Photoplethysmogram-based Cognitive Load Assessment Using Multi-Feature Fusion Model

XIAO ZHANG, Beijing University of Technology, China and Tsinghua University, China
YONGQIANG LYU, TONG QU, and PENGFEI QIU, Tsinghua University, China
XIAOMIN LUO, Beijing Genomics Institute, China
JINGYU ZHANG, Chinese Academy of Sciences, China
SHUNJIE FAN, Siemens Ltd., China, China
YUANCHUN SHI, Tsinghua University, China

Cognitive load assessment is crucial for user studies and human–computer interaction designs. As a noninvasive and easy-to-use category of measures, current photoplethysmogram- (PPG) based assessment methods rely on single or small-scale predefined features to recognize responses induced by people's cognitive load, which are not stable in assessment accuracy. In this study, we propose a machine-learning method by using 46 kinds of PPG features together to improve the measurement accuracy for cognitive load. We test the method on 16 participants through the classical n-back tasks (0-back, 1-back, and 2-back). The accuracy of the machine-learning method in differentiating different levels of cognitive loads induced by task difficulties can reach 100% in 0-back vs. 2-back tasks, which outperformed the traditional HRV-based and single-PPG-feature-based methods by 12–55%. When using "leave-one-participant-out" subject-independent cross validation, 87.5% binary classification accuracy was reached, which is at the state-of-the-art level. The proposed method can also support real-time cognitive load assessment by beat-to-beat classifications with better performance than the traditional single-feature-based real-time evaluation method.

CCS Concepts: • **Human-centered computing → Human computer interaction (HCI)**; • **Computing methodologies → Cognitive science**; **Perception**;

Additional Key Words and Phrases: Cognitive load, photoplethysmogram, multi-feature fusion, real-time assessment

ACM Reference format:
Xiao Zhang, Yongqiang Lyu, Tong Qu, Pengfei Qiu, Xiaomin Luo, Jingyu Zhang, Shunjie Fan, and Yuanchun Shi. 2019. Photoplethysmogram-based Cognitive Load Assessment Using Multi-Feature Fusion Model. *ACM Trans. Appl. Percept.* 16, 4, Article 19 (September 2019), 17 pages.
https://doi.org/10.1145/3340962

We gratefully acknowledge the grant from National Key R&D Program of China (Grant No. 2017YFB0403404).
Authors' addresses: X. Zhang, Beijing University of Technology, 100 Pingyue Park, Chaoyang District, Beijing 100124, China; email: zhangxiaohappier@163.com; Y. Lyu (corresponding author), T. Qu, P. Qiu, and Y. Shi, Tsinghua University, 30 Shuangqing Road, Haidian District, Beijing 100084, China; emails: luyq@tsinghua.edu.cn, qutong7@foxmail.com, qpf15@mails.tsinghua.edu.cn, shiyc@tsinghua.edu.cn; X. Luo, Tibet Branch, Beijing Genomics Institute, 189 Jinzhu Road, Lhasa 850032, China; email: luo.xiaomin@139.com; J. Zhang, Chinese Academy of Sciences, 16 Lincui Road, Chaoyang District, Beijing 100101, China; email: zhangjingyu@psych.ac.cn; S. Fan, Siemens Ltd., China, 7 Wangjing Zhonghuan Nanlu, Chaoyang District, Beijing 100102, China; email: shunjie.fan@siemens.com.

1544-3558/2019/09-ART19 $15.00
https://doi.org/10.1145/3340962

1 INTRODUCTION

Cognitive load is the load that uses people's limited cognitive processing capacity to perform a particular task [39, 40]. In a user-centered system, minimizing users' cognitive load can effectively free up mental resources to perform better [38]. Such systems can also help to reduce cognitive load in overload learning scenarios and guide the design of multimedia learning [33]. Real-time cognitive load assessment in driving improves the driving performance and safety [25, 45]. Therefore, it is necessary to measure cognitive load in smart human-centered HCI system.

Various methods have been proposed to measure cognitive load, such as subjective rating scales [39], electrocardiogram (ECG) [12, 20], electroencephalography (EEG) [48, 49], and the galvanic skin response (GSR) [43]. However, most of them have the problem of invasive and reliability and cannot give real-time cognitive load level.

Nowadays, the use of PPG is more and more popular for its non-invasive properties in cognitive load assessment. In 2015, Lyu et al. [32] proposed sVRI, a time-domain characteristic, which has been proved as an effective parameter in statistical results. McDuff et al. used three frequency-domain indices derived from PPG to measure cognitive load.

In this article, we develop a multi-feature-based model to measure cognitive load using PPG, which achieved 100% accuracy in differentiating different levels of cognitive loads induced by task difficulties. The PPG signal is measured from a sensor on a finger clip. The original signal is then segmented into single-pulse waveforms. For every single waveform, we extract 46 features from either the time domain or frequency domain. Finally, a voting-based fusion model can be trained for each individual. We also compare the result with the features of the ECG signal obtained simultaneously, which, in the end, showed that PPG performed much better.

In summary, we focus on the PPG-based multi-feature fusion model to access the user's cognitive load. The main contributions of this article are as follows:

- We collected a set of 46 features extracted from both the frequency domain and time domain of the PPG waveform or its first- and second-order derivatives based on previous literature.
- Based on the 46 features, we build eight cognitive load assessment models using seven popular machine-learning algorithms separately and a voting method over the seven algorithms. The classical n-back task was used to cause different cognitive loads of participants.
- For a single-pulse form, we analyze the real-time property of the proposed models and prove the potential of PPG as a real-time cognitive load assessment.

The article is organized as follows: Section 2 presents previous research in this field, Section 3 introduces the 46 PPG features in assessing cognitive load, and Section 4 describes the experiment. We present and compare the cognitive load assessment result in Section 5 and Section 6. Limitations and feature work are discussed in Section 7.

2 RELATED WORK

2.1 Cognitive Load Measurement Method

The cognitive load measurement methods can be divided into two categories: subjective methods and objective methods [32]. Subjective methods contain self-report scales [10] and questionnaires, which heavily rely on subjective recall after the task and are not real-time measures. Objective methods can be divided into behavioral methods and physiological methods. Behavioral methods include behavior measurement and behavioral performance measurements, such as touch-screen behaviors [14], speech-based indices [8], physical factors of the body [11], and dual-task-based performance measurement [10]. Behavioral methods overcome the problems of intrusiveness and determine the user's cognitive load in real time in spite of having problems in sensitivity and accuracy. The dual-task method is based on the assumption of the existence of limited cognitive resources

that can be allocated flexibly to the primary task and the secondary task performed simultaneously. Dual-task necessitates dedicated design and control of the tasks, and its sensitivity and reliability vary due to variable participants' attention [7]. Behavioral measurements may not reflect direct changes on cognitive load, but rather they require further research on psycho-physiological mechanisms.

A range of physiological signals has been used since the early 1960s [27]. Frequently used signals include EEG [3, 30], ECG, GSR, and PPG. EEG measures electrical activity generated by the brain. The EEG-based method may be the most direct method to assess cognitive load, but it requires electrodes to be placed on the scalp, which might be too intrusive. Haapalainen et al. [17] found that the ECG median absolute deviation and median heat flux measurements can distinguish different cognitive loads accurately. HRV derived from ECG is another widely used index. Different research works have different trends of HRV features during cognitive load [12, 20], and it demands at least 2 minutes of data for analysis. The effect of this method needs further verification. Electrodermal activity (EDA) is a signal controlled only by the sympathetic nervous system. It is easily influenced by sweat and environmental temperature variations. Some medical methods are also utilized in cognition-related research, including neuroimaging techniques (e.g., positron-emission tomography [PET] and functional magnetic resonance imaging [fMRI]) [19] and biomarker techniques (e.g., cortisol or adrenaline examination through saliva or blood samples [31]). Breathing is also one of the most important physiological cues for cognitive load recognition [9]. Skin temperature–based methods are also proposed [1, 37].

2.2 PPG-based Cognitive Load Measurement

In recent years, PPG has been gradually used as a vital signal in cognitive load assessment. In 2015, Lyu et al. [32] proposed PPG-based stress-induced vascular response index (sVRI) to measure cognitive load and stress. sVRI has the potential to be a real-time parameter, but no research has been done. In 2011, Poh et al. [42] extracted PPG from a basic webcam and got indices such as heart rate (HR), respiratory rate, and HRV from the camera PPG. In 2014, McDuff et al. [34] used the indices to measure cognitive load. Then, in 2016, they [36] compared the three indices and came to the conclusion that HRV is a more discriminative indicator of cognitive load. However, as they use HRV indicators, which need at least 2 minutes of data to calculate, they cannot give beat-to-beat cognitive load.

2.3 Multi-Feature Fusion Combining for Cognitive Load Measurement

Using a combination of variables associated with different aspects of cognitive load is expected to improve cognitive load assessment [21]. The combination has two hierarchies; one is among different physiological signals and the other is among various characters of one physiological signals. Hogervorst et al. [21] combined EEG, skin conductance, respiration, ECG, pupil size. and eye blinks for assessment of mental workload. The result showed that combining variables from different sensors did not significantly improve workload assessment over using EEG alone. In that article, they also compared combinations of features from a single sensor with the best-performing single feature. Only a few combinations performed better than one best feature. Wang et al. [50] combined features from ECG, EOG, RSP, GSR, and PPG to measure cross-task mental workload. There were 5 features from PPG and 37 features from the other signals. However, there is no study to our knowledge that has systematically studied only a PPG-based multi-feature (more than 10) fusion model for cognitive load measurement.

3 MULTI-FEATURE-BASED COGNITIVE LOAD MEASUREMENT METHOD

3.1 Filter

As we can see in Table 1, most of the features are morphological features. The shape influences the feature's value. The movement of hands can cause motion artifacts. That is, noise directly affected the availability of the algorithm. In that case, precise filtering is needed. The original PPG signals were first smoothed by an FIR bandpass filter (0.5–5Hz) and then periodically segmented into single-PPG-pulse waveforms. After period segmentation,

Table 1. Definition of PPG-based Features (PPG-45)

#	Feature	Description	#	Feature	Description
1	x	Systolic peak	2	y	Diastolic peak
3	z	Dicrotic notch	4	t_{pi}	Pulse interval
5	y/x	Augmentation index	6	$(x-y)/x$	Relative augmentation index
7	z/x		8	$(y-z)/x$	
9	t_1	Systolic peak time	10	t_2	Diastolic peak time
11	t_3	Dicrotic notch time	12	ΔT	Time between systolic and diastolic peaks
13	w	Time between half systolic peak points	14	$A_3/(A_1+A_2)$	Inflection point area ratio
15	$(A_2+A_3)/A_1$	Stress-Induced Vascular Response Index (sVRI)	16	t_1/x	Systolic peak rising slope
17	$y/(t_{pi}-t_3)$	Diastolic peak falling slope	18	t_1/t_{pi}	
19	t_2/t_{pi}		20	t_3/t_{pi}	
21	$\Delta T/t_{pi}$		22	t_{a1}	
23	t_{b1}		24	t_{e1}	
25	t_{f1}		26	b_2/a_2	
27	e_2/a_2		28	$(b_2+e_2)/a_2$	
29	t_{a2}		30	t_{b2}	
31	t_{a1}/t_{pi}		32	t_{b1}/t_{pi}	
33	t_{e1}/t_{pi}		34	t_{f1}/t_{pi}	
35	t_{a2}/t_{pi}		36	t_{b2}/t_{pi}	
37	$(t_{a1}+t_{a2})/t_{pi}$		38	$(t_{b1}+t_{b2})/t_{pi}$	
39	$(t_{e1}+t_2)/t_{pi}$		40	$(t_{f1}+t_3)/t_{pi}$	
41	f_{base}	Fundamental component frequency	42	$\lvert s_{base}\rvert$	Fundamental component magnitude
43	f_{2nd}	Second-harmonic frequency	44	$\lvert s_{2nd}\rvert$	Second-harmonic magnitude
45	f_{3rd}	Third-harmonic frequency	46	$\lvert s_{3rd}\rvert$	Third-harmonic magnitude

we use the Least-Mean-Square (LMS) filter. However, after two filtering processes, some waves still have great errors, which need to be checked for eligibility. These waveforms are screened according to the following criteria:

(1) the pulse interval t_{pi} was between 0.5s and 1.2s,
(2) the systolic peak time t_1 was less than half of the pulse interval $t_{pi}/2$,
(3) the minimum pulse value only occurred on the starting point or the ending point of the waveform segment,
(4) the first derivative of the rising edge (between the starting point and the systolic peak point) was greater than zero, i.e., monotonically increasing,
(5) the amplitude difference between the starting point and the ending point was not greater than 1/10 of the overall amplitude of the waveform segment.

Only when the above five conditions are met at the same time can the waveform be used as a qualified waveform for feature extraction in the next step, feature extraction.

3.2 Feature Set Extraction

In this section, we empirically selected a set of PPG-based features to assess cognitive load (through mental effort) and stress. As the PPG is used for cognitive load measurement only in recent years, there is little feature

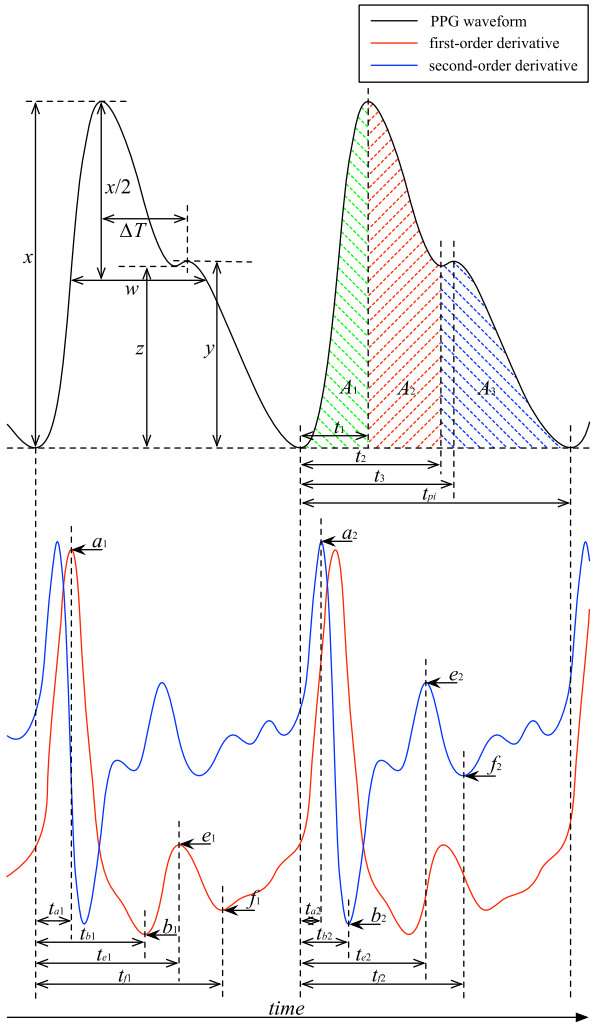


Fig. 1. Considered feature points on PPG pulse waveform and its corresponding first-order and second-order derivatives [28].

used in cognitive research. Therefore, we review a wide range of features not only in the field of cognition but also relevant for biomedical science, biometric authentication, and so on. A good starting point is given in a biometric recognition study by Resit et al. [28], which we extended by several important works in the cognitive load measurement. We generally distinguish features in the time domain and frequency domain. At last, for each valid PPG pulse waveform, a total of 46 features are considered, including 39 time-domain features and 6 frequency-domain features. Definitions of all 46 PPG-based features (PPG-46) are summarized in Table 1, where #1–40 are time-domain features and #41–46 are frequency-domain features.

3.2.1 Time-Domain Features. The time-domain features are extracted from the PPG pulse waveform and its corresponding first-order and second-order derivatives.

Lyu et al. [32] have divided the pulse area into two parts at the systolic peak and defined sVRI as the rate of the two areas (known as $(A_2 + A_3)/A_1$ in Figure 1). sVRI has been proven as a reliable index of cognitive load.

Similarly to sVRI, the IPA is defined as the ratio of the two areas that divided the pulse area at the dicrotic notch (known as $A_3/(A_1 + A_2)$ in Table 1), and it is used as an indicator of total peripheral resistance [47]. The systolic peak (known as x in Table 1) derived from the pulse waveform, as an indicator of vasoconstriction in peripheral blood circulation, was usually analyzed as the effect of peripheral sympathetic nerve activation in some literature [13, 18, 22]. The systolic peak time (known as t_1 in Table 1) was proven to be a useful feature for cardiovasular disease classification [2]. The peak–peak interval correlates closely with the R-R interval in ECG signals [26].

Since the PPG pulse waveform is collected as discrete data, the first-order and second-order derivatives are achieved by calculating its first and second discrete differences, respectively. The PPG pulse waveform is denoted as

$$x(n). \tag{1}$$

Therefore, the first-order derivative is

$$y_1(n) = x(n+1) - x(n), \tag{2}$$

and the second-order derivative is

$$y_2(n) = x(n+2) - 2x(n+1) + x(n). \tag{3}$$

The selected feature points on the curves of the pulse waveform and its corresponding first-order and second-order derivatives are illustrated in Figure 1, where

(1) on the pulse waveform curve: x is the systolic peak point, y is the diastolic peak point, z is the dicrotic notch point, t_{pi} is the pulse interval, ΔT is the time between systolic and diastolic peaks, w is the time between half systolic peak points, and A_1, A_2 denote the corresponding marked areas.
(2) on the first-order derivative curve: a_1 and b_1 are the first maximum and minimum points, respectively; e_1 and f_1 are the first maximum and minimum points after the dicrotic notch point, respectively.
(3) on the second-order derivative curve: a_2 and b_2 are the first maximum and minimum points, respectively; e_2 and f_2 are the first maximum and minimum points after b_1, respectively.

3.2.2 Frequency-Domain Features. A total of six frequency-domain features are obtained by performing FFT of the PPG pulse waveform, which are the frequency of fundamental component f_{base}, the magnitude of fundamental component $|s_{base}|$, the frequency of the second harmonic f_{2nd}, the magnitude of the second harmonic $|s_{2nd}|$, the frequency of the third harmonic f_{3rd}, and the magnitude of the third harmonic $|s_{3rd}|$.

4 METHODOLOGY

This study employed the n-back task to impose cognitive load or, more specifically, memory load [29] on the participants. This task had low requirements for learning [6]. While the participants were performing the tasks, their performance data and physiological data (PPG and ECG) were recorded simultaneously. Questionnaires from the participants about their experience were also collected.

4.1 Participants

To reduce the effects on performance caused by individual capabilities or backgrounds [6], we recruited 16 university students with similar educational backgrounds and memory capabilities. None of the participants had prior experience with the experimental content. Participants were aged between 19 and 27 years old (mean age 22.81), 8 female and 8 male. All participants were healthy and did not use any medications. All were right-handed and had a normal or corrected-to-normal vision. The experiment was performed in accordance with the local ethics guidelines. Before experiment participants were gave written informed consent. Each participant was paid 150 RMB.

4.2 Materials

PPG and ECG of the participants were measured throughout the experiment. PPG was measured using an HKG 07C infrared digital pulse sensor (Hefei Huake Electronic Technology Research Institute, Hefei, China) at a sampling rate of 200 Hz. For each participant, the PPG sensor was placed on the left index finger. ECG was measured using a BIOPAC MP150 device (BIOPAC Systems Inc., USA) at a sampling rate of 500 Hz. For ECG measurement, self-adhesive 1 1/2″ electrodes with 7% chloride wet gel were attached to the participant's chest in a standard configuration of leads.

Stimuli (letters), subjective workload scales and announcements about the type of the n-back task to follow were presented on a Surface Pro tablet device. Feedback about task performance was presented through the device's speakers in the form of beeps. Participants used the Surface Type Cover keyboard to indicate whether presented letters were targets or non-targets—the left arrow key was used for "target" and the right arrow key for "non-target." Participants used the Surface Type Cover touchpad to rate subjective workload on a scale (Rating Scale Mental Effort (RSME)) between the stimulus blocks.

We used the RSME [51] scale to measure subjectively experienced mental effort. This scale runs from 0 to 150 with higher values reflecting higher workload. It has nine descriptors along the axis, e.g., "absolutely no effort" at value 2 and "rather much effort" at value 57. This simple one-dimensional scale is more sensitive than the often-used NASA-TLX [46].

The experiment took place in a sound-attenuated, temperature-controlled, and electrically shielded room. Room temperature and humidity during the experiment were held constant.

4.3 Task

Participants viewed letters, successively presented on a screen. For each letter, they pressed a button to indicate whether the letter was a target or a non-target. In the 0-back condition, the letter x is the target. In the 1-back condition, a letter is a target when it is the same as the one before. In the 2-back condition, a letter is a target when it is the same as two letters before. With this version of the n-back task, the level of workload is varied without varying visual input or frequency and type of motor output (button presses). A 3-back condition was not used, due to evidence that many participants find it too difficult and tend to give up [4, 23].

Participants were informed after every button press whether it was a correct decision by a high (correct) or a low (incorrect) pitched tone. This was intended to help the participant, who in our experiment switched rather often between n-back conditions, and to increase the likelihood that participants would decide to invest effort, since the participant knew the experiment leader would hear the sounds as well.

4.4 Stimuli

The letters used in the n-back task were black (font style: sans serif, approximately 3cm high) and were presented on a white background. The letters were presented for 500 ms followed by a 2,000 ms inter-stimulus interval during which the letter was replaced by a fixation cross. In all conditions, 33% of letters were targets. Except for the letter x in the 0-back task, letters were randomly selected from English consonants. Vowels were excluded to reduce the likeliness of participants developing chunking strategies that reduce mental effort, as suggested in Reference [16].

4.5 Design

The three conditions (0-back, 1-back, 2-back) were presented in 2-min blocks divided across four sessions. Each session consisted of one repetition of each of the three blocks. Thus, for each of the three conditions participants performed 4 sessions $\times$ 1 repetitions = 4 blocks. In each block, 48 letters were presented, 16 of which were targets. The blocks were presented in pseudorandom order. Before each session was a baseline block of 5 minutes in which the participant quietly fixated a cross on the screen. The experiment protocol is shown in Figure 2.

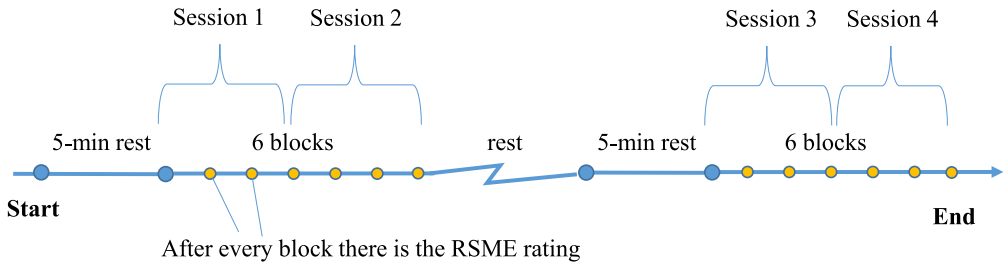


Fig. 2. The experiment protocol.

Table 2. Sensors and Analyzed Features

Sensor	Feature	Dimension
PPG finger clip	PPG-46	46^*N_{pulse}
	sVRI	1^*N_{pulse}
ECG Electrodes	Heart rate (RRI)	1
	RMSSD	1
	Mid-frequency HRV	1
	High-frequency HRV	1

4.6 Procedure

After entering the lab, participants read and were explained about the experimental procedure. They then signed an informed consent form. The three conditions were practiced up to the point that the participant was familiar with the task. Regardless of this, all participants completed at least one block of the 2-back task to also practice the RSME rating that appeared at the end of the block. It was stressed that the 2-back task could be difficult, but that even when the participant thought it was too difficult he or she should keep trying to do as well as possible. Participants were asked to avoid movement as much as possible while performing the task but they can use the breaks in between the blocks to make necessary movements. Before the start of each block, the participant was informed about the nature of the block (rest, 0-back, 1-back, or 2-back) via the tablet screen. After each block, the RSME scale was presented and the participant rated subjective mental effort by clicking the appropriate location on the scale using the mouse. The next block started after the participant indicated to be ready by pressing a button. Between sessions, participants had longer breaks, chatting with the experiment leader or having a drink.

4.7 Analysis

All the analysed features are shown in Table 2. For each valid PPG pulse waveform, the PPG-46-feature set was extracted. Among the features, the PPG-based Stress-Induced Vascular Response Index (sVRI) is suggested to be a sensitive, reliable, and usable physiological measure for assessing cognitive load and stress [32]. sVRI was used as a single feature for analysis.

As a measure of heart rate, we determined the mean RRI for each block. RRI is the interval between successive heartbeats or, more precisely, the interval between subsequent R-peaks in the ECG. Three measures of heart rate variability were computed. The root mean squared successive difference (RMSSD) between the RRIs reflects high frequency heart rate variability [15]. High-frequency heart rate variability was also computed as the power in the high-frequency range (0.15–0.5Hz) of the RRI over time using Welch's method applied after spline interpolation; similarly, for mid-frequency heart rate variability the power in the frequency range of 0.07–0.15Hz was used.

4.8 Classification

The first three sessions, containing three blocks of each n-back condition, were used to train the model parameters to individual participants. The last session was used to evaluate the model's classification accuracy. In that case, there were no data both in the trainning set and the test set. Then it can avoid over-fitted problems. As a default, the classification models were trained and applied to distinguish between 0- and 2-back blocks, each containing 2 minutes of data or 48 trials (letters). Average classification performance (*fraction correct* in the last session) over all participants was used as a measure of model performance. As all the PPG features have strict time synchrony, feature level fusion was used. Feature vectors were constructed for each of the PPG pulse waveforms. For instance, the feature vectors used for the model that includes all PPG-46 value over 2-min blocks of data contains $46 \times N_{pulse}$ features $\times$ 4 blocks (4 sessions $\times$ 1 block) (see Table 2, first row), where N_{pulse} denotes the number of valid PPG pulse waveforms in a 2-min block, usually at the range of 120–200. The data from the first three sessions were used to train a classifier model for each individual participant. The features were standardized to have mean 0 and standard deviation 1 on the basis of data from the training set. The same standardization transformation was applied to the test data (the data of the last session). After training the model using the training data (the first 3 blocks of $46 \times N_{pulse}$ features in the example above), the classification was applied to the test data(the last block of $46 \times N_{pulse}$ features), and the performance score of each of the individual models was determined. Finally, overall performance is calculated by taking the average score overall individual models.

Classification accuracy was determined for a range of models differing in the (types of) features that were included in the model, differing in the type of classifier and differing in the fusion rule that was used. As our datasets are relatively small, simple classifier may achieve better performance than many complex models, because complex models use too many assumptions, resulting in under ftting. We used simple version of logistic regression (LR), support vector machine (SVM), Gaussian naïve bayes (GNB), decision tree (DT), random forest (RF), adaboost, and gradient boosting (GB) as representing standard models. And, to obtain confidence measures that can be used to fuse information, we used a voting model. Classification was performed using scikit-learn [41].

5 RESULTS

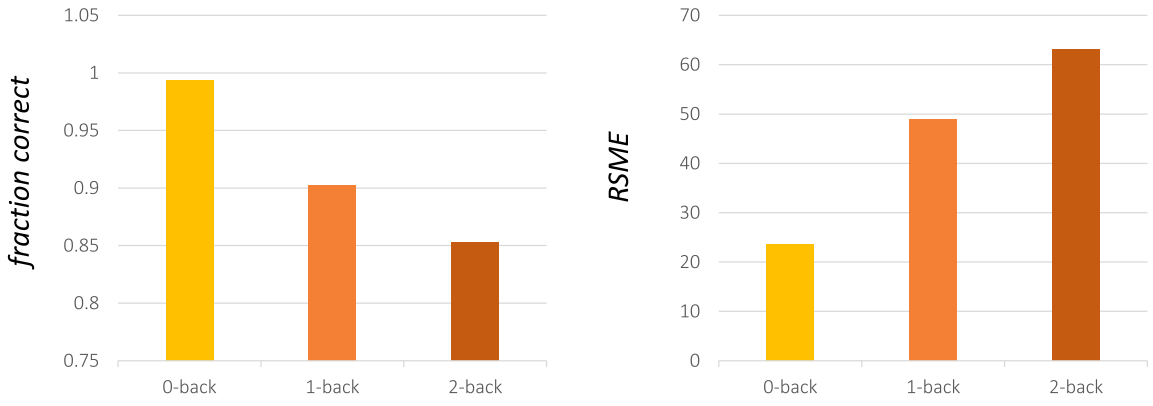
PPG signals were collected from 16 participants. We analyzed the statistical results and classification results.

5.1 Statistical Results

5.1.1 Performance and Subjective Rating Data of the Subjects. First, the performance data and subjective rating data were analyzed to confirm the validity of the n-back task. The result was shown in Figure 3. We used Brouwer's [5] *fraction correct* parameter to represent the behavioral performance. *fraction correct* was defined as the total number of right judgment divided by the total number of stimuli. The mean of the *fraction correct* was maximal for the 0-back task condition (M = 0.99, SD = 0.01), intermediate for the 1-back task condition (M = 0.90, SD = 0.15), and minimal for the 2-back task condition (M = 0.85, SD = 0.16), $F(2,30) = 8.527$, $p = 0.001$. The result showed that all three levels are significantly different from each other.

Subjective rating mental effort was measured by RSME as mentioned in Section 4.2. The average scores of RSME increased with memory load of 23.54, 49.03, and 63.15 for the 0-back, 1-back, and 2-back conditions, respectively, $F(1.133, 16.997) = 24.167$, $p < 0.001$, which reconfirmed the significant difference among different conditions.

5.1.2 Statistical Results of the Traditional Parameters. Table 3 shows the mean and variance of sVRI, RRI, HRV LF, and HRV HF for all the participants during each of the conditions. In addition, Table 4 show the effects of the task difficulty of those features. The average sVRI was significantly different in any case, showing that sVRI alone may be a very discriminative indicator of cognitive load. Among the features of ECG, only RRI showed significant difference during different n-back task. The main reason was that every n-back task lasts 2 minutes and it was too short for HRV data analysis. Usually, HRV analysis needs 3–5 minutes of data. That is, ECG was not a suitable signal for real-time cognitive load measurement.



(a) fraction correct (b) RSME

Fig. 3. Performance and subjective rating result.

Table 3. Results of Physiological Measures for 0-back, 1-back, and 2-back, Respectively

Feature	0-back	1-back	2-back
PPG: sVRI	0.75±0.11	0.77±0.12	0.80±0.12
ECG:RRI	0.78±0.09	0.77±0.09	0.75±0.09
ECG:LF(n.u.)	44.72±15.24	41.75±12.34	43.30±15.03
ECG:HF(n.u.)	45.51±10.94	47.08±10.11	44.00±12.70

Table 4. Within-subject Effects of the Task Difficulty by 3*1 Repeated-measures ANOVA: p and F Values for the Main Effects of Task Difficulty, η^2 for the Effect Sizes

Feature	p	F	df1,df2	η^2
PPG: sVRI**	0.000	29.911	2,30	0.666
ECG:RRI*	0.001	9.189	2.30	0.380
ECG:LF(n.u.)	0.505	0.639	1.627,24.041	0.041
ECG:HF(n.u.)	0.464	0.717	1.539,23.086	0.046

5.2 Classification Results

The first three session data were used to train the model and the last session data were used as the test set for each participant's model. That is, no test data were in the training data, so this method can be used to avoid the overfitting problem. Seven types of classical classifiers and one voting classifier were used.

5.2.1 Real-time Result. PPG signals were composed of a heartbeat pulse waveform. Every waveform can give out the parameters. Therefore, we can use one waveform for real-time cognitive load assessment. The last group of conditions was used as a test set, and the former groups were used to train the models. Every pulse waveform gives a prediction of the current cognitive load level. And the accuracy was the average rate of right predicted

Table 5. Real-time Classification Results (0-back vs. 1-back tasks)

Feature	LR	SVM	GNB	DT	RF	AdaBoost	GB	Voting
PPG-46	56.30%	52.82%	54.21%	52.55%	53.84%	54.23%	53.64%	54.55%
sVRI	56.70%	55.17%	54.04%	54.03%	53.25%	54.56%	54.66%	53.96%

Table 6. Real-time Classification Results (0-back vs. 2-back tasks)

Feature	LR	SVM	GNB	DT	RF	AdaBoost	GB	Voting
PPG-46	71.16%	70.01%	69.41%	66.97%	73.05%	71.24%	71.35%	71.84%
sVRI	58.71%	61.34%	60.42%	61.35%	60.06%	59.48%	60.39%	60.82%

Table 7. Real-time Classification Results (1-back vs. 2-back tasks)

Feature	LR	SVM	GNB	DT	RF	AdaBoost	GB	Voting
PPG-46	60.59%	59.51%	60.30%	57.59%	58.37%	59.57%	58.80%	59.01%
sVRI	53.50%	52.69%	52.83%	52.76%	51.76%	52.54%	51.64%	52.78%

Table 8. Two-minute Classification Results (0-back vs. 1-back tasks)

Feature	LR	SVM	GNB	DT	RF	AdaBoost	GB	Voting
PPG-46	68.75%	50.00%	62.50%	56.25%	68.75%	62.50%	68.75%	62.50%
sVRI	56.25%	62.50%	56.25%	62.50%	50.00%	68.75%	62.50%	50.00%

Table 9. Two-minute Classification Results (1- vs. 2-back tasks)

Feature	LR	SVM	GNB	DT	RF	AdaBoost	GB	Voting
PPG-46	68.75%	68.75%	68.75%	68.75%	62.50%	62.50%	68.75%	62.50%
sVRI	68.75%	68.75%	62.50%	75.00%	68.75%	62.50%	62.50%	68.75%

pulse to all test pulse number. Table 5, Table 6, and Table 7 show the classification accuracy of the features for two classes, separately for the seven classifiers and a voting model. The 0-back vs. 1-back tasks are shown in Table 5, the 1-back vs. 2-back tasks are shown in Table 7, and the 0-back vs. 2-back tasks are shown in Table 6. These are two-class cases with balanced class sizes, and therefore a random chance prediction would be 50%.

In the three conditions, multi-features models always performed better than single-feature models. In 0-back vs. 2-back tasks classification, the multi-features models got the best accuracy (73.05%). The accuracy of the classifiers of the 0-back vs. 1-back was lower than the accuracy of 1-back vs. 2-back both in the multi-feature models and in the single-feature model.

5.2.2 Two-minute Result. In the 2-minute classify condition, we used all the two-minute data to judge which n-back task was done in the 2 minutes. Every pulse waveform parameters could give a judgment, and the judgment with the largest number of votes would be selected as the judgment for the two-minute n-back task. Tables 8 to 10 show the cognitive load classification accuracies of the features for two classes, separately for the seven classifiers and one voting model. There are three conditions, 0-back vs. 1-back tasks shown in Table 8, 1-back vs. 2-back tasks shown in Table 9, and 0-back vs. 2-back tasks shown in Table 10. These are two-class cases with balanced class sizes, and therefore a random chance prediction would be 50%.

Table 10. Two-minute Classification Results (0- vs. 2-back tasks)

Feature	LR	SVM	GNB	DT	RF	AdaBoost	GB	Voting
PPG-46	100.0%	93.75%	93.75%	93.75%	100.0%	100.0%	93.75%	93.75%
sVRI	93.75%	87.50%	75.00%	87.50%	81.25%	81.25%	87.50%	81.25%

The best classification of 0-back and 1-back was 68.75%, which was achieved by 46 PPG features using LR. Between 1 and 2 back the best accuracy was 75%, achieved by sVRI. Between 0 and 2, many methods can achieve 100%. sVRI achieved a good result in statistics, but the classification result was low. tThe multi-feature model performed better. As we can see, in all three conditions, models based on the features of ECG perform poorly.

The classification results of 1- vs. 2 back tasks was higher than the results of 0- vs. 1-back tasks. And the 0-vs. 2-back tasks got the best performance. The results were consistent with the statistical results. It shows that in 2-back conditions, participants need to pay much more cognitive resources on the task than 0-back and 1-back conditions. And the requirement of 0-back and 1-back tasks were close. Many participants also expressed the same feeling after the experiments.

On the whole, the one-pulse waveform performed worse than 2-min data but have similar trends.

6 DISCUSSION

6.1 Subject-independent Characteristic

Subject independent characteristic was also being tested using the leave-one-subject-out cross-validation method. The training sets and testing sets were participant independent so that nobody in the testing set was also in the training set. The person-independent testing was performed by withholding the data for one subject in the test set and leaving all the remaining data for training. We repeated this 16 times, once for each participant. The classification accuracies of all rounds were averaged. The voting classifier prediction accuracy was very good even with a challenging person-independent training scheme; 87.5% of sessions were correctly labeled as 0-back or 2-back. The result of the 1-back vs. 2-back was a little worse, only 68.7% with the random forest classifier. It is hard to classify 0-back vs. 1-back with an accuracy of 55%, similarly as in the subject-dependent result.

6.2 PPG Feature Selection and Analysis

In some conditions, not all the features have a positive effect on the classification. What was worse, too many features can easily cause the over-fitting problem. In that case, we use feature selection SVM-Recursive Feature elimination (SVM-RFE) with the class of feature selection RFECV in Python Sklearn to run on the real-time dataset. SVM was used as a base model for providing information about feature importance. In each eliminates, the lease weight feature was abandoned and evaluates the current feature subset using 10-fold cross-validation method. At last, the features sub with the highest accuracy is retained as the feature subset of the final selection.

Figure 4 shows the cross-validation accuracy score on the training set when the features of different numbers (0∼46) are selected respectively in three conditions. In the 1-vs. 2-back condition, in the stage of 1∼9, the classification accuracy rate has been significantly improved. Then the accuracy rate tended to be slightly higher and stable. The 0- vs. 1-back condition was similar with the 1- vs. 2-back condition and they both got the best classification result with 46-features models. In 0 vs. 2-back condition, it is not difficult to find that in the stage of 1∼10, the classification accuracy rate has been significantly improved. In the 10∼18 stage, the classification accuracy increased in a small range, while in the 18∼34 stage, the accuracy rate of classification tended to be slightly lower. In 34∼46 stage, the accuracy tended to be stable. The feature subset, which has the highest accuracy rate (82.9%) of cross validation, includes 18 features. With the 46-feature model, the result was (82.6%), which was close to the best result. Considering the overall feature of the number is far less than the number of samples ($46 \ll 27{,}600$), therefore, we used all 46 features, in the final classification.

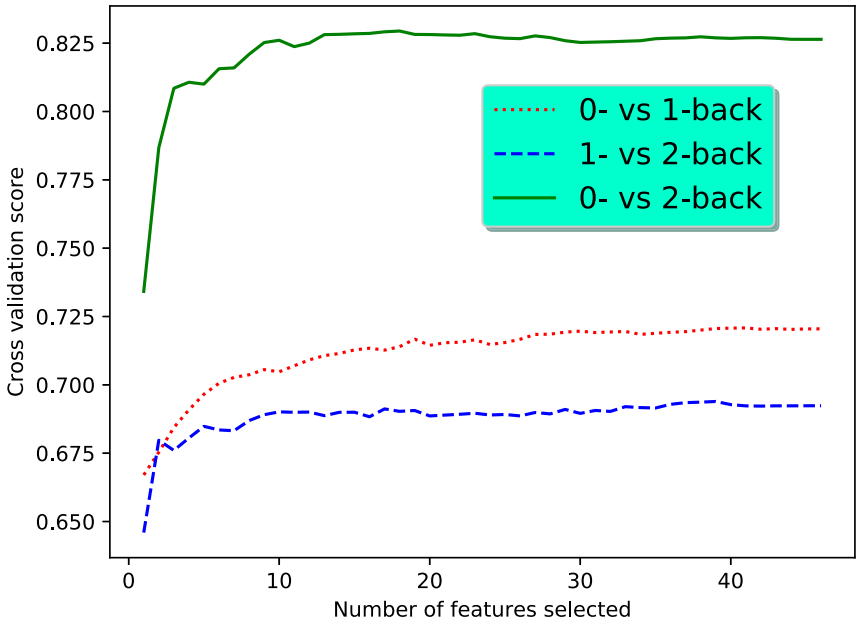


Fig. 4. Cross validation scores of different sizes of feature set.

Table 11. 10 Features Ranked Top for the Best Performance in the 0-vs 2-back Conditions

Numbers	features	Physical description
7	z/x	
27	e_2/a_2	
8	$(y-z)/x$	
1	x	Systolic peak
16	t_1/x	Systolic peak rising slope
40	$(t_{f1}+t_3)/t_{pi}$	
42	$\lvert s_{base}\rvert$	Fundamental component magnitude
15	$(A_2+A_3)/A_1$	Stress-Induced Vascular Response Index (sVRI)
3	z	Dicrotic notch
29	t_{a2}	

We also ranked the 46 features using model-based ranking. Ten folders cross-validation SVM classifier was used for every feature. The top 10 features are shown in Table 11.

6.3 Comparison with the Previous Study

The experimental design of this article is almost entirely based on Reference [21]. The main difference is that the experiment only collects 4 blocks for each n-back, and the experiment in Reference [21] collects 8 blocks for each n-back. The EEG, skin electricity, breathing, ECG, pupil size and blink parameters of 14 subjects were measured in Reference [21]. The data of the first 6 blocks were used for training, and the last 2 blocks were used for testing.

In 0-back and 2-back condition, the EEG signal performed best. The classification accuracy was 86%. The rest of the signals performed poorly. Although the data in this article are half the time of Reference [21], the accuracy of multi-feature classification based on PPG signals is much higher than that of Reference [21]. In the same case, the accuracy of this article is 100% using SVM classification as shown in Table 10. Compared with skin electricity, breathing, and eye movement, PPG's multi-feature measurement's advantage is more obvious. Although EEG directly measures brain activity, the signal is noisy, and many components of EEG are not very clear. Only 966 features can be extracted from the 2-minute EEG data, but the 2-minute PPG data extract $46 \times (2 \times (60\sim100))$, which is an order of magnitude higher than EEG and is, therefore, more suitable for machine-learning classification of cognitive load. More importnatly, compared with EEG, PPG is low invasive, easy to use, and low cost. It is more suitable for the ubiquitous environment.

6.4 Real-time Property

In the real-time processing, there are two main computing periods, the preprocess period and the predicting period. The preprocess period includes filter, segment, and feature extracting. In the model predicting process, we take the SVM algorithm as an example to analyze the complexity of the method. We used linear SVM. If the number of examples is n and each example has N features, then the training time is $O(nN)$ for classification problems [24]. For our problem, $n = 3block * 2min * Heartrate(pulse/min)$, N = 46, compared with other problems (n in the millions), such as text classification and word-sense disambiguation, and our data size is much smaller.

In real-time cognitive load assessment, we used a 5s-window for the data process and measured cognitive load in every second. The average execution time for every window data was 0.005991s on a ThinkPad T470p laptop equipped with an i7-7700HQ processor and 8G memory. The delay was so short that they would not affect our assessment at a 1s granularity. Thus, according to Shin and Ramanathan [44], results of the algorithm are available in real time.

Other real-time cognitive load assessment methods usually use several physiological signals, and the data amount was larger, especially when using EEG signals, and, therefore, the processing time was usually much longer than our method, for example, the execution time was 1.67s in Reference [50] .

7 CONCLUSION

This article showed the performance of PPG-based multi-feature fusion model for cognitive load measurement. The classic n-back task was used to induced different cognitive load. Forty-six PPG-based features were extracted from PPG waveform, the first-order derivative and the second-order derivative. We compared the 46-feature fusion model with the one-PPG-feature model, and the former performed much better. In the 0-back vs. 2-back mode, the classification results even reached 100%. The result also showed that the PPG features performed much better than ECG.

There are some limitations of our method that need to be focused on in future work. First, we only tested the method on young students. More experiments need to be done for people of different ages. Second, our method still needs PPG sensors to be clamped on the finger, which may bring some discomfort to the participants. Since the approach of non-contact acquisition of PPG has been proposed in some literature [35, 42], we will test it in a noninvasive way in the future.

As future work, we will focus on the current limitations of this method and improve the usability for the pervasive environment, such as real-time remote cognitive load measurement.

ACKNOWLEDGMENTS

We thank all the volunteers for their participation in our experiment. We thank all the anonymous reviewers for their insightful comments and helpful suggestions for improving the initial draft of this article.

REFERENCES

[1] Yomna Abdelrahman, Eduardo Velloso, Tilman Dingler, Albrecht Schmidt, and Frank Vetere. 2017. Cognitive heat: Exploring the usage of thermal imaging to unobtrusively estimate cognitive load. *Proc. ACM Interact. Mobile Wear. Ubiq. Technol.* 1, 3 (2017), 33.

[2] S. R. Alty, N. Angaritajaimes, S. C. Millasseau, and P. J. Chowienczyk. 2007. Predicting arterial stiffness from the digital volume pulse waveform. *IEEE Trans. Bio-med. Eng.* 54, 12 (2007), 2268–75.

[3] Pavlo Antonenko, Fred Paas, Roland Grabner, and Tamara Van Gog. 2010. Using electroencephalography to measure cognitive load. *Educ. Psychol. Rev.* 22, 4 (2010), 425–438.

[4] Hasan Ayaz, Meltem Izzetoglu, Scott Bunce, Terry Heiman-Patterson, and Banu Onaral. 2007. Detecting cognitive activity related hemodynamic signal for brain computer interface using functional near infrared spectroscopy. In *Proceedings of the 3rd International IEEE/EMBS Conference on Neural Engineering 2007 (CNE'07)*. IEEE, 342–345.

[5] Annemarie Brouwer, Maarten A. Hogervorst, Jan B. F. Van Erp, Tobias Heffelaar, Patrick H Zimmerman, and Robert Oostenveld. 2012. Estimating workload using EEG spectral power and ERPs in the n-back task. *J. Neur. Eng.* 9, 4 (2012), 045008.

[6] Anne-Marie Brouwer, Maarten A. Hogervorst, Michael Holewijn, and Jan B. F. van Erp. 2014. Evidence for effects of task difficulty but not learning on neurophysiological variables associated with effort. *Int. J. Psychophysiol.* 93, 2 (2014), 242–252.

[7] Roland Brunken, Jan L. Plass, and Detlev Leutner. 2003. Direct measurement of cognitive load in multimedia learning. *Educ. Psychol.* 38, 1 (2003), 53–61.

[8] Fang Chen, Natalie Ruiz, Eric Choi, Julien Epps, M. Asif Khawaja, Ronnie Taib, Bo Yin, and Yang Wang. 2012. Multimodal behavior and interaction as indicators of cognitive load. *ACM Trans. Interact. Intell. Syst.* 2, 4, Article 22 (Jan. 2012), 36 pages. DOI: https://doi.org/10.1145/2395123.2395127

[9] Y. Cho, N. Bianchi-Berthouze, and S. J. Julier. 2017. DeepBreath: Deep learning of breathing patterns for automatic stress recognition using low-cost thermal imaging in unconstrained settings. In *Proceedings of the 2017 7th International Conference on Affective Computing and Intelligent Interaction (ACII'17)*. IEEE, 456–463. DOI: https://doi.org/10.1109/ACII.2017.8273639

[10] Krista E. DeLeeuw and Richard E. Mayer. 2008. A comparison of three measures of cognitive load: Evidence for separable measures of intrinsic, extraneous, and germane load.*J. Educ. Psychol.* 100, 1 (2008), 223.

[11] Jack Dennerlein, Theodore Becker, Peter Johnson, Carson Reynolds, and Rosalind W. Picard. 2003. Frustrating computer users increases exposure to physical factors. In *Proceedings of the International Ergonomics Association*.

[12] Gautier Durantin, J.-F. Gagnon, Sébastien Tremblay, and Frédéric Dehais. 2014. Using near infrared spectroscopy and heart rate variability to detect mental overload. *Behav. Brain Res.* 259, 1 (2014), 16–23.

[13] M. Elgendi. 2012. On the analysis of fingertip photoplethysmogram signals. *Curr. Cardiol. Rev.* 8, 1 (2012).

[14] Yuan Gao, Nadia Bianchiberthouze, and Hongying Meng. 2012. What does touch tell us about emotions in touchscreen-based gameplay? *ACM Trans. Comput.-Hum. Interact.* 19, 4 (2012), 1–30.

[15] Annebet D. Goedhart, Sophie Van Der Sluis, Jan H. Houtveen, Gonneke Willemsen, and Eco J. C. De Geus. 2007. Comparison of time and frequency domain measures of RSA in ambulatory recordings. *Psychophysiology* 44, 2 (2007), 203–215.

[16] David Grimes, Desney S. Tan, Scott E. Hudson, Pradeep Shenoy, and Rajesh P. N. Rao. 2008. Feasibility and pragmatics of classifying working memory load with an electroencephalograph. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'08)*. ACM, New York, NY, 835–844. DOI: https://doi.org/10.1145/1357054.1357187

[17] Eija Haapalainen, SeungJun Kim, Jodi F. Forlizzi, and Anind K. Dey. 2010. Psycho-physiological measures for assessing cognitive load. In *Proceedings of the 12th ACM International Conference on Ubiquitous Computing (UbiComp'10)*. ACM, New York, NY, 301–310. DOI: https://doi.org/10.1145/1864349.1864395

[18] C. W. Harris, J. L. Edwards, A Baruch, W. A. Riley, B. E. Pusser, W. J. Rejeski, and D. M. Herrington. 2000. Effects of mental stress on brachial artery flow-mediated vasodilation in healthy normal individuals. *Am. Heart J.* 139, 3 (2000), 405–411.

[19] Joshua Harrison, Kurtuluş İzzetoğlu, Hasan Ayaz, Ben Willems, Sehchang Hah, Ulf Ahlstrom, Hyun Woo, Patricia A. Shewokis, Scott C. Bunce, and Banu Onaral. 2014. Cognitive workload and learning assessment during the implementation of a next-generation air traffic control technology using functional near-infrared spectroscopy. *IEEE Trans. Hum.-Mach. Syst.* 44, 4 (2014), 429–440.

[20] Nis Hjortskov, Dag Rissén, Anne Katrine Blangsted, Nils Fallentin, Ulf Lundberg, and Karen Søgaard. 2004. The effect of mental stress on heart rate variability and blood pressure during computer work. *Eur. J. Appl. Physiol.* 92, 1–2 (2004), 84–89.

[21] Maarten A. Hogervorst, Anne-Marie Brouwer, and Jan B. F. van Erp. 2014. Combining and comparing EEG, peripheral physiology and eye-related measures for the assessment of mental workload. *Front. Neurosci.* 8, 1 (2014), 322. DOI: https://doi.org/10.3389/fnins.2014.00322

[22] C. Iani, D. Gopher, and P. Lavie. 2004. Effects of task difficulty and invested mental effort on peripheral vasoconstriction. *Psychophysiology* 41, 5 (2004), 789–798.

[23] Meltem Izzetoglu, Scott C. Bunce, Kurtulus Izzetoglu, Banu Onaral, and Kambiz Pourrezaei. 2007. Functional brain imaging using near-infrared technology. *IEEE Eng. Med. Biol. Mag.* 26, 4 (2007), 38.

[24] Thorsten Joachims. 2006. Training linear SVMs in linear time. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'06)*. ACM, New York, NY, 217–226. DOI: https://doi.org/10.1145/1150402.1150429

[25] Mishel Johns, Srinath Sibi, and Wendy Ju. 2014. Effect of cognitive load in autonomous vehicles on driver performance during transfer of control. In *Adjunct Proceedings of the 6th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI'14)*. ACM, New York, NY, 1–4. DOI: https://doi.org/10.1145/2667239.2667296

[26] W. M. Jubadi and S. F. A. Mohd Sahak. 2009. Heartbeat monitoring alert via SMS. In *Proceedings of the IEEE Symposium onIndustrial Electronics & Applications, 2009. ISIEA 2009.* IEEE, 1–5.

[27] J. W. H. Kalsbeek and J. H. Ettema. 1963. Continuous recording of heart rate and the measurement of perceptual load. *Ergonomics* 6 (1963), 306–307.

[28] A. Reşit Kavsaoğlu, Kemal Polat, and M. Recep Bozkurt. 2014. A novel feature ranking algorithm for biometric recognition with PPG signals. *Comput. Biol. Med.* 49, 1 (2014), 1–14. DOI: https://doi.org/10.1016/j.compbiomed.2014.03.005

[29] Wayne K. Kirchner. 1958. Age differences in short-term retention of rapidly changing information. *J. Exp. Psychol.* 55, 4 (1958), 352.

[30] Naveen Kumar and Jyoti Kumar. 2016. Measurement of cognitive load in HCI systems using EEG power spectrum: An experimental study. *Proc. Comput. Sci.* 84, 1 (2016), 70–78.

[31] L. Luo, L. Xiao, D. Miao, and X. Luo. 2012. The relationship between mental stress induced changes in cortisol levels and vascular responses quantified by waveform analysis: Investigating stress-dependent indices of vascular changes. In *Proceedings of the 2012 International Conference on Biomedical Engineering and Biotechnology*. IEEE, 929–933. DOI: https://doi.org/10.1109/iCBEB.2012.437

[32] Yongqiang Lyu, Xiaomin Luo, Jun Zhou, Chun Yu, Congcong Miao, Tong Wang, Yuanchun Shi, and Ken-ichi Kameyama. 2015. Measuring photoplethysmogram-based stress-induced vascular response index to assess cognitive load and stress. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI'15)*. ACM, New York, NY, 857–866. DOI: https://doi.org/10.1145/2702123.2702399

[33] Richard E. Mayer and Roxana Moreno. 2003. Nine ways to reduce cognitive load in multimedia learning. *Educ. Psychol.* 38, 1 (2003), 43–52.

[34] D. McDuff, S. Gontarek, and R. Picard. 2014. Remote measurement of cognitive stress via heart rate variability. In *Proceedings of the 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2957–2960. DOI: https://doi.org/10.1109/EMBC.2014.6944243

[35] D. Mcduff, S. Gontarek, and R. W. Picard. 2014. Improvements in remote cardiopulmonary measurement using a five band digital camera. *IEEE Trans. Bio-med. Eng.* 61, 10 (2014), 2593–601.

[36] Daniel J. McDuff, Javier Hernandez, Sarah Gontarek, and Rosalind W. Picard. 2016. COGCAM: Contact-free measurement of cognitive stress during computer tasks with a digital camera. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI'16)*. ACM, New York, NY, 4000–4004. DOI: https://doi.org/10.1145/2858036.2858247

[37] Calvin K. L. Or and Vincent G. Duffy. 2007. Development of a facial skin temperature-based methodology for non-intrusive mental workload measurement. *Occupat. Ergon.* 7, 2 (2007), 83–94.

[38] Sharon Oviatt. 2006. Human-centered design meets cognitive load theory: Designing interfaces that help people think. In *Proceedings of the 14th ACM International Conference on Multimedia (MM'06)*. ACM, New York, NY, 871–880. DOI: https://doi.org/10.1145/1180639.1180831

[39] Fred Paas, Juhani E. Tuovinen, Huib Tabbers, and Pascal W. M. Van Gerven. 2003. Cognitive load measurement as a means to advance cognitive load theory. *Educ. Psychol.* 38, 1 (2003), 63–71.

[40] Fred G. W. C. Paas and Jeroen J. G. Van Merriënboer. 1994. Instructional control of cognitive load in the training of complex cognitive tasks. *Educ. Psychol. Rev.* 6, 4 (1994), 351–371.

[41] 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. 2011. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* 12, 1 (2011), 2825–2830.

[42] Ming Zher Poh, Daniel J. Mcduff, and Rosalind W. Picard. 2011. Advancements in noncontact, multiparameter physiological measurements using a webcam. *IEEE Trans. Bio-med. Eng.* 58, 1 (2011), 7.

[43] Yu Shi, Natalie Ruiz, Ronnie Taib, Eric Choi, and Fang Chen. 2007. Galvanic skin response (GSR) as an index of cognitive load. In *Extended Abstracts on Human Factors in Computing Systems (CHI EA'07)*. ACM, New York, NY, 2651–2656. DOI: https://doi.org/10.1145/1240866.1241057

[44] K. G. Shin and P. Ramanathan. 1994. Real-time computing: A new discipline of computer science and engineering. *Proc. IEEE* 82, 1 (Jan. 1994), 6–24. DOI: https://doi.org/10.1109/5.259423

[45] Erin T. Solovey, Marin Zec, Enrique Abdon Garcia Perez, Bryan Reimer, and Bruce Mehler. 2014. 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)*. ACM, New York, NY, 4057–4066. DOI: https://doi.org/10.1145/2556288.2557068

[46] Willem B. Verwey and Hans A. Veltman. 1996. Detecting short periods of elevated workload: A comparison of nine workload assessment techniques.*J. Exp. Psychol. Appl.* 2, 3 (1996), 270–285.

[47] L. Wang, Emma Pickwell-Macpherson, Y. P. Liang, and Y. T. Zhang. 2009. Noninvasive cardiac output estimation using a novel photoplethysmogram index. In *Proceedings of the International Conference of the IEEE Engineering in Medicine & Biology Society*. IEEE, 1746.

[48] Shouyi Wang, Jacek Gwizdka, and W. Art Chaovalitwongse. 2016. Using wireless EEG signals to assess memory workload in the *n*-back task. *IEEE Trans. Hum.-Mach. Syst.* 46, 3 (2016), 424–435.
[49] Jianhua Zhang, Zhong Yin, and Rubin Wang. 2015. Recognition of mental workload levels under complex human–machine collaboration by using physiological features and adaptive support vector machines. *IEEE Trans. Hum.-Mach. Syst.* 45, 2 (2015), 200–214.
[50] G. Zhao, Y. J. Liu, and Y. Shi. 2018. Real-time assessment of the cross-task mental workload using physiological measures during anomaly detection. *IEEE Trans. Hum.-Mach. Syst.* 48, 2 (Apr. 2018), 149–160. DOI: https://doi.org/10.1109/THMS.2018.2803025
[51] Ferdinand Rudolf Hendrikus Zijlstra. 1993. *Efficiency in Work Behaviour: A Design Approach for Modern Tools*. Ph.D. Dissertation. Delft University, Delft.

Received July 2018; revised June 2019; accepted June 2019