The Mental Machine: Classifying Mental Workload State from Unobtrusive Heart Rate-Measures Using Machine Learning

Roderic H. L. Hillege[1,2(✉)], Julia C. Lo[1,3], Christian P. Janssen[4], and Nico Romeijn[4]

[1] ProRail B.V., Utrecht, The Netherlands
[2] Ordina N.V., Nieuwegein, The Netherlands
roderic.hillege@gmail.com
[3] Faculty of Technology, Policy and Management, Delft University of Technology, Delft, The Netherlands
[4] Experimental Psychology & Helmholtz Institute, Utrecht University, Utrecht, The Netherlands

Abstract. This paper investigates whether mental workload can be classified in an operator setting using unobtrusive psychophysiological measures. Having reliable predictions of workload using unobtrusive sensors can be useful for adaptive instructional systems, as knowledge of a trainee's workload can then be used to provide appropriate training level (not too hard, not too easy). Previous work has investigated automatic mental workload prediction using biophysical measures and machine learning, however less attention has been given to the level of physical obtrusiveness of the used measures. We therefore explore the use of color-, and infrared-spectrum cameras for remote photoplethysmography (rPPG) as physically unobtrusive measures. Sixteen expert train traffic operators participated in a railway human-in-the-loop simulator. We used two machine learning models (AdaBoost and Random Forests) to predict low-, medium- and high-mental workload levels based on heart rate features in a leave-one-out cross-validated design. Results show above chance classification for low- and high-mental workload states. Based on infrared-spectrum rPPG derived features, the AdaBoost machine learning model yielded the highest classification performance.

Keywords: Mental workload classification · Machine learning · Remote photoplethysmography · Adaptive Instructional Systems

1 Introduction

The concept of mental workload is recognized as a critical component in the management of operational work. It is also one of the most widely used concepts within the field of cognitive engineering, human factors & ergonomics next to situation awareness (e.g. [1,2]). Besides the development of various measurement

R. A. Sottilare and J. Schwarz (Eds.): HCII 2020, LNCS 12214, pp. 330–349, 2020.
https://doi.org/10.1007/978-3-030-50788-6_24

tools to identify the mental workload of operators, researchers have focused on the application of adaptive automation for the management of operator workload [3–5]. More recent developments also focus on the support of novice operators by providing individual tailored feedback through Adaptive Instructional Systems (AIS) dynamically [6,7]. These systems aim to adapt the environment or problem difficulty based on the capacity of a student in real-time [8].

The use of psychophysiological measures in adaptive automation and AIS has proven useful, particularly by their ability to present continuous data and potential real-time assessment of mental workload [9]. Previous research has explored various psychophysiological measurement instruments, such as Electro-EncephaloGraphy (EEG), electrocardiogram (ECG) and functional Near InfraRed spectroscopy (fNIRS) [10–12].

An open question is whether other metrics can also successfully detect the mental workload of operators. In particular, can these measures help to reliably differentiate low and high mental workload conditions? Moreover, can this be detected through sensors that are less obtrusive to wear compared to typical clinical research instruments? Having less obtrusive, yet reliable sensors available would be valuable, as it would allow for measurement in more mobile and social settings.

Given these objectives, we explore the use of a psychophysiological measure using remote photo-plethysmography (PPG) in the color-, and infrared-spectrum, based on camera data. Mental workload measures are obtained in a railway traffic human-in-the-loop simulator, in which 16 professional expert train traffic controllers participated. Within the scenarios train traffic controllers operate under low-, medium-, and high-workload conditions, as identified by training experts. The question is then whether unobtrusive, objective, psychophysiological measures can also detect these three workload levels. To find patterns in the measures that can separate the mental workload levels, a machine learning model will be used. Machine learning is chosen due to its flexibility in finding relations in high dimensional spaces compared to statistical modeling, yet offering some degree of explainability compared to (deep) neural nets, which inner workings are a blackbox [13]. Furthermore, machine learning has been successfully used in previous work where mental workload was classified using multimodal input [14–17]. By looking at the features that contribute to the performance of the model, more can be learned about the underlying mechanisms that underlie mental workload.

1.1 Mental Workload

A universal definition of mental workload has not been agreed upon. Various definitions can be found in the literature where some recurring components can be deduced, for example, external task demand, internal competence, and the internal capacity to deal with the task [2,18,19]. Since internal capacity has substantial impact on task performance [20], having a better grasp of its state could significantly boost the prediction of task performance.

Current methods for measuring mental workload include self-reports like the NASA-TLX [21], expert observations, and physiological measurements (i.e. EEG, ECG and so on). A detailed overview of measures (and their obtrusiveness) to capture mental workload can be found in Alberdi, et al. [22]. We will summarize a couple of the key metrics below.

Self-report questionnaires require the subject at set intervals to report on their mental state, while performing a task. However, a disadvantage is that such reporting is hard to do fully in parallel with task performance, thereby impacting performance and clouding the workload measure [23]. Expert observations require manual classification of mental workload, which makes it expensive and not scalable to actual work settings such as that of train traffic controllers. Heart rate features, among others, are often used as physiological signals. Other physiological means are for example EEG, and functional magnetic resonance imaging (fMRI). The traditional apparatus to obtain these measures are obtrusive, requiring static task-, or controlled (lab-) environments [24]. Advances in wearable sensors reduce this obtrusiveness; however, true unobtrusiveness and data quality remain a challenge [25–27].

In summary, the traditional measures lack practical applicability outside of lab environments since they interrupt the workflow, physically limit or restrict the freedom of movement due to attached sensors, are expensive, require expert judgments, or have a combination of these factors. However, the new trend of the quantified self brings opportunities for physically less- or even unobtrusive psychophysiological mental workload measures [28]. An example is camera-based remote photo-plethysmography (rPPG), which can detect heart features [29–31] and requires no physical contact. This metric in turn can be used to determine the inter-beat-interval or heart rate variability, which can be used to classify mental workload [2,32].

The aforementioned metrics can be used in experiments, to post-hoc test whether different levels of workload can be detected. However, for real-time use in an actual operator work context, ad-hoc workload assessment is of added value. Current ad-hoc workload classification models based on automated and high-frequency sampled metrics have already been developed e.g., Martinez, et al. [16], Gosh, et al. [17], and Lopez et al. [14]. These studies reported on models that utilize unobtrusive features to classify mental workload. All three studies used skin conductance- and heart rate-features measured at the wrist, in conjunction with machine learning models, and were able to classify various levels of mental workload states. Van Gent et al., [15] conducted a multilevel mental workload classification experiment in a driving simulator. Using heart rate features extracted from a finger-worn photo-plethysmography-sensor and machine learning, a multi-level mental workload classifier was built. It achieved a group generalized classification score of 46% correct if miss-classification-by-one-level was allowed.

These studies show the potential of automated and timely fine granular mental workload classification models using sensors that should be physically worn and could be perceived as obtrusive. It is currently unknown if non-invasive

measures, complemented by other machine learning models can improve classification and practicality in daily use. The current study builds further upon previous mental workload classification studies, and contributes by exploring the use of cameras as an unobtrusive measure to develop a mental workload prediction model in a railway traffic control setting.

2 Methods

2.1 Experiment Setup

We used a human-in-the-loop simulator to collect a dataset of psychophysiological responses by expert train operators that worked on a scenario with varying levels of mental workload. The study was conducted in a Dutch regional railway traffic control center in Amsterdam. This data was collected to train and test three machine learning algorithms to classify mental workload levels.

2.2 Participants

Sixteen ProRail train traffic controller operators (four female, $M = 13.44$, $SD = 10.00$ years of working experience) were recruited to voluntarily participate in the study. The setup of the study followed the guidelines set out in the Declaration of Helsinki. All participants were informed about the goal of the study beforehand and provided written informed consent.

2.3 Experimental Design

A within-subjects design that manipulated workload as being low, medium, or high was used. The workload scenarios were part of one larger scenario developed by five subject matter experts (see Fig. 1 for a schematic overview). The task of the operator was to execute their regular job: manage the train traffic while adhering to the correct safety protocols. The events in the scenario started at set times; however, the duration of each scenario varied depending on the chosen strategy and efficiency applied by the operator. Overlap of tasks from one scenario to the next was minimized due to the largely serial nature of the workflow (e.g., the fire department is not involved until they are called by the operator, in which case the operator needs to wait for clearance from the fire department before continuing their work).

Mental workload was manipulated through the complexity and number of activities the operator had to act on. In the lowest workload condition, train traffic operated according to plan and only passive monitoring was required from the operator. In the medium workload condition, monitoring and occasional input were required (e.g., removing obstructions, setting permissions for trains to move – but no bi-directional communication with other parties). In the high workload condition, an emergency call was received requiring direct input-,

communication-, and decision-making from the operator (e.g., gather information regarding the event, make a decision on what protocol is applicable, and execute the actions associated with the applicable protocol).

Four possible scenarios were drafted in which each scenario consisted of a slight variation in the emergency event that occurred. Due to time constraints, each operator conducted two scenarios, pseudo-randomly chosen. The scenarios were validated by five subject matter experts to be comparable in expected mental workload. The duration of a session varied between 15 and 35 min, dependent on the execution and efficiency of the plan deployed by the operator.

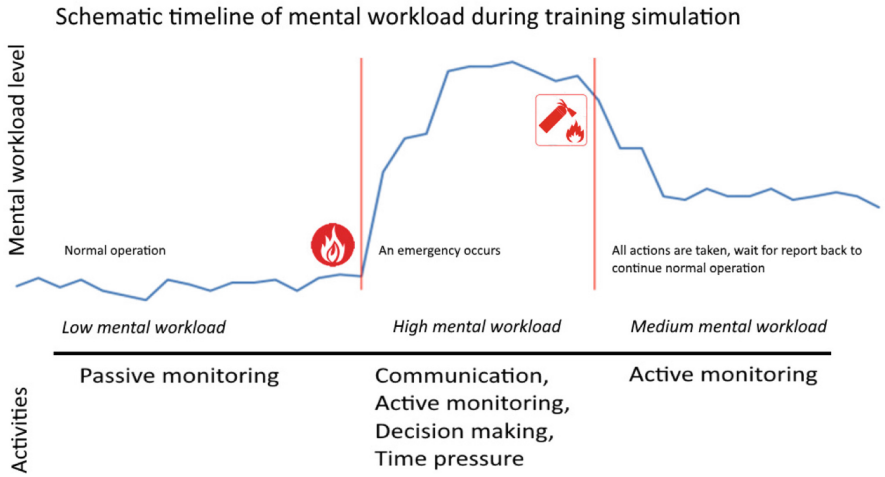

Fig. 1. A schematic timeline of the mental workload scenarios. The first third of the scenario starts with all traffic according to plan. The second third a fire alarm is given: communication and actions are required. The last third, all necessary input from the operator is done and active monitoring for updates is required.

2.4 Aparatus

Heart rate features were recorded from the color- and infrared spectrum using remote camera-based photoplethysmography.

The participants were recorded in the color spectrum with a GoPro Hero Black 7 (GoPro Inc., San Mateo, CA, USA; see Fig. 2B) with a resolution of 1280×720 at 59.94 frames per second, and in the infrared spectrum with a Basler acA640-120gm (Basler AG, An der Strusbek, Ahrensburg, Germany) with a 8 mm $f/1.4$ varifocal lens with a resolution of 659×494 at 24 frames per second (see Fig. 2C). Many factors influence the quality of rPPG [33,34]. In the next section, measures related to the quality of the frame recording taken to this end are discussed. Prior to data storage, the image streams were compressed.

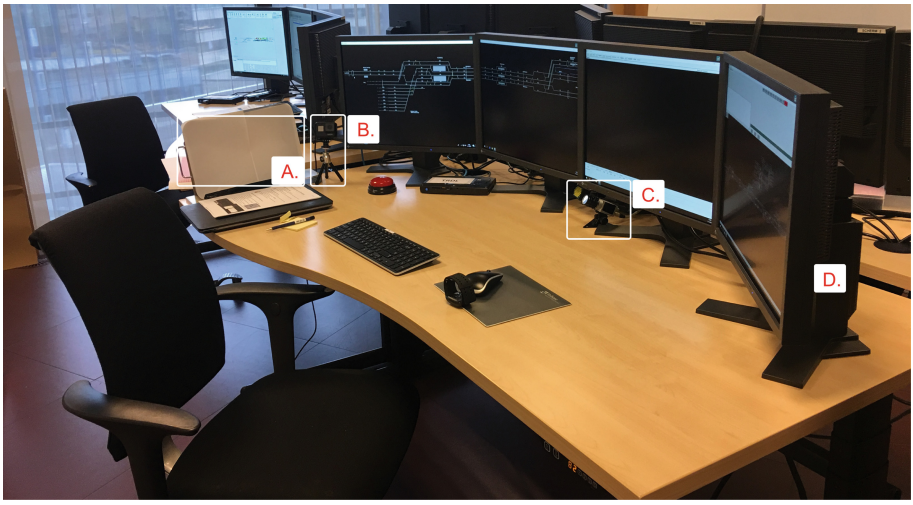

Fig. 2. The following components were used: (A.) The LED light with a CRI rating of 95+ (B.) GoPro Hero 7 Black mounted on a tripod. (C.) Basler acA640-120gm Infrared camera and (D) four 24 in. HP monitors with a resolution of 1920× 1080 at 60 Hz, displaying the railway simulator.

Image stream compression reduces the amount of data per time unit, which is favorable for the storage and throughput of an image stream. However, this negatively affects rPPG quality, which relies on color fluctuation between frames. With heavy compression, these fluctuations are lost. For an optimal result, raw or very lightly-compressed image streams (at least $4.3{*}10^4$ kb/s for random motion) are needed [34]. The GoPro supported a maximum image stream compression of $4 * 10^4$. The proprietary Basler "Pylon Viewer 5.2.0" package supports either raw $200 * 10^5$ kb/s, or compressed MPEG-4 image streams at $1.9 * 10^3$ kb/s. Due to storage-, and video container-limitations handling the uncompressed frame stream, the compressed stream was used.

Since the GoPro image sensor can only capture light that is reflected from a surface, a LED lamp with a color temperature of 3000 K and a Color Rating Index (CRI) of 95% was used to illuminate the left front of the operator (see Fig. 2A). The infrared spectrum was lighted with an integrated two watt infrared flasher from Smart-Eye, which was synchronized with the shutter speed of the sensor to provide optimal lighting.

After the completion of simulation sessions, an informal survey was recorded. The expert operators were asked to rate their subjective experienced mental workload during the simulation. "On a scale from 1 to 7, with 7 being the highest possible score, what grade would you give the workload you experienced during the experiment?"

3 Data Analysis and Model Construction

All data processing was done using Python 3.7 [35] and the Scikit-learn package [36]. Figure 3 summarizes the data processing steps. There were three main steps that are described next: (1) pre-processing, (2) feature extraction, and (3) model construction.

Pre-processing. The first step was to detect the face on each frame. To this end, on each frame from the color- and infrared-spectrum recordings of the operator, a deep neural net face detector was applied to extract 68 facial landmarks, see the red dots in Fig. 4A for an example [37,38].

We then identified a patch of skin on the forehead, and extracted the mean color intensity from it as input for the rPPG algorithm [39]. This forehead region of interest spanned the space between the facial landmarks 20, 21, 24, and 25. The horizontal distance between 21 and 24 was used to vertically shift 21 and 24 upwards, creating a patch on the forehead between those points (see the black patch in Fig. 4A). The forehead was chosen as region of interest because, compared to the cheeks, the lighting was more evenly distributed and under vertical head movements, it remained in-frame for a larger proportion of the time [40].

For each frame where facial landmarks could be detected, the averaged pixel values from the region of interest of the three color channels (red-, green-, blue) and the one infrared channel were calculated. The results from a sequence of frames formed a time series.

To filter noise sources from the color time series the amplitude selective filtering algorithm developed by Wang et al. [41] was used and rewritten for python implementation. The amplitude selective filtering algorithm uses the known reflective properties of the skin to assess signal change frequency, and to remove frequencies that are outside the expected heart rate frequency band (e.g., head movement, reflections of light) from the color channels. These filtered color channels were then used as input for the rPPG plane orthogonal skin response algorithm, developed by Wang et al. [42]. This resulted in a one dimensional PPG signal which was then band-pass filtered between 0.9 Hz and 2.5 Hz, corresponding to a minimum and maximum heart rate of 54 and 150 beats per minute.

To remove noise from the infrared signal, visual inspection was used. This was done due to the one-dimensional nature of the infrared signal which the amplitude selective filtering algorithm can not process. The amplitude selective filtering requires three color channels to remove noise. The obtained filter after visual inspection was a high-pass filter at 0.9 Hz and low-pass filter at 2.5 Hz. Visually, this resulted in a PPG-like signal, however containing more amplitude variations than the amplitude selective filtered color signal.

The preprocessed rPPG data was split into temporal windows (see Fig. 4B). Each window overlaps with the previous one with a specific overlap factor, where the size of the overlap was equal between the rPPG measures (color-

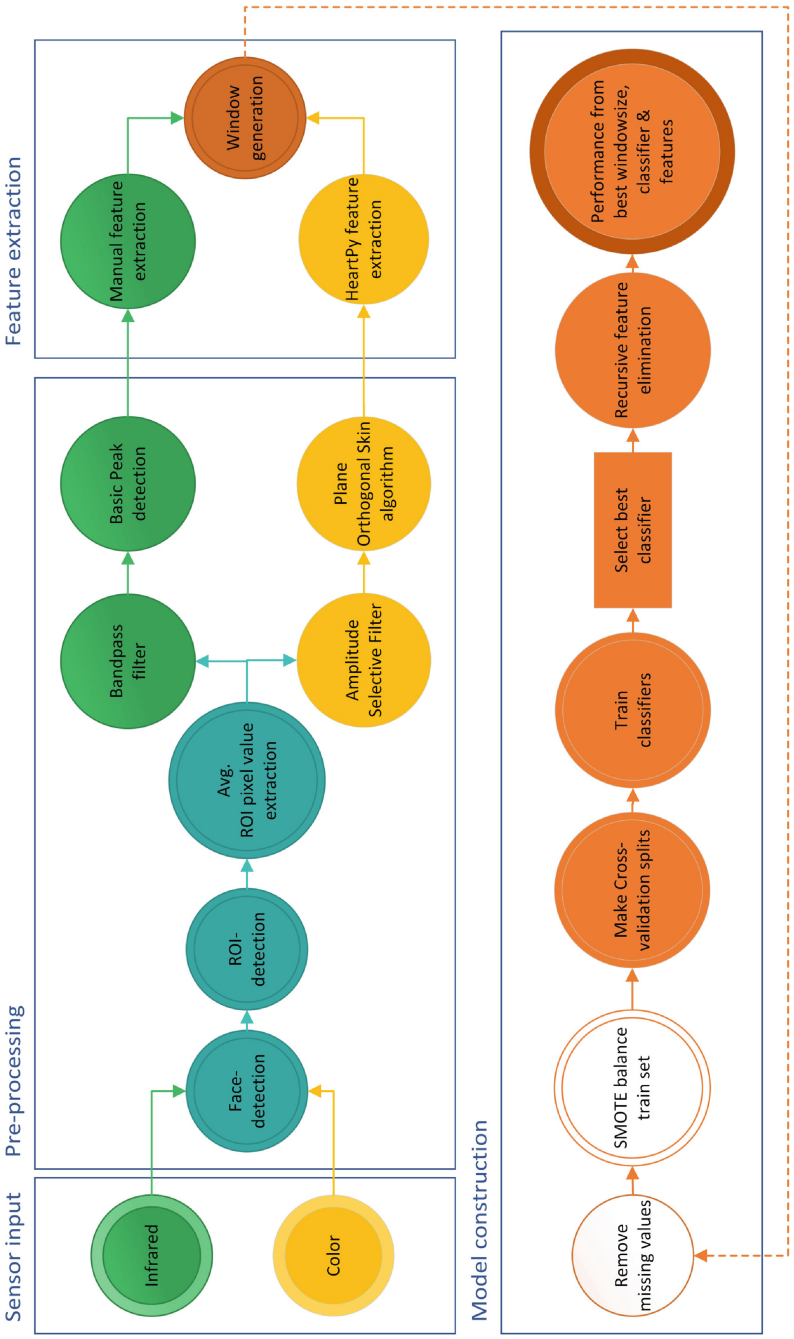

Fig. 3. An overview of the data pre-processing, feature extraction, and model construction.

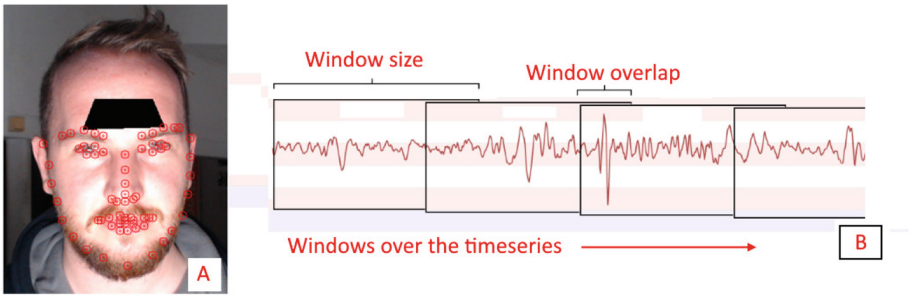

Fig. 4. A) An example of the facial landmark points (red dots) and the region of interest (black square) extracted from it. (B) A schematic overview of the sliding window approach for rPPG-derived heart rate calculation. (Color figure online)

and infrared). The temporal-step size between a window and its succeeding window was equal for all sensors. Heart rate feature calculations are sensitive to the temporal length of windows and the shared overlap between windows they are calculated over. For heart rate features, time-domain features were reliably found from 20-s windows, and frequency domain features from 120-s windows [43,44].

To explore the effect of window sizes on resulting calculated heart rate, two sets with varying window-sizes but identical step sizes were created for the rPPG sources. The "small window" consisted of a 45 s time span with an overlap of 95% resulting in 2.25 s step sizes. The "large window" consisted of a 60 s time span with 95% overlap, resulting in 3 s step sizes. Missing samples in a heart rate window that did not exceed two seconds were, due to the gradual change over time of heart rate features [45], interpolated using Pandas 24.0 interpolate function [46]. In all other cases, windows containing missing values were removed from the dataset.

3.1 Feature Extraction

Heart rate features were calculated over each window. The filtered infrared signal was analyzed for heart rate features, using the basic 'find peaks'-function from the scipy signal package [47]. The filtered color signal was analyzed for heart rate features using the 'find peaks'-scipy signal function (marked in Fig. 5 and Table 2 as "basic"') in addition to the HeartPy toolbox [15]. The following features were extracted from both signals: Beats per minute (BPM)[1,2], Inter beat interval (IBI)[1,2], Mean absolute difference (MAD)[1], Standard deviation of intervals between adjacent beats (SDNN)[1,2], Standard deviation of successive differences (SDSD)[1,2], Proportion of differences greater than 20 ms between beats (pNN20)[1,2], Proportion of differences greater than 50 ms between beats (pNN50)[1,2] ([15][1], [48][2]).

3.2 Machine Learning Datasets

Based on the part of the scenario that participants were performing, mental workload levels could be assigned to each time frame and its associated features. This labelled data set could then be used for a supervised machine learning model that aims to classify workload level based on feature observations.

Data Bias. Absolute heart rate characteristics can identify individuals, especially given the small sample size in our dataset. To avoid model overfitting on absolute heart rate values, and since the heart rate features rely on relative changes over time, the heart rate values were normalized within participant [62]. Overfitting on the training-data caused by unbalanced within-participant proportions of the workload levels was reduced by applying the synthetic minority over-sampling technique (SMOTE) on the training set [49,50]. SMOTE was used because, compared to random oversampling, it preserves some of the variances in the oversampled instances. Auto-correlation is an inherent risk in human physiological data [15]. To avoid such leaking of information, leave one out cross-validation was used. The data of two participants was withheld from the training set and used as the test set. The test- (and resulting training-) set composition were used as input for the model to iteratively run over all possible unique combinations (k) of one and two, from the total number (n) of nine participants $\frac{n!}{(k!(n-k)!)}$ for a total of 28 cross-validation train-test sets.

To create a performance baseline, and test for data bias, the classifiers were run with randomly shuffled train-set labels.

3.3 Models and Classifiers

Out of all the options for machine learning models we used two types of classifiers, random forest (100 trees) and AdaBoost- (60 estimators). Random Forest was chosen as it is frequently used in similar mental workload classification studies [14–17]. AdaBoost falls under the same ensemble learner family as Random Forest, and shares a lot of similarities - however, depending on the data it performs better in some cases [51–55]. The feature importance was determined using Scikit-learn's cross-validated recursive feature elimination ranking [56]. Using the identified best features, a new model was built with only these features. Scikit-learn's "ROC-AUC-CURVE" performance evaluation [36], for the average area under the receiver-operator-characteristic curve of all cross-validated models was used to evaluate the performance [57]. Each workload condition was evaluated in a one- vs. other-mental workload classification manner, resulting in three mean cross-validated AUC-ROC curves.

4 Results

In this results section, a brief description of the empirical data is given first. This is followed by the perceived mental workload, performance of the classifiers and the feature importances. Finally the AUC-ROCs results obtained from using the best performing classifier, window size and features are given.

Sample Selection. From the sixteen participants, data of only nine participants could be used for analysis purposes. Six participants were excluded due to data logging problems, another participant was excluded as the data preprocessing left less than 40 usable samples in both the low- and medium mental workload condition, which is too few to train a classifier on.

For an overview of the samples per workload condition after removing missing values, see Table 1.

Table 1. Number of samples per sensor, before-, and after-SMOTE oversampling.

	Workload level	Color		Infrared		Color and infrared	
Small window:							
Raw	Low		3300		3169		3064
	Medium		2946		2896		2740
	High		2205		2111		1971
SMOTE	Low	+26%	4153	+30%	4106	+27%	3890
	Medium	+41%	4153	+42%	4106	+42%	3890
	High	+88%	4153	+95%	4106	+97%	3890
Large window:							
Raw	Low		2450		2327		2231
	Medium		2191		2126		2012
	High		1630		1513		1413
SMOTE	Low	+26%	3090	+29%	3012	+28%	2861
	Medium	+41%	3090	+42%	3012	+42%	2861
	High	+90%	3090	+99%	3012	+102%	2861

Perceived Mental Workload. The perceived difficulty score recorded from the survey was $M = 3.75$, $SD = 1.13$ for the first-, and $M = 4.00$, $SD = 1.67$ for the second-scenario. The ROC-AUC curves from a model trained on randomly shuffled labels returned chance level performance for all mental workload levels (low $M = .50$, $SD = 0.07$, medium $M = .50$, $SD = 0.05$ and high $M = .51$, $SD = 0.04$). Confirming that there is no data bias that the model could exploit in its classification process and a baseline performance at chance level.

Performance. Considering both window sizes, per-mental workload level AdaBoost outperformed Random Forest for low- ($M = .64$, $SD = 0.09$) and medium- ($M = .54$, $SD = 0.05$) mental workload (see Table 2, bold scores). Random forest scored best for high mental workload ($M = .61$, $SD = 0.09$). See Table 2 for an overview.

Table 2. Model performance for AdaBoost & RandomForest classifier, small & large windows, color-, infrared- and color & infrared data. Italic marking the best scores per classifier & window combination. Bold scores marking the overall best score per workload level.

Workload level	Sensor	Small window		Large window	
		AdaBoost	Random forest	AdaBoost	Random forest
		AUC-ROC (*SD*)	AUC-ROC (*SD*)	AUC-ROC (*SD*)	AUC-ROC (*SD*)
Low	Color	0.54 (0.06)	0.54 (0.05)	0.54 (0.05)	0.54 (0.04)
	Infrared	0.62 (0.09)	*0.61 (0.08)*	**0.64 (0.09)**	*0.62 (0.08)*
	Color and infrared	*0.63 (0.08)*	*0.61 (0.08)*	0.63 (0.09)	0.62 (0.09)
Medium	Color	0.52 (0.01)	0.52 (0.03)	0.51 (0.03)	0.52 (0.03)
	Infrared	**0.54 (0.05)**	*0.54 (0.06)*	0.52 (0.07)	*0.54 (0.07)*
	Color and Infrared	0.52 (0.03)	*0.54 (0.06)*	*0.52 (0.06)*	0.53 (0.06)
High	Color	0.54 (0.04)	0.54 (0.03)	0.53 (0.04)	0.53 (0.03)
	Infrared	*0.57 (0.06)*	*0.56 (0.06)*	*0.58 (0.08)*	**0.61 (0.10)**
	Color and infrared	*0.57 (0.06)*	*0.56 (0.06)*	0.58 (0.08)	0.61 (0.08)
Average best sensor per classifier		0.58 (0.06)	0.57 (0.07)	0.58 (0.08)	0.59 (0.08)

The recursive-cross-validated feature elimination of one vs. other mental workload states using the AdaBoost classifier and AUC-ROC performance scoring, found that: (1) the best performing low-mental workload window size is large, (2) the best performing medium-mental workload window is small, and (3) the best performing high-mental workload window size is large (see Table 2, values in bold).

Feature Elimination. Recursive feature elimination was used to inspect the relative feature-performance contribution. The used window sizes were large, small, large for respectively low-, medium- and high-mental workload. For an overview of the best features after recursive feature elimination, see Table 3. For an overview of the cumulative contribution of the best features to the AUC-ROC score for respective best workload level-window size combination, see Fig. 5.

Using the AdaBoost classifier three models were created, one for each mental workload condition containing the best performing features (Table 3) and window size (Table 2). AUC-ROC scores of low- ($M = .67$, $SD = 0.07$ AUC-ROC), medium- ($M = .55$, $SD = 0.05$ AUC-ROC) and high- ($M = .57$, $SD = 0.08$ AUC-ROC) were found. See Fig. 6 for the resulting AUC-ROC plots. As the AUC-ROC score is a continuum of the true-positive (y-axis) vs. false-positive (x-axis) rate, the standard deviation (grey are) represents the variation in classification performance given different train- and test sets. Since the test sets are comprised of two individuals, the mean variation should be taken as an indicator for model performance, where standard deviation crossing chance does not invalidate the results as is the case in statistical modelling.

Table 3. Feature importance obtained from Scikit learn's cross-validated recursive feature elimination, given the best window size per mental workload level & using the AdaBoost classifier.

Infrared features:	Mental workload levels		
	Low	Medium	High
Inter beat interval (IBI)	0.58	0.33	0.20
Std. dev. of intervals between adjacent beats (SDNN)	0.42	0.27	0.13
Proportion of diff. greater than 20 ms between beats (pNN20)		0.40	0.10
Std. dev. of successive differences (SDSD)			0.20
Beats per minute (BPM)			0.10
Color features:			
Mean Abs. difference (MAD)			0.08
Std. dev. of intervals between adjacent beats (SDNN)			0.05
Inter beat interval (IBI)			0.05
Beats per minute (BPM)			0.03
Proportion of diff. greater than 50 ms between beats (pNN50)			0.03
Std. dev. of intervals between adjacent beats (SDNN)			0.02

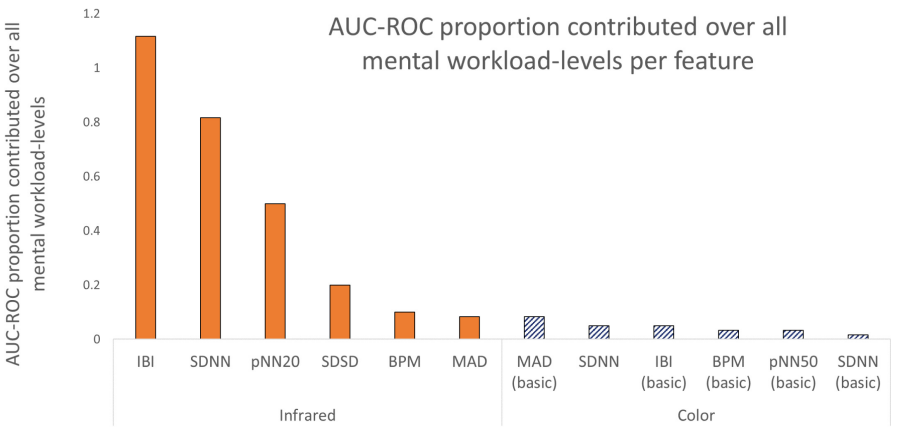

Fig. 5. The cumulative contribution of each feature towards classification performance for all mental workload levels, using the best window size per workload level. The "basic" label indicates use of basic scipy signal peak detection & filtering during (pre-) processing. The features are: Inter beat interval (IBI), Standard deviation of intervals between adjacent beats (SDNN), Proportion of differences greater than 20 ms between beats (pNN20), Standard deviation of successive differences (SDSD), Mean absolute difference (MAD), Proportion of differences greater than 50 ms between beats (pNN50). (basic) denotes basic Scipy 'find peak' filtering

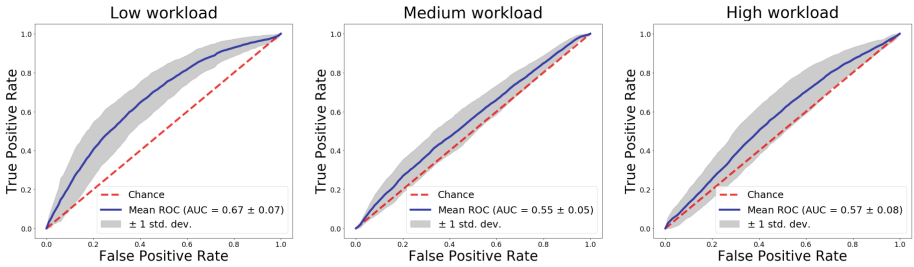

Fig. 6. The AdaBoost cross-validated AUC-ROC curves of the best features and best window size per workload vs. others classification. The red line indicates chance performance, the blue line the mean and the grey the standard deviation received from the cross validations. A large standard deviation indicates large classification variance between different train- and test-sets. The standard deviation is an indicator of the generalizability of the classification. (Color figure online)

5 Discussion

The main objective of this research was to determine to what extent cameras, based on data from the color-, and infrared-spectrum, can differentiate mental workload levels in a human-in-the-loop simulator setting. The measures were taken using remote photoplethysmography, which can be used to detect heart rate. We used an AdaBoost and a Random Forest machine learning model to train a mental workload classifier. We found that low- and high mental workload could be classified above chance. For both low- and high mental workload, classification was best using a large window (i.e., 60 s timespan), regardless of classifier and (combination of) color spectrum (see Table 2). We found the performance of AdaBoost to be on par with RandomForest. Where AdaBoost achieved the best classification score for the low- and medium mental workload levels, Random Forest achieved the best classification score for the high mental workload level.

Looking at the color and infrared spectra and the combination of both, infrared was found to achieve the best classification performance. When decomposing what features the model uses to achieve its performance, the inter beat interval (IBI), standard deviation of intervals between adjacent beats (SDNN) and the proportion of differences greater than 20 ms between beats (pNN20) contribute significantly across classification of the three mental workload levels (cf. [15,48]).

6 Limitations and Future Work

Our scenarios were developed by subject matters experts, with the goal to reflect low, medium and high workload in the expert operators. However, perceived mental workload survey outcomes and debriefing indicated that participants experienced at most moderate workload. Therefore, subjective experience might not have aligned with intended workload. A major factor of this subjective overall

low experienced workload was ascribed by participants to the lack of communication (e.g. with train operators, fire departments) that they otherwise encounter in their job. The communication in this experiment was fewer-, less varied- and serial in nature because the experiment leader was limited in simulating communication from all different stakeholders by him/herself. These limitations suggest workload levels found in the field might be even more pronounced.

Furthermore, the transition between levels of mental workload was modeled as instantaneous. During the labeling of the data, the trigger of an event resulted in an immediate mental workload change (e.g., from low to high). However, the psychophysical mental workload change is typically more gradual [58]. Because of this more gradual psychophysiological change, data sections spanning these transitions are of ambiguous mental workload state. To combat this mixing of states, a solution could be finer grained levels of mental workload to capture the mental workload transition states as was done by Van Gent et al. [15]. Furthermore, it would be interesting to see informed data selection around an event, as is typical for EEG event-related research [59].

Further improvements can be made during the processing and classification of the data. The preprocessing, feature extraction, and workload state labeling can contribute to a better model. Better performance of the 60 s window size compared to the 45 s window size was observed in this study. Heart rate features have been reliably extracted from segments spanning this temporal size in earlier studies [43,44], thus perhaps the quality of the rPPG required longer spanning windows to pick up on a pattern - or find a reliable pattern. Other research that uses heart rate features for mental workload detection has even used windows of up to five minutes [26]. Thus it is interesting to explore the use of larger window sizes and its effect on classification.

Combining the infrared and color channel, and using this merged signal as input for the amplitude selective filter algorithm, could be another improvement. This would effectively allow one to make use of both the infrared- and color channels, similar to what Trumpp et al. [39] have done. The color channels can be used to remove non-heart rate related frequencies, and the infrared for the heart rate related frequency. To sustain temporal synchronization one would need to control for the facial landmark tracking, time synchronization, and the horizontal camera vantage point between the infrared- and color-spectrum recording.

As the rPPG algorithm relies on color changes from artery reflections, improving the frame capture is expected to yield a stronger rPPG signal. The quality of the rPPG signal can be improved on three fronts; (1) by the used hardware, (2) the used lighting, and (3) recording compression-settings.

From the literature the low performance of the color spectrum was not expected, as green light (around 550 nm wavelength) reflects on the arteries [42,60]. When looking at the setup used, some pointers for the observed performance difference between color and infrared can be theorized. The infrared camera came with a dedicated light-source, where for the color spectrum, a CRI95+ rated LED light was used. Furthermore, whilst head-on lighting was possible for the infrared camera, head-on lighting for color camera was not possible without

obtrusively blinding the participants due to the intensity of the LED lamp. The direct versus orthogonal lighting resulted in a better illuminated infrared image stream compared to the color image stream. Furthermore, the default wide-angle lense on the color camera is suboptimal for focusing the light reflected from the expert operator. The 8 mm f/1.4 lense of the infrared camera was much better equipped for this purpose of focusing the light reflecting from the expert operator on the image sensor. Given the comparatively lower performance of the color spectrum, these differences in illumination (head-on dedicated vs from the side) and camera setup warrant further research. We suggest recording the color spectrum using a camera with a dedicated lens for indoor use, to produce more detailed frames.

The compression of the image stream can also be improved. Due to the restraints of the proprietary Basler software, the recordings we used were moderate- to strongly compressed. McDuff et al. [34] show that using raw, uncompressed recordings yields a much cleaner PPG signal with a significantly higher signal-to-noise ratio. Preliminary testing on sub-two minute recordings using the same infrared Basler camera, confirmed this finding of very clean rPPG signal.

For practical and technical reasons, the images in this study were recorded and analyzed for features post-hoc. However, should future studies incorporate (industrial) cameras that allow direct access to their raw image stream, preprocessing could be done on-line, which removes the need to encode, store, decode, and then separately process the images. Given powerful enough hardware, processing and classification of mental workload state could possibly be done on-line, thereby enabling real time access to the mental workload state.

7 Conclusion

Ideally this mental workload model can be used as a tool for real-time mental workload feedback of novice operators during their education program. This insight can be used to provide novice operators and/or instructors feedback in terms of their mental workload development in relation to conducted tasks. Many other domains have the potential to use such knowledge as well. In particular, over the last decades there has been a rise of settings and domains in which humans interact with automation, including use by non-professional users [61]. This includes for example monitoring semi-automated vehicles, drones, and health applications. These domains require novel models of attention management [61]. Our work can contribute to this, by providing methods to automatically detect human workload (and potentially underload and overload).

Acknowledgments. We would like to thank J. Mug & E. Sehic, for their feedback on the experimental design, W. L. Tielman & T. Kootstra for their contribution to the data analysis and from the ProRail Amsterdam train traffic control center the train traffic controllers that participated in this study. This work was supported by ProRail.

References

1. Parasuraman, R., Sheridan, T.B., Wickens, C.D.: Situation awareness, mental workload, and trust in automation: viable, empirically supported cognitive engineering constructs. J. Cogn. Eng. Decis. Making **2**(2), 140–160 (2008)
2. Young, M.S., Brookhuis, K.A., Wickens, C.D., Hancock, P.A.: State of science: mental workload in ergonomics. Ergonomics **58**(1), 1–17 (2015)
3. Brookhuis, K.A., Waard, D.D.: On the assessment of (mental) workload and other subjective qualifications. Ergonomics **45**(14), 1026–1030 (2002)
4. Kaber, D.B., Endsley, M.R.: The effects of level of automation and adaptive automation on human performance, situation awareness and workload in a dynamic control task. Theor. Issues Ergon. Sci. **5**(2), 113–153 (2004)
5. Parasuraman, R.: Adaptive automation for human-robot teaming in future command and control systems. Int. C2 J. **1**(2), 43–68 (2007)
6. Park, O., Lee, J.: Adaptive instructional systems. In: Jonassen, D.H. (ed.) Handbook of Research on Educational Communications and Technology. Simon & Schuster, New York (1996)
7. Bruder, A., Schwarz, J.: Evaluation of diagnostic rules for real-time assessment of mental workload within a dynamic adaptation framework. In: Sottilare, R.A., Schwarz, J. (eds.) HCII 2019. LNCS, vol. 11597, pp. 391–404. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-22341-0_31
8. Lane, H.C., D'Mello, S.K.: Uses of physiological monitoring in intelligent learning environments: a review of research, evidence, and technologies. In: Parsons, T.D., Lin, L., Cockerham, D. (eds.) Mind, Brain and Technology. ECTII, pp. 67–86. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-02631-8_5
9. Byrne, E.A., Parasuraman, R.: Psychophysiology and adaptive automation. Biol. Psychol. **42**(3), 249–268 (1996)
10. Ayaz, H., Shewokis, P.A., Bunce, S., Izzetoglu, K., Willems, B., Onaral, B.: Optical brain monitoring for operator training and mental workload assessment. Neuroimage **59**(1), 36–47 (2012)
11. Prinzel III, L.J., Freeman, F.G., Scerbo, M.W., Mikulka, P.J., Pope, A.T.: Effects of a psychophysiological system for adaptive automation on performance, workload, and the event-related potential P300 component. Hum. Fact. **45**(4), 601–614 (2003)
12. Taylor, G., Reinerman-Jones, L., Cosenzo, K., Nicholson, D.: Comparison of multiple physiological sensors to classify operator state in adaptive automation systems. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 54, no. 3, pp. 195–199 (2010)
13. Goebel, R., et al.: Explainable AI: the new 42? In: Holzinger, A., Kieseberg, P., Tjoa, A.M., Weippl, E. (eds.) CD-MAKE 2018. LNCS, vol. 11015, pp. 295–303. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-99740-7_21
14. Suni Lopez, F., Condori-Fernandez, N., Catala, A.: Towards real-time automatic stress detection for office workplaces. In: Lossio-Ventura, J.A., Muñante, D., Alatrista-Salas, H. (eds.) SIMBig 2018. CCIS, vol. 898, pp. 273–288. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-11680-4_27
15. Van Gent, P., Melman, T., Farah, H., van Nes, N., van Arem, B.: Multi-level driver workload prediction using machine learning and off-the-shelf sensors. Transp. Res. Rec. **2672**(37), 141–152 (2018)
16. Martinez, R., Irigoyen, E., Arruti, A., Martín, J.I., Muguerza, J.: A real-time stress classification system based on arousal analysis of the nervous system by an F-state machine. Comput. Methods Programs Biomed. **148**, 81–90 (2017)

17. Ghosh, A., Danieli, M., Riccardi, G.: Annotation and prediction of stress and workload from physiological and inertial signals. In: 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 1621–1624. IEEE, August 2015
18. Gaillard, A.W.K.: Comparing the concepts of mental load and stress. Ergonomics **36**(9), 991–1005 (1993)
19. Welford, A.T.: Mental work-load as a function of demand, capacity, strategy and skill. Ergonomics **21**(3), 151–167 (1978)
20. Staal, M.A.: Stress, cognition, and human performance: a literature review and conceptual framework (2004)
21. Hart, S.G., Staveland, L.E.: Development of NASA-TLX (Task Load Index): results of empirical and theoretical research. In: Advances in Psychology, vol. 52, pp. 139–183, North-Holland (1988)
22. Alberdi, A., Aztiria, A., Basarab, A.: Towards an automatic early stress recognition system for office environments based on multimodal measurements: a review. J. Biomed. Inform. **59**, 49–75 (2016)
23. Mitchell, J.P., Macrae, C.N., Gilchrist, I.D.: Working memory and the suppression of reflexive saccades. J. Cogn. Neurosci. **14**(1), 95–103 (2002)
24. Hogervorst, M.A., Brouwer, A.M., Van Erp, J.B.: Combining and comparing EEG, peripheral physiology and eye-related measures for the assessment of mental workload. Front. Neurosci. **8**, 322 (2014)
25. Yu, H., Cang, S., Wang, Y.: A review of sensor selection, sensor devices and sensor deployment for wearable sensor-based human activity recognition systems. In: 2016 10th International Conference on Software, Knowledge, Information Management & Applications (SKIMA), pp. 250–257. IEEE, December 2016
26. Lo, J.C., Sehic, E., Meijer, S.A.: Measuring mental workload with low-cost and wearable sensors: insights into the accuracy, obtrusiveness, and research usability of three instruments. J. Cogn. Eng. Decis. Making **11**(4), 323–336 (2017)
27. Lux, E., Adam, M.T., Dorner, V., Helming, S., Knierim, M.T., Weinhardt, C.: Live biofeedback as a user interface design element: a review of the literature. Commun. Assoc. Inf. Syst. **43**(1), 257–296 (2018)
28. Swan, M.: Sensor mania! the Internet of Things, wearable computing, objective metrics, and the quantified self 2.0. J. Sens. Actuator Netw. **1**(3), 217–253 (2012)
29. Verkruysse, W., Svaasand, L.O., Nelson, J.S.: Remote plethysmographic imaging using ambient light. Opt. Express **16**(26), 21434–21445 (2008)
30. Takano, C., Ohta, Y.: Heart rate measurement based on a time-lapse image. Med. Eng. Phys. **29**(8), 853–857 (2007)
31. Huelsbusch, M., Blazek, V.: Contactless mapping of rhythmical phenomena in tissue perfusion using PPGI. In: Medical Imaging 2002: Physiology and Function from Multidimensional Images, vol. 4683, pp. 110–117. International Society for Optics and Photonics, April 2002
32. Charles, R.L., Nixon, J.: Measuring mental workload using physiological measures: a systematic review. Appl. Ergon. **74**, 221–232 (2019)
33. Zaunseder, S., Trumpp, A., Wedekind, D., Malberg, H.: Cardiovascular assessment by imaging photoplethysmography - a review. Biomedical Engineering/Biomedizinische Technik **63**(5), 617–634 (2018)
34. McDuff, D.J., Blackford, E.B., Estepp, J.R.: The impact of video compression on remote cardiac pulse measurement using imaging photoplethysmography. In: 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), pp. 63–70. IEEE, May 2017

35. Van Rossum, G.: Python tutorial, Technical Report CS-R9526, Centrum voor Wiskunde en Informatica (CWI), Amsterdam (1995)
36. Pedregosa, F., et al.: Scikit-learn: machine learning in Python. J. Mach. Learn. Res. **12**(Oct), 2825–2830 (2011)
37. Bulat, A., Tzimiropoulos, G.: How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3D facial landmarks). In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1021–1030 (2017)
38. King, D.E.: Dlib-ml: a machine learning toolkit. J. Mach. Learn. Res. **10**(Jul), 1755–1758 (2009)
39. Trumpp, A., et al.: Camera-based photoplethysmography in an intraoperative setting. Biomed. Eng. Online **17**(1), 33 (2018)
40. Lempe, G., Zaunseder, S., Wirthgen, T., Zipser, S., Malberg, H.: ROI selection for remote photoplethysmography. In: Meinzer, H.P., Deserno, T., Handels, H., Tolxdorff, T. (eds.) Bildverarbeitung für die Medizin, pp. 99–103. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-36480-8_19
41. Wang, W., den Brinker, A.C., Stuijk, S., de Haan, G.: Amplitude-selective filtering for remote-PPG. Biomed. Opt. Express **8**(3), 1965–1980 (2017)
42. Wang, W., den Brinker, A.C., Stuijk, S., de Haan, G.: Algorithmic principles of remote PPG. IEEE Trans. Biomed. Eng. **64**(7), 1479–1491 (2016)
43. Salahuddin, L., Cho, J., Jeong, M.G., Kim, D.: Ultra short term analysis of heart rate variability for monitoring mental stress in mobile settings. In: 2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 4656–4659. IEEE (2007)
44. McNames, J., Aboy, M.: Reliability and accuracy of heart rate variability metrics versus ECG segment duration. Med. Biol. Eng. Compuy. **44**(9), 747–756 (2006)
45. Borst, C., Wieling, W., Van Brederode, J.F., Hond, A., De Rijk, L.G., Dunning, A.J.: Mechanisms of initial heart rate response to postural change. Am. J. Physiol.-Heart Circulatory Physiol. **243**(5), H676–H681 (1982)
46. McKinney, W.: Data structures for statistical computing in python. In: Proceedings of the 9th Python in Science Conference, vol. 445, pp. 51–56, June 2010
47. Jones, E., Oliphant, T., Peterson, P.: SciPy: Open source scientific tools for Python, 2001 (2016)
48. Heart rate variability: standards of measurement, physiological interpretation, and clinical use. In: Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology. Circulation, vol. 93, pp. 1043–1065 (1996)
49. Cawley, G.C., Talbot, N.L.: On over-fitting in model selection and subsequent selection bias in performance evaluation. J. Mach. Learn. Res. **11**(Jul), 2079–2107 (2010)
50. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. **16**, 321–357 (2002)
51. Wyner, A.J., Olson, M., Bleich, J., Mease, D.: Explaining the success of adaboost and random forests as interpolating classifiers. J. Mach. Learn. Res. **18**(1), 1558–1590 (2017)
52. Scikit-learn: scikit-learn.org. Choosing the right estimator. https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html. Accessed 2 Oct 2019
53. Head, T., et al.: scikit-optimize/scikit-optimize: v0.5.2 (Version v0.5.2). Zenodo. https://doi.org/10.5281/zenodo.1207017
54. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
55. Freund, Y., Schapire, R., Abe, N.: A short introduction to boosting. J.-Jpn. Soc. Artif. Intell. **14**(771–780), 1612 (1999)

56. Guyon, I., Weston, J., Barnhill, S., Vapnik, V.: Gene selection for cancer classification using support vector machines. Mach. Learn. **46**(1–3), 389–422 (2002)
57. Huang, J., Ling, C.X.: Using AUC and accuracy in evaluating learning algorithms. IEEE Trans. Knowl. Data Eng. **17**(3), 299–310 (2005)
58. Kim, H.G., Cheon, E.J., Bai, D.S., Lee, Y.H., Koo, B.H.: Stress and heart rate variability: a meta-analysis and review of the literature. Psychiatry Invest. **15**(3), 235 (2018)
59. Luck, S.J.: An Introduction to the Event-related Potential Technique. MIT Press (2014)
60. van Gastel, M., Stuijk, S., de Haan, G.: Motion robust remote-PPG in infrared. IEEE Trans. Biomed. Eng. **62**(5), 1425–1433 (2015)
61. Janssen, C.P., Donker, S.F., Brumby, D.P., Kun, A.L.: History and future of human-automation interaction. Int. J. Hum Comput Stud. **131**, 99–107 (2019)
62. Dietterich, T.G., Kong, E.B.: Machine learning bias, statistical bias, and statistical variance of decision tree algorithms. Technical report, Department of Computer Science, Oregon State University (1995)