooper, and D. L. Strayer, "This is your brain on autopilot: Neural indices of driver workload and engagement during partial vehicle automation," *Hum. Factors*, Aug. 2021, Art. no. 187208211039091. [Online]. Available: http://journals.sagepub.com/doi/10.1177/00187208211039091, doi: 10.1177/00187208211039091.

[49] A. L. Müller, N. Fernandes-Estrela, R. Hetfleisch, L. Zecha, and B. Abendroth, "Effects of non-driving related tasks on mental workload and take-over times during conditional automated driving," *Eur. Transp. Res. Rev.*, vol. 13, p. 16, Dec. 2021.

[50] A. Gomaa, A. Alles, E. Meiser, L. H. Rupp, M. Molz, and G. Reyes, "What's on your mind? A mental and perceptual load estimation framework towards adaptive in-vehicle interaction while driving," in *Proc. 14th Int. Conf. Automot. User Interfaces Interact. Veh. Appl.* New York, NY, USA: Association for Computing Machinery, Sep. 2022, pp. 215–225.

[51] J. Stapel, F. A. Mullakkal-Babu, and R. Happee, "Automated driving reduces perceived workload, but monitoring causes higher cognitive load than manual driving," *Transp. Res. F, Traffic Psychol. Behaviour*, vol. 60, pp. 590–605, Jan. 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1369847818301335

[52] A. Calvi, F. D'Amico, L. B. Ciampoli, and C. Ferrante, "Evaluation of driving performance after a transition from automated to manual control: A driving simulator study," *Transp. Res. Proc.*, vol. 45, pp. 755–762, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352146520301514

[53] C. J. Normark, P. Tretten, and A. Gärling, "Do redundant head-up and head-down display configurations cause distractions?" in *Proc. 5th Int. Driving Symp. Human Factors Driver Assessment, Training, Vehicle Design Driving Assessment*, vol. 5, 2009, pp. 398–404.

[54] A. R. Kumar, S. Ho, X. Wu, and T. Misu, "How do different cueing strategies affect drivers' perceived workload," in *Proc. Hum. Factors Ergonom. Soc. Annu. Meeting*, vol. 66, no. 1. Los Angeles, CA, USA: SAGE, 2022, pp. 1696–1700.

[55] M. Vengust, B. Kaluža, K. Stojmenova, and J. Sodnik, "NERVteh compact motion based driving simulator," in *Proc. 9th Int. Conf. Automot. User Interfaces Interact. Veh. Appl. Adjunct*, 2017, pp. 242–243.

[56] T. Sweden. (2020). *Pro Glasses 2 Eye Tracker*. Accessed: May 19, 2023. [Online]. Available: https://www.tobii.com/products/discontinued/tobii-pro-glasses-2

[57] Empatica. (2022). *Pro Glasses 2 Eye Tracker*. Accessed: May 19, 2023. [Online]. Available: https://www.empatica.com/research/e4/

[58] G. Van Rossum and F. L. Drake, *Python 3 Reference Manual*. Scotts Valley, CA, USA: CreateSpace, 2009.

[59] R. Vallat, "Pingouin: Statistics in Python," *J. Open Source Softw.*, vol. 3, no. 31, p. 1026, Nov. 2018.

[60] S. Raschka, "MLxtend: Providing machine learning and data science utilities and extensions to Python's scientific computing stack," *J. Open Source Softw.*, vol. 3, no. 24, p. 638, Apr. 2018. [Online]. Available: https://joss.theoj.org/papers/10.21105/joss.00638

[61] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Oct. 2011.

[62] C. H. Lubba, S. S. Sethi, P. Knaute, S. R. Schultz, B. D. Fulcher, and N. S. Jones, "*catch22*: Canonical time-series characteristics: Selected through highly comparative time-series analysis," *Data Mining Knowl. Discovery*, vol. 33, no. 6, pp. 1821–1852, Nov. 2019.

[63] D. Makowski, T. Pham, Z. J. Lau, J. C. Brammer, F. Lespinasse, H. Pham, C. Schölzel, and S. H. A. Chen, "NeuroKit2: A Python toolbox for neurophysiological signal processing," *Behav. Res. Methods*, vol. 53, no. 4, pp. 1689–1696, Aug. 2021, doi: 10.3758/s13428-020-01516-y.

[64] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2017, pp. 4765–4774. [Online]. Available: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

[65] J. Xu, Y. Wang, F. Chen, and E. Choi, "Pupillary response based cognitive workload measurement under luminance changes," in *Proc. IFIP Conf. Hum.-Comput. Interact.* Cham, Switzerland: Springer, 2011, pp. 178–185.

GREGOR STRLE received the Ph.D. degree in cognitive science from the University of Nova Gorica, in 2012. He is currently a Research Fellow with the Faculty of Electrical Engineering, University of Ljubljana, and the Scientific Research Centre of Slovenian Academy of Sciences and Arts (ZRC SAZU). He is also an Assistant Professor in philosophy of AI with the Postgraduate School ZRC SAZU and a member of the User-Adapted Communications and Ambient Intelligence Laboratory. His research interests include user modeling based on psychophysiology and machine learning, affective computing, and human–computer interaction in general.

ANDREJ KOŠIR (Senior Member, IEEE) received the B.Sc. degree in mathematics and the Ph.D. degree in electrical engineering from the University of Ljubljana, in 1999. Since 2014, he has been a Full Professor with the Faculty of Electrical Engineering, University of Ljubljana, and the Head of the User-Adapted Communications and Ambient Intelligence Laboratory. He is active in a broad research fields including: user modeling and personalization (user models and recommender systems), user interfaces (machine learning method design and statistical analysis), and social signal processing.

JAKA SODNIK (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from the University of Ljubljana, in 2007. He is currently an Associate Professor in electrical engineering with the Faculty of Electrical Engineering, University of Ljubljana. As a member of the ICT Department and the Laboratory for Information Technologies. He is also an active researcher and a supervisor in the fields of human–machine interaction, web technologies, and acoustics. He has extensive experience and references in the field of information and communication technologies. His research interests include acoustics, telecommunication networks, web technologies, and human–machine interaction, especially driver-vehicle interaction.

KRISTINA STOJMENOVA PEČEČNIK (Member, IEEE) received the Ph.D. degree in electrical engineering from the Faculty of Electrical Engineering, University of Ljubljana, Slovenia, in 2018. She is currently a Research Associate with the Information and Communications Technology Department, Faculty of Electrical Engineering, University of Ljubljana. Her research interests include human–computer interaction and information technologies, focusing on in-vehicle information systems and driving state and behavior assessment.