Towards Real-Time Recognition of Users' Mental Workload Using Integrated Physiological Sensors Into a VR HMD

Tiffany Luong*[†][‡] Nicolas Martin*[†] Anais Raison[†] Ferran Argelaguet[‡] Jean-Marc Diverrez[†]
Anatole Lécuyer[‡]

[†]IRT b<>com, Cesson-Sevigne, France
[‡]Univ Rennes, Inria, CNRS, IRISA, Rennes, France

Abstract

This paper describes an "all-in-one" solution for the real-time recognition of users' mental workloads in virtual reality through the customization of a commercial HMD with physiological sensors. First, we describe the hardware and software solution employed to build the system. Second, we detail the machine learning methods used for the automatic recognition of the users' mental workload, which are based on the well-known Random Forest algorithm. In order to gather data to train the system, we conducted an extensive user study with 75 participants using a VR flight simulator to induce different levels of mental workload. In contrast to previous works which label the data based on a standardized task (e.g. n-back task) or on a pre-defined task-difficulty, participants were asked about their perceived mental workload level along the experiment. With the data collected, we were able to train the system in order to classify four different levels of mental workload with an accuracy up to 65%. In addition, we discuss the role of the signal normalization procedures, the contribution of the different physiological signals on the recognition accuracy and compare the results obtained with the sensors embedded in the HMD with commercial grade systems. Preliminary results show our pipeline is able to recognize mental workload in real-time. Taken together, our results suggest that such all-in-one approach, with physiological sensors directly embedded in the HMD, is a promising path for VR applications in which the real-time or off-line estimation of Mental Workload assessment is beneficial.

Index Terms: Human-centered computing—Human computer interaction (HCI)—Interaction paradigms—Virtual reality

1 Introduction

Using virtual reality (VR), complex virtual environments (VEs) can be simulated in order to study human behaviour and psychological states. In addition, VR has the ability to create reproducible and sophisticated protocols in a safe way, which might be not feasible or too expensive in real-life. It can immerse and engage users by making them feel "*present*" in the virtual world [71]. For those reasons, VR has been used extensively to design training applications, and to test system design. On the other hand, mental workload (MW), which can refer to the "*ratio of demand to allocated resources*" [22], has long been recognized as an important factor in these fields [14,27,55,82]. It was shown to have an effect on workers' well-being and work performances [27,82].

The most common methods to assess MW in VR rely on the use of questionnaires (e.g., NASA TLX [33]) that are administered punctually (e.g., at the end of the experience), or physiological monitoring which can be done continuously [18,47,48,60,81]. While the administration of the questionnaires can disrupt the VR experience, thus degrading users' immersion, physiological monitoring is non-invasive and if done in real-time allows to customize the training/experience based on the users' psychological states.

However, due to the complexity of physiological data, machine learning (ML) methods are typically required in order to extract the users' mental workloads. A particular challenge remains the training of the ML models as they require the gathering of labeled physiological data and specific induction protocols to ensure that different levels of MW are generated. For now, induction protocols have mainly used single standardized tasks (e.g., n-back task) or relied on task-difficulty metrics to label MW data. However, single tasks tend to stimulate limited pools of cognitive resources depending on the nature of the stimuli [90], which can influence the physiological signals. Moreover, although a number of different physiological measurements has been used to classify MW, it still remains unclear which are the ones that provide better classification rates in VR and whether they can be used in real-time recognition scenarios. Furthermore, the VR context implies different constraints than in the real-world, such as the amount of interaction, the cumbersomeness of VR equipment, the lack of visual feedback of the real world, and the effect of cybersickness on users' responses. While a wide range of commercial physiological sensors exist (e.g., bracelets, torso belts, electrodes patch) refereed hereinafter "Commercial Grade Systems" (CGS), a number of recent initiatives started to embed VR HMD with physiological sensors [1, 6, 7, 36]. Such setups proved to be convenient to reduce the potential cumbersomeness added by sensors, but this type of setups also implies different signals shapes and sensitivity compared to CGS sensors, which can interfere with the MW classification accuracy.

In this paper, we present an all-in-one solution to assess users' MW in real-time in VR. Section 2 presents an overview of related work. The all-in-one solution is depicted in Section 3 considering the hardware and software perspectives. Section 4 describes the study, which was conducted to assess the physiological (i.e., ocular activity, electrodermal activity (EDA), and cardiac activity) and task performance measures of 75 participants in the context of a VR flight simulator, where users had to perform different sub-tasks. The users self-reported their subjective MW levels throughout the experiment, and this measure was then used to label the dataset to train models, using the random forest (RF) algorithm, to classify 4 levels of MW. The classification performances are compared in Section 5, considering the HMD sensors setup and the CGS setup, as well as the different types of measure, and normalization methods. The analysis results show overall similar classification performances between the HMD setup and the CGS setup, reaching a classification accuracy of 65%. This supports the use of the physiological sensors integrated into VR HMDs for the recognition of users' MW in VR. Ocular activity features were especially important, followed by EDA, cardiac activity, and task performance features. Moreover,

*Both authors have contributed equally to this work.
[†]e-mail: firstname.lastname@b-com.com
[‡]e-mail: firstname.lastname@inria.fr

978-1-7281-8508-8/20/$31.00 ©2020 IEEE
DOI 10.1109/ISMAR50242.2020.00068

preliminary results show that our solution pipeline is able to do MW recognition in real-time. Finally, these results are further discussed in Sections 6 and 7; and Section 8 provides the concluding remarks.

There are 2 main contributions provided by this work: (1) a technical contribution with an all-in-one solution to recognize MW in VR in real-time using physiological sensors directly integrated into a VR HMD; (2) an experimental contribution which evaluates the approach using HMD sensors in comparison to CGS sensors, and which proposes an original evaluation based on classification performances. The experimental results also provide insights on the methods of normalization and the contribution of each type of measure and sensors on the classification accuracy.

2 Related Work

2.1 Mental Workload Measurement

O'Donnell classified the methods to measure MW in 3 categories [55]: subjective measures, task performances, and physiological measures. In this paper, our main focus is on physiological measures as they can provide a continuous assessment of the users' states and are less task-dependent as performance measurements [16, 40]. Changes in the user's physiological state can reflect processes in the Autonomic Nervous System (ANS) and in the Central Nervous System (CNS) [20, 35]. Nevertheless, subjective measurements (e.g., questionnaires [87]) are typically needed for data labeling purposes in supervised machine learning methods [54].

In this way, Heart Rate (HR) and Heart Rate Variability (HRV) seem to be impacted by MW (e.g., [11, 28, 32, 51]). For example, De Rivecourt et al. [21] showed that the HR is sensitive to task complexity and mental effort during a simulated flight. The HRV seems also to be sensitive to MW and usually decreases when MW increases [11]. Second, EDA revealed to be sensitive to MW according to prior research (e.g., [32, 50, 53, 73]). For instance, in a driving simulation, the addition of a visual stimulus to the driving task significantly influences EDA, suggesting that EDA is sensitive to MW in realistic tasks [50]. Third, many features from ocular activity have been related to MW, ranging from blink rate and blink frequency, to pupil diameter (e.g., [16, 19, 46, 85]). For example, pupil diameter has been shown to be sensitive to errors in a nuclear power plant simulation [52], suggesting that this measure is sensitive to MW. Finally, Recarte and Nunes observed significant changes in pupil diameters when a verbal output is added to a real driving task [64]. Other physiological signals such as electroencephalography (EEG) (e.g., [81]) or functional near-infrared spectroscopy (fNIRS) (e.g., [60]) have been explored in the context of MW measurement. Nevertheless, EEGs are cumbersome [12] and their installation can be tricky, and the exploitation of fNIRS in real-time can be challenging [76].

The exploitation of the relationship between physiological signals and MW is not straightforward. Typical signal processing methods such as normalization [54], feature extraction and classification via ML algorithms (e.g., [3, 73, 74, 87]) are required in order to tackle the inter-individual variability [79] and the non-linearity between physiological signals and MW [92]. Regarding the classification methods, the RF model has shown great results in the physiological computing field [18, 26]. Mostly based on a supervised approach, these ML methods require the collection of labeled data under specific experimental conditions that VR can offer.

In summary, physiological signals can offer a suitable solution to evaluate MW, but their exploitation requires complex solutions involving ML and labeled datasets.

2.2 Mental Workload Recognition in VR

VR implies different constraints than those in the real-world. Cybersickness is still a major issue in VR, and the literature shows that it influences users' physiological responses when using an HMD [23, 65]. Other findings support that breaks in presence can also influence physiological signals [70], and that presence can modulate the intensity of physiological responses [49]. Users also behave differently in VR compared to in the real world due to the interactions and the cumbersomeness of VR equipment, which can impact measures linked to their psychological states. For these reasons, assessing MW can differ in VR compared to in the real world.

Recently, a number of works have explored the recognition of MW in VR using EEG [81], fNIRS [60], EDA [18] or cardiac signals [18]. They labeled their physiological features based on task difficulty levels [60, 81] which were sometimes inferred using users' overall performances [18], and used different protocols of induction to elicit several levels of mental workload in VR. Most focused on a standardized single task. For instance, the n-back task is a cognitive method, where the user is asked to react (e.g., by pressing a button, or by interacting with virtual objects [60, 81]) if the presented stimulus is the same as the n-th previous one. It relies on the working memory paradigm and has shown to induce different levels of mental workload depending on its level of difficulty (as n increases, the difficulty increases) [4]. In another study, Collins et al. used a spatial rotation task to classify 3 levels of difficulty based on the participants' overall performances [18]. Their task consisted in rotating an hypercube (a four-dimensional ”cube”) to match the rotation of a static hypercube. They manipulated the difficulty in function of the combinations of 4D rotational planes rotated and the extent of the rotation.

However, there was no VR study which tried to classify MW levels using the users' subjective response to label the data and which elicited mental workload in a multitask context.

2.3 Physiological Sensors Integrated in VR HMDs

From a practical perspective, the use of physiological monitoring in VR poses several challenges. Physiological sensors are cumbersome therefore particularly troublesome when using VR HMDs. As users cannot see their real bodies nor the physiological sensors, their motion could be disrupted by them (e.g., cables). This can disturb the sensors functioning, generating artifacts in the recordings, and reduce users's immersion and engagement in the VEs. In addition, each physiological sensor often has to be installed and calibrated individually which increases the setup time and between-individual differences on the sensors placement.

In fact, physiological signal quality and shape greatly depends on the sensor's positions [91]. As a consequence, training classification models using a sensor located at a specific place will not necessarily work with the same sensor at another place. Fixing sensors directly on the headset allows to help maintaining them at the same place overtime, between sessions and between users. This is a great advantage to improve the signal quality and when building a dataset based on physiological signals for classification purposes. Moreover, the literature shows that the face location can be relevant to assess some physiological data related to users' psychological states [7, 91].

For all these reasons, some companies and laboratories focused on the integration of physiological sensors directly into the VR headsets. As such, some commercialized headsets directly integrates eye-tracking and gaze-tracking technologies, such as the Vive Pro Eye or the FOVE headsets [39]. Pupil Labs also offers add-on solutions for virtual and augmented reality headsets. Other headsets were developed to integrate EEG, such as LooxidVR [1] and Neurable [36] HMDs. Finally, the MIT Fluid Interface laboratory developed HMDs which integrates EDA and PPG sensors [6], as well as EEG, electromyography, and electrooculography sensors [7], with applications targeting mainly facial and emotion recognition.

However, none of these HMDs compared their sensors to CGS sensors in the same study, especially in the context of the real-time recognition of users' MW in VR.

3 Our "all-in-one" Approach to Assess User's Mental Workload in VR in Real-Time

In this section, we present our all-in-one solution to assess MW in VR in real-time. First, the hardware components concerning the sensor integration in the HMD will be addressed. Next, the software components presenting the solution proposed for the data synchronization and the real-time recognition pipeline will be addressed.

3.1 Integrated Hardware

Several physiological dimensions relevant to MW measurement were mentioned in Section 2.1. We chose to focus on cardiac activity, EDA, and oculomotor activity as those have proved to be influenced by MW [40], are non-intrusive, and can easily be positioned inside and on VR HMDs. The main efforts focused on the integration of cardiac and EDA into the VR HMD.

The cardiac activity was monitored via a PPG sensor: the Maxim MAX30102, which was fixed on a small clip to assess data on one of the user's earlobes (see Fig. 1). This location was chosen following the recommendation of the literature for PPG sensors, as blood vessels are close to the surface of the skin and light can readily be detected [77, 91].

For the EDA, pairs of electrodes were made out of a flexible printed circuit board. The latter was chosen based on previous works (e.g., [7]) and for its various qualities. It is robust enough to weld electrical wires on it to make the connection to the electronic card. As for the conductive material, gold was chosen as it is stainless and biocompatible. The prefrontal area has been found to be relevant in order to measure EDA [86, 91], so we chose to place one pair of electrodes on the foam in contact with the forehead part of the headset (see Fig. 1). The dimensions of electrodes were chosen to be thin (i.e., 100 µm) to not mark the skin, and large to palliate the reduced presence of sweat glands in the prefrontal area. Those were spaced a few inches apart to let a weak current flow through the skin, and the EDA was given as the difference in potential between the 2 electrodes.

All these sensors were plug in a custom-made electronic card (see Fig. 1), based on SOM Variscite i.MX8M Mini, which was powered by a 5V powerbank. The later were placed in a designed 3D printed case, which was positioned in the back of the HMD using the vertical strap of the headset (see Fig. 1). The electronic card main features are that (i) it can collect physiological data from multiple sensors, (ii) it allows the aggregation and time stamping of all samples, (iii) it has the capacity to process AI models and algorithms (not currently considered), and (iv) it can transmit the data to a computer either by a wired medium, using ethernet, or in a wireless way, via wifi connection. Further slots were available on the electronic card to plug in more sensors if necessary.

As for the ocular activity assessment, Section 2.3 presents a few HMDs which already integrate eye-tracking solutions and can be used as a base to input the remaining sensors (i.e., EDA and PPG here). We chose to use the Vive Pro Eye HMD.

3.2 Data Processing and MW Assessment

The software component aims at recognizing users' MW level in real-time using multiple sources of measurements. It is composed of several steps depicted in the Figure 2 which can be divided in 2 parts: the training phase and the real-time phase.

3.2.1 Data collection

The recording of data coming from various sensors could be tricky especially due to problems such as time synchronization, or data format. Moreover, sensors usually come with their own software, and dealing with all of them to record participants' signals can be tedious. For these reasons, a middleware, called LibSTR, has been developed for the collection of physiological and behavioral data. It allows to collect, synchronize and distribute in real-time the data coming from various sensors by abstracting the capture of data. It is composed of three components: hub, sensors wrappers and listeners. The hub collects and synchronizes the data from the sensors wrappers, and exposes them to the listeners. Developed in C and without software dependency, it works on multiple platforms.

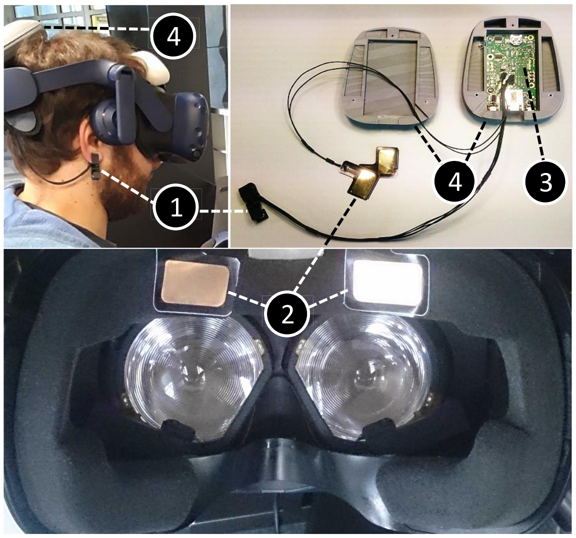

Figure 1: Hardware solution. The sensors are placed on a Vive Pro Eye HMD, which has eye-tracking. (1) PPG sensor, (2), electrodes to assess the EDA, (3) electronic card, (4) 3D printed case.

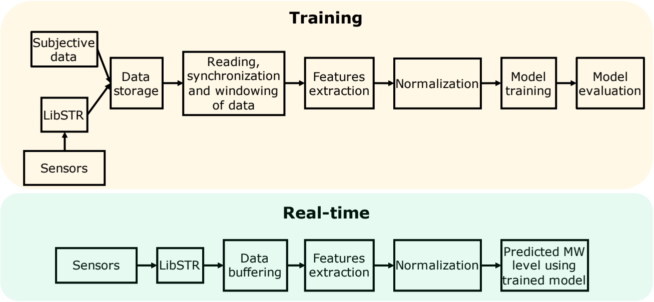

Figure 2: Software processing chain of the real-time recognition of users' MW. There are 2 main steps: the training of the recognition model and the real-time use of the model.

3.2.2 Data windowing

The training of ML models in a supervised way requires labeled data [54] (i.e., the labels correspond to the subjective measures of the user's state and the baseline in our case). For this purpose, a fixed-size window of the collected data (e.g., blood volume pulse (BVP), performance) are extracted before each label. Based on previous studies [73], a window size of 30 seconds was selected as it seems an appropriate compromise between performance and real-time use. Thus, for each label, the 30 seconds of data preceding the label timestamp are used to calculate the said-label dataset features [29] (see Fig. 3).

3.2.3 Features extraction

The exploitation of physiological signals requires the extraction of specific physiological features [2]. Based on the window of signals described above, common features from the relevant state of the art

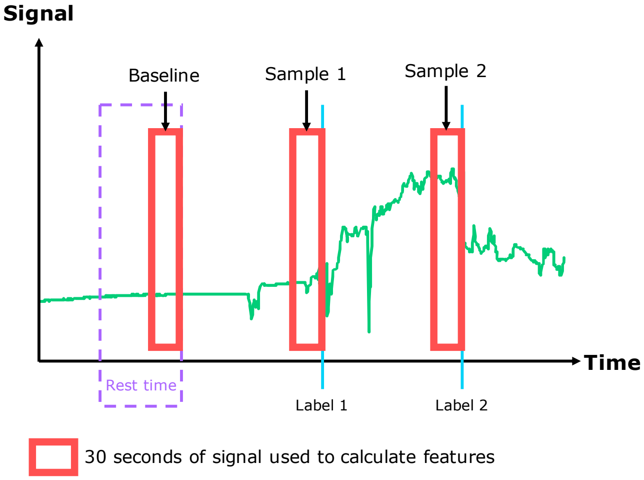

Figure 3: Illustration of a data windowing in the case of a physiological signal (in green). The time window (in red) is used to extract the physiological features based on given labels.

were used. It should be noted that performance measures depend on the context of the application and of the tasks users have to perform. A summary of all the features extracted using our setup can be found in Table. 1. The windowing of data as well as the extraction of features was implemented in Python using numpy [56], pandas [57] and scipy [89].

Cardiac activity. 6 time-domain and 5 frequency-domain features related to the Heart Rate Variability (HRV) were extracted [41]. These features are based on the InterBeat Interval (IBI). For this purpose, a bandpass filter was firstly applied (cutoff frequency $=$ $[0.66;3.33]$ Hz, order $=$ 3). It allows to reduce noise such as user's motions [2] and to assess the cardiac activity situated between 40 and 200 beats per minute. Then, the peaks on the BVP signal are detected using a threshold (arbitrarily set) and an estimation of local minima/maxima [80], allowing to estimate the IBI and to calculate the related features (see Table 1).

Electrodermal activity. The EDA signal is composed of 2 components: the phasic part and the tonic part [8]. The phasic part (also called Skin Conductance Level - SCL) corresponds to slow changes in the EDA while the tonic part (also called Skin Conductance Responses - SCR) corresponds to the rapid physiological responses to a stimulus. The extraction of those two components from the raw signal is composed of several steps. First, a low-pass filter (cutoff frequency $=$ 1Hz, order $=$ 3) is applied to reduce noise in the signal [8]. Second, a low-pass filter (cutoff frequency $=$ 0.05Hz, order $=$ 3) is applied on previously filtered signals to extract the tonic part of the EDA [9]. Lastly, the phasic part is obtained by subtracting the tonic signal to the filtered signal. For the phasic part, 9 features were extracted based on the estimated EDA peaks[1]. For the tonic part, 29 features were calculated (see Table 1). Inspired by research from other domains, some are computed based on the shape of the signal [44], others from the data-driven signal decomposition (i.e., Empirical Mode Decomposition - EMD) [34] and finally some are EDA components from the frequency-domain [67, 69].

Ocular activity. 18 features have been extracted from the pupil diameter, the dynamic of pupil diameter, and the dilatation and constriction of the pupil (see Table 1). The signal was cleaned by taking into account the data only when the pupil was detected.

Table 1: Extracted physiological and performance features

Features on cardiac activity
Heart Rate
Average of NN intervals
Standard deviation of NN intervals (SDNN)
Root mean square of successive differences between NN intervals (RMSQ)
Number of interval differences of successive NN intervals greater than 50 ms
Percentage of interval differences of successive NN intervals greater than 50 ms
Very low frequency (0.003 to 0.004 Hz)
Low frequency (0.04 to 0.15 Hz)
High frequency (0.15 to 0.4 Hz)
Ratio of low frequency and high frequency
Total spectral power
Features on tonic EDA
Max, range, inter-quartile range, root mean square error, mean, SD, skewness and kurtosis of the signal
Mean absolute value of 1st differences and mean absolute value of 2nd differences of the signal
Mean absolute value of the 1st differences and mean absolute value of the 2nd differences of the standardized signal
Mean, SD, min and max of 3 Intrinsic Mode Functions
Very low frequency (0 to 0.1 Hz)
Low frequency (0.1 to 0.2 Hz)
Middle frequency (0.2 to 0.3 Hz)
High frequency (0.3 to 0.4 Hz)
Very high frequency (0.4 to 0.5 Hz)
Features on phasic EDA
Number of peaks
Mean, SD, min and max of peak amplitude
Mean, SD, min and max of half of recovery time of peaks
Features on ocular activity
Min, max, range, mean and SD on pupil diameter
Min, max, range, mean and SD of the pupil amplitude
Min, max, range, mean and SD of the pupil constriction and dilatation speed
Features on task performance
Min, max, range, mean and SD on distance to the center of the circle

3.2.4 Normalization of physiological features

Considering the inter-individual variability is a key point in research dealing with physiological data [5, 29], several methods have been proposed in the literature to reduce its influence on the recognition accuracy and to make data comparable between participants. One of the most common approaches is to collect data during a rest time (i.e., the baseline) and to subtract the mean value of the data collected during this rest time from the whole signal [9]. This approach can be effective on non-periodic signals such as the EDA signal, but is not compatible with periodic signals (e.g., BVP [2]). As such, the methods of normalization at feature level seem interesting (e.g., subtraction of feature values during rest time from other feature values) as they can be applied to all signals. However, as no *de facto* normalization method exists, the most common approaches used in the literature will be evaluated.

3.2.5 Model training using Machine Learning

Models are trained, using supervised ML algorithms, to recognize MW based on the extracted physiological features, task performance measures, and related subjective responses (i.e., the subjective measures of users' state). In this way, the function between the input data (e.g., extracted physiological and performance features) and

[1] An EDA peak is characterized by the amplitude (the height of the peak) and the recovery time (time to return to the level of EDA before the peak) [8, 72].

[2] Contrarily to the EDA signal, the BVP signal is periodic. Therefore, subtracting the mean value of the BVP signal collected during the baseline to the signal will only bring the BVP signal to the same amplitude level (roughly the same mean). However, the features related to the BVP signal are time-based, which makes such normalization not relevant.

output data (i.e., subjective responses) is automatically inferred [54]. The objective is to be able to detect users' subjective MW level using only objective measures (i.e., physiological responses and/or task performance) without requesting any evaluation from the users, which could disrupt their experience [61]. RF [10] was selected as it presented the best performance in similar contexts (e.g., [18, 26]). The number of trees as well as the number of randomly selected predictors at each cut in the tree were tuned during the training. As the evaluation of feature selection (i.e., principal component analysis) showed no improvement of the recognition accuracy and as the tree-based models are generally robust against unhelpful features, all the features per sensor were considered. All trainings were realized using R [63] and the caret library [42].

3.2.6 Real-time recognition

The real-time pipeline adopts a similar processing chain as the training part. However, there are some adaptations.

Data buffering. As the data is progressively captured, it is necessary to store it in a buffer. Indeed, 30 seconds windows of data are required to calculate the features. The buffer starts empty and is progressively filled with available data. When the 30 seconds of data are reached, the data are fed to the rest of the processing chain. Then, whenever new data is available, it is added to the buffer by pushing the oldest data at a 1 second step.

Normalization. The recording of data during a baseline is required to normalize the calculated features, regardless of the method used. The normalization should be done in the same way as in the training part.

Real-time prediction. The predicted MW level and related physiological signals are displayed in an interface (see Fig. 4). Wrappers were written to communicate the estimated MW level output from Python to other environments, such as C# for Unity3D VR environments.

In order to have a unified processing chain, the best ML configuration was implemented in Python using scikit-learn [58] for real-time purpose.

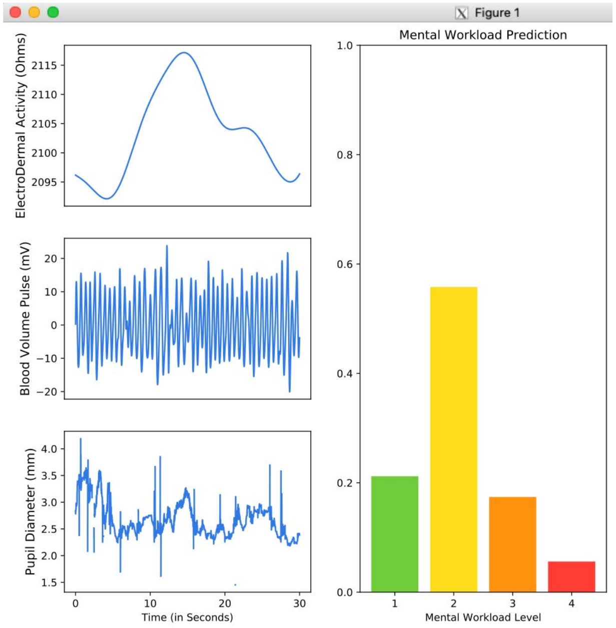

Figure 4: Interface for mental workload recognition in real-time. It depicts on the left: the EDA, BVP, and pupil diameter signals overtime; on the right: the predicted mental workload level (i.e., the probability that a user is at a particular mental workload level).

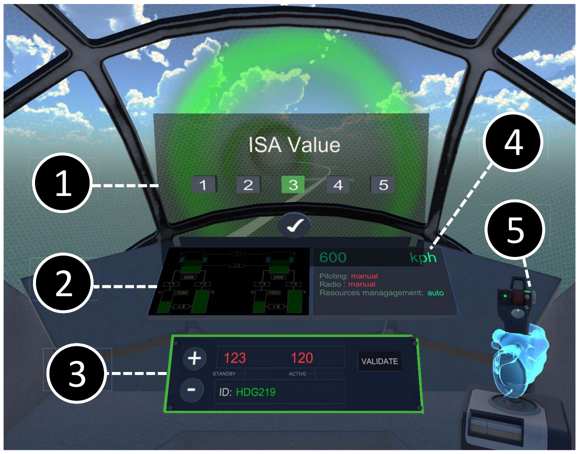

Figure 5: Virtual cockpit view. (1) Instantaneous Self Assessment (ISA) interface. (2) Resources management task interface (deactivated); when activated, the interface lit up (with a green outline). (3) Communication task interface (activated); when deactivated, the interface lit off (no green outline). (4) Informative panel which gives information about which task is activated or not at the current time. (5) Virtual representation of the joystick used to pilot the aircraft and of the right hand. The left hand is represented in the same way, but tracked by a Vive Controller and animated depending on the interaction.

4 Data Acquisition Protocol

A user study was conducted to test the viability of our solution to classify MW levels. The objective was to create a dataset to train the recognition model. Three major constraints were enforced. First, a real-life task which required the user to perform different sub-tasks. Second, the possibility to modulate the task difficulty ensuring that a wide range of mental workload levels could be induced. Third, enforce a good balance between the different levels of mental workload induced in order to ensure an optimal classification model. In addition, we compare our solution to CGS sensors, in regards to the MW classification performances.

4.1 Tasks Design

The Multi-Attribute Task Battery II (MATB-II) was originally developed to study human performances in a multi-task context [68]. It has been used extensively to study mental workload and to train users to situations where they might be overloaded. Three tasks of the MATB-II [68] were adapted in VR to induce different levels of subjective MW: the tracking task, the communication task, and the resources management task. The VR cockpit is depicted in Fig. 5.

The **tracking task** of the MATB-II was adapted into a piloting task in VR. Users could orientate the aircraft using a joystick, but they could not accelerate nor decelerate. They were asked to follow the green line which went through all circles centers as closely as possible (see Fig. 5). Three different difficulty levels were considered: easy, medium, and hard. Those were manipulated by modulating the speed of the aircraft, and the number of circles users could see at a time. For the **communication task**, users could hear a voice in the headset asking a specific aircraft to turn the radio on a given frequency in french. Users had to pay attention to determine if the message targeted their aircraft or not and could click on the "+" and "−" buttons in the VR cockpit to change the radio frequency (see Fig. 5). Two difficulty levels were considered here: activated (audio messages) or deactivated (no audio message). Finally, for the **resources management task**, users were asked to maintain 2 tanks levels in the blue zone (see Fig. 6) by activating or deactivating 8 different pumps buttons. There were 2 difficulty levels: activated

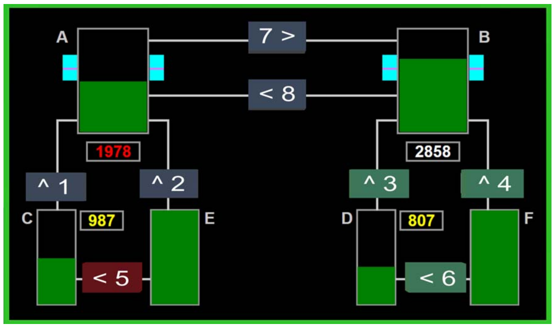

Figure 6: Resources management task interface. A and B are the main tanks; their fuel levels are indicated below the tanks. C and D are supply tanks; their fuel levels are indicated on their right side. E and F are supply tanks with unlimited capacities. The buttons numbered from 1 to 8 are pumps button. Pumps 3, 4 and 6 are activated, pump 5 is failed, and all other grey pump buttons are deactivated.

or deactivated. When the task was activated, the two tanks levels started to decrease, otherwise, the resources management task interface was unlit (see Fig. 5) and frozen. More details about the tasks instructions can be found in [45] and [68].

All these tasks levels could be associated to form 12 tasks levels associations (TLAs) (3 piloting levels × 2 communication levels × 2 resources management levels). For the sake of clarity, each TLA was labeled using 3 digits, one for each task. The first, second, and third digits represent respectively the piloting task difficulty level (0-easy, 1-medium, 2-hard), the radio task level (0-deactivated, 1-activated), and the resources management task level (0-deactivated, 1-activated). For instance, in the TLA ”110”, the difficulty of the piloting task is set to ”1-medium”, the radio task is ”1-activated”, and the resources management task is ”0-deactivated”. Each TLA was set to last 107s on average.

4.2 Apparatus

The participants were installed on a cockpit, which was assembled for the experiment (see Fig. 7). The virtual cockpit was modeled in 3D and calibrated so it matched the real one in position and size (see Fig. 5).

Users were equipped with the customized Vive Pro Eye (see Section. 3.1) and the Shimmer3 GSR+ [13] sensors (i.e., the CGS sensors) (see Fig. 7). The Shimmer wristband was disposed on users' left wrist. Its EDA sensors were placed on the middle phalanx of the users' left ring and middle fingers, and the PPG sensor ear clip, on users' left earlobe. The data were collected using our libSTR middleware (see Section 3.2.1) with homogeneous timestamp formats at different frequencies (Vive Pro Eye-tracker: 250Hz, HMD-PPG:100Hz, HMD-EDA: 50Hz, Shimmer sensors: 60Hz).

Users could pilot the virtual aircraft using the Logitech 3M X52 Pro joystick with their right hand. They were also equipped with a Vive Controller with their left hand to interact with the virtual interfaces. In the VE, both hands were represented by transparent virtual hands. The right hand was placed on the virtual joystick (see Fig. 5), which moved when the user was interacting with the real one, and the left hand was tracked with the Vive Controller. The users did not need to use any of the buttons on the Vive Controller or on the joystick to perform the tasks. The interactive virtual objects were highlighted when the participants advanced their virtual hands in their direction, and the animation turned into a pointing index upon approaching. A haptic pulse feedback on the Vive controller informed the participants that their action had been carried out. Audio instructions were provided using the audio headset supplied with the Vive Pro Eye Headset during the flight simulation.

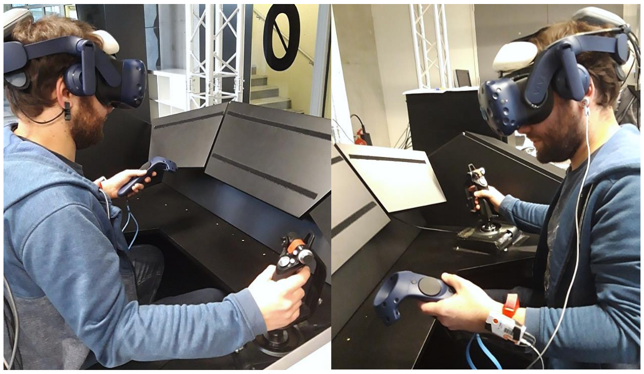

Figure 7: Multiple views of the experimental setup. The user is wearing a Vive Pro Eye equipped with sensors on the headset, a Shimmer3 GSR+ on his left hand, and using a joystick and a Vive Controller on the cockpit.

The support application was developed in Unity 3D, and run with the recording on a computer equipped with an Intel(R) Core(TM) i9-7900X CPU, a Nvidia Titan V graphic card, and 16 GB Random-Access Memory.

4.3 Collected Data

The assessed data were: self-report (i.e., the subjective MW reported by participants), physiological, and tasks performance measures.

4.3.1 Self-Report

We wanted to assess users' MW while they were performing other tasks. Therefore, the focus was set on the Instantaneous Self-Assessment (ISA) [78], which rates the MW level using 5 different ratings (1-underutilized, 2-relaxed, 3-comfortable, 4-high, 5-excessive) and has been especially used during air traffic control tasks [78]. The meaning of each rating of the ISA is described in [78] and was explained to each participant before the experiment. They were asked to report the MW level they experienced during the last 30s, when a screen appeared in front of them with the 5 buttons corresponding to the ISA ratings (see Fig. 5). First, they were asked to push the button corresponding to their MW level, and then, to click on the validate button on the same screen to make it disappear.

4.3.2 Physiological Measures

Multiple types of physiological signals were assessed during the experiment: ocular activity via the eye-tracking cameras present in the HMD, and cardiac activity and EDA via our hardware solution integrating sensors into the HMD (see Section 3.1) and via the Shimmer3 GSR+ (i.e., the CGS sensors). The features extracted from these sensors are depicted in Table 1.

4.3.3 Task Performance Measures

Some performance indicators of all 3 tasks were assessed throughout the experiment. However, among the three tasks used in the experiment, only one was always present across the different TLAs: the piloting task. Thus, only this measure is considered as a performance measure.

Users were given indications on how to align the aircraft with the green line optimally before the experiment, and the distance to each circle centre when they passed it was recorded throughout the experiment.

4.4 Experimental Design

In a previous study [45], 38 participants did the 12 TLAs in a row in a randomized order and reported their subjective MW level using the ISA scale [78], 3 times per TLA. Since unbalanced datasets can lead to poor recognition performance for minority classes [25], the experimental protocol was designed to induce the highest number of different subjective MW levels in the most balanced way possible.

The constraints were the following: a TLA could not appear more than 3 times in total, and the total duration of the experiment was set not to last more than 25 min, conducting to a number of 10 TLAs. Following these constraints and based on the ISA responses reported during the first study [45], the best subjective MW level distribution was given for the following 10 TLAs: "000 – 000 – 000 – 100 – 111 – 111 – 201 – 211 – 211 – 211". Those were selected for the new experimental study, which results and analysis are depicted and discussed in this paper.

4.5 Participants

77 healthy participants, who were completely naive to the experiment, were recruited through an external cabinet. They were paid 30€ for their participation to the study. Two of the users were excluded from the study due to motion sickness, resulting in a final sample of 75 participants with ages ranging from 18 to 64 (40 females, 38 males; M = 38.69, SD = 13.54).

There were some inclusion criteria: the participants had to be fluent in french. They should not have taken any medication that could influence their physiological responses. They were also asked not to consume coffee and/or tea in the 2 hours preceding the experiment. Moreover, variables such as their experience in VR, games, flight simulator, aircraft piloting, vision state, and dominant hand were controlled. One participant reported having a great experience in VR, 4 reported having few experience in VR, and all others (i.e., 70 participants), none. As for the gaming experience, 64% of the participants were novice, 20% played occasionally, and 16% regularly. All of them reported having no experience in flight simulator and aircraft piloting.

In accordance with ethical principles, participants were required to complete an informed consent form, advising them of their right to withdraw at any time from the study, of the preservation of their anonymity and about the potential side effects of VR.

4.6 Experimental Procedure

The experiment lasted around 1h and was subdivided into the following steps:

Written Consent and Instructions: Users completed a consent form, prior to the experiment. They were then instructed with the nature of the experiment, the equipment used, the data recorded (which was anonymized), and the tasks instructions. Participants were also asked to fill a questionnaire (experience with VR, video games, and piloting an aircraft, dominant hand, level of alertness, state of vision, demographic information, simulator sickness questionnaire (SSQ) [37]) to gather information about their background and their state before the start of the experiment.

Training: Users were then equipped with the Shimmer sensors, a Vive Controller and the Vive Pro Eye HMD with the sensors. The eye-tracker of the Headset was first calibrated following the instructions given in the headset. They were then immersed in the virtual cockpit environment. Once they got used to the VE, they were asked to breath normally and to remain still for 1 minute, to record their physiological signals in a neutral state (i.e., for the physiological baseline). Users were then asked to interact with the buttons of the tasks interfaces to familiarized themselves with the interactions. Then, they travelled the TLAs following this path: "000 – 010 – 011 – 001 – 101 – 201 – 211", which gave them a good overview of each task and their levels. Users were then invited to ask any question they may have had.

Experiment: In the experiment part, users were first asked to do a 1-min baseline again. Then, they travelled the 10 TLAs mentioned in Section. 4.4 (i.e., "000 – 000 – 000 – 100 – 111 – 111 – 201 – 211 – 211 – 211") in a pseudo-randomized order (two identical TLAs could not appear twice in a row). The TLAs were set to last 110 seconds with 4 communication calls and 3 ISA calls (spaced in time of 30s).

Debriefing: At the end of the experiment, they were asked to fill the SSQ again, debriefed and invited to ask questions.

4.7 Resulting Data

3265 subjective responses were collected. As unbalanced datasets can lead to poor recognition performance for minority classes [25], we chose to retain 4 classes of subjective MW level, using the following data distribution (see Table. 2):

Table 2: Contingency table of subjective responses.

MW level	1	2	3	4
ISA value	1	2	3	4 & 5
N	624	822	1143	676
Perc.	19 %	25 %	35 %	21 %

The results for other data splittings (2 classes, 3 classes, and 5 classes) based on the reported ISA levels are presented in supplementary materials.

5 Results

This section presents an overview of the classification accuracy of the proposed system, considering the 4 levels of MW chosen (see Table 2). Table 3 presents the classification results using the EDA and PPG from the CGS sensors, and Table 4 presents the classification results using the EDA and PPG from our custom sensors integrated into the VR HMD (see Section 3.1). Ocular and task performance measures are the same in these 2 groups. The normalization method, as well as the different combinations of sensors using either the Shimmer sensors (i.e., CGS sensors) or our setup (i.e., HMD sensors) are provided. In order to facilitate the interpretation of results, the performance of a naive model was calculated (i.e., model always predicting the most represented class in the training dataset).

Classification accuracy was computed using a 10-fold cross-validation method [54]. To reduce the potential problems linked to random splitting with cross-validation, 10 independent 10-fold cross-validations were performed for each trained model.

5.1 Normalization of Physiological Features

As previously explained, the physiological signals are subject to inter-individual variability, which can affect recognition performance. Thus, three types of normalization approaches have been evaluated in the context of MW detection:

- **Normalization by subtraction of features (NSF)**: For each participant, subtracting the features estimated for each subjective response with the features estimated during the baseline [17, 38, 75];

- **Normalization by adding features (NAF)**: For each participant, adding the features estimated during the baseline to the dataset (i.e., the feature space) [59, 84].

- **Normalization by Min-Max**: For each participant, each feature is normalized using the following formula:

$$\frac{X - min}{max - min}$$

The max corresponds to the feature value during the first trial with the highest ISA score and the min corresponds to the feature value during the first trial with the lowest ISA score.

5.2 CGS Sensors

The recognition performances based on data collected with the Shimmer sensors are presented in Table 3. According to the results of the training, the best accuracy using the CGS sensors is 64.2%. It is obtained with a combination of performance and all physiological signals, using either NSF or NAF normalization.

Taking sensors and task performance measure individually, the best classification performance is achieved either with the ocular activity (in the cases of no-normalization and NAF normalization) or the EDA (in the cases of NSF and min-max normalizations), followed by the cardiac activity, and the task performance measures.

There is a mean drop in accuracy of only 1.33% when considering only the EDA and ocular signals instead of all physiological signals and performance measure, using any normalization method. A confusion matrix using all data and the NSF normalization is given in Fig. 8.

5.3 Sensors Integrated Into the HMD

The recognition performances based on data collected with the sensors integrated into the HMD are presented in Table 4. Similarly to results with CGS sensors, the best accuracy is 65%. It was obtained when all data are considered, and with NSF normalization.

Taking sensors and task performance measure individually, the ocular activity lead to the best classification performance, followed by the EDA, the cardiac activity, then the task performance measure.

Similarly to results with CGS sensors, the combination of EDA and ocular signals lead to a mean drop of only 2.1% compared to the best configuration requiring more sensors.

A confusion matrix using all data and the NSF normalization is given in Fig. 9.

5.4 Mental Workload Recognition in Real-Time

The recognition in real-time was simulated using the collected data and the configuration offering the best performance (i.e., input: all physiological sensors and performance, NSF normalization, HMD sensors; see Tables 3 & 4).

To simulate real-time, a sliding window of 30s with 1s step was applied to the physiological data. Then, these windows of data were fed every second to the recognition processing chain until the last available data (see Fig. 1).

For illustrative purposes, the real-time was simulated on one participant (the stages of the processing chain are identical between participants). The predicted MW as well as the ground truth (i.e., true MW level, see Table 2) are depicted in Fig. 10. As subjective responses are only provided punctually, the MW level was linearly interpolated between the different subjective mental workload responses in order to have a continuous ground truth. A moving average (on 10s) was also used to smooth the prediction and limit the effect of misclassified MW level. The prediction latency did not exceed 0.20s.

6 Discussion

In this paper, the recognition of MW using physiological signals and task performance data was explored. Based on data from 75 participants, the trained models were able to classify 4 levels of MW with an accuracy up to 65%. Yet, it should be noted that the misclassified levels are mostly contained in the adjacent classes, as shown in Figures 8 and 9. As such, the classification accuracy reaches 95.5% using NSF normalization when considering the classes which are adjacent to the predicted mental workload level (see Fig. 9). This highlights the good performance of our approach. Moreover, the recognition accuracy considering CGS sensors and sensors integrated in the HMD were compared. The normalization method, as well as each sensors and combination of sensors were also tested in regards to the classification accuracy.

Unlike previous research exploring the recognition of MW in VR [18, 60, 81], the data were labeled using subjective responses. This approach is novel in VR, as previous work labeled their physiological data based on the task difficulty levels [60, 81] or based on task performances [18]. While it is true task difficulty has shown to be correlated to users' mental workload [4], it does not take into account the users' subjective impressions. The same task can induce different MW levels to users because of individual differences [83]. In the

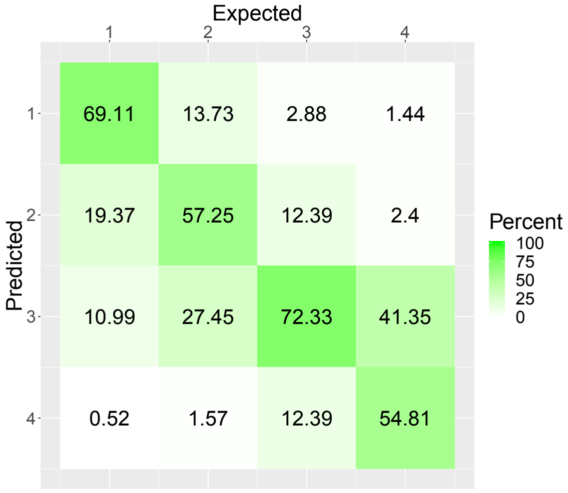

Figure 8: Confusion matrix for "CGS sensors" (Expected and predicted classes) in the "All physio+perf" using NSF normalization setup.

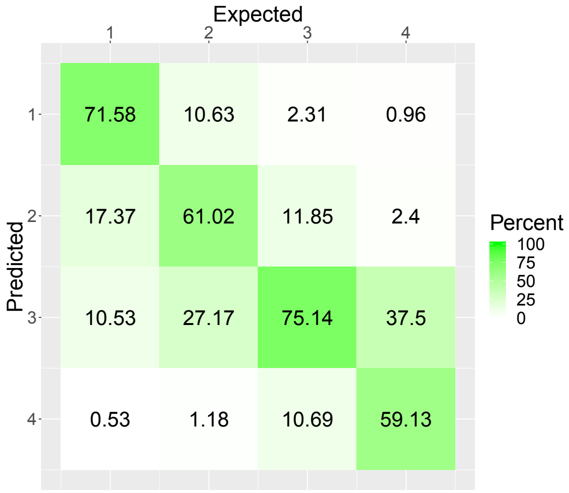

Figure 9: Confusion matrix for "HMD sensors" (Expected and predicted classes) in the "All physio+perf" using NSF normalization setup.

Table 3: Mental workload classification accuracy results (in %) using the Shimmer sensors (i.e., CGS sensors) in function of the normalization methods and of the type of measure. “Perf” corresponds to the task performance. Task performances and ocular activity are common to the CGS and HMD sensors.

Normalization	**Input**								
	Perf	Ocular	**Shimmer Sensors (CGS)**						
			Cardiac	EDA	Cardiac + EDA	Cardiac + Ocular	EDA + Ocular	All physio	Physio + Perf
Naive model	*19.1*								
No Normalization	41.4	44.4	36.7	41.1	44.8	50.5	49.9	52.5	57.6
NSF	41.5	52.3	42.9	54.2	56.4	57.3	62.2	62.5	**64.2**
NAF	45.8	60.5	50.8	56.9	56.3	61.1	62.1	62.3	**64.2**
Min-Max	45.3	53.6	45.8	56.6	58.2	57.8	63.3	62.5	63.2

Table 4: Mental workload classification accuracy results (in %) using the sensors integrated into the VR HMD (i.e., HMD sensors) in function of the normalization methods and of the type of measure. “Perf” corresponds to the task performance. Task performances and ocular activity are common to the CGS and HMD sensors.

Normalization	**Input**								
	Perf	Ocular	**HMD Sensors**						
			Cardiac	EDA	Cardiac + EDA	Cardiac + Ocular	EDA + Ocular	All physio	Physio + Perf
Naive model	*19.1*								
No Normalization	41.4	44.4	36.5	44.1	49.7	51.0	53.6	57.7	61.2
NSF	41.5	52.3	44.9	50.7	54.5	59.1	62.0	62.8	**65.0**
NAF	45.8	60.5	49.9	52.9	56.3	60.3	62.6	62.1	63.8
Min-Max	45.3	53.6	45.6	51.6	54.5	56.9	61.3	61.8	63.4

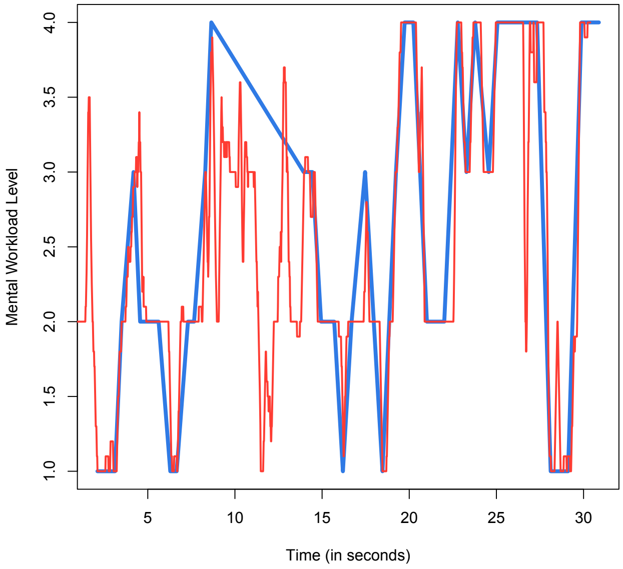

Figure 10: Simulation of the real-time prediction of mental workload using the data of one of the participants and the HMD sensors in the “All Physiosiological signals + task performance” setup with NSF normalization. The blue line represents the ground truth (i.e., the mental workload level reported by the participant); the red line represent the predicted mental workload using our trained recognition model.

same way, while performance measures have shown to be correlated to users’ mental workload in some studies [55], some users might show similar performances while experiencing different levels of mental workload [24, 82]. In addition, relying on task difficulty or task performance for the labeling could result in a recognition based on features linked to the task specificities, and not linked to the users’ psychological state. Labeling data based on users’ subjective MW levels via self-report responses provides an efficient way to capture the actual state of the user as they show greater face validity [54]. Furthermore, subjective measures are task agnostic, as opposed to task performance measures and task difficulty. This allows to use models trained using subjective measures in different contexts than those in which they were trained.

A special attention was given to the data acquisition protocol. It differs from previous work as it was done in a multitask context. Usual protocols, which gather data for classification purposes, focus on a single standardized task which stimulates limited pools of cognitive resources depending on the nature of the stimuli (e.g., visual, auditory) and information processing (e.g., perception, action) [90]. This has for consequences to influence physiological signals in a way that might prevent the generalization of the recognition models.Knowing MW is mostly studied in complex contexts where users have to perform multiple tasks in parallel, we chose to assess MW data in a VR flight simulator, with different tasks and stimuli natures.

Although the integration of sensors in HMDs was already explored in the literature [6, 7, 36], to the best of our knowledge, the recognition of MW based on ML using multiple physiological sensors integrated into a VR HMD was yet not investigated. Thus, we conducted a user study to compare CGS and HMD sensors for MW recognition. The current results showed overall similar recognition performances between CGS and integrated sensors in a HMD (see Table 3 & Table 4). The best classification performance was achieved with the HMD sensors when using NSF normalization (i.e., +0.8% compared to CGS sensors with NSF). Given the advantage brought by the integration of sensors in the HMD in terms of cumbersomeness, this encourages the use of this kind of setup to monitor users’ psychological state in VR.

Moreover, multiple types of data were gathered to classify MW level: cardiac, electrodermal, and ocular activities, as well as task performance measures. The contribution of each signal was explored. When taking sensors individually, ocular activity lead overall to the highest classification accuracy (except using CGS sensors and NSF or Min-Max normalization, where it goes second to EDA), followed by EDA, cardiac activity, and finally performance measures. Otherwise, the combination of sensors improved the recognition accuracy with maximum performance when all physiological sensors are included as well as performance data. Performance measures have been used extensively in the literature as an indicator of users’ MW level [55]. While it has the disadvantage to be task-dependent and not to be generic, its impact was expected to be major in the classification accuracy. Therefore, it is encouraging to observe that taken individually, physiological signals, which are less task-dependant, contributed more than performance features in the classification accuracy of MW levels. With a normalization, the mean gain of the ”All Physio + Perf” configuration compared to the ”All physio” configuration is low (i.e., +1.6%) (see Tables 3 & 4), which is why we would advise not to use performance measures to make the recognition model more generic to other contexts.

From a features perspective, 68 were calculated on the cardiac, EDA and ocular signals, ranging from conventional ones such as heart rate or mean pupil diameter to less explored ones such as EMD ones [34] (see Table 1). In particular, the introduction of frequency-domain EDA measurements features [44, 67, 69] appears promising. Additional analysis showed that they strongly contributed to the MW recognition according to the estimated features importance [3]. Four (CGS sensors) and two (integrated sensors) of the five EDA frequency features appeared among the 20 most important variables.

The exploitation of physiological signals commonly lead to the well-know problem of inter-individual variability [79]. To minimize it, three normalization methods have been evaluated: subtraction of features calculated on baseline, adding features calculated on baseline, and min-max normalization. Overall, normalizing the data improved the classification accuracy compared to no normalization (i.e., with a mean gain of +8.2%). When taking sensors or performance individually, the best normalization methods was found to be NAF with a mean gain of +11.64% compared to the no-normalization case. As for the min-max method, the results are inconsistent between the different sensors associations. It has the disadvantage to necessitate to record the users' data when they are experiencing the lowest and highest mental workload possible beforehand to train the classification model, which makes it a bad normalization candidate for a real-time use. In the “Physio+Perf” measure configuration (i.e., the one which had the best results), the approach by subtraction was found to be the most effective, with a mean gain of +5.2% in recognition accuracy compared to the no-normalization case. In addition, it has the advantage to be compatible with real-time applications.

A preliminary study allowed to demonstrate the ability of our processing chain to classify MW in real-time through a simulation. It showed that the whole pipeline using multiple sensors is compatible with a real-time use, which is presented for the first time in the context of MW recognition in VR. This result paves the way to new HMDs with integrated sensors facilitating the real-time adaptation of VR environments based on detected MW levels.

7 Limitations and Future Work

Some limits can be pointed out. While efforts have been made to try to balance the 4 MW level classes, our model tends to predict more often the class 3 due to its over-representation compared to the other classes (see Table 2). The classification performance of subject-independent models (i.e., generalization of recognition on unseen participants) was not explored in this experiment. The current approaches (i.e., feature extraction followed by model training) seem not adapted to this rarely explored problem [88]. Recent advances in deep learning seem to provide a solution and offer unmatched performances in various fields [43]. Nevertheless, this promising approach requires in particular very large datasets [31]. Thus, even if the number of participants in the current paper is similar to comparable studies (e.g., [30, 73]), the number of labeled data is small compared to datasets in some other areas (e.g., ImageNet for object recognition [66]). Collecting such large physiological datasets is very complex, in particular due to the cost of data collecting. Moreover, some research pointed out the interest of other sensors such as EEG [81] or fNIRS [60]. This type of measure will also be valuable to qualify the nature of the detected state. In fact, only peripheral data were collected in the present study, which did not make it possible to ensure the distinction between an activity of the sympathetic system and the MW. However, as prior research showed an influence of MATB-II tasks on EEG signals (e.g., [15]), we can hypothesize that the measured changes in the peripheral data were related to the stimulus induction. Future work could focus on integrating this type of sensors in our setup and to compare their contribution in the MW recognition compared to other sensors.

Our methodology relies on subjective assessment, which makes it application agnostic when not taking into account task performance. A further direction would be to test if the recognition of MW can be shared between different applications (e.g., training the model using a standardized task, and using it in another application). The methodology was demonstrated in a situation were users were seated. Evaluating our method in a context requiring full-body motions to study the robustness of integrated HMD sensors compared to CGS setup to motion artefacts could be interesting. While this paper explores the normalization of physiological signals, further works could also focus on the normalization of self-reported scores, addressing users' scale perception. Another path would be to use our recognition model in a VR application, and to propose a VEs adaptation model based on users' MW.

8 Conclusion

This paper proposes an all-in-one approach to assess users' MW in VR in real-time, using physiological sensors directly embedded into the headset and the Random Forest algorithm. The hardware and software solutions employed to build the system are depicted, and a user study with 75 participants was conducted to train the system to recognize 4 MW levels using physiological and performance measures. Contrarily to previous work which focused on single standardized tasks to elicit MW, users performed different tasks on a VR flight simulator and reported their subjective MW level during the experiment, which was then used to label the dataset. Moreover, the contribution of different normalization procedures, and of different types of measure and sensors, considering our solution integrating sensors into the HMD and CGS, are compared in regards to the recognition accuracy. Results show similar recognition performances between the HMD sensors and the CGS sensors with an accuracy up to 65%. Because of its advantages in terms of cumbersomeness, this encourages the use of physiological sensors integrated into VR HMDs to monitor users' psychological state in VR. Normalizing the dataset features also greatly improved the classification performance. As for the type of measures, ocular activity features were found to be especially important, followed by EDA, cardiac activity, and task performance features. Preliminary results demonstrate the ability of our pipeline to recognize mental workload in real-time. Taken together, the results support that our all-in-one approach is promising for real-time MW recognition in VR.

Acknowledgments

This study was carried out within b<>com, an institute of research and technology dedicated to digital technologies. It received support from the Future Investments program of the French National Research Agency (grant no. ANR-07-A0-AIRT).

References

[1] Looxidvr. `https://looxidlabs.com/looxidvr/`. Accessed: 2020-03-02.

[2] M. Ali, A. H. Mosa, F. A. Machot, and K. Kyamakya. *Emotion Recognition Involving Physiological and Speech Signals: A Comprehensive Review*, vol. 109, pp. 287–302. Springer International Publishing, 2018. doi: 10.1007/978-3-319-58996-1_13

[3] A. Appriou, A. Cichocki, and F. Lotte. Towards Robust Neuroadaptive HCI: Exploring Modern Machine Learning Methods to Estimate Mental Workload From EEG Signals. In *Extended Abstracts of the*

[3] The importance of features was estimated based on the impurity decrease. It corresponds to the mean decrease in impurity averaged over all nodes where that feature was used to split the node [62].

2018 CHI Conference on Human Factors in Computing Systems - CHI '18, pp. 1–6. ACM Press, Montreal QC, Canada, 2018. doi: 10.1145/3170427.3188617

[4] J. Au, E. Sheehan, N. Tsai, G. J. Duncan, M. Buschkuehl, and S. M. Jaeggi. Improving fluid intelligence with training on working memory: a meta-analysis. *Psychonomic bulletin & review*, 22(2):366–377, 2015.

[5] M. Benedek and C. Kaernbach. A continuous measure of phasic electrodermal activity. *Journal of Neuroscience Methods*, 190(1):80–91, Jun 2010. doi: 10.1016/j.jneumeth.2010.04.028

[6] G. Bernal and P. Maes. Emotional beasts: Visually expressing emotions through avatars in VR. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pp. 2395–2402. ACM, 2017.

[7] G. Bernal, T. Yang, A. Jain, and P. Maes. PhysioHMD: a conformable, modular toolkit for collecting physiological data from head-mounted displays. In *Proceedings of the 2018 ACM International Symposium on Wearable Computers*, pp. 160–167, 2018.

[8] W. Boucsein. *Electrodermal Activity*. Springer-Verlag New York Inc., 2nd ed. ed., Sep 2011.

[9] J. J. Braithwaite, D. G. Watson, R. Jones, and M. Rowe. *A Guide for Analysing Electrodermal Activity (EDA) & Skin Conductance Responses (SCRs) for Psychological Experiments*. 2015.

[10] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, Oct 2001. doi: 10.1023/A:1010933404324

[11] K. A. Brookhuis and D. de Waard. Monitoring drivers' mental workload in driving simulators using physiological measures. *Accident; Analysis and Prevention*, 42(3):898–903, May 2010. doi: 10.1016/j.aap.2009.06.001

[12] A.-M. Brouwer, M. A. Hogervorst, J. B. Van Erp, T. Heffelaar, P. H. Zimmerman, and R. Oostenveld. Estimating workload using EEG spectral power and ERPs in the n-back task. *Journal of neural engineering*, 9(4):045008, 2012.

[13] A. Burns, B. R. Greene, M. J. McGrath, T. J. O'Shea, B. Kuris, S. M. Ayer, F. Stroiescu, and V. Cionca. Shimmer™–a wireless sensor platform for noninvasive biomedical research. *IEEE Sensors Journal*, 10(9):1527–1534, 2010.

[14] B. Cain. A review of the mental workload literature. Technical report, Defence Research And Development Toronto (Canada), 2007.

[15] S. Chandra, G. Sharma, K. Verma, A. Mittal, and D. Jha. EEG based cognitive workload classification during NASA MATB-II multitasking. *International Journal of Cognitive Research in Science, Engineering and Education*, 3(1), 2015.

[16] R. L. Charles and J. Nixon. Measuring mental workload using physiological measures: A systematic review. *Applied Ergonomics*, 74:221–232, Jan. 2019. doi: 10.1016/j.apergo.2018.08.028

[17] I. C. Christie and B. H. Friedman. Autonomic specificity of discrete emotion and dimensions of affective space: a multivariate approach. *International Journal of Psychophysiology*, 51(2):143–153, Jan 2004. doi: 10.1016/j.ijpsycho.2003.08.002

[18] J. Collins, H. Regenbrecht, T. Langlotz, Y. S. Can, C. Ersoy, and R. Butson. Measuring cognitive load and insight: A methodology exemplified in a virtual reality learning context. In *2019 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 351–362. IEEE, 2019.

[19] M. Coral. *Analyzing Cognitive Workload through Eye-Related Measurements: A Meta-Analysis*. PhD thesis, Wright State University, Dayton, Ohio, Jan. 2016.

[20] W. Damm, M. Franzle, A. Ludtke, J. W. Rieger, A. Trende, and A. Unni. Integrating Neurophysiological Sensors and Driver Models for Safe and Performant Automated Vehicle Control in Mixed Traffic. In *2019 IEEE Intelligent Vehicles Symposium (IV)*, pp. 82–89. IEEE, Paris, France, June 2019. doi: 10.1109/IVS.2019.8814188

[21] M. De Rivecourt, M. N. Kuperus, W. J. Post, and L. J. Mulder. Cardiovascular and eye activity measures as indices for momentary changes in mental effort during simulated flight. *Ergonomics*, 51(9):1295–1319, Sept. 2008. doi: 10.1080/00140130802120267

[22] D. De Waard. *The measurement of drivers' mental workload*. Groningen University, Traffic Research Center Netherlands, 1996.

[23] M. S. Dennison, A. Z. Wisti, and M. D'Zmura. Use of physiological signals to predict cybersickness. *Displays*, 44:42–52, 2016.

[24] A. Dey, A. Chatourn, and M. Billinghurst. Exploration of an EEG-based cognitively adaptive training system in virtual reality. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 220–226. IEEE, 2019.

[25] Q. Dong, S. Gong, and X. Zhu. Imbalanced Deep Learning by Minority Class Incremental Rectification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2018. doi: 10.1109/TPAMI.2018.2832629

[26] C. Elkin and V. Devabhaktuni. Comparative Analysis of Machine Learning Techniques in Assessing Cognitive Workload. In H. Ayaz, ed., *Advances in Neuroergonomics and Cognitive Engineering*, Advances in Intelligent Systems and Computing, pp. 185–195. Springer International Publishing, Cham, 2020. doi: 10.1007/978-3-030-20473-0_19

[27] M. R. Endsley and M. D. Rodgers. Distribution of attention, situation awareness, and workload in a passive air traffic control task: Implications for operational errors and automation. Technical report, FEDERAL AVIATION ADMINISTRATION WASHINGTON DC OFFICE OF AVIATION MEDICINE, 1997.

[28] M. Fallahi, M. Motamedzade, R. Heidarimoghadam, A. R. Soltanian, and S. Miyake. Effects of mental workload on physiological and subjective responses during traffic density monitoring: A field study. *Applied Ergonomics*, 52:95–103, Jan. 2016. doi: 10.1016/j.apergo.2015.07.009

[29] A. Fortin-Côté, C. Chamberland, M. Parent, S. Tremblay, N. Beaudoin-Gagnon, A. Campeau-Lecours, J. Bergeron-Boucher, and L. Lefebvre. Predicting video game players' fun from physiological and behavioural data. In K. Arai, S. Kapoor, and R. Bhatia, eds., *Advances in Information and Communication Networks*, Advances in Intelligent Systems and Computing, pp. 479–495. Springer International Publishing, 2018.

[30] L. Fridman, B. Reimer, B. Mehler, and W. T. Freeman. Cognitive Load Estimation in the Wild. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pp. 652:1–652:9. ACM, New York, NY, USA, 2018. event-place: Montreal QC, Canada. doi: 10.1145/3173574.3174226

[31] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. The MIT Press, Nov. 2016.

[32] T. Gruden, K. Stojmenova, J. Sodnik, and G. Jakus. Assessing Drivers' Physiological Responses Using Consumer Grade Devices. *Applied Sciences*, 9(24):5353, Dec. 2019. doi: 10.3390/app9245353

[33] S. G. Hart. Nasa-task load index (nasa-tlx); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting*, vol. 50, pp. 904–908. Sage Publications Sage CA: Los Angeles, CA, 2006.

[34] L. He, D. Jiang, L. Yang, E. Pei, P. Wu, and H. Sahli. Multimodal affective dimension prediction using deep bidirectional long short-term memory recurrent neural networks. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, AVEC '15, pp. 73–80. ACM, 2015. doi: 10.1145/2808196.2811641

[35] T. Heine, G. Lenis, P. Reichensperger, T. Beran, O. Doessel, and B. Deml. Electrocardiographic features for the measurement of drivers' mental workload. *Applied Ergonomics*, 61:31–43, May 2017. doi: 10.1016/j.apergo.2016.12.015

[36] J. Jantz, A. Molnar, and R. Alcaide. A brain-computer interface for extended reality interfaces. In *ACM SIGGRAPH 2017 VR Village*, pp. 1–2. 2017.

[37] R. S. Kennedy, N. E. Lane, K. S. Berbaum, and M. G. Lilienthal. Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness. *The international journal of aviation psychology*, 3(3):203–220, 1993.

[38] K. H. Kim, S. W. Bang, and S. R. Kim. Emotion recognition system using short-term monitoring of physiological signals. *Medical and Biological Engineering and Computing*, 42(3):419–427, May 2004. doi: 10.1007/BF02344719

[39] G. A. Koulieris, K. Akşit, M. Stengel, R. K. Mantiuk, K. Mania, and C. Richardt. Near-eye display and tracking technologies for virtual and augmented reality. In *Computer Graphics Forum*, vol. 38, pp. 493–519. Wiley Online Library, 2019.

[40] A. F. Kramer. Physiological metrics of mental workload: A review of recent progress. *Multiple-task performance*, pp. 279–328, 1991.

[41] S. D. Kreibig. Autonomic nervous system activity in emotion: a review. *Biological Psychology*, 84(3):394–421, Jul 2010. doi: 10.1016/j.biopsycho.2010.03.010

[42] M. Kuhn, J. Wing, S. Weston, A. Williams, C. Keefer, A. Engelhardt, T. Cooper, Z. Mayer, B. Kenkel, t. R. C. Team, M. Benesty, R. Lescarbeau, A. Ziem, L. Scrucca, Y. Tang, C. Candan, and T. Hunt. *caret: Classification and Regression Training*. 2018.

[43] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, May 2015.

[44] D. Lopez-Martinez and R. Picard. Multi-task neural networks for personalized pain recognition from physiological signals. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pp. 181–184. IEEE, Oct 2017. doi: 10.1109/ACIIW.2017.8272611

[45] T. Luong, F. Argelaguet, N. Martin, and A. Lecuyer. Introducing mental workload assessment for the design of virtual reality training scenarios. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 662–671, 2020.

[46] K. Mandrick, V. Peysakhovich, F. Rémy, E. Lepron, and M. Causse. Neural and psychophysiological correlates of human performance under stress and high mental workload. *Biological Psychology*, 121:62–73, Dec. 2016. doi: 10.1016/j.biopsycho.2016.10.002

[47] J. Marín-Morales, J. L. Higuera-Trujillo, A. Greco, J. Guixeres, C. Llinares, C. Gentili, E. P. Scilingo, M. Alcañiz, and G. Valenza. Real vs. immersive-virtual emotional experience: Analysis of psychophysiological patterns in a free exploration of an art museum. *PloS one*, 14(10), 2019.

[48] J. Marín-Morales, J. L. Higuera-Trujillo, A. Greco, J. Guixeres, C. Llinares, E. P. Scilingo, M. Alcañiz, and G. Valenza. Affective computing in virtual reality: emotion recognition from brain and heartbeat dynamics using wearable sensors. *Scientific reports*, 8(1):1–15, 2018.

[49] M. Meehan, B. Insko, M. Whitton, and F. P. Brooks Jr. Physiological measures of presence in stressful virtual environments. *Acm transactions on graphics (tog)*, 21(3):645–652, 2002.

[50] B. Mehler, B. Reimer, J. F. Coughlin, and J. A. Dusek. Impact of Incremental Increases in Cognitive Workload on Physiological Arousal and Performance in Young Adult Drivers. *Transportation Research Record: Journal of the Transportation Research Board*, 2138(1):6–12, Jan. 2009. doi: 10.3141/2138-02

[51] 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 *Proceedings of the 6th International Driving Symposium on Human Factors in Driver Assessment, Training, and Vehicle Design : driving assessment 2011*, pp. 590–597. University of Iowa, Olympic Valley-Lake Tahoe, California, USA>, 2011. doi: 10.17077/drivingassessment.1451

[52] A. Murata and H. Iwase. Evaluation of mental workload by fluctuation analysis of pupil area. In *Proceedings of the 20th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. Vol.20 Biomedical Engineering Towards the Year 2000 and Beyond (Cat. No.98CH36286)*, vol. 6, pp. 3094–3097. IEEE, Hong Kong, China, 1998. doi: 10.1109/IEMBS.1998.746146

[53] N. Nourbakhsh, Y. Wang, and F. Chen. GSR and Blink Features for Cognitive Load Classification. In P. Kotzé, G. Marsden, G. Lindgaard, J. Wesson, and M. Winckler, eds., *Human-Computer Interaction – INTERACT 2013*, Lecture Notes in Computer Science, pp. 159–166. Springer, Berlin, Heidelberg, 2013. doi: 10.1007/978-3-642-40483-2_11

[54] D. Novak, M. Mihelj, and M. Munih. A survey of methods for data fusion and system adaptation using autonomic nervous system responses in physiological computing. *Interacting with Computers*, 24(3):154–172, May 2012. doi: 10.1016/j.intcom.2012.04.003

[55] R. D. O'Donnell and F. T. Eggemeier. *Workload assessment methodology*, pp. 1–49. John Wiley & Sons, 1986.

[56] T. E. Oliphant. *A guide to NumPy*, vol. 1. Trelgol Publishing USA, 2006.

[57] T. pandas development team. pandas-dev/pandas: Pandas, Feb. 2020. doi: 10.5281/zenodo.3509134

[58] 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. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[59] R. W. Picard, E. Vyzas, and J. Healey. Toward machine emotional intelligence: analysis of affective physiological state. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(10):1175–1191, Oct 2001. doi: 10.1109/34.954607

[60] F. Putze, C. Herff, C. Tremmel, T. Schultz, and D. J. Krusienski. Decoding mental workload in virtual environments: a fnirs study using an immersive n-back task. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 3103–3106. IEEE, 2019.

[61] S. Putze, D. Alexandrovsky, F. Putze, S. Höffner, J. D. Smeddinck, and R. Malaka. Breaking the experience: Effects of questionnaires in vr user studies. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–15, 2020.

[62] Y. Qi. Random Forest for Bioinformatics. In C. Zhang and Y. Ma, eds., *Ensemble Machine Learning*, pp. 307–323. Springer US, Boston, MA, 2012. doi: 10.1007/978-1-4419-9326-7_11

[63] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018.

[64] M. A. Recarte and L. M. Nunes. Mental workload while driving: Effects on visual search, discrimination, and decision making. *Journal of Experimental Psychology: Applied*, 9(2):119–137, 2003. doi: 10.1037/1076-898X.9.2.119

[65] W. K. Roberts and J. J. Gallimore. A physiological model of cybersickness during virtual environment interaction. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 49, pp. 2230–2234. SAGE Publications Sage CA: Los Angeles, CA, 2005.

[66] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y

[67] A. Sano, W. Chen, D. L. Martinez, S. Taylor, and R. W. Picard. Multimodal ambulatory sleep detection using lstm recurrent neural networks. *IEEE Journal of Biomedical and Health Informatics*, pp. 1–1, 2018. doi: 10.1109/JBHI.2018.2867619

[68] Y. Santiago-Espada, R. R. Myer, K. A. Latorella, and J. R. Comstock Jr. The multi-attribute task battery ii (matb-ii) software for human performance and workload research: A user's guide. 2011.

[69] Y. Shimomura, T. Yoda, K. Sugiura, A. Horiguchi, K. Iwanaga, and T. Katsuura. Use of frequency domain analysis of skin conductance for evaluation of mental workload. *Journal of Physiological Anthropology*, 27(4):173–177, June 2008. doi: 10.2114/jpa2.27.173

[70] M. Slater, C. Guger, G. Edlinger, R. Leeb, G. Pfurtscheller, A. Antley, M. Garau, A. Brogni, and D. Friedman. Analysis of physiological responses to a social situation in an immersive virtual environment. *Presence: Teleoperators and Virtual Environments*, 15(5):553–569, 2006.

[71] M. Slater, M. Usoh, and A. Steed. Depth of presence in virtual environments. *Presence: Teleoperators & Virtual Environments*, 3(2):130–144, 1994.

[72] Society for Psychophysiological Research Ad Hoc Committee on Electrodermal Measures. Publication recommendations for electrodermal measurements: Publication standards for EDA. *Psychophysiology*, 49(8):1017–1034, Aug. 2012. doi: 10.1111/j.1469-8986.2012.01384.x

[73] E. T. Solovey, M. Zec, E. A. Garcia Perez, B. Reimer, and B. Mehler. Classifying driver workload using physiological and driving performance data: two field studies. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems - CHI '14*, pp. 4057–4066. ACM Press, Toronto, Ontario, Canada, 2014. doi: 10.1145/2556288.2557068

[74] J. Son, H. Oh, and M. Park. Identification of driver cognitive workload using support vector machines with driving performance, physiology and eye movement in a driving simulator. *International Journal of Precision Engineering and Manufacturing*, 14(8):1321–1327, Aug. 2013. doi: 10.1007/s12541-013-0179-7

[75] C. L. Stephens, I. C. Christie, and B. H. Friedman. Autonomic specificity of basic emotions: Evidence from pattern classification and cluster analysis. *Biological Psychology*, 84(3):463–473, Jul 2010. doi: 10.1016/j.biopsycho.2010.03.014

[76] M. Strait and M. Scheutz. What we can and cannot (yet) do with functional near infrared spectroscopy. *Frontiers in Neuroscience*, 8, May 2014. doi: 10.3389/fnins.2014.00117

[77] T. Tamura, Y. Maeda, M. Sekine, and M. Yoshida. Wearable photoplethysmographic sensors—past and present. *Electronics*, 3(2):282–302, 2014.

[78] A. J. Tattersall and P. S. Foord. An experimental evaluation of instantaneous self-assessment as a measure of workload. *Ergonomics*, 39(5):740–748, 1996.

[79] G. Teo, G. Matthews, L. Reinerman-Jones, and D. Barber. Adaptive aiding with an individualized workload model based on psychophysiological measures. *Human-Intelligent Systems Integration*, Nov. 2019. doi: 10.1007/s42454-019-00005-8

[80] B. S. Todd and D. C. Andrews. The identification of peaks in physiological signals. *Computers and Biomedical Research*, 32(4):322–335, Aug 1999. doi: 10.1006/cbmr.1999.1518

[81] C. Tremmel, C. Herff, T. Sato, K. Rechowicz, Y. Yamani, and D. J. Krusienski. Estimating cognitive workload in an interactive virtual reality environment using eeg. *Frontiers in Human Neuroscience*, 13, 2019.

[82] P. S. Tsang and M. A. Vidulich. Mental workload and situation awareness. 2006.

[83] M. L. Turner and R. W. Engle. Is working memory capacity task dependent? *Journal of memory and language*, 28(2):127–154, 1989.

[84] E. L. van den Broek, V. Lisý, J. H. Janssen, J. H. D. M. Westerink, M. H. Schut, and K. Tuinenbreijer. Affective Man-Machine Interface: Unveiling Human Emotions through Biosignals. In A. Fred, J. Filipe, and H. Gamboa, eds., *Biomedical Engineering Systems and Technologies*, Communications in Computer and Information Science, pp. 21–47. Springer Berlin Heidelberg, 2010.

[85] P. van der Wel and H. van Steenbergen. Pupil dilation as an index of effort in cognitive control tasks: A review. *Psychonomic Bulletin & Review*, 25(6):2005–2015, Dec. 2018. doi: 10.3758/s13423-018-1432-y

[86] M. van Dooren, J. H. Janssen, et al. Emotional sweating across the body: Comparing 16 different skin conductance measurement locations. *Physiology & behavior*, 106(2):298–304, 2012.

[87] P. van Gent, H. Farah, N. V. Nes, and B. van Arem. Towards Real-Time, Nonintrusive Estimation of Driver Workload: A Simulator Study. *Road Safety and Simulation*, 2017.

[88] P. van Gent, T. Melman, H. Farah, N. van Nes, and B. van Arem. Multi-Level Driver Workload Prediction using Machine Learning and Off-the-Shelf Sensors. *Transportation Research Record: Journal of the Transportation Research Board*, 2672(37):141–152, Dec. 2018. doi: 10.1177/0361198118790372

[89] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. Jarrod Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. Carey, İ. Polat, Y. Feng, E. W. Moore, J. Vand erPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and S. . . Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2

[90] C. D. Wickens. Multiple resources and mental workload. *Human factors*, 50(3):449–455, 2008.

[91] C. Zeagler. Where to wear it: functional, technical, and social considerations in on-body location for wearable technology 20 years of designing for wearability. In *Proceedings of the 2017 ACM International Symposium on Wearable Computers*, pp. 150–157, 2017.

[92] J. Zhang, Z. Yin, and R. Wang. Nonlinear Dynamic Classification of Momentary Mental Workload Using Physiological Features and NARX-Model-Based Least-Squares Support Vector Machines. *IEEE Transactions on Human-Machine Systems*, 47(4):536–549, Aug. 2017. doi: 10.1109/THMS.2017.2700631