Personal and Ubiquitous Computing
https://doi.org/10.1007/s00779-020-01455-7

ORIGINAL ARTICLE

A framework to estimate cognitive load using physiological data

Muneeb Imtiaz Ahmad[1,2] · **Ingo Keller**[1] · **David A. Robb**[1] · **Katrin S. Lohan**[3,4]

Received: 27 November 2019 / Accepted: 5 September 2020

Abstract

Cognitive load has been widely studied to help understand human performance. It is desirable to monitor user cognitive load in applications such as automation, robotics, and aerospace to achieve operational safety and to improve user experience. This can allow efficient workload management and can help to avoid or to reduce human error. However, tracking cognitive load in real time with high accuracy remains a challenge. Hence, we propose a framework to detect cognitive load by non-intrusively measuring physiological data from the eyes and heart. We exemplify and evaluate the framework where participants engage in a task that induces different levels of cognitive load. The framework uses a set of classifiers to accurately predict low, medium and high levels of cognitive load. The classifiers achieve high predictive accuracy. In particular, Random Forest and Naive Bayes performed best with accuracies of 91.66% and 85.83% respectively. Furthermore, we found that, while mean pupil diameter change for both right and left eye were the most prominent features, blinking rate also made a moderately important contribution to this highly accurate prediction of low, medium and high cognitive load. The existing results on accuracy considerably outperform prior approaches and demonstrate the applicability of our framework to detect cognitive load.

Keywords Cognitive load · Framework · Physiological data · Human-computer interaction

1 Introduction

In the past few decades, cognitive load (CL) has been shown to negatively impact human performance in various tasks demanding a high amount of mental effort [4]. In general, CL refers to the load placed on the user's working memory, also viewed as short-term memory, during a task [53]. The significance of measuring CL has been well described in the past due to its application under various contexts such as problem-solving, instructional design, multimedia, aircraft, and automation [42]. CL can be monitored in real time as a method to capture the automation experience [15, 57]. Accurate measurement of CL can be used to apply mitigation strategies, such as the adaptation of the user interface in response to changes in CL [36]. One approach is to present information differently for a naive user vs. an expert user. This is needed because an expert user may view the task as trivial, and this can cause boredom which may induce cognitive under-load [57]. The purpose of such strategies is to improve performance, operational efficiency, and operational safety, while reducing failures [50]. For example, a driver in an autonomous vehicle needs to monitor and supervise automation to achieve operational safety. However, drivers may experience cognitive under-load over time. This raises concern about their ability to consistently monitor automation, possibly resulting in an accident. Other application areas include the deployment of robots in extreme environments and in automated environments such as smart

✉ Muneeb Imtiaz Ahmad
m.ahmad@hw.ac.uk

Ingo Keller
i.keller@hw.ac.uk

David A. Robb
d.a.robb@hw.ac.uk

Katrin Lohan
k.lohan@hw.ac.uk

1 Edinburgh Center for Robotics, Heriot-Watt University, Edinburgh, UK

2 Department of Computer Science, Swansea University, Swansea, UK

3 Department of Mathematical and Computer Science, Heriot-Watt University, Edinburgh, UK

4 EMS Institute for Development of Mechatronic Systems, NTB University of Applied Sciences in Technology, Buchs, Switzerland

Published online: 27 September 2020

factories, where supervisors observe autonomous operations [2, 20]. As a result, intelligent user interfaces are needed to provide situation awareness to the supervisors. This will help them to observe, analyse, and supervise autonomous operations safely and efficiently by managing their CL [35].

CL has been classified into three different types: (1) intrinsic load, (2) extraneous load, and (3) germane load [52]. While intrinsic load stems from the complexity of the task and its association with the user, extraneous load is caused by the presentation style of the material. Lastly, germane load refers to the ability of the user to fully understand the material. We believe that both extraneous load and germane load are relevant factors affecting the operators' interaction. For example, an interface presenting data in a particular manner can result in an increase in both extraneous and germane load, which could induce high CL. Consequently, we need to reduce CL through creation of intelligent user interfaces that measure CL in real time and adjust the presentation accordingly. However, to the best of our knowledge, it remains a challenge to measure CL in a robust and non-intrusive manner.

To address this challenge of classifying CL, we designed a framework (Fig. 1) that applies machine learning to the physiological data gathered from available state-of-the-art sensing technologies. The rationale for calling the framework *generic* lies in the concept of avoiding task- or stimuli-dependent physiological behaviour. In principle, the framework can be applied across different settings. For example, it could be used to monitor a driver's CL in an autonomous vehicle or a supervisor's CL in a control room to either ensure operational safety or to reduce mistakes. The framework incorporates *machine learning* to understand the relevant features in a range of physiological behaviour data while automatically taking the task into account. Our contributions are threefold:

- We demonstrate a ***generic framework*** to classify low, medium and high levels of CL.
- We present the results of an evaluation through the creation of a ***novel task*** to test our framework using eye- and heart-based data. To promote reuse, we make our task and sensor application code in addition to the scripts for generation of stimuli and data analysis available.
- We make the ***dataset*** publicly available for community to use in order to classify CL. We further show that the evaluation of the framework using the dataset achieves high predictive accuracy on the exemplar task.

While we have used eye- and heart-based data in this current work, we understand that data collected from multiple sources synchronously can further improve the robustness and general applicability of our framework for different kinds of stimuli.

2 Related work

Historically, CL was introduced in the context of problem-solving and instructional design to understand its effects on learning [51]. However, CL was later studied as a construct of operators' (e.g. pilots') mental workload. It was found that increased CL has an adverse effect on operator

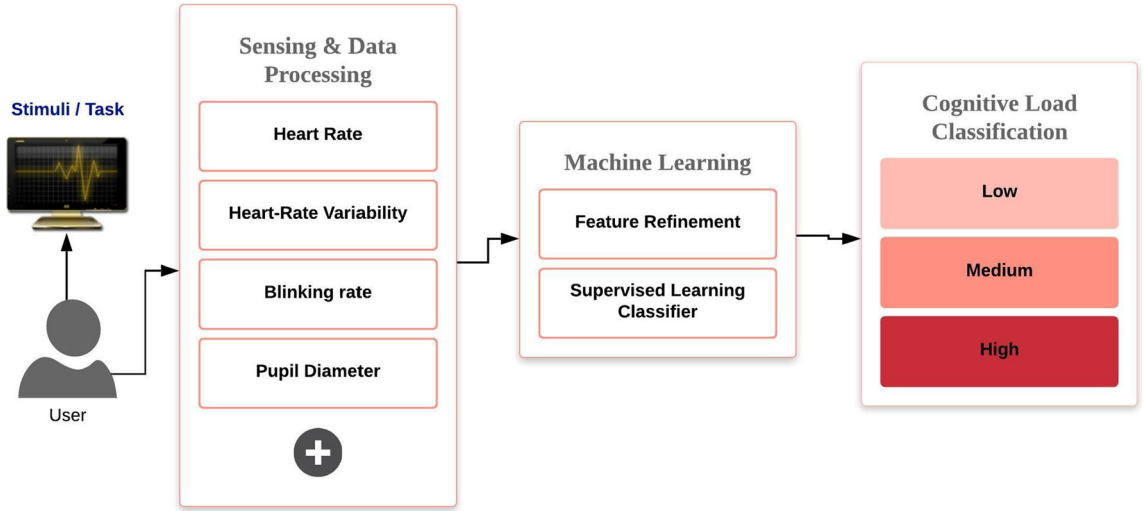

Fig. 1 A framework to estimate cognitive load using physiological data

performance [39]. The term CL, has been referred to under various application contexts, with terminology such as work load and mental load. However, these all refer to the same general concept [14]. In the rest of this section, we set out methods used in previous research to measure CL in humans in experimental settings.

2.1 Subjective questionnaire-based methods

Subjective rating questionnaires have been the most popular mechanism to measure CL, perhaps due to their ease of use [42]. The NASA Task Load Index is one of the most commonly used questionnaires to measure subjective CL [19]. The questionnaire consists of rating six constructs: (1) mental demand, (2) physical demand, (3) temporal demand, (4) effort, (5) performance and (6) frustration, each being rated from low to high [19]. Other known questionnaires include the *Cognitive Load Component* survey that separates the three classifications of load: intrinsic, extraneous, and germane load. This self-report-based measure is particularly relevant to instructional settings [31]. It is, however, interesting to note that rating scales are not generally regarded as reliable measures of CL. Researchers are critical of using CL subjectively at the end of the task because changes in CL are momentary. Therefore, they should be estimated in real time [37].

2.2 Performance-based methods

Prior findings suggest a relationship between CL and a user's task performance; therefore, CL can be measured by measuring performance [14]. Pass et al. divided performance into sub-classes: one refers to the task performance and the other refers to the performance measures derived from task performance (i.e. response time, error-rate, or accuracy). Pass et al. believe that these factors can contribute to estimating CL and are highly sensitive and reliable [42]. However, while we understand that such factors can be used in real time, these are task-dependent measures which cannot be used generally in the case of adapting technology.

2.3 Speech-based methods

Human speech behaviours are observable as measures of CL [58]. These include language-based and emotion-based behaviours. The language-based behaviours include hesitations, increased use of pauses, decreased articulation rate, decreased speech rate, self-corrections and several others [3, 26, 29, 34]. The emotion-based behaviours include increased use of negative emotions, decreased use of positive emotions, and several other indicators [27, 28]. We understand that identifying CL based on these behaviours is, perhaps, relevant but is also task dependent. Linguistic behaviours, while non-intrusive, are only relevant in settings where speech input is used. We believe that these behaviours can be used in our framework because speech input can be collected unobtrusively. However, we do not use them in the current demonstration of the framework due to the nature of the task.

2.4 Physiological behaviour-based methods

The most commonly used method to measure CL is to observe and to report on the changes in humans' physiological behaviours [42]. These physiological behaviours are mostly based on changes in the measurements taken from four different human organs: (1) brain , through measuring neurological activity via an electroencephalogram (EEG); (2) heart, through measuring heart rate (HR), or heart rate variability (HRV); (3) skin, through measuring galvanic skin conductance (GSR); and (4) eyes, through measuring eye movements, mean pupil diameter change (MPDC), or blinking rate (BR).

Prior findings on the variation of HRV during a range of computerized tasks have shown that reduction in HRV is attributed to higher CL [40]. Cranford et al. [10] reported that there was an increase in HR with an increase in task difficulty. This suggests that as task difficulty increases, in other words, as CL grows, it results in an increase in HR and reduction in HRV. We also found several experiments in the literature that report an increase in MPDC in a situation demanding higher mental workload [46, 47]. Past findings also indicate that BR decreases in the case of higher mental load [23]. For GSR, it has been found that GSR readings increase with an increase in CL [48]. Similarly, the EEG theta wave activity increases in the frontal region with an increase in the CL [16]. In summary, there is empirical evidence from past research implying that the changes in the measurements of physiological behaviours can be attributed both to CL, and to various levels of mental processing. It is also important to note that the existing sensing technologies work well and provide an accurate representation of the particular behaviours [24, 54]. Furthermore, with the advancement of design and technology, solutions are now available to collect such data in less invasive ways.

In relation to our work, the existing literature shows that changes in one physiological behaviour may be related to another observable physiological behaviour. For instance, Siegle et al. [49] showed that there is a relationship between MPDC and BR in a digit-sorting task. Therefore, in this paper, we collected data on a variety of physiological behaviours and used the data to classify low, medium and

high levels of CL. The rationale for classifying three levels of CL is grounded in one of the most recent works on predicting CL in the wild [14].

Machine learning to estimate CL In the past, researchers have used machine learning-based approaches to estimate CL [18, 41, 56, 59]. Zhang et al. [59] proposed an adaptive support vector machine (SVM)-based method to classify operator mental workload by using electroencephalogram (EEG), electrocardiogram and electrooculography signals in a simulated human-machine system. However, three of the aforementioned papers on these techniques used data from brain-specific EEG sensors [41, 56, 59].

Recently, in 2018, Heard et al. [21] published a survey on workload assessment algorithms. These algorithms use a range of machine learning methods to predict different levels of workload. In particular, the survey enlisted 24 workload assessment algorithms that achieved an accuracy between 60 and 90%. The assessment algorithms established a *baseline*, or ground truth, for the measurement of CL by enabling participants to stare at the screen or to involve them in a task inducing low CL such as adding two numbers. However, it is important to note that the term "workload" is vast in its scope and has seven different decompositions, CL or cognitive workload is one of them. Our work differs from prior work due to the following reasons. Firstly, the accuracy results are based on small user groups (on average, 8–10 with lows of 3 and 4), which are known to have low statistical power when developing techniques for wider use. Secondly, only three studies were carried out on CL-based tasks to estimate CL. Thirdly, within those 24 algorithms, the algorithms achieving an accuracy over 90% cannot be applied in real time because they use the NASA TLX questionnaire data. Others were based on predicting only two-levels of CL (low vs high). Lastly, prior work uses SVM and suffers from overfitting. These highlighted aspects of the previous work indicate that we should create datasets based on a high number of participants to create robust measurements of CL. In addition, there is a need to train a model to predict three levels of CL in real time. Moreover, more research is needed with features based on the combination of heart and eye data as this is a niche area that has not been researched strongly [21]. Furthermore, we need to use a range of physiological behaviours to collect data in a non-intrusive way to extend their use in real-settings. In the real world, which is the focus of our work (e.g. applications to supervise and plan missions for robots in extreme environments or to manage autonomous systems operations in smart factories), the use of EEG sensors are impractical, therefore we do not use them. Most of the prior work described above has focused on them (21/24 and 3/3 for CL tasks), which also sets us apart [21].

We also see studies that have been conducted to collect data from one of the physiological behaviours such as eye-based measures, GSR measures, or speech measures to classify CL [7]. Although the results from these studies are encouraging, the accuracy is relatively low. The most recent work on CL estimation in a driving task was based on using deep learning to extract the pupil size from a video. It then used a classification algorithm to classify CL as low, medium and high [14]. This work is closest to our approach in terms of predicting three levels of CL, however, the method only takes pupil size as input. We understand that this method is suitable for driving but may not be suitable for other tasks, because the existing literature on the CL measurement indicates that measurements of some physiological behaviours may not be suitable for some tasks [42].

Our approach encourages the use of a range of features (physiological behaviours collected non-intrusively) and later applies feature elimination methods to determine the indicative ones in the given context. We show that this approach can yield better accuracy on a dataset that is based on a large cohort of participants.

3 Setting and method

3.1 Cognitive load framework

The framework (Fig. 1) has three modules: (1) stimuli, (2) sensing and data collection and processing, and (3) applying machine learning for detecting CL.

3.1.1 Stimuli

The first module consists of an external stimulus or a task to induce CL. As previously highlighted, physiological behaviours tend to be task- or stimuli-dependent. Hence, the framework does not propose a specific stimulus. Instead, it removes the context-dependency to classify CL. This suggests that the framework can, in principle, be applied regardless of the task.

3.1.2 Sensing and data processing

In the sensing module, we used state-of-the-art sensing technology to collect eye- and heart-based data. It is important to note that our sensing module is not limited to only two measurements and can accommodate other physiological behaviours.

We capture the data from various sensing devices consisting of raw signals. Data from these sources need to be synchronized and cleaned. In this module, we first handle data synchronicity through applying time-frames to our raw signals and later apply various widely used techniques to clean and filter our data.

3.1.3 Machine learning for detecting cognitive load

This module consists of two sub-modules: (1) feature refinement and (2) supervised learning classification. The feature refinement module selects the best or worst performing features. Various algorithms such as *SelectKBest*, recursive feature elimination, correlation-based feature selection, and others can be used for this. It is important to use feature refinement methods as physiological behaviours tend to be task dependent. For instance, HR and HRV are insensitive to the instantaneous load caused by the fluctuations every time someone works on a task, hence they can be task dependent [42]. Similarly, pupil diameter (PD) is sensitive to changes in light and also varies with age [33]. Additionally, PD may also be unsuitable for some tasks. Consequently, the feature refinement sub-module is needed and can help improve the robustness of CL detection. Following the feature refinement step, the framework uses a range of supervised learning methods such Naive Bayes, Logistical Regression, Support Vector Machine, and others to classify low, medium and high levels of CL.

3.2 Applying the CL detection framework

Our **stimuli** had three phases: (1) *the rest phase*, (2) *the trial phase*, and (3) *the task phase*. These were used to generate three levels of CL, *Low*, *Medium* and *High* respectively. The data collected from these three phases were used later to train and test the machine learning classifiers. It is important to note that the duration of each phase was different, therefore, we normalized all the physiological behaviour data to avoid any bias. One of the initial steps was to perform calibration with the eye tracking device. In the rest phase, once the basic calibration was performed, the participants viewed a changing full-screen display of white, black, and grey colors while their data on eye and heart activity was recorded. In the trial phase, participants first read an introduction. Then, they were asked to undertake simpler versions of stimuli items than those which they would meet in the main task phase. This was done to familiarize them with the nature of the task. Lastly, the task phase was a simple game-based task to recognize correct and made-up words, and correct and incorrect sentences. Our task had six different item types as shown in Table 1. Each item type had 20 words or sentences. The task was designed to induce two components from cognitive load theory: (1) the complexity of the task was inherently difficult, inducing intrinsic load, and (2) the presentation of the words and sentences in an arbitrary order induced extraneous load.

We used the list of words from the British National Corpus [8] to create the word-based task items. We developed a simple script in python to select 20 words (nouns) of length 10 with frequency ranging from 1013 to 1026 in the corpus. We also looked into the movie review dataset [44] to prepare the sentence-based task items. We developed another script to select sentences containing 10 words each. We removed sentences from the dataset containing words having apostrophes, quotes, numbers, etc. Also, we removedv sentences having very short words such as "a" or "I". Finally, we selected the first 20 out of the remaining 54 sentences. It is important to add that the rationale for our choice of words was based on our understanding of a recent study that indicated that more surprising words take longer to read and result in increasing pupil sizes [13]. Therefore, as one of our features to estimate CL is pupil sizes, this justifies our choice of the task. Furthermore, the rationale for the character length of the word was based on maintaining the difficulty of the words at a certain level.

> We emphasise that our task is novel in terms of its use in inducing CL and this new task helps to demonstrate the framework by the creation of a dataset to classify three levels of CL.

Table 1 Task item overview

Item type	Content	Example
1	a correct English word	reluctance
2	as 1 but with the middle letters switched	relu**tc**ance
3	as 1 but with scrambled letters	anctuce**rel**
4	an arbitrary mnemonic word	lcwvcdkxob
5	a correct English sentence from a movie review dataset [44]	the only problems come during the first and third acts
6	as 5 but with rearranged words rendering them incorrect	the only problems come **first and third acts the during**

Bold letters refer to the changes in stimulus for each item type used in the task

The **sensing** module used the state-of-the-art eye tracker - Tobii Pro Glasses 2 Eye Tracker (Eye Tracker) to collect on PD and BR. In addition, we used the EliteHRV CorSense device [25] to collect data on HR and HRV. The sensing devices can be seen in Fig. 3.

Measuring pupil diameter We used the following steps to clean the data collected from the Eye Tracker as described in prior literature [30]. In the first step, we prepared the raw data on pupil size for the left and right eyes in a standard format. The instances where the sizes contained negative values were removed. In the second step, we filtered the raw data by removing three types of the most frequently occurring invalid pupil size samples: (1) dilation speed outliers and edge artifacts, (2) trend-line deviation outliers, and (3) temporally isolated samples. Dilation Speed outliers refer to data that consists of large pupil sizes relative to their adjacent samples. We used median absolute deviation (MAD), a commonly used technique [32] as represented in (1) and (2), to detect outliers and later remove them from our sample. Once removed, we identified trend-line deviation outliers, mostly due to the gaps in the data that may have been caused by blinks. We later removed these gaps using the same MAD technique. Finally, we removed the temporally isolated samples containing noise due to a momentary eye tracker glitch. We used a sparsity filter that splits any pupil size signal that has a gap greater than 40ms and then rejects any resulting section that is less than 50ms. We also removed pupil size values that were not inside the range of 1.5 to 9 mm. In the third step, once our raw data samples were filtered, we performed data sectioning and conducted our analysis. The code for the filtering can be found on Github using a link provided at the end of this section.

$$\text{MAD} = \text{median}(|X_i - \tilde{X}|) \tag{1}$$

$$\tilde{X} = \text{median}(X) \tag{2}$$

To apply the process described above, we recorded the whole session, including the basic calibration with the Eye Tracker followed by an additional step that presents a changing full-screen display of white, black and grey to establish a first estimation for minimum and maximum values of PDs for both left and right eyes. We tracked PD during this calibration, during the explanation of the task, and during the task itself. Afterwards, we manually annotated the start and end of the task by finding the corresponding frames from the front-view camera stream. Segmentation of the pupil data was done by converting the frame IDs to time stamps. We used these to determine the start and the end of the task segment in the PD readings as provided by the glasses. To account for different pupil sizes, we extracted the raw data for both eyes. We applied the previously described cleaning, and three step filtering method, to clean and filter the raw data. We later computed the MPDC, for each task phase, as the ratio between the overall mean PD (over all three task phases), and the mean PD while performing each of the individual phases of the task. Our method to compute the MPDC is grounded in literature as it follows the approach applied by Palinko et al. [43].

Measuring blink rate To calculate BR, we used the Eye Tracker to record the eye stream of the full session. We reused the aforementioned manual annotation to get the task segment by finding the correct frames in the front-view stream and calculating the corresponding frame IDs for the eye stream. To detect the total number of blinks, we applied the following mechanism: Firstly, we converted each frame into grey-scale and applied a Gaussian blur to it. Secondly, we applied a binary threshold to the frame and used the blurred frame to find contours in it. The convex hull was calculated for all contours. Lastly, we computed the ratio between the squared circumference and the area of the convex hull to remove all non-spherical hulls. We used a threshold of 150 to 1200 as a limit for the area and values from 10 to 17 for the ratio to exclude non-pupil hulls. Mathematically, the ratio value should be $4\pi \approx 12.57$, but due to noise in the data, we had to widen the ratio range. The code for detecting blinks can be found via a link at the end of this section. We also normalized the data by computing the number of blinks per minute for each phase. We did this because the duration for each phase varied between individuals.

Measuring heart rate and heart rate variability For the computation of HRV, we collected data from the CorSense device. HRV refers to the millisecond changes in duration between successive heartbeats. These are termed, the R-R intervals. We used the interquartile range method, a function that removes outliers, to clean and filter the data [55]. Afterwards, we applied a Root Mean Square of Successive Differences (RMSSD) calculation to the R-R intervals. Finally, a natural log(ln) is applied to the RMSSD [24]. To compute HR, we divide $60 * 1000$ by the mean of the R-R intervals [1].

To make sure that the data is collected synchronously, we compared time stamps and used them to compute the values of PD, BR, HR, and HRV while the participant is performing a specific task. Once the data was cleaned and corrected, we created three levels in our dataset based on the previously indicated three phases in our stimuli. These levels were indicative of the low, medium and high CL.

In the **machine learning module**, we applied a feature elimination method. In this step, we conducted statistical tests to select those features that have the strongest

relationship with the classifications (Low, Medium and High). This suggested that the features that were not statistically significant to our classification could be dropped from the feature set. To achieve this, we manually conducted a one-way analysis of variance (ANOVA) with the feature set (MPDC, BR, HR, HRV) as the list of dependent variables and the classification of CL as the independent variable. We then performed feature selection using *SelectKBest* and removed all but the *k* best performing features. We used the *chi-squared* test to choose the top performing features. In essence, this identifies the features that have the strongest relationship with the output variable. Once our feature set was finalized, we used the following classification algorithms (or classifiers): AdaBoost (AB), Decision Tree (DT), Naive Bayes (NB), Logistical Regression (LR), Random Forest (RF), Support Vector Machine (SVM) and *k*-Nearest Neighbour (*k*NN). The goal of the classification algorithm was to predict the trait class, i.e. to predict the low, medium and high levels of CL. To apply the classifiers, we first used *Stratified KFold* to create five different splits in our dataset. Later, we applied all of the classifiers, one after the other, to compute their accuracy in predicting the level of CL. Lastly, we used a classification report to generate F1-scores.

3.3 Data collection and evaluation setting

3.3.1 Research aims

Our research attempted to answer the following questions (Q):

- Q1 - Did our experiment's main task induce CL as evidenced by participants' subjective ratings and task performance?
- Q2 - Did we observe differences for the three task phases for physiological behaviours (MPDC for left and right eyes, BR, HR, & HRV)?
- Q3 - Which classification method should be used to detect CL?
- Q4 - Which features are predictive of each level (low, medium and high).

3.3.2 Participants and procedure

We conducted our study with 41 participants (demographics as shown in Table 2). We asked participants about any reading difficulties and if they were native English language speakers as the task was based on reading. It is important to note that all the participants were attending university in an English speaking country, hence, they were highly proficient in English language. As our participants were required to wear eye tracking glasses, we asked if they usually wear glasses. We were not able to capture eye tracking data for one of the participants, therefore, we are reporting analysis of 40 participants.

Table 2 Participant demographics

Participants	41
Gender	20 female/21 male
Age	18–37 (mean: 23.3, two unreported) SD: 4.53
Native english speakers	Yes: 23, No: 18
Reading difficulties	Yes: 2, No: 39
Wear glasses	Yes: 15, No: 26

The study was conducted in the following steps:

1. Participant reads an information sheet and completes a consent form.
2. Participant completes (a) a questionnaire to report information on age, number of languages, and whether they have reading difficulties and (b) a physical activity questionnaire [12] to control for any bias in HR and HRV measurements.
3. Participant puts the CorSense Heart-Rate device on their finger (ring finger of left hand) and wears the Eye Tracker.
4. Participant performs the task consisting of three phases.
   (a) In the one-minute first phase, participant views the black, grey and white color changing screen.
   (b) In the two-minute second phase, participant spots the correct and incorrect trial words such as "which", "lagrat", "should", "aryst" and others.
   (c) In the five-minute third phase, participant performs the main task of playing the spot the correct or incorrect (made-up) words and sentences game task (see Table 1 for the task examples and description).
5. The physiological data was recorded using Eye Tracker (BR, PD), Heart Rate Monitor (HR, HRV), and Webcam facing the participant throughout the complete task.
6. Participant completes the NASA TLX Questionnaire to record their subjective ratings of CL. It is important to note that the participants were asked to give their subjective rating specifically and only about the third phase (4(c) above).

Participants were entered in a prize draw for shopping vouchers as a reward for participation. Ethical approval was obtained from our institution.

Fig. 2 Setup—a participant ready to perform the task

3.3.3 Setup and materials

The setup (shown in Fig. 2) involved a participant performing the word game task on a computer screen while wearing Tobii eye tracking glasses along with CorSense Heart-Rate device.

We used PsychoPy,[1] an open-source application, to programme our experiment. To collect data on changes in eye and heart behaviour, we used Eye Tracker pro[2] and a CorSense HRV device[3] respectively. We also used an external webcam to collect additional data on the BR as shown in Fig. 3. However, in the end, we did not use the recorded videos to calculate BR due to low quality of the recorded data and unsatisfactory rate of robustly detecting blinks. Instead, as previously described, we used the Eye Tracker's eye stream for both BR and PD analysis. During the task, we also collected data on the task performance of the participants (a score based on the number of correctly or incorrectly categorized words and sentences) to investigate the relationship between their task performance and their subjective rating of CL.

The NASA Task Load Index questionnaire[4] [19] was used to collect subjective ratings of the amount of CL generated by the task [42]. In addition, we used the International Physical Activity Questionnaire (IPAQ)[5] [9] to get relevant data on health-related physical activity. We collected this data on physical activity because the literature suggests that participants' physical activity index can create an experimental bias [17], bringing heart rate results into question if it is not taken into account. This IPAQ data provided us with reassurance that none of the participants were involved in highly physical activity or training before performing the task.

To analyse the data collected from the Eye Tracker, we created application program interface software that eases the access to the data and allows running the same analysis over all participants.

> The software can be found on GitHub at https://github.com/BrutusTT/tobii_api. The scripts for the generation of the stimuli and analysis of the data in this paper can be found at https://github.com/BrutusTT/ml_study/tree/master/ml_study/stimuli. Additionally, the dataset with three levels of CL can also be found on GitHub at https://github.com/BrutusTT/ml_study/tree/master/ml_study/modal. More details can be found in the included Readme files.

3.4 Summary of measurements

In summary, we collected the following measures during the experiment. *Physiological measures*: These were BR, PD, HR, and HRV. These are used in the framework. *Validation measures*: These were Physical Activity index (IPAQ pre-task), task performance score, and subjective task load (NASA TLX post-task). These were used as controls and for validating that the task induced CL.

4 Results

4.1 Did the main task induce CL?

To answer Q1 we made use of the performance scores, which were collected from the third phase of the task, and the NASA TLX ratings which we asked participants to provide specifically about their subjective task load during that same third phase. Thus, we have a set of *Subjective CL* ratings and an associated set of *Performance* scores. Based on the empirical evidence in literature, as CL increases, we would expect performance to reduce (see Subsection 2.2 in Section 2). Therefore, we would expect there to be a negative correlation between third phase *Performance*, which should vary with CL, and third phase *Subjective CL*, from the TLX ratings.

We ran a *Pearson* correlation between *Subjective CL* and *Performance*. We found that there was a negative correlation between *Subjective CL* and *Performance*, $r(41) = -.456$, $p < .00$. This is between a medium and a

[1] PsychoPy - https://www.psychopy.org

[2] Tobii eye tracking glasses pro - https://www.tobiipro.com/product-listing/tobii-pro-glasses-2/

[3] CorSense Elite HRV Device - https://elitehrv.com/corsense

[4] NASA Task Load Index Questionnaire - https://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/20000021488.pdf

[5] International Physical Activity Questionnaire - https://www.sdp.univ.fvg.it/sites/default/files/IPAQ_English_self-admin_long.pdf

Fig. 3 Tobii Eye Tracking Glasses (left), Webcam (middle), and Cor Sense Heart-rate Monitor (right)

large effect, but closer to a large effect [11].[6] The *M* and SD values for NASA TLX and performance are *M*: 50.96 and SD: 16.96 and *M*: 111.62 and SD: 5.77, respectively. This correlation result motivated us to conduct a simple linear regression of *Subjective CL* with *Performance*. A significant regression model was found ($F_{1,39} = 10.162$, $p < .00$), with an $R^2 = .207$, adjusted $R^2 = .186$, $\beta = -.456$. Thus, for our task a higher participants' *Subjective CL* rating does predict lower *Performance*. We understand that it has also been shown in numerous studies that a high CL adversely impacts user's task performance [4, 14, 39].

Hence, we conclude that this negative correlation between *Subjective CL* and *Performance* in our experiment, demonstrates and validates that our task does in fact induce CL. This positively answers Q1.

4.2 Descriptive statistics for the features to classify CL

We present descriptive statistics based on the three phases of our stimuli for all the 40 participants in Table 3. The table shows overall minimum and maximum values (range) along with mean (*M*) and standard deviation (SD) in the complete dataset based on all the phases of the stimulus. From the data gathered in the three individual stimulus phases, we were able to define the three classifications (low, medium and high). This was based on the significant difference, presented in the next subsection for the physiological behaviours observed in the three phases. We also show the ranges for all the features, for these three classes, along with their *M* and SD in the data. It can be seen that MPDC for both left and right eye increase according to each classification. On the contrary, BR declines from high to low level. HR did not show differences for all the levels from the three stimulus phases. However, HRV slightly declined across all the levels (this is discussed further in Section 4.5). It is also notable that the range for BR (between 0 and 52) in our data is in line with previous work which suggests that the mean BR is generally between 2 and 50 [38].

4.3 Did we observe changes in physiological behaviours across the three levels?

To answer Q2, we conducted a one-way, between-subjects, ANOVA to compare the effect of the three phases on all the physiological behaviours (MPDC for left and right eye, BR, HR, and HRV) in low, medium and high conditions. There was a significant effect of the phases on MPDC for left ($F_{2,119} = 139.75$, $p < .00$) and right eyes ($F_{2,119} = 149.50$, $p < .00$), and BR ($F_{2,119} = 3.475$, $p < .04$). We did not observe a significant effect of HR ($F_{2,119} = .09$, $p < .91$) and HRV ($F_{2,119} = 1.364$, $p < .26$).

We also conducted a post hoc test to observe the significant difference among the three phases. We found that the MPDC for both left and right eyes were statistically significant ($p < .00$) for all three levels of CL. This suggests that MPDC for both left and right eyes increased significantly from low to medium, and from medium to high levels of CL as indicated in Table 3. On the other hand, BR was marginally significant ($p < .07$) between low and medium, and between low and high levels of CL. This suggests that as the CL increased, there was a decrease in the rate of blinking. The participants blinked the least while under high CL.

The above analysis shows that our results here are in line with the findings reported in prior literature. That is, there is an increase in PD with an increase in the level of CL [46, 47] and BR declines with an increase in the level of CL [23]. We conjecture that, although we did not find a significant difference for HR and HRV, nonetheless, HRV marginally declined as CL increased [40] and HR marginally increased as CL increased [10].

[6] Field [11] suggests 0.3 is a medium effect while 0.5 is a large effect.

Table 3 Descriptive statistics for overall measures and the ranges of the three classifications corresponding to the three stimulus phases (the rest phase (Low), the trial phase (Medium), and the task phase (High))

Features	Range	Mean (*M*)	SD	Levels of CL		
				Low (range, *M*, SD)	Medium (range, *M*, SD)	High (range, *M*, SD)
MPDC left eye	[0.39,1.71]	1.01	0.26	[0.39, 1.08], *M*: 0.75, SD: 0.16	[0.79, 1.19], *M*: 0.97, SD: 0.07	[0.97, 1.72], *M*: 1.28, SD: 0.17
MPDC right eye	[0.42,1.67]	1.01	0.27	[0.42, 1.13], *M*: 0.75, SD: 0.26	[0.69, 1.12], *M*: 0.96, SD: 0.07	[1.04, 1.67], *M*: 1.29, SD: 0.16
BR	[0.0,53.0]	14.76	11.60	[0.0, 42.0], *M*: 18.63, SD: 13.82	[1.5, 39.0], *M*: 12.81, SD: 8.83	[0.0,53.0], *M*: 12.85, SD: 11.15
HR	[52.16,125.82]	80.07	13.06	[53.48, 125.81], *M*: 79.86, SD: 13.43	[52.84, 125.38], *M*: 79.53, SD: 13.31	[52.16, 119.67], *M*: 80.78, SD: 13.19
HRV	[2.44,4.74]	3.74	0.46	[2.64, 4.48], *M*: 3.78, SD: 0.44	[2.43,4.5], *M*: 3.77, SD: 0.45	[2.73, 4.74], *M*: 3.63, SD: 0.45

Revisiting Q2, we observed significant differences for the three phases for MPDC for left and right eyes. We observed marginally significant differences for low to medium and low to high for BR for left and right eyes.

4.4 Classifier performance

Addressing Q3, to investigate the most suitable classification method, we used the seven different classifiers to compare their performance to predict the low, medium and high levels of CL in our dataset. We used *Stratified KFold*, created five different splits and, finally, computed the mean accuracy of all seven classifiers. These are shown in Table 4. In general, we obtained a high predictive accuracy for most of the classifiers. However, RF outperformed the others with a mean accuracy of 91.66% followed by NB, DT, and SVM. We also found that LR and AB performed moderately, with a mean accuracy of 77.5% and 69.16% respectively. On the contrary, we obtained substantially lower performance from the KNN with a mean accuracy of 45.83%. Table 5 shows the classification report for the seven classifiers and also illustrates the F1-scores for each class (low, medium, high) of CL. It can be seen that the RF classifier resulted in the highest F1-score for each class followed by NB. We also observed a relatively lower F1-score for the medium class of CL compared with the other two classes. Nonetheless, in general, a high F1-score was achieved.

The results show that RF and NB performed well, with the feature selection based on *SelectKBest*. This selects the features that have a strong relationship with the CL level. The basic idea behind RF is that it operates as an ensemble. The algorithm creates trees (models) that output a class prediction. The model is predicted based on the class with the most votes. The key to better performance lies in the low correlation between the trees. We understand that the high predictive accuracy of DT reflected on the RF performance, as the trees created by the RF, as an ensemble, may have enhanced the predictive performance of the classifier. On the other hand, one reason for relatively low accuracy of AB could have been the presence of outliers in one of the features, as it can be seen that for BR, we had a wide range of data in our dataset.

Table 4 The mean accuracy (%) for the seven classifiers to predict CL

Classifier	AB	NB	DT	SVM	LR	RF	KNN
Accuracy	69.16	**85.83**	**85.00**	82.50	77.50	**91.66**	45.83

Bold values signify the classifier that achieved high accuracy in detecting Cognitive load

Table 5 F1-scores for the seven classifier to predict low, medium and high levels of CL

Cognitive Load	Classifier	F1-score
Low	AB	0.62
	NB	0.87
	DT	0.84
	SVM	0.83
	LR	0.80
	RF	**0.91**
	KNN	0.48
Medium	AB	0.71
	NB	0.81
	DT	0.80
	SVM	0.77
	LR	0.64
	RF	**0.85**
	KNN	0.51
High	AB	0.75
	NB	0.90
	DT	0.84
	SVM	0.88
	LR	0.85
	RF	**0.95**
	KNN	0.37

Bold RF is the classifier that achieves the highest accuracy

It is recognized that NB performs best if the input features are independent of each other, i.e. are less correlated with each other. It is also known that NB performs relatively better than LR and similar models when that is true. Consequently, we speculate that these are the reasons for the better performance of NB here. On the other hand, KNN demonstrated the lowest performance. We understand that KNN is a clustering-based method, and the results suggest that the dataset did not find proper clusters. In other words, the data was relatively hard to separate in the case of the KNN classifier.

In general, these classifier results suggest that our approach yielded promising findings.

To answer Q3, we conclude *Random Forest* as the most appropriate of the classifiers to classify the three levels of CL.

4.5 Feature importance per level of CL

To investigate which features are predictive of each class of CL in our dataset, we used one feature at a time and later computed the F1-score for each level of CL. The rationale was to investigate the effectiveness of all the features individually towards accurately classifying a certain level of CL in our dataset. Below, we only report the feature importance per level of CL for the RF classifier, because it was the best performing classifier to predict the levels of CL.

Figure 4 illustrates the feature importance for the RF classifiers based on the F1-scores of each level of CL. In general, it highlights that MPDC for left and right eyes, were the best performing features for all levels of CL for the RF classifier. In the case of RF, we observed that, for the low CL class, BR was relatively important as it generated a relatively good F1-score. Lastly, HRV and HR were found to be the least important features for the RF classifier to predict the low class of CL.

Looking at the predictions of the medium level of CL, we observe that BR was deemed a fairly significant feature for the RF classifier. We also recognize that HR and HRV were comparatively less important features to predict medium CL than to predict low CL. Overall, they both were found to be less critical as compared with other features (MPDC, BR). Lastly, to predict high CL, BR was found to be less important. In general, however, HR and HRV features were relatively more influential in low and medium levels of CL.

We understand that perhaps due to the nature of our task, HR and HRV were not among the critically important features in our dataset. The time pressures from the presentation of tasks or stimuli or the display of data at a faster rate may have induced differences in the heart data [22]. We compared the F1-scores of individual features, with the case where the F1-scores were computed through using all the features, as shown in Table 5. We observe that the F1-scores (91%, 85%, and 95%) to predict the low, medium and high levels of CL using all the features were higher than the F1-score of the individual features for each of the low, medium and high levels of CL (as shown in Fig. 4). It is also important to note that for high CL, the F1-score, after using only MPDC for the right eye, was 94% and 87% respectively. This shows that for the high level of CL, MPDC was the most important feature in our dataset. In other words, it shows that other features were not important for the high level of CL; however, we want to emphasise that all features in different ways played a role to achieve a high F1-score for a low and medium level of CL. None the less, this all shows the potential of the idea of using various physiological behaviours as features in our framework because it can make the detection of CL less dependent on the task. Also, it highlights the need to conduct more studies in the future and shows that it can indeed improve the accuracy of predicting the three levels of CL.

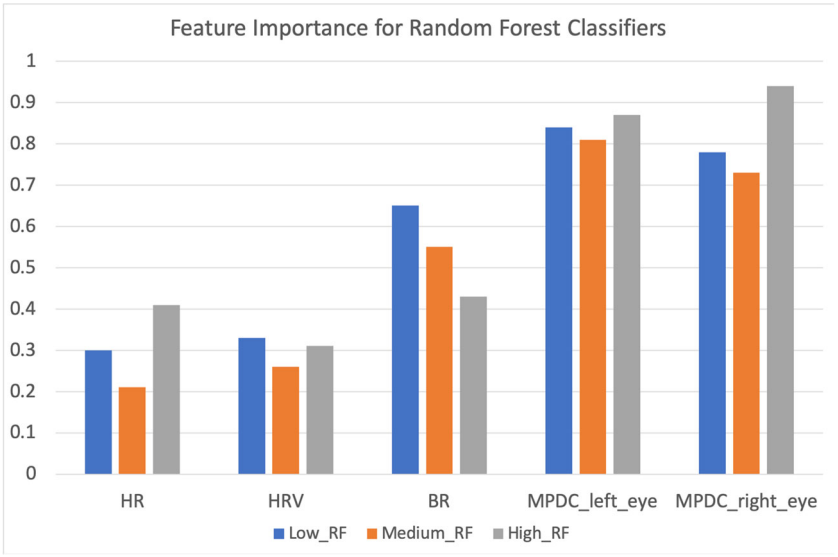

Fig. 4 Feature importance for the RF classifier based on the F1-scores for each level of CL. The *x*-axis shows all the physiological behaviours while the *y*-axis shows the accuracies achieved by each physiological behaviour as one feature to predict low, medium and high levels of CL

Answering Q4, we conclude that MPDC for both right and left eye were the most notable features, and BR was also viewed as moderately important for predicting low, medium and high CL.

5 Discussion

Our work presents a framework to detect three levels of CL by analysing physiological data based on eyes and heart when exposed to a task. Our findings show that the RF classification algorithm, in combination with univariate feature selection, considerably outperformed other classification algorithms. Overall, by using the RF classification, low and high levels of CL were predicted with F1-scores higher than 90% and the medium level of CL was predicted with F1-score of 85%. Other classifiers such as NB and DT also predicted the three levels of CL with a high F1-score of over 80%. As stated earlier, prior work has classified two and four levels of CL and they have used one kind of physiological behaviour based on either eyes, or skin conductance, or brain. Nonetheless, it is important to compare our results with the past classifications of CL [7, 14]. For instance, Chen et al. [7] have reported several studies to classify CL using eye- and GSR-based measurements individually. Their classification accuracy, based on two and four levels of task difficulty, on a dataset based on the pupillary response, was able to achieve an accuracy of 79.3% and 45% respectively [5]. Additionally, the accuracy achieved on the dataset based on GSR data for classifying two and four levels of CL, was equal to 71.2% and 40.4% respectively [6]. Other recent work [14] on classifying low, medium and high levels of CL during a driving task, achieved a considerably high accuracy of 86.1% on a dataset based on pupil sizes. However, to the best of our knowledge, our CL classification framework, which gathers data from various physiological measures synchronously, and achieves an accuracy of nearly 92%, has notably outperformed previous classification accuracies. Furthermore, our dataset had a larger number of subjects than the previous works [21]. Beside the notable performance, we emphasise that there is a need to conduct more demonstrations of the framework with a variety of tasks, and under different settings, to further establish the robustness and value of the framework. We plan to demonstrate the framework in different setups in which drivers monitor autonomous vehicle operations and in which supervisors or operators monitor and observe the autonomous operations of robots deployed in offshore environments [20] and smart factories [57]. We conjecture that the framework can estimate CL robustly and accurately under different settings in principle. Hence, it can potentially be applied in a number of domains such as the aerospace domain. In the aerospace domain, it can help to create a system that dynamically adapts the workflow and facilitates the automatic assignment of tasks to supervisors that operate in the control rooms of space stations based on their CL. We believe such a system can enhance the productivity of the supervisors and consequently reduce errors that could, potentially, have vastly expensive consequences [21]. Furthermore, such a system has a wider implication in maintaining the supervisor's mental health and well-being, as a result of managing their CL. In summary, the described framework can, potentially, be applied in different settings in a non-intrusive manner, while collecting the physiological data from high

definition cameras to record the heart-based [45] and eye-based data [54].

In summary, the framework to detect CL works in principle. We further show that the synchronous collection of data based on various physiological behaviours , is the key to its performance in terms of classification accuracy for low and high levels of CL.

6 Conclusion

In this paper, we present our work on the detection of cognitive load (CL) using physiological responses based on eye- and heart-related data. In particular, we introduced a framework that consists of the following steps to detect CL. First, we collect physiological measurements with state-of-the-art, off-the-shelf, sensing technologies, during a task. Second, we apply supervised machine learning algorithms, along with feature elimination. We applied our framework during an experimental setting, in which participants played a game to spot correct and incorrect words and sentences while we collected their eye and heart measurements. Our work confirms that CL was detected with high accuracy. This suggests its potential use in various practical applications, particularly for the purpose of the adaptation of interfaces, to improve user experience, user performance, and to help reduce human errors.

Considering our research questions, we conclude that (1) our task was able to induce CL in the participants, (2) we found that mean pupil diameter change for both left and right eyes increases with each level of CL, (3) the blinking rate decreases from low to high levels of CL , (4) Random Forest was the most accurate classification method, and (5) mean pupil diameter changes for left and right eyes and blinking rate were among the most important features to classify low, medium and high levels of CL. Our findings achieved better accuracy to classify CL in comparison with previous work.

7 Limitations and future work

Our work has the following limitations. In general, our method to detect blinks is effective, however, we intend to improve blink detection when there is more noise in the data. Also, we understand that we had a shorter data gathering time for the rest phase in this study and we believe that recording data for a longer duration in that phase would yield better accuracy.

We demonstrated the framework in the lab under a controlled environment. Therefore, more testing under various tasks is needed to further establish the conformity of the framework. Furthermore, the pool of participants had both native and non-native English language speakers. Although they were all university students, who had passed an English language test and achieved an appropriate standard to gain university admission, we might get different results with all native, or all non-native participant groups for this task. Nonetheless, the paper investigated the framework to classify CL and we recognize that more testing is needed to further establish its robustness.

Our future work is focused on the following aspects. Firstly, we plan to gather data during more diverse tasks. These tasks could consist of playing games. Additionally, it could be an interface showing data in different visualizations and asking individuals to perform various tasks on them. The idea is to make the dataset rich enough to classify CL robustly in various settings. Secondly, we intend to measure physiological behaviours based on skin and brain, to further improve the robustness of the framework to detect CL. We also intend to use speech-based features in the framework. Thirdly, we also intend to address the limitations noted above in our future work. Lastly, our long-term goal is to adapt systems based on the measurement of CL in real time. Therefore, we plan to employ our measurement of CL to create adaptive interfaces, which manages user CL and help improve user performance in various environments.

Acknowledgements The authors would like to thank and acknowledge the reviewers for their insightful comments that have certainly improved the quality of the work described in this paper.

Funding This work received financial support from the ORCA Hub EPSRC (EP/R026173/1, 2017-2021) and consortium partners.

Compliance with ethical standards

Conflict of interest The fourth author is one of the recipients of the EPSRC grant. The first three authors declare that they have no conflict of interest.

References

1. Aga A, Chan A, Narasimhan R (2019) Measuring psychological stress from cardiovascular and activity signals. US Patent App. 10/213,146
2. Bahrin MAK, Othman MF, Azli NN, Talib MF (2016) Industry 4.0: a review on industrial automation and robotic. Jurnal Teknologi 78(6-13):137–143
3. Berthold A, Jameson A (1999) Interpreting symptoms of cognitive load in speech input. In: UM99 user modeling. Springer, pp 235–244
4. Chandler P, Sweller J (1996) Cognitive load while learning to use a computer program. Applied Cognitive Psychology 10(2):151–170
5. Chen F, Zhou J, Wang Y, Yu K, Arshad SZ, Khawaji A, Conway D (2016) Eye-based measures. Springer International Publishing, Cham, pp 75–85. https://doi.org/10.1007/978-3-319-31700-7_4
6. Chen F, Zhou J, Wang Y, Yu K, Arshad SZ, Khawaji A, Conway D (2016) Galvanic skin response-based measures. Springer International Publishing, Cham, pp 87–99. https://doi.org/10.1007/978-3-319-31700-7_5
7. Chen F, Zhou J, Wang Y, Yu K, Arshad SZ, Khawaji A, Conway D (2016) Robust multimodal cognitive load measurement. Springer, Berlin
8. Consortium BNC et al (2007) British national corpus version 3 (bnc xml edition). Distributed by Oxford University Computing Services on behalf of the BNC Consortium. Retrieved February 13, 2012
9. Craig CL, Marshall AL, Sjöström M, Bauman AE, Booth ML, Ainsworth BE, Pratt M, Ekelund U, Yngve A, Sallis JF et al (2003) International physical activity questionnaire: 12-country reliability and validity. Medicine & Science in Sports & Exercise 35(8):1381–1395
10. Cranford KN, Tiettmeyer JM, Chuprinko BC, Jordan S, Grove NP (2014) Measuring load on working memory: the use of heart rate as a means of measuring chemistry students' cognitive load. J Chem Educ 91(5):641–647
11. Field A (2009) Discovering statistics using SPSS, 3rd edn. Sage, London
12. Fogelholm M, Malmberg J, Suni J, Santtila M, Kyröläinen H, Mäntysaari M, Oja P (2006) International physical activity questionnaire: validity against fitness. Medicine and Science in Sports and Exercise 38(4):753–760
13. Frank S, Thompson R (2012) Early effects of word surprisal on pupil size during reading. In: Proceedings of the annual meeting of the cognitive science society, vol 34
14. Fridman L, Reimer B, Mehler B, Freeman WT (2018) Cognitive load estimation in the wild. In: Proceedings of the 2018 CHI conference on human factors in computing Systems. ACM, p 652
15. Fröhlich P, Baldauf M, Meneweger T, Erickson I, Tscheligi M, Gable T, de Ruyter B, Paternò F (2019) Everyday automation experience: non-expert users encountering ubiquitous automated systems. In: Extended abstracts of the 2019 CHI conference on human factors in computing systems, pp 1–8
16. Gevins A, Smith ME, Leong H, McEvoy L, Whitfield S, Du R, Rush G (1998) Monitoring working memory load during computer-based tasks with eeg pattern recognition methods. Human Factors 40(1):79–91
17. Gregoire J, Tuck S, Hughson RL, Yamamoto Y (1996) Heart rate variability at rest and exercise: influence of age, gender, and physical training. Can J Appl Physiol 21(6):455–470
18. Haapalainen E, Kim S, Forlizzi JF, Dey AK (2010) Psycho-physiological measures for assessing cognitive load. In: Proceedings of the 12th ACM international conference on ubiquitous computing. ACM, pp 301–310
19. Hart SG (2006) Nasa-task load index (nasa-tlx); 20 years later. In: Proceedings of the human factors and ergonomics society annual meeting, vol 50. Sage Publications Sage CA, Los Angeles, pp 904–908
20. Hastie H, Robb DA, Lopes J, Ahmad M, Bras PL, Liu X, Petrick R, Lohan K, Chantler MJ (2019) Challenges in collaborative hri for remote robot teams. arXiv:1905.07379
21. Heard J, Harriott CE, Adams JA (2018) A survey of workload assessment algorithms. IEEE Transactions on Human-Machine Systems 48(5):434–451
22. Hjortskov N, Rissén D., Blangsted AK, Fallentin N, Lundberg U, Søgaard K (2004) The effect of mental stress on heart rate variability and blood pressure during computer work. European Journal of Applied Physiology 92(1-2):84–89
23. Holland MK, Tarlow G (1972) Blinking and mental load. Psychol Rep 31(1):119–127
24. HRV E (2018) How do you calculate the hrv score? Webpage. https://help.elitehrv.com/article/54-how-do-you-calculate-the-hrv-score
25. HRV E (2019) Corsense heart rate variability. Webpage. https://elitehrv.com/corsense
26. Jameson A, Kiefer J, Müller C, Großmann-hutter B, Wittig F, Rummer R (2010) Assessment of a user's time pressure and cognitive load on the basis of features of speech. In: Resource-adaptive cognitive processes. Springer, pp 171–204
27. Khawaja MA, Chen F, Marcus N (2010) Using language complexity to measure cognitive load for adaptive interaction design. In: Proceedings of the 15th international conference on intelligent user interfaces. ACM, pp 333–336
28. Khawaja MA, Chen F, Marcus N (2012) Analysis of collaborative communication for linguistic cues of cognitive load. Human factors 54(4):518–529
29. Khawaja MA, Ruiz N, Chen F (2008) Think before you talk: an empirical study of relationship between speech pauses and cognitive load. In: Proceedings of the 20th Australasian conference on computer-human interaction: designing for habitus and habitat. ACM, pp 335–338
30. Kret ME, Sjak-Shie EE (2019) Preprocessing pupil size data: guidelines and code. Behavior Research Methods 51(3):1336–1342
31. Leppink J, Paas F, Van der Vleuten CP, Van Gog T, Van Merriënboer JJ (2013) Development of an instrument for measuring different types of cognitive load. Behavior Research Methods 45(4):1058–1072
32. Leys C, Ley C, Klein O, Bernard P, Licata L (2013) Detecting outliers: do not use standard deviation around the mean, use absolute deviation around the median. J Exp Soc Psychol 49(4):764–766
33. Lobato-Rincón LL, Cabanillas-Campos MDC, Bonnin-Arias C, Chamorro-Gutiérrez E, Murciano-Cespedosa A, Sánchez-Ramos Roda C (2014) Pupillary behavior in relation to wavelength and age. Frontiers in Human Neuroscience 8:221
34. Lopes J, Lohan K, Hastie H (2018) Symptoms of cognitive load in interactions with a dialogue system. In: Proceedings of the workshop on modeling cognitive processes from multimodal data. ACM, p 4
35. Lopes J, Robb DA, Ahmad M, Liu X, Lohan K, Hastie H (2019) Towards a conversational agent for remote robot-human teaming. In: 2019 14th ACM/IEEE international conference on human-robot interaction (HRI). IEEE, pp 548–549
36. Mayer RE, Moreno R (2003) Nine ways to reduce cognitive load in multimedia learning. Educational Psychologist 38(1):43–52
37. Mital A, Govindaraju M (1999) Is it possible to have a single measure for all work? International Journal of Industrial Engineering-Theory Applications and Practice 6(3):190–195

38. Monster A, Chan H, O'Connor D (1978) Long-term trends in human eye blink rate. Biotelemetry and Patient Monitoring 5(4):206–222
39. Morris CH, Leung YK (2006) Pilot mental workload: how well do pilots really perform? Ergonomics 49(15):1581–1596
40. Mukherjee S, Yadav R, Yung I, Zajdel DP, Oken BS (2011) Sensitivity to mental effort and test–retest reliability of heart rate variability measures in healthy seniors. Clin Neurophysiol 122(10):2059–2066
41. Noel JB, Bauer KW Jr, Lanning JW (2005) Improving pilot mental workload classification through feature exploitation and combination: a feasibility study. Computers & Operations Research 32(10):2713–2730
42. Paas F, Tuovinen JE, Tabbers H, Van Gerven PW (2003) Cognitive load measurement as a means to advance cognitive load theory. Educational Psychologist 38(1):63–71
43. Palinko O, Kun AL, Shyrokov A, Heeman P (2010) Estimating cognitive load using remote eye tracking in a driving simulator. In: Proceedings of the 2010 symposium on eye-tracking research & applications. ACM, pp 141–144
44. Pang B, Lee L (2004) A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In: Proceedings of the ACL
45. Rapczynski M, Werner P, Al-Hamadi A (2019) Effects of video encoding on camera-based heart rate estimation. IEEE Trans Biomed Eng 66(12):3360–3370
46. Reilly J, Kelly A, Kim SH, Jett S, Zuckerman B (2018) The human task-evoked pupillary response function is linear: implications for baseline response scaling in pupillometry. Behavior Research Methods, pp 1–14
47. Sabyruly Y, Broz F, Keller I, Lohan K (2015) Gaze and attention during an hri storytelling task. In: Artificial intelligence for human-robot interaction: AAAI 2015 fall symposium series, AI-HRI 2015; Conference date: 12-11-2015 through 14-11-2015
48. Shi Y, Ruiz N, Taib R, Choi E, Chen F (2007) Galvanic skin response (gsr) as an index of cognitive load. In: CHI '07 extended abstracts on human factors in computing systems, CHI EA '07. ACM, New York, pp 2651–2656. https://doi.org/10.1145/1240866.1241057
49. Siegle GJ, Ichikawa N, Steinhauer S (2008) Blink before and after you think: blinks occur prior to and following cognitive load indexed by pupillary responses. Psychophysiology 45(5):679–687
50. Stapel J, Mullakkal-Babu FA, Happee R (2019) Automated driving reduces perceived workload, but monitoring causes higher cognitive load than manual driving. Transportation Research Part F: Traffic Psychology and Behaviour 60:590–605
51. Sweller J (1988) Cognitive load during problem solving: effects on learning. Cognitive Science 12(2):257–285
52. Sweller J (1994) Cognitive load theory, learning difficulty, and instructional design. Learning and instruction 4(4):295–312
53. Sweller J, Van Merrienboer JJ, Paas FG (1998) Cognitive architecture and instructional design. Educational Psychology Review 10(3):251–296
54. Tommaso DD (2018) Tobii pro glasses 2 python controller. Webpage. https://github.com/ddetommaso/TobiiProGlasses2_PyCtrl
55. Wan X, Wang W, Liu J, Tong T (2014) Estimating the sample mean and standard deviation from the sample size, median, range and/or interquartile range. BMC Medical Research Methodology 14(1):135
56. Wilson GF, Russell CA (2003) Real-time assessment of mental workload using psychophysiological measures and artificial neural networks. Human factors 45(4):635–644
57. Wurhofer D, Meneweger T, Meschtscherjakov A, Gerdenitsch C, Tscheligi M Experiencing automation in the factory and automotive domain: differences, similarities, and challenges workshop proceedings everyday automation experience'19 in conjunction with chi'19, May 5th, 2019, Glasgow, UK website: http://everyday-automation.tech-experience.at
58. Yin B, Chen F, Ruiz N, Ambikairajah E (2008) Speech-based cognitive load monitoring system. In: 2008 IEEE international conference on acoustics, speech and signal processing. IEEE, pp 2041–2044
59. Zhang J, Yin Z, Wang R (2015) Recognition of mental workload levels under complex human–machine collaboration by using physiological features and adaptive support vector machines. IEEE Transactions on Human-Machine Systems 45(2):200–214