Mind Spy:

Unobtrusive Mental Workload State Detection Exploiting Heart Rate- and Posture Movement- Features using Machine Learning

Roderic Hillege

Supervisors:
Drs. Julia Lo (ProRail), Dr. Nico Romeijn (UU) & Dr. Chris Janssen (UU)

A thesis presented for the degree of
Master of Science

Utrecht University

Artificial Intelligence
Utrecht University
28-10-2019

Table of Contents

Introduction **2**
Contribution to science 3

Experiment setup **3**
Participants 3
Apparatus 3
Experimental design 4

Data analysis & model construction **4**
Pre-processing 4
Feature extraction 5
Machine learning datasets 7

Results **7**

Discussion & Conclusion **8**

Limitations & Future work **12**
Acknowledgments 13

References **13**

Appendix **16**
Appendix 1: GoPro and Basler Infrared Settings 16
Appendix 2a: Classifier performances on Infrared and Gyroscope and Infrared, for all mental workload states and window sizes 19
Appendix 2b: Classifier performances on all sensor combinations, for all mental workload states and window sizes . 21
Appendix 3: An overview of the best performing window size- and features per workload level on the AUC-ROC score. 24

Mind Spy:
Unobtrusive Mental Workload State Detection Exploiting Heart Rate and Posture Movement Features using Machine Learning

Roderic Hillege
Student number: 5551242

Abstract

This paper explores mental workload classification using unobtrusive psychophysiological measures. Posture movements on a chair, a wrist-worn heart rate sensor, color- and infrared-spectrum remote photoplethysmography were recorded from sixteen expert train traffic operators in a railway human-in-the-loop simulator with low-, medium-, and high- mental workload conditions. Normalized heart rate- and posture movement- features were extracted and used as input for nearest-neighbor and ensemble machine learning classifiers. The classifiers were trained using a cross-validated, leave-one-out, and between-subject design. Results show that the classifiers can distinguish low-, and high- mental workload states of the expert operators above chance. Posture movements and heart rate variability measures from the infrared spectrum and yielded the highest performance and, combined with the properties of no physical contact to the subject in case of the gyroscope, and the invisibility of infrared light to the human eye, these measures make for the least obtrusive mental workload classification sensor-setup tested in this paper.

Keywords: Mental workload; remote photoplethysmography; posture movements; machine learning; unobtrusive

Introduction

In digital labor, especially in a safety-critical environment (e.g., nuclear plant operators, train-, and air- traffic controllers), the mental workload is arguably a critical cognitive component (Harteis, 2018). Despite the lack of a universal definition, an indication of what mental workload is comprised of can be found in the literature. A set of components can be distilled to external task demand, internal competence, and the capacity to deal with the task which determines the level of mental workload (Young, Brookhuis, Wickens, & Hancock, 2015; Gaillard, 1993; Welford, 1978).

Current methods for measuring mental workload include self-report, observations, and physiological measurements. Self-report methods like the NASA-TLX (Hart & Staveland, 1988) require the subject during set intervals to report on their mental state while performing a task that can not be done in parallel (Mitchell, Macrae, & Gilchrist, 2002). Mental workload assessments by observation require an expert to classify the mental workload manually. This makes it expensive and not scalable. Physiological measures often include heart rate features, among others, the inter-beat-interval and resulting heart rate variability (Young et al., 2015; Charles & Nixon, 2019). These can be acquired via electroencephalograms (EEG), electrocardiograms (ECG), and functional magnetic resonance imaging (fMRI). The traditional means to obtain these measures are obtrusive as static task-, or controlled (lab-) environments are required (Hogervorst, Brouwer, & Van Erp, 2014). Advances in wearable sensors reduce the obtrusiveness of these physiological measures; however, true unobtrusiveness and data quality remain a challenge (Yu, Cang, & Wang, 2016; Lo, Sehic, & Meijer, 2017; Lux et al., 2018). Traditional measures are not practical in a production environment since they are physically obtrusive (i.e., physically limiting or restricting the freedom of movement due to attached sensors or interrupting the workflow), expensive, or both.

The new trend of the quantified self brings physically less- or even un-obtrusive physiological- mental workload measures (Swan, 2012). For instance, camera-based remote photo-plethysmography (rPPG), which can detect heart features (Verkruysse, Svaasand, & Nelson, 2008; Takano & Ohta, 2007; Huelsbusch & Blazek, 2002) and requires no physical contact. Another example is body sway, where a relation between body sway and cognitive functioning was found (Andersson, Hagman, Talianzadeh, Svedberg, & Larsen, 2002). Posture movements during a task, measured by frontal-, lateral- and side-to-side movement on a chair, were also found to discriminate cognitive load (Arnrich, Setz, La Marca, Tröster, & Ehlert, 2010). Like camera-based observations, measuring posture movements on a chair does not require physical contact with the subject, and in that respect is an unobtrusive mental workload measure. Furthermore, the omnipresence of chairs in a monitoring setting makes for a convenient means. For a recent overview of mental workload measures and their corresponding obtrusiveness, see Alberdi, Aztiria, and Basarab (2016).

Furthermore, automated mental workload classification models using relatively unobtrusive measures as input have been made. For example, Martinez, Irigoyen, Arruti, Martín, and Muguerza (2017) and Lopez, Condori-Fernandez, and Catala (2018) created models that utilize unobtrusive features to classify mental workload in an automated fashion. Using skin conductance- and heart rate- features measured at the wrist, in conjunction with a machine learning model, they were able to classify low-, medium- and high-mental workload states. Van Gent, Farah, Nes, and Arem (2018) 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, they successfully built a multi-level mental workload classifier. Ghosh, Danieli, and Riccardi (2015) used physiological signals (skin conductance, skin temperature, heart rate features, and body movement) recorded with a wristband and self-reports gathered via an app as input for a machine learning model, to successfully classify the stress levels reported by the subjects.

These studies show the potential of automated mental workload classification models. The current study builds further upon these findings by exploring the use of sensors with a higher degree of unobtrusiveness, i.e., using sensors to measure mental workload, which require no direct physical contact between the subject and the sensor.

Contribution to science

This paper aims to investigate the extent to which it is possible to classify mental workload state unobtrusively. The required conditions for such unobtrusive mental workload detection in a human-in-the-loop simulation environment will be explored.

The research question of this paper is: to what extent is it possible to classify mental workload states with unobtrusive behavioral and psychophysical measures as input features for a machine learning model. Following from this main question, the sub-question is what features contribute to this classification and what combination of those is the least obtrusive. This approach is bottom-up and data-driven, as no prior hypothesis exists as to how each signal feature or combination thereof contributes to mental workload classification. It is also top-down theory-driven, as the type of input signals and resulting features extracted will be based on the mental workload literature.

An experiment will be conducted where data of three different mental workload levels; low-, medium-, and high is collected. Remote photo-plethysmography in the color-, and infrared- spectrum, and posture movement measures are used as input due to their minimal obtrusive nature. A machine learning model will be used to search for patterns that correspond to the mental workload states. Furthermore, the features contributing to correct classification workload levels will be inspected to gain insight on what features contribute to classifying mental workload in an operator monitoring setting.

This research contributes to the literature, and in particular, to the field of artificial intelligence by exploring a new-, automated-, and more refined physically unobtrusive- method for the detection of mental workload. The ability to physically unobtrusively measure mental workload will enable more symbiotic human-computer interaction. When mental workload levels are known, more granular (artificial) human resource management can be conducted in situations where alertness is critical. Work can be distributed dynamically among the available workforce based on the mental workload levels of individuals. This opens up the possibility for a safer, more efficient, and better working conditions, and reduced mental under- and overload.

Experiment setup

In the ProRail Amsterdam Train Traffic Control Centre, data under known and varying levels of mental workload from expert train traffic controllers was recorded in a human-in-the-loop simulator. In the following section, the experimental setup is discussed.

Participants

Sixteen ProRail train traffic controller operators (four female, *M* = 13.44, *SD* = 10.00 working experience in years) were recruited voluntarily. Operators were informed about the goal of the study beforehand, and the setup of the study is following the guidelines set out in the Declaration of Helsinki.

Apparatus

Posture movements were recorded using a gyroscope mounted to a chair, and to record heart rate features, color- and infrared spectrum remote camera-based photoplethysmography and a wrist-worn sensor were used.

Gyroscope and wrist-worn sensor On the back of a regular office chair with roller wheels, a metal- screw-in phone mount was attached, holding a Samsung Galaxy S4 with Android 5.0.1 (see figure 1a). The app "SensorRecord" (version 2.3.0), recorded the pitch- (x-axis), roll- (z-axis), and jaw (y-axis) from the internal inertial measurement unit (Mourcou, Fleury, Franco, Klopcic, & Vuillerme, 2015). A sampling rate of 20Hz was used (Khusainov, Azzi, Achumba, & Bersch, 2013). A wrist-worn Empatica^{TM} E4 (see figure 1e), attached to the non-dominant wrist of the operator, was used to extract heart rate features (Lo et al., 2017; McCarthy, Pradhan, Redpath, & Adler, 2016).

Color and Infrared rPPG For the rPPG measures, the subjects were recorded in the color spectrum with a GoPro Hero Black 7 (see figure 1c) with a resolution of 1280 x 720 at 59.94 frames per second, and in the infrared spectrum with a Basler acA640-120gm with a 8mm f/1.4 varifocal lens at 659x494 and 24 frames per second (see figure 1d). For an overview of the GoPro and acA640-120gm settings, see appendix 1.

Quality of rPPG is influenced by many factors (Zaunseder, Trumpp, Wedekind, & Malberg, 2018; McDuff, Blackford, & Estepp, 2017). The measures taken related to frame recording are discussed in the next section. The measures taken that are related to data handling and feature extraction will be discussed in the data analysis and model construction section.
Compressing video streams reduces the amount of data that is favorable for storage and throughput, but since rPPG relies on color fluctuation between frames, which is lost

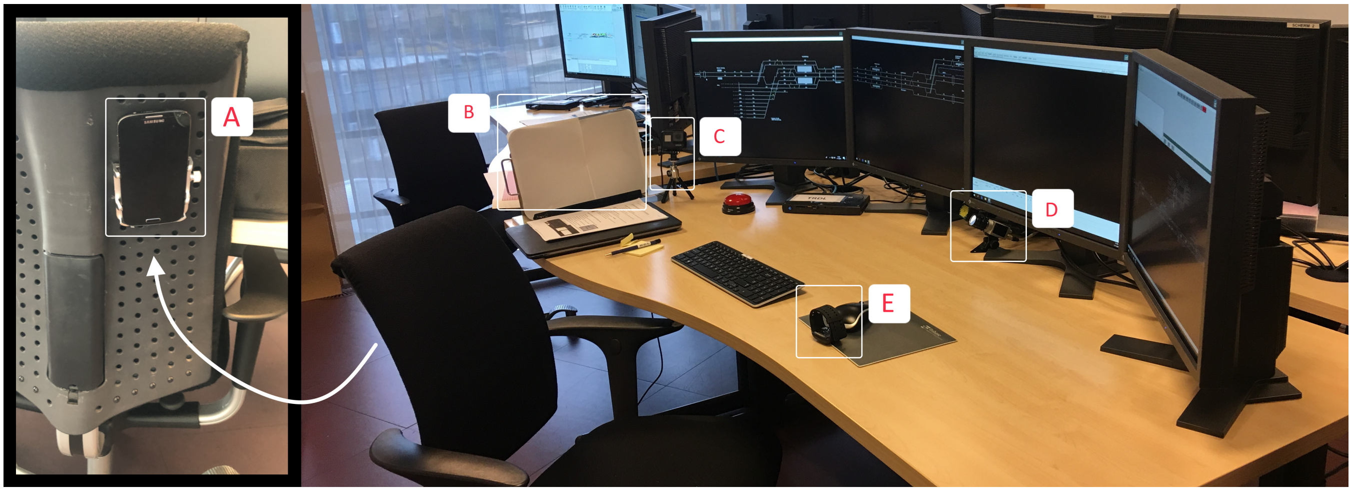


Figure 1: The simulator setup consists of four 24 inch HP monitors with a resolution of 1920 x 1080 at 60Hz, displaying the railway simulator "Amsterdam Schiphol Tunnel", the following components were used: (a) the Samsung S4, mounted on the back of the office chair with a metal-screw on mount. (b) The LED light with CRI 95+ (c) GoPro Hero 7 Black mounted on a tripod. (d) Basler acA640-120gm Infrared camera (e) Empatica E4 wristband.

with (heavy) compression, raw or very lightly-compressed frame streams (at least 4.3*10^4 kb/s for random motion) are needed (McDuff et al., 2017). The GoPro supported a minimum compression of 4*10^4, the proprietary Basler "Pylon Viewer 5.2.0" software supports raw 200*10^5 kb/s uncompressed, and compressed 1.9*10^3 recording modes. Due to storage limitations handling the uncompressed frame stream, the compressed MPEG-4 stream was used. To be captured by the image sensor, light reflecting from the skin and blood vessels has to be emitted by a light source, making lighting another essential consideration for rPPG. For the color spectrum, an LED lamp with a color temperature of 3000 Kelvin and a Color rating index (CRI) of 95%+ was used to light the left front of the operator (see figure 1b). The infrared spectrum was lighted with a dedicated "Smarteye" two watts infrared flasher, which is synchronized with the shutter speed of the sensor – providing optimal lighting.

Experimental design

Five domain experts drafted a simulation scenario consisting of three sections with varying workload levels (see figure 2a for a schematic overview). The task of the operator was to manage the 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. The workload was manipulated with the number of activities the operator had to act on. In the lowest workload condition, passive monitoring was required, and train traffic operated according to plan. In the medium workload condition, active 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 the resulting actions to take, etc.). Each operator conducted two of these sessions – where each session consisted of a slight variation in the emergency event that occurred. Activities required and the resulting mental workload in the variations of the scenarios was comparable. The duration of a session varied between 15 and 35 minutes, dependent on the execution and efficiency of the plan deployed by the operator.

Data analysis & model construction

All data processing was done using Python 3.7 (Van Rossum & Drake Jr, 1995) and the Scikit-learn package (Pedregosa et al., 2011). The data were preprocessed, preparing them for the next step where features were extracted. The data were then compiled into datasets that could be used to build the machine learning models. For an overview of the pipeline, see figure 3.

Pre-processing

The gyroscope data consisted of the three-axis (x,y, and z) with a precision of nine decimals. Due to the accuracy and low noise of the inertial measurement unit of the Samsung S4, the data is processed raw (Mourcou et al., 2015). The Empatica E4 wristband was processed with Empatica's proprietary software, returning processed inter-beat-interval data.

On each frame from the recordings of the operator in the color- and infrared- spectrum, a deep neural net face detector with a capacity to detect faces under a wide variation of head orientations was applied (Bulat & Tzimiropoulos, 2017; King, 2009). The face detector extracts a fixed set of 68 fa-

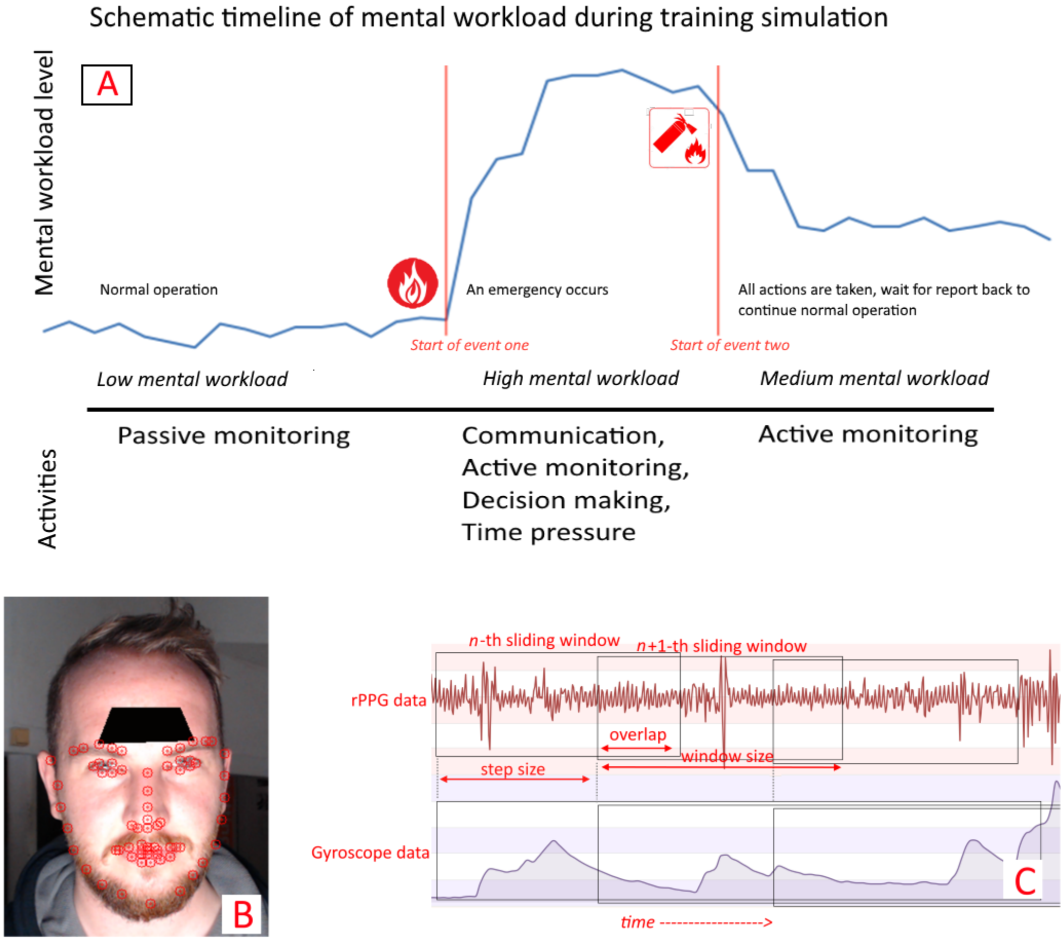


Figure 2: (a) A schematic timeline of the scenarios. The first third of the scenario starts with all traffic according to plan. The second third a fire alarm is given: communication and actions are required. The last third, all necessary input from the operator is done and active monitoring for updates is required. (b) An example of the facial landmark points and the region of interest extracted from it (black square). (c) A schematic overview of the sliding window approach.

cial landmarks from detected faces (see figure 2b, the red dots). rPPG requires a region of interest of the skin to extract color information from (Trumpp et al., 2018). Despite the fact that the forehead performs worse compared to the cheeks, the forehead was chosen to extract the mean color channel pixel values from because under vertical head movements it remained in-frame for longer and lighting was more evenly distributed (Lempe, Zaunseder, Wirthgen, Zipser, & Malberg, 2013). This forehead region of interest spanned the space between the facial landmarks 20, 21, 24, and 25, where the horizontal distance between 21 and 24 was used to vertically shift 21 and 24 up (see figure 2a, the black square). For each frame where facial landmarks could be detected, the average red-, green-, blue- and infrared pixel values were calculated. The resulting time series dataset contained for each frame where facial landmarks were found the mean values for the red-, green-, blue- and infrared channels. A python implementation of Wang, Brinker, Stuijk, and Haan (2017) amplitude selective filtering algorithm was written to filter this time series. The amplitude selective filtering algorithm exploits known reflective properties of the skin, to remove frequencies that are outside the expected heart rate frequency band (e.g., head movement, reflections of light, etc.) from the color channels. These filtered color channels were then used as input for the rPPG *plane orthogonal skin response* algorithm, developed by Wang, Brinker, Stuijk, and Haan (2016), resulting in a one dimensional PPG signal. This PPG signal was band-pass filtered between 0.9 and 2.5Hz, corresponding to a minimum and maximum heart rate of 54 and 150 beats per minute, based on the Empatica E4 heart rate measurements. The infrared channel was, after visual inspection, high-pass filtered on 0.9Hz and low-pass filtered on 2.5Hz.

Feature extraction

The preprocessed rPPG and gyroscope data are split into temporal windows. Each window overlaps with the previous one with a specific overlap factor, where the size of the over-

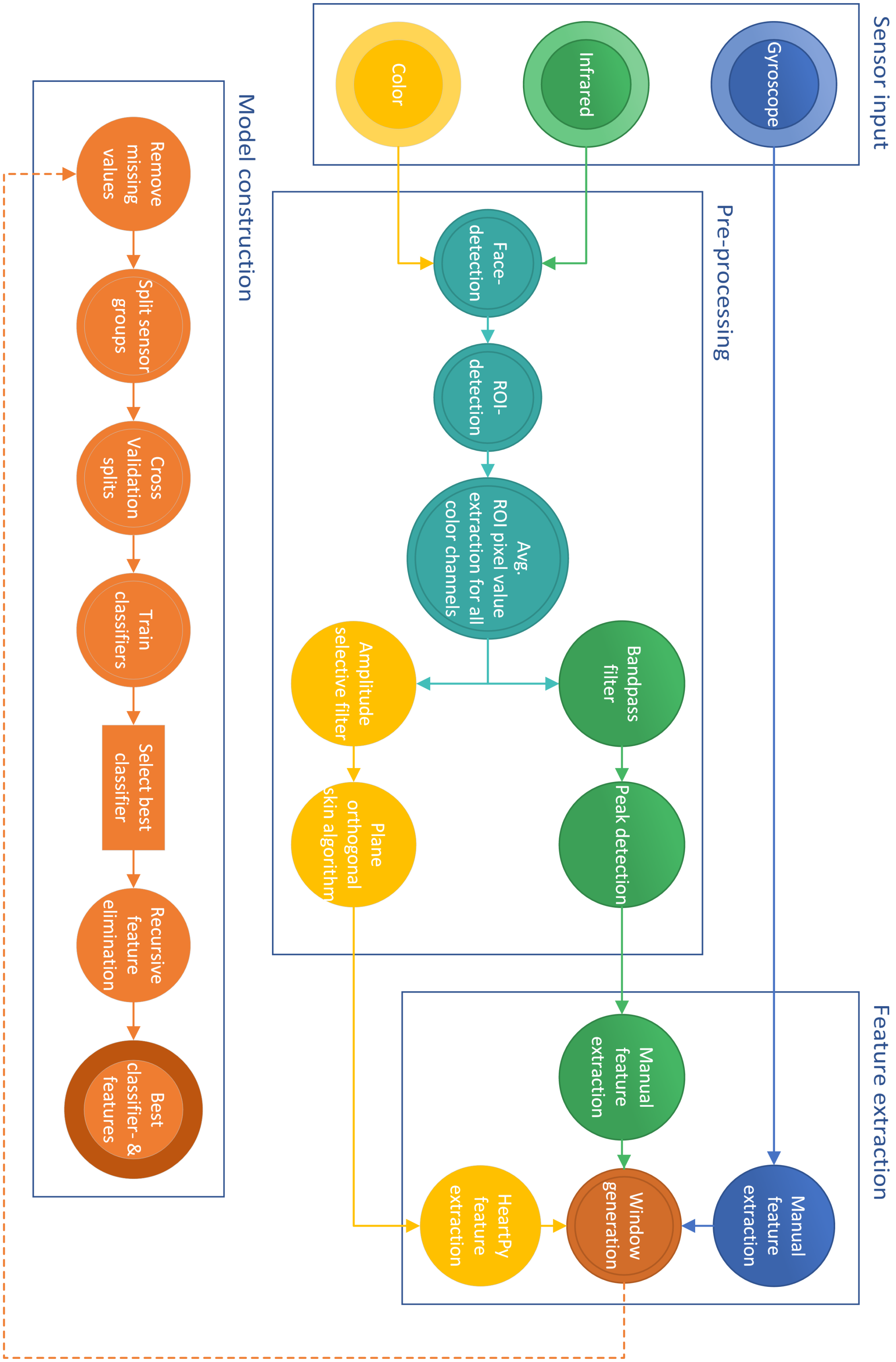


Figure 3: An overview of the data pre-processing, feature extraction and model construction

lap was synchronized between the rPPG measures (color- and infrared) and the gyroscope measures (see figure 2c). The temporal-step size between a window and its succeeding window was equal for all sensors. Heart rate and posture movement features are sensitive to the temporal length- and shared overlap between- windows they are calculated over. For heart rate features, time-domain features were reliably found from 20-second windows, and frequency domain features from 120-second windows (Salahuddin, Cho, Jeong, & Kim, 2007; McNames & Aboy, 2006). For posture movement features, temporal lengths starting from 2.5-seconds and overlap starting from 50% have been found to be reliable sizes (Khusainov et al., 2013). To explore the effect of window sizes, two sets with varying window-sizes but identical step sizes were created for rPPG and gyroscope. The first consisted of 45 seconds for rPPG and 5 seconds for gyroscope (the "small" window set), and the second of 60 seconds for rPPG and 6 seconds for gyroscope (the "large" window set). See Table 1 for the window- and overlap sizes used and their respective resulting step-size. 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 (Borst et al., 1982), interpolated using Pandas 24.0 interpolate function (McKinney et al., 2010). In all other cases, windows containing missing values were removed from the dataset.

Over each window, sensor-specific features are calculated. The HeartPy toolbox (Van Gent et al., 2018) was used to calculate the color-channel heart rate features. Scipy signal's "find peaks" function was used to analyze the infrared channel (Jones, Oliphant, Peterson, et al., 2001). See table 2 for an overview of features per sensor.

Machine learning datasets

The features that were calculated from the preprocessed sensor windows are prepared as datasets for use in a machine learning model.

Data bias Because posture movement- and heart rate features for mental workload classification rely on changes over time rather than absolute differences, and absolute differences between participants are data structure patterns a machine learning model could exploit, all features are within-participant normalized. Model overfitting on the training-data caused by unbalanced within-participant proportions of the workload levels is reduced by applying the synthetic minority over-sampling technique (SMOTE) on the training set (Cawley & Talbot, 2010; Chawla, Bowyer, Hall, & Kegelmeyer, 2002). SMOTE is used because, compared to random oversampling, it preserves some of the variances in the oversampled instances. To avoid leaking of information from the training- into the test- set due to the autocorrelational risk inherent in human physiological data (Van Gent et al., 2018), the test set consisted of the data of three participants, which was withheld from the training set. The test- (and resulting training-) set composition was cross-validated with leave-one-out, by running the model iteratively 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.

Models & Classifiers KNeighbours-, AdaBoost-, and Random Forest- classifiers in their default form were evaluated (Head et al., 2018; Breiman, 2001; Freund, Schapire, & Abe, 1999; Cover & Hart, 1967). Bayesian hyper-parameter optimization using a Gaussian process from the Scikit-optimize package (Head et al., 2018) was used as suggested by Shahriari, Swersky, Wang, Adams, and De Freitas (2015). Note that for unbiased performance estimation, a validation set should be used. This was not done due to the limited influence of hyperparameter tuning on model performance, the added complexity to implement, and the scope of this research. The feature importance was determined using Scikit-learn's cross-validated recursive feature elimination ranking (Guyon, Weston, Barnhill, & Vapnik, 2002). A final model was built using the previously found features, and Scikit-learn's "AUC-ROC-CURVE" performance evaluation (Pedregosa et al., 2011), for the average area under the receiver-operator-characteristic curve of all cross-validated models (Huang & Ling, 2005). Each workload condition was evaluated in a one- vs. other-mental workload classification manner, resulting in three mean cross-validated AUC ROC curves.

Results

A brief description of the data is given, followed by the validation and performance characteristics for sensor combinations, the used classifiers, the importance of features, and the model performance with the best performing sensor, classifier, and features.

Descriptive Statistics From the sixteen participants, six were excluded due to recording problems, and one due to less than 40 samples in both the low- and medium workload condition. For an overview of the samples after removing missing values for both the small- and large-window sizes, and before and after oversampling, see table 3 and figure 4. The average heart rate measured by the Empatica E4 was *M* = 78.74 bpm, *SD* = 4.45 bpm, with average max heart rate *M* = 115 bpm *SD* = 19 bpm, where the individual max measured heart rate was 143 bpm. The average minimum heart rate was *M* = 63.18 bpm, *SD* = 4.62 bpm, and the individual minimum heart rate was 56 bpm. Because the results from the Empatica E4 both before and after oversampling were too few to successfully train a classifier on, they were discarded from the analysis.

Validation To create a baseline and to test for data bias, all classifier–sensor combinations were run with randomly shuffled train-set labels. The resulting AUC-ROC curves returned chance level performance for all mental workload levels, confirming no data bias the model could exploit – and

Table 1

Window- and step sizes used for scenario one and two

Sensor	Window size in sec.		Overlap in %		Step size in sec.	
	Small	Large	Small	Large	Small	Large
Color- & Infrared- spectrum	45	60	95	95	2.25	3
Gyroscope	5	6	55	50	2.25	3

Table 2

Overview of features used

Sensor	Features
Color- & Infrared- spectrum:	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 20ms between beats (pNN20)[1,2], Proportion of differences greater than 50ms between beats (pNN50)[1,2]
Gyroscope for each axis:	Min[3], Max [3,4], Mean [3,4], Variance [3,4], Standard deviation (SD) [3], Skew[3,4], Kurtosis[3,4], Root mean square (RMS) [3,4], Zero-crossings (ZCR)[3], absolute difference (ABSDIFF)[3], first five fast Fourier transform frequencies (FFT)[3,4], Uniformity[3]

[1](Van Gent et al., 2018)
[2](Rawenwaaij-Arts, Kallee, Hopman, et al., 1993)
[3](Figo, Diniz, Ferreira, & Cardoso, 2010)
[4](Atallah, Lo, King, & Yang, 2010)

chance level baseline performance.
An informal survey was recorded after the completion of simulation sessions. On the question “On a scale from 1 to 7, with 7 being the highest, what grade would you give to the workload of the experiment?” the training operators responded with *M* = 3.75, *SD* = 1.13 for the first experiment and *M* = 4.00, *SD* = 1.67 for the second.

Performance All sensor combinations were evaluated using three classifiers with default parameters (Random Forest with 100 trees, AdaBoost with 60 estimators, and KNeighbours with 3 classes). For an overview of the AUC-ROC performance for each classifier, see appendix 2a. Adaboost and Random Forest both outperformed KNeighbours, and AdaBoost outperformed Random Forest for all workload levels on the *Infrared* set by one percent, and on the *Gyroscope and Infrared* combination by one percent lower standard deviation (see appendix 2a). For an overview of the results, see appendix 2b. The recursive- cross-validated feature elimination of one vs. other mental workload states using the AdaBoost classifier and AUC-ROC performance scoring, found the best performing low- mental workload window size is large, the best performing medium- mental workload is small, and the best performing high- mental workload window size is large (see figure 6 for an overview).

Three final models were created, one for each mental workload condition containing the best performing features found with the cross-validated recursive feature elimination using the AdaBoost classifier. For an overview of the resulting AUC-ROC curves, see figure 5. Low- (*M* = 0.66, *SD* = 0.10 AUC-ROC), medium- (*M* = 0.53, *SD* = 0.09 AUC-ROC) and high- (*M* = 0.62, *SD* = 0.11 AUC-ROC) vs. the other mental workload conditions could be classified above chance.

Feature elimination The features of the low- mental workload, large window; medium mental workload, small window and high mental workload large window were inspected for relative performance contribution. For an overview of the best features per workload level, see appendix 4. For an overview of the total contribution the best features made to the AUC-ROC score for respective best workload level- window size combination, see figure 7. For an overview of the total contribution of the sensors to the AUC-ROC scores for the respective best workload level- window size combinations, see figure 8.

Discussion & Conclusion

It was found that the various physically unobtrusive heart rate- and posture- features used to train different machine learning models were able to classify mental workload states. From the three constructed levels of mental workload, the low- and high workload levels were found to be distinguishable best. Although the medium workload could be classified above chance, its performance was found to be weak. A possible explanation for this weak performance of the medium mental workload classification could be that it

Table 3

Sensor sets with corresponding window-samples raw, and after removing missing and SMOTE oversampling, for both the small- and large windows.

		Small window		Large window	
Sensor	Features	Raw	Oversampled	Raw	Oversampled
Color	14	8451	12459	6271	9270
Infrared	7	8176	12318	5966	9036
Gyroscope	48	9289	13857	7002	10446
Empatica	8	2004	2865	1647	2352

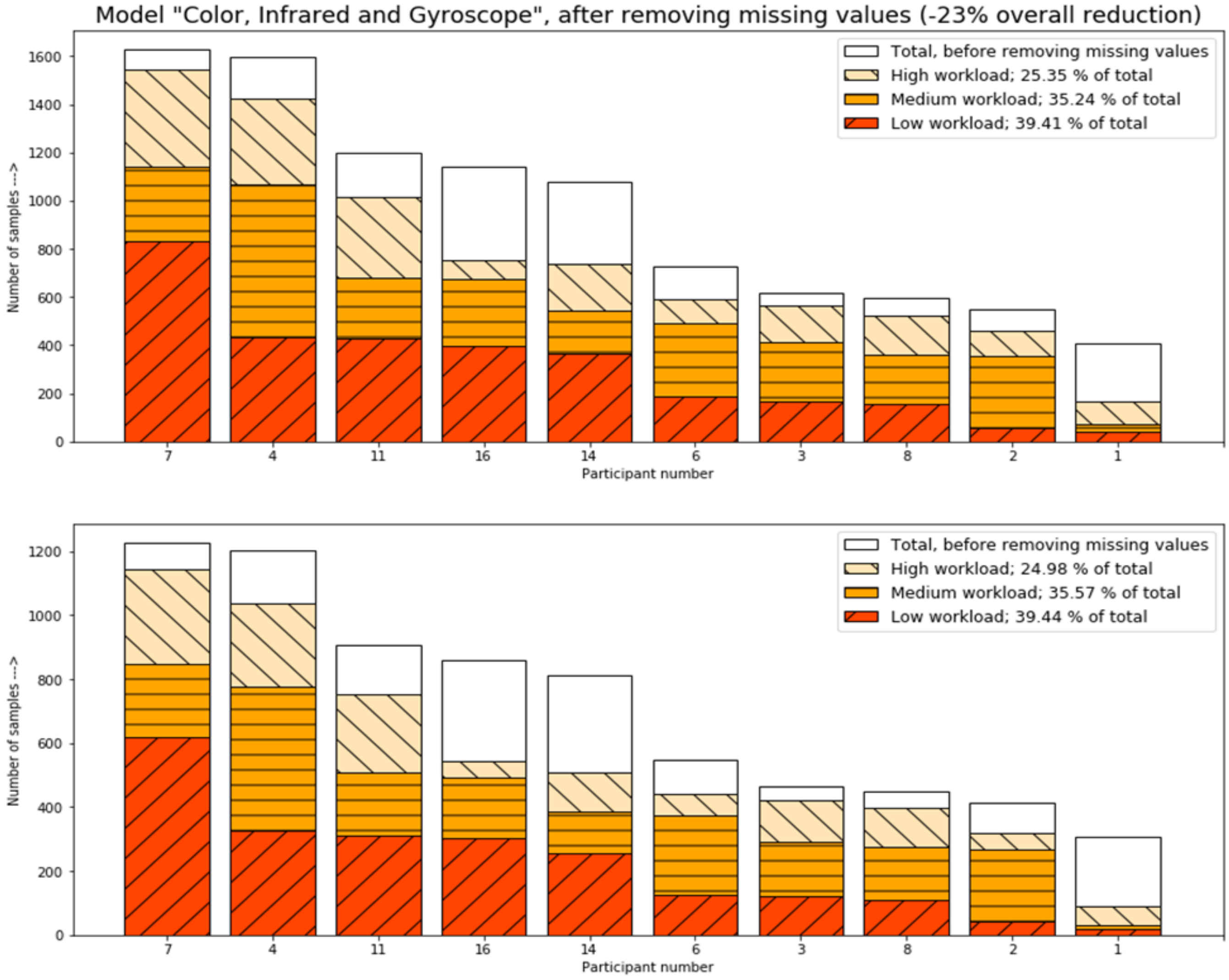


Figure 4: An overview of the raw window samples per participant for the Color, Infrared and Gyroscope sensors in the small (top) and large (bottom) window conditions.

was not pronounced enough to be differentiated from either low- or high mental workload.

The measures were taken with a gyroscope, a wrist-worn sensor, and remote photoplethysmography in the color-, and infrared-spectrum, and feature importance were determined using cross-validated recursive feature elimination. The movement features provided information to classify mental workload states, where a substantial contribution from the fast Fourier frequencies was found, confirming previous findings from the literature (Khusainov et al., 2013). The wrist-worn Empatica E4 sensor did not provide enough data samples to be used in the explored machine learning models. From the color- and infrared spectrum rPPG, heart rate features from the infrared spectrum were found to contribute

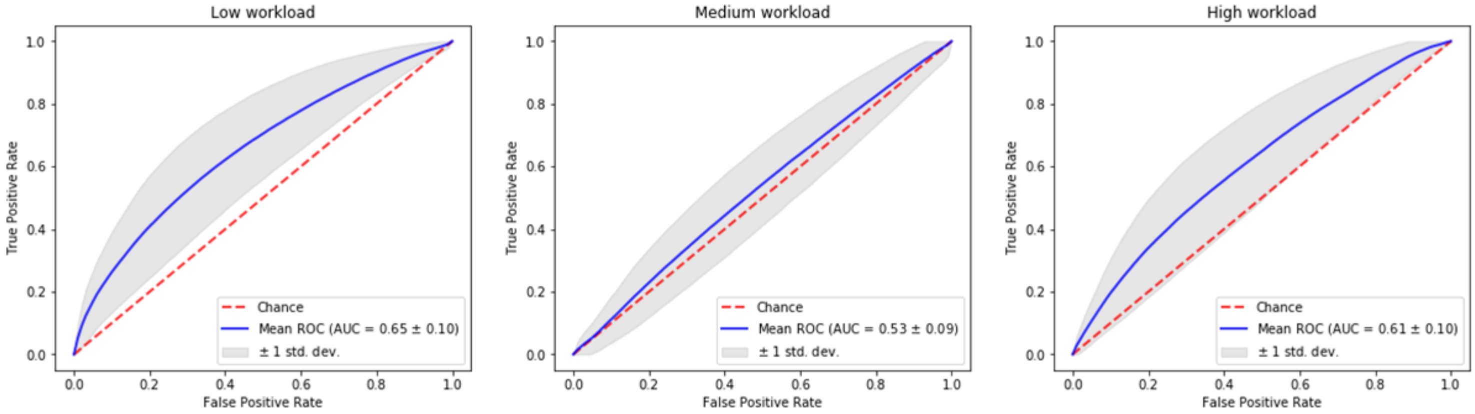


Figure 5: The cross-validated area under receiver-operator curves of the Color-, Infrared and Gyroscope model for low-, medium- and high- workload vs. others classification. With on the x-axis the false positive rate, on the y-axis the true positive rate. The red line is chance performance, the blue 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

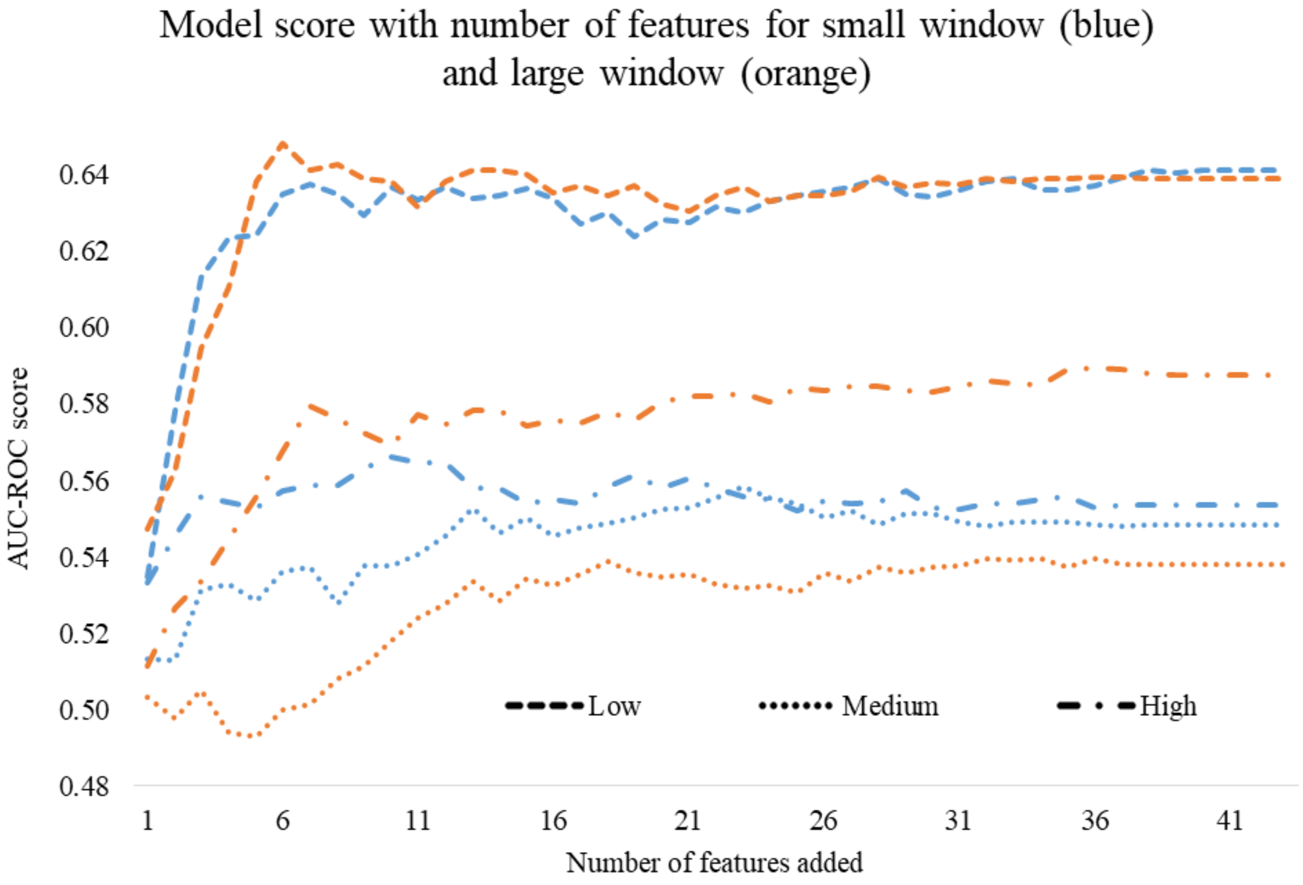


Figure 6: AUC-ROC model performance for the number of features added, where the blue lines are small window size, and the orange lines large window sizes.

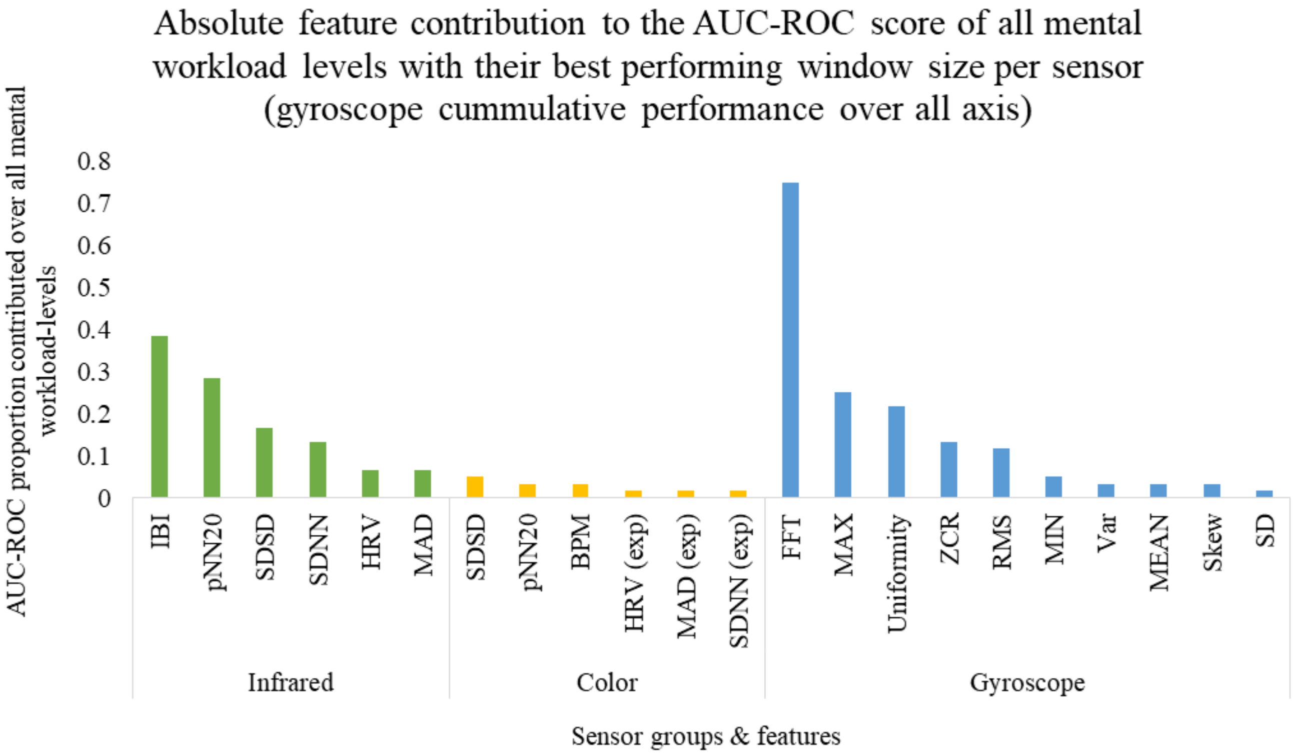


Figure 7: AUC-ROC model performance for the number of features added, where the blue lines are small window size, and the orange lines large window sizes.

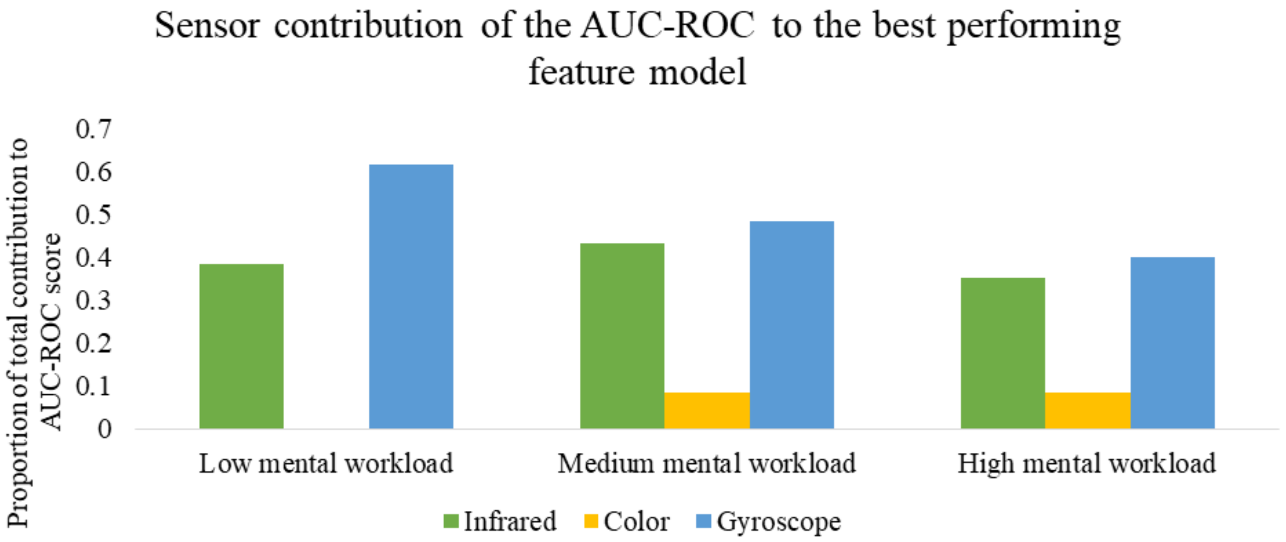


Figure 8: Absolute AUC-ROC score contribution per best performing feature.

most. The high contribution of heart rate features is in line with findings of the mental workload literature (Gastel, Stuijk, & Haan, 2015). It is interesting to see the higher performance of infrared rPPG compared to the color spectrum. A possible explanation might be the quality difference in lighting, although fundamental lighting properties could play a role too Wang et al. (2016).

Since the findings show that heart rate measures can classify mental workload states, and fundamentally both color- and infrared measures rely on the same rPPG principle, cross-sensor compatibility is suspected. This could mean that various other measures for extracting heart rate features can be used in a machine learning model to classify mental workload state, opening up the possibility to choose the measure most practical given the environment it is used in.

In previous research, the quality of the data of the wrist-worn Empatica E4 is found to be sufficient for mental workload classification, and since it also measures heart rate features – it is interesting to see that it failed to contribute in this experiment (Lo et al., 2017). A factor that might explain this contradiction with previous findings is the duration of the data used in this experiment. Both the total duration of each session was relatively short (between 15 and 30 minutes), as were the used window sizes of 45 and 60 seconds. Perhaps using similar overlap- but larger window- sizes of 300 seconds as used by Lo et al. (2017), would result in a coarser temporal resolution of the mental workload measurement, but better performance for the wrist-worn sensor.

The main objective of this research was to determine to what extent unobtrusive mental workload classification in an operator simulation setting is possible. The resulting area-under receiver-operator -curve scores show that it is possible to classify mental workload states. Furthermore, posture movements and heart rate features from the infrared rPPG contribute most to the classification. The properties of no physical contact to the subject in case of the gyroscope and the invisibility to the human eye in case of the infrared light camera make these measures the least obtrusive mental workload classification sensor-setup tested in this paper.

Limitations & Future work

A few considerations should be noted to explain the findings. With regards to internal and ecological validity, expert operators had to act on an uncommon and critical situation in the simulation, where the exact extent of elicited mental workload is unknown. Domain experts constructed the scenario intending to include as much mental workload variation as possible. The fidelity of the simulator and the simulated communication are factors that hampered an ecologically valid setting. The simulator fidelity is moderate; since it is built to practice procedures, it lacks more advanced functionality for fine-grained adjustments compared to reality. The missing fine-grained functionality limits the operator's flexibility, and when routine operations are not possible can cause confusion. The communication with various train operators is another aspect, which – as also indicated by the field experts and operators themselves - is a major contributing factor to the mental workload experienced. Because the experiment leader emulated all communication, it was fewer-, less varied- and serial. These limitations suggest workload levels in the field are expected to be more pronounced.

rPPG Signal capture The quality of the infrared- and color rPPG signal can be improved by the hardware used, lighting and compression settings of the recording.

The quality of a camera lens influences how the light reaches the image sensor and influences the resulting image detail, where a high-grade lens will improve sharpness and color. A general-purpose outdoor sports camera, the GoPro Hero 7 black, with its default lens, was used to record the color spectrum. A situation-specific image capture device with a dedicated lens for indoor use under relatively static conditions will produce more detailed frames in which the rPPG algorithm can retrieve a stronger PPG signal. The infrared camera came with a dedicated light-source for optimal lighting, which could illuminate the subject's face frontally. For the color spectrum recording, a CRI95+ rated LED light was, which, due to the intensity of the LED lamp, could not light the subject frontally without obtrusively blinding. This raises the question if the difference in lighting conditions could be a factor for the comparatively lower color-spectrum performance. Another color-related improvement is adjusting the color settings of the recording manually. By increasing the dynamic range of the colors, amplitude differences should become more distinct, allowing for better discrimination of signals. Due to the restraints of the proprietary Basler software, uncompressed $2*10^5$ kbit/s recordings of more than two minutes could not be made. Using a raw and uncompressed image stream McDuff et al. (2017) showed that compared to a compressed stream, the raw stream yields a much cleaner signal with a significantly higher signal-to-noise ratio PPG signal. Preliminary testing on small sub-two minute segments recorded with the Basler camera in the infrared spectrum confirmed this finding. This indicates that applying on-line processing on this stream yields much better rPPG signals while eliminating the storage needs.

Data processing & classification The preprocessing, feature extraction and workload state labeling can contribute to a better model.

The amplitude selective filter algorithm could be tweaked by making use of both infrared- and color channels as Trumpp et al. (2018) have done. In this case, the infrared and color channels would be merged into one stream, using the color channels to remove non-heart rate related frequencies, and the infrared for the heart rate related frequency. To sustain temporal synchronization, the facial landmark tracking, time synchronization, and the horizontal camera vantage point between the infrared- and color-spectrum recording would need to be controlled for. Top-down, literature driven feature extraction was used in this paper. (Autoregressive) convolutional neural nets are another way for feature extraction, which have shown promising results for time series data (Bińkowski, Marti, & Donnat, 2017; Yang, Nguyen, San, Li, & Krishnaswamy, 2015). These autoregressive convolutional neural nets could be used both for posture movement- and heart rate- feature extraction.

The transition between levels of mental workload is modeled instantaneous; the trigger of an event results in immediate mental workload change in the labeling of the data; however, the psychophysical mental workload change is more gradual (Kim, Cheon, Bai, Lee, & Koo, 2018). Because of this more gradual psychophysiological change, sections spanning these transitions are of ambiguous mental workload state. Finer grained levels of mental workload to capture the mental workload transition states (Van Gent et al., 2018), or informed data selection around an event, as is typical for EEG event-related research, could be a solution (Luck, 2014).

Acknowledgments

The author would like to thank Julia Lo, Nico Romeijn & Chris Janssen for their supervision; Emdzad, Wilco, Timo & Jurjen for their shared expertise, and from the ProRail Amsterdam Train Traffic control center; Jeffry, Sjoerd, Amanda, and the train traffic controllers that participated in this study for their help.

References

Alberdi, A., Aztiria, A., & Basarab, A. (2016). Towards an automatic early stress recognition system for office environments based on multimodal measurements: A review. *Journal of biomedical informatics*, *59*, 49–75.

Andersson, G., Hagman, J., Talianzadeh, R., Svedberg, A., & Larsen, H. C. (2002). Effect of cognitive load on postural control. *Brain research bulletin*, *58*(1), 135–139.

Arnrich, B., Setz, C., La Marca, R., Tröster, G., & Ehlert, U. (2010). What does your chair know about your stress level? *IEEE Trans. Information Technology in Biomedicine*, *14*(2), 207–214.

Atallah, L., Lo, B., King, R., & Yang, G.-Z. (2010). Sensor placement for activity detection using wearable accelerometers. In *2010 international conference on body sensor networks* (pp. 24–29).

Bińkowski, M., Marti, G., & Donnat, P. (2017). Autoregressive convolutional neural networks for asynchronous time series. *arXiv preprint arXiv:1703.04122*.

Borst, C., Wieling, W., Van Brederode, J., Hond, A., De Rijk, L., & Dunning, A. (1982). Mechanisms of initial heart rate response to postural change. *American Journal of Physiology-Heart and Circulatory Physiology*, *243*(5), H676–H681.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32.

Bulat, A., & Tzimiropoulos, G. (2017). 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).

Cawley, G. C., & Talbot, N. L. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, *11*(Jul), 2079–2107.

Charles, R. L., & Nixon, J. (2019). Measuring mental workload using physiological measures: a systematic review. *Applied ergonomics*, *74*, 221–232.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321–357.

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE transactions on information theory*, *13*(1), 21–27.

Figo, D., Diniz, P. C., Ferreira, D. R., & Cardoso, J. M. (2010). Preprocessing techniques for context recognition from accelerometer data. *Personal and Ubiquitous Computing*, *14*(7), 645–662.

Freund, Y., Schapire, R., & Abe, N. (1999). A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence*, *14*(771-780), 1612.

Gaillard, A. (1993). Comparing the concepts of mental load and stress. *Ergonomics*, *36*(9), 991–1005.

Gastel, M. van, Stuijk, S., & Haan, G. de. (2015). Motion robust remote-ppg in infrared. *IEEE Transactions on Biomedical Engineering*, *62*(5), 1425–1433.

Ghosh, A., Danieli, M., & Riccardi, G. (2015). 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).

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine learning*, *46*(1-3), 389–422.

Hart, S. G., & Staveland, L. E. (1988). Development of nasa-tlx (task load index): Results of empirical and theoretical research. In *Advances in psychology* (Vol. 52, pp. 139–183). Elsevier.

Harteis, C. (2018). Machines, change and work: An educational view on the digitalization of work. In *The impact of digitalization in the workplace* (pp. 1–10). Springer.

Head, T., MechCoder, Louppe, G., Shcherbatyi, I., fcharras, Vinícius, Z., et al. (2018, March). *scikit-optimize/scikit-optimize: v0.5.2.* Zenodo. Available from `https://doi.org/10.5281/zenodo.1207017`

Hogervorst, M. A., Brouwer, A.-M., & Van Erp, J. B. (2014). Combining and comparing eeg, peripheral physiology and eye-related measures for the assessment of mental workload. *Frontiers in neuroscience*, *8*, 322.

Huang, J., & Ling, C. X. (2005). Using auc and accuracy in evaluating learning algorithms. *IEEE Transactions on knowledge and Data Engineering*, *17*(3), 299–310.

Huelsbusch, M., & Blazek, V. (2002). 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).

Jones, E., Oliphant, T., Peterson, P., et al. (2001). Scipy: Open source scientific tools for python.

Khusainov, R., Azzi, D., Achumba, I. E., & Bersch, S. D.

(2013). Real-time human ambulation, activity, and physiological monitoring: Taxonomy of issues, techniques, applications, challenges and limitations. *Sensors*, *13*(10), 12852–12902.

Kim, H.-G., Cheon, E.-J., Bai, D.-S., Lee, Y. H., & Koo, B.-H. (2018). Stress and heart rate variability: A meta-analysis and review of the literature. *Psychiatry investigation*, *15*(3), 235.

King, D. E. (2009). Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, *10*(Jul), 1755–1758.

Lempe, G., Zaunseder, S., Wirthgen, T., Zipser, S., & Malberg, H. (2013). Roi selection for remote photoplethysmography. In *Bildverarbeitung für die medizin 2013* (pp. 99–103). Springer.

Lo, J. C., Sehic, E., & Meijer, S. A. (2017). Measuring mental workload with low-cost and wearable sensors: Insights into the accuracy, obtrusiveness, and research usability of three instruments. *Journal of cognitive engineering and decision making*, *11*(4), 323–336.

Lopez, F. S., Condori-Fernandez, N., & Catala, A. (2018). Towards real-time automatic stress detection for office workplaces. In *Annual international symposium on information management and big data* (pp. 273–288).

Luck, S. J. (2014). *An introduction to the event-related potential technique*. MIT press.

Lux, E., Adam, M. T., Dorner, V., Helming, S., Knierim, M. T., & Weinhardt, C. (2018). Live biofeedback as a user interface design element: A review of the literature. *Communications of the Association for Information Systems*, *43*(1), 257–296.

Martinez, R., Irigoyen, E., Arruti, A., Martín, J. I., & Muguerza, J. (2017). A real-time stress classification system based on arousal analysis of the nervous system by an f-state machine. *Computer methods and programs in biomedicine*, *148*, 81–90.

McCarthy, C., Pradhan, N., Redpath, C., & Adler, A. (2016). Validation of the empatica e4 wristband. In *2016 ieee embs international student conference (isc)* (pp. 1–4).

McDuff, D. J., Blackford, E. B., & Estepp, J. R. (2017). 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).

McKinney, W., et al. (2010). Data structures for statistical computing in python. In *Proceedings of the 9th python in science conference* (Vol. 445, pp. 51–56).

McNames, J., & Aboy, M. (2006). Reliability and accuracy of heart rate variability metrics versus ecg segment duration. *Medical and Biological Engineering and Computing*, *44*(9), 747–756.

Mitchell, J. P., Macrae, C. N., & Gilchrist, I. D. (2002). Working memory and the suppression of reflexive saccades. *Journal of cognitive neuroscience*, *14*(1), 95–103.

Mourcou, Q., Fleury, A., Franco, C., Klopcic, F., & Vuillerme, N. (2015). Performance evaluation of smartphone inertial sensors measurement for range of motion. *Sensors*, *15*(9), 23168–23187.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in python. *Journal of machine learning research*, *12*(Oct), 2825–2830.

Rawenwaaij-Arts, C., Kallee, L., Hopman, J., et al. (1993). Task force of the european society of cardiology and the north american society of pacing and electrophysiology. heart rate variability. standards of measurement, physiologic interpretation, and clinical use. circulation 1996; 93: 1043-1065. *Intern. Med*, *118*, 436–447.

Salahuddin, L., Cho, J., Jeong, M. G., & Kim, D. (2007). 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).

Shahriari, B., Swersky, K., Wang, Z., Adams, R. P., & De Freitas, N. (2015). Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE*, *104*(1), 148–175.

Swan, M. (2012). Sensor mania! the internet of things, wearable computing, objective metrics, and the quantified self 2.0. *Journal of Sensor and Actuator networks*, *1*(3), 217–253.

Takano, C., & Ohta, Y. (2007). Heart rate measurement based on a time-lapse image. *Medical engineering & physics*, *29*(8), 853–857.

Trumpp, A., Lohr, J., Wedekind, D., Schmidt, M., Burghardt, M., Heller, A. R., et al. (2018). Camera-based photoplethysmography in an intraoperative setting. *Biomedical engineering online*, *17*(1), 33.

Van Gent, P., Farah, H., Nes, N. van, & Arem, B. van. (2018). Analysing noisy driver physiology real-time using off-the-self sensors: Heart rate analysis software from the taking the fast lane project. *Journal of Open Research Software*.

Van Rossum, G., & Drake Jr, F. L. (1995). *Python tutorial*. Centrum voor Wiskunde en Informatica Amsterdam.

Verkruysse, W., Svaasand, L. O., & Nelson, J. S. (2008). Remote plethysmographic imaging using ambient light. *Optics express*, *16*(26), 21434–21445.

Wang, W., Brinker, A. C. den, Stuijk, S., & Haan, G. de. (2016). Algorithmic principles of remote ppg. *IEEE Transactions on Biomedical Engineering*, *64*(7), 1479–1491.

Wang, W., Brinker, A. C. den, Stuijk, S., & Haan, G. de. (2017). Amplitude-selective filtering for remote-ppg. *Biomedical optics express*, *8*(3), 1965–1980.

Welford, A. (1978). Mental work-load as a function of demand, capacity, strategy and skill. *Ergonomics*, *21*(3), 151–167.

Yang, J., Nguyen, M. N., San, P. P., Li, X. L., & Krishnaswamy, S. (2015). Deep convolutional neural networks on multichannel time series for human activity recognition. In *Twenty-fourth international joint conference on artificial* *intelligence.*

Young, M. S., Brookhuis, K. A., Wickens, C. D., & Hancock, P. A. (2015). State of science: mental workload in ergonomics. *Ergonomics*, *58*(1), 1–17.

Yu, H., Cang, S., & Wang, Y. (2016). 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).

Zaunseder, S., Trumpp, A., Wedekind, D., & Malberg, H. (2018). Cardiovascular assessment by imaging photoplethysmography–a review. *Biomedical Engineering/Biomedizinische Technik*, *63*(5), 617–634.

Appendix

Appendix 1: GoPro and Basler Infrared Settings

Table 1

Color settings GoPro Hero 7 Black

Setting	*Value*
Aspect Ratio	16:9
Resolution	720 x 1280
Frames per second	60 (59.94 actual)
Field of View	Linear
Video Stabilisation	Off
Low Light	Off
Short CI	Off
Protune	On
Shutter	1/120 (should be: 1/ (framerate* 2))
ISO min	100
ISO max	400
Whitebalance	4000 K
Sharpness	Medium
Color	Flat
Raw Audio Track	Off
AAMG	Auto

Table 2

Basler Infrared Settings

Setting	Value
SequenceSetTotalNumber	2
SequenceSetIndex	0
SequenceSetExecutions	1
SequenceAdvanceMode	Auto
GainAuto	Off
GainSelector	All
GainRaw	600
GainSelector	All
BlackLevelSelector	All
BlackLevelRaw	64
BlackLevelSelector	All
GammaEnable	1
GammaSelector	sRGB
DigitalShift	0
PixelFormat	Mono12Packed
ReverseX	0
TestImageSelector	Off
Width	659
Height	494
OffsetX	0
OffsetY	0
CenterX	0
CenterY	0
BinningHorizontal	1
BinningVertical	1
TriggerSelector	AcquisitionStart
TriggerMode	Off
TriggerSelector	FrameStart
TriggerMode	On
TriggerSelector	FrameStart
TriggerSelector	AcquisitionStart
TriggerSource	Line1
TriggerSelector	FrameStart
TriggerSource	Line1
TriggerSelector	FrameStart
TriggerSelector	AcquisitionStart
TriggerActivation	RisingEdge
TriggerSelector	FrameStart
TriggerActivation	RisingEdge
TriggerSelector	FrameStart
TriggerSelector	AcquisitionStart
TriggerDelayAbs	0
TriggerSelector	FrameStart
TriggerDelayAbs	0
TriggerSelector	FrameStart
ExposureMode	TriggerWidth
ExposureAuto	Off
ExposureOverlapTimeMaxRaw	0
AcquisitionFrameRateEnable	1
AcquisitionFrameRateAbs	30.0003
LineSelector	Line1
LineMode	Input
LineSelector	Out1
LineMode	Output
LineSelector	Out1
LineSelector	Line1
LineFormat	OptoCoupled
LineSelector	Out1
LineFormat	OptoCoupled
LineSelector	Out1
LineSelector	Out1
LineSource	UserOutput
LineSelector	Out1
LineSelector	Line1
LineInverter	0
LineSelector	Out1
LineInverter	0
LineSelector	Out1
LineSelector	Line1
LineDebouncerTimeRaw	10000
LineSelector	Out1
UserOutputValueAll	0
CounterSelector	Counter1
CounterEventSource	FrameTrigger
CounterSelector	Counter2
CounterEventSource	FrameStart
CounterSelector	Counter1
CounterSelector	Counter1
CounterResetSource	Off
CounterSelector	Counter2
CounterResetSource	Off
CounterSelector	Counter1
LUTSelector	Luminance
LUTEnable	0
LUTSelector	Luminance
LUTSelector	Luminance
LUTValueAll	
LUTSelector	Luminance
GevStreamChannelSelector	StreamChannel0
GevSCPSPacketSize	9000
GevStreamChannelSelector	StreamChannel0
GevStreamChannelSelector	StreamChannel0
GevSCPD	0
GevStreamChannelSelector	StreamChannel0
GevStreamChannelSelector	StreamChannel0
GevSCFTD	0
GevStreamChannelSelector	StreamChannel0
GevStreamChannelSelector	StreamChannel0
GevSCBWR	10
GevStreamChannelSelector	StreamChannel0
GevStreamChannelSelector	StreamChannel0
GevSCBWRA	1
GevStreamChannelSelector	StreamChannel0
AutoTargetValue	1280
AutoGainRawLowerLimit	300
AutoGainRawUpperLimit	600
AutoExposureTimeAbsLowerLimit	4
AutoExposureTimeAbsUpperLimit	1e+06

Setting	Value
AutoFunctionProfile	GainMinimum
AutoFunctionAOISelector	AOI1
AutoFunctionAOIWidth	659
AutoFunctionAOISelector	AOI2
AutoFunctionAOIWidth	659
AutoFunctionAOISelector	AOI1
AutoFunctionAOISelector	AOI1
AutoFunctionAOIHeight	494
AutoFunctionAOISelector	AOI2
AutoFunctionAOIHeight	494
AutoFunctionAOISelector	AOI1
AutoFunctionAOISelector	AOI1
AutoFunctionAOIOffsetX	0
AutoFunctionAOISelector	AOI2
AutoFunctionAOIOffsetX	0
AutoFunctionAOISelector	AOI1
AutoFunctionAOISelector	AOI1
AutoFunctionAOIOffsetY	0
AutoFunctionAOISelector	AOI2
AutoFunctionAOIOffsetY	0
AutoFunctionAOISelector	AOI1
UserDefinedValueSelector	Value1
UserDefinedValue	0
UserDefinedValueSelector	Value2
UserDefinedValue	0
UserDefinedValueSelector	Value3
UserDefinedValue	0
UserDefinedValueSelector	Value4
UserDefinedValue	0
UserDefinedValueSelector	Value5
UserDefinedValue	0
UserDefinedValueSelector	Value1
ParameterSelector	Gain
RemoveLimits	0
ParameterSelector	Framerate
RemoveLimits	0
ParameterSelector	Gain
ChunkModeActive	0
EventSelector	ExposureEnd
EventNotification	Off
EventSelector	FrameStartOvertrigger
EventNotification	Off
EventSelector	AcquisitionStartOvertrigger
EventNotification	Off
EventSelector	FrameStart
EventNotification	Off
EventSelector	AcquisitionStart
EventNotification	Off
EventSelector	EventOverrun
EventNotification	Off
EventSelector	ExposureEnd

Appendix 2a: Classifier performances on Infrared and Gyroscope and Infrared, for all mental workload states and window sizes

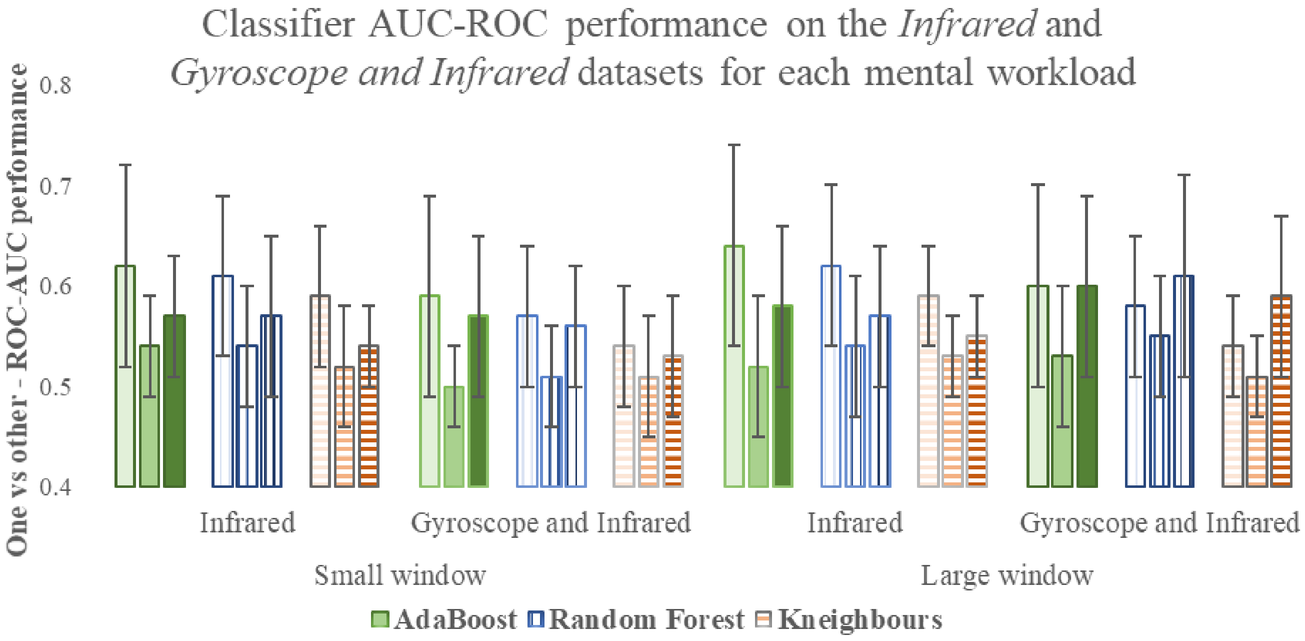
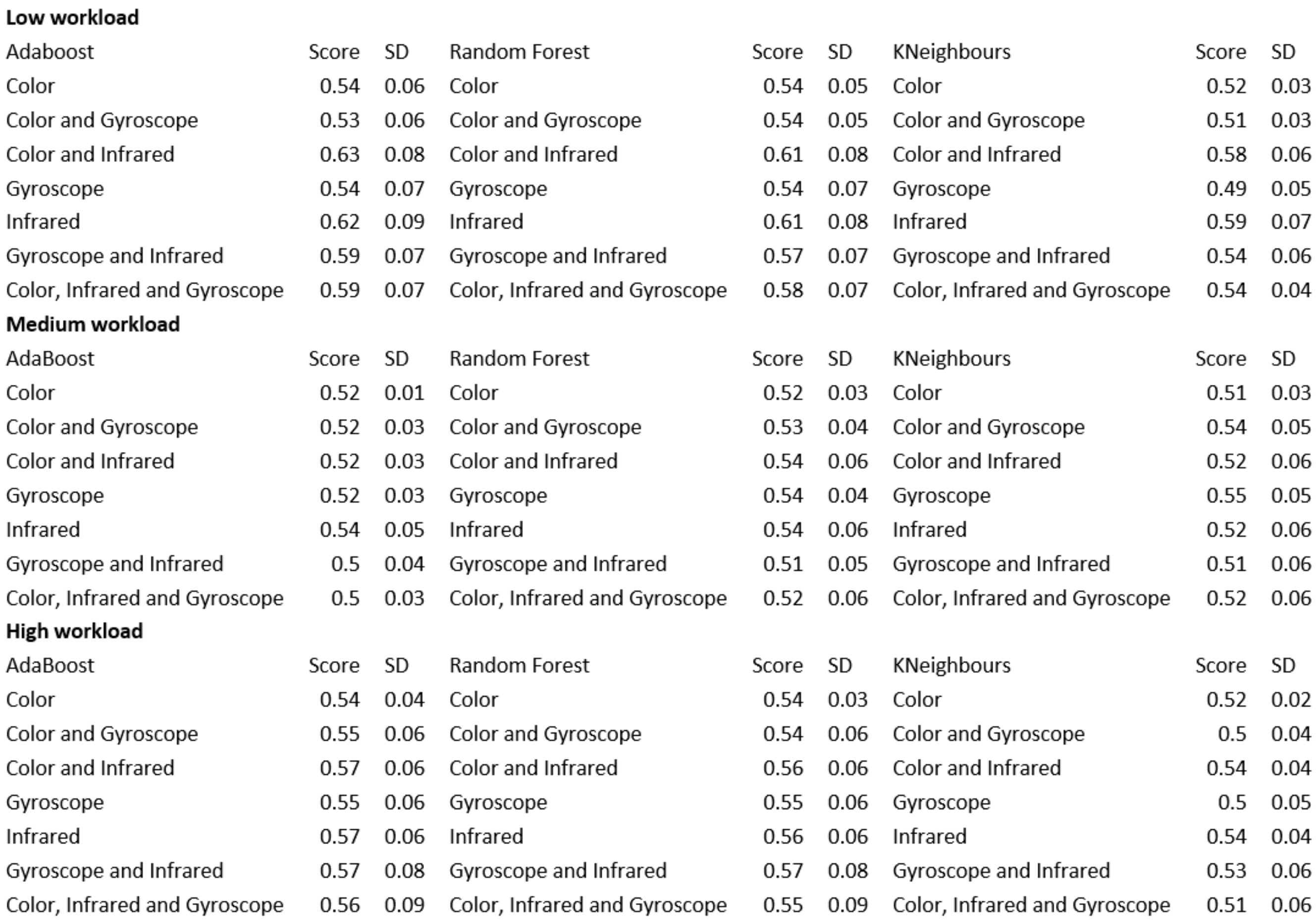


Figure 9: An overview of the classifier ROC-AUC score for the different sensor groups and window sizes and mental workload levels.

Appendix 2b: Classifier performances on all sensor combinations, for all mental workload states and window sizes



Low workload

Adaboost	Score	SD	Random Forest	Score	SD	KNeighbours	Score	SD
Color	0.54	0.06	Color	0.54	0.05	Color	0.52	0.03
Color and Gyroscope	0.53	0.06	Color and Gyroscope	0.54	0.05	Color and Gyroscope	0.51	0.03
Color and Infrared	0.63	0.08	Color and Infrared	0.61	0.08	Color and Infrared	0.58	0.06
Gyroscope	0.54	0.07	Gyroscope	0.54	0.07	Gyroscope	0.49	0.05
Infrared	0.62	0.09	Infrared	0.61	0.08	Infrared	0.59	0.07
Gyroscope and Infrared	0.59	0.07	Gyroscope and Infrared	0.57	0.07	Gyroscope and Infrared	0.54	0.06
Color, Infrared and Gyroscope	0.59	0.07	Color, Infrared and Gyroscope	0.58	0.07	Color, Infrared and Gyroscope	0.54	0.04

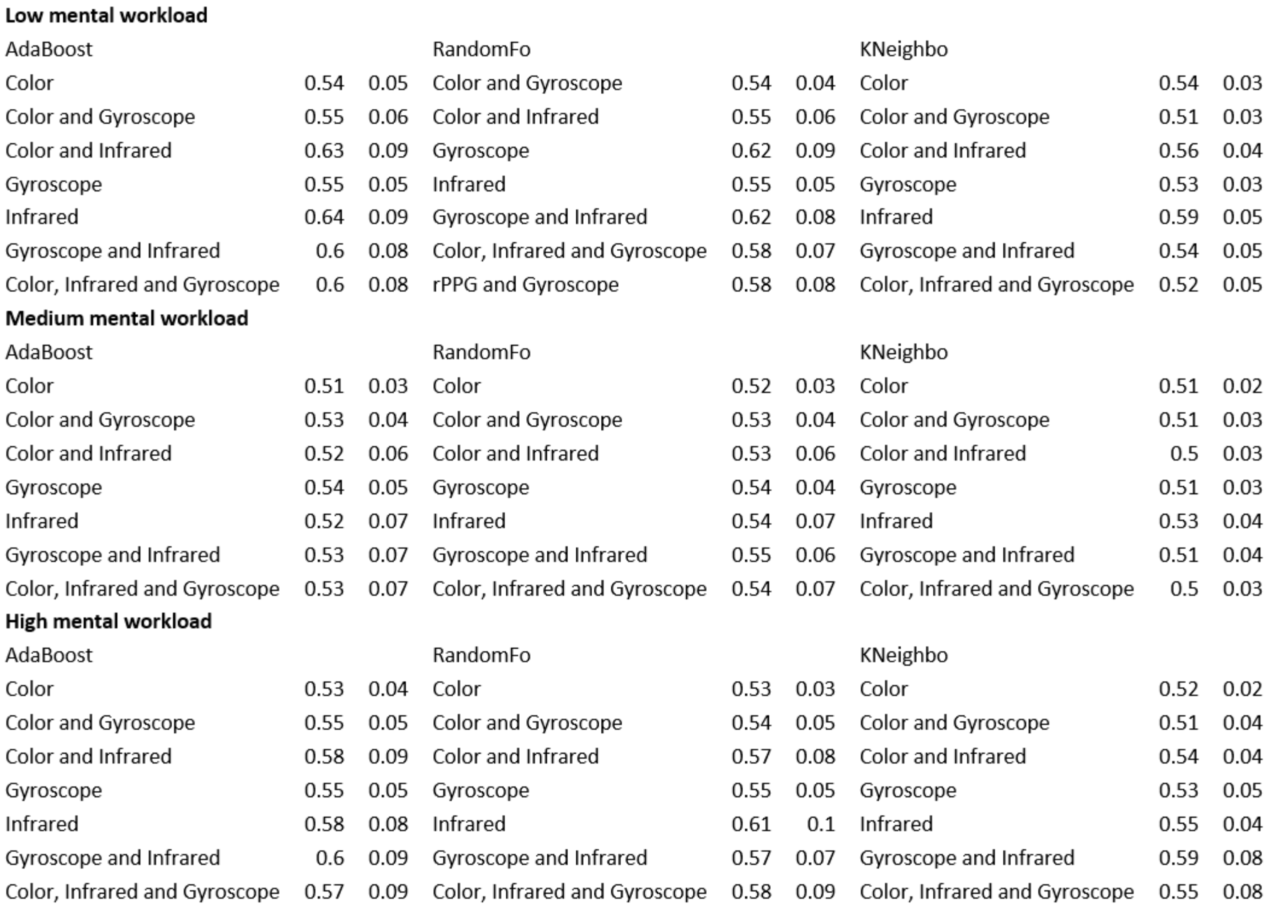
Medium workload

AdaBoost	Score	SD	Random Forest	Score	SD	KNeighbours	Score	SD
Color	0.52	0.01	Color	0.52	0.03	Color	0.51	0.03
Color and Gyroscope	0.52	0.03	Color and Gyroscope	0.53	0.04	Color and Gyroscope	0.54	0.05
Color and Infrared	0.52	0.03	Color and Infrared	0.54	0.06	Color and Infrared	0.52	0.06
Gyroscope	0.52	0.03	Gyroscope	0.54	0.04	Gyroscope	0.55	0.05
Infrared	0.54	0.05	Infrared	0.54	0.06	Infrared	0.52	0.06
Gyroscope and Infrared	0.5	0.04	Gyroscope and Infrared	0.51	0.05	Gyroscope and Infrared	0.51	0.06
Color, Infrared and Gyroscope	0.5	0.03	Color, Infrared and Gyroscope	0.52	0.06	Color, Infrared and Gyroscope	0.52	0.06

High workload

AdaBoost	Score	SD	Random Forest	Score	SD	KNeighbours	Score	SD
Color	0.54	0.04	Color	0.54	0.03	Color	0.52	0.02
Color and Gyroscope	0.55	0.06	Color and Gyroscope	0.54	0.06	Color and Gyroscope	0.5	0.04
Color and Infrared	0.57	0.06	Color and Infrared	0.56	0.06	Color and Infrared	0.54	0.04
Gyroscope	0.55	0.06	Gyroscope	0.55	0.06	Gyroscope	0.5	0.05
Infrared	0.57	0.06	Infrared	0.56	0.06	Infrared	0.54	0.04
Gyroscope and Infrared	0.57	0.08	Gyroscope and Infrared	0.57	0.08	Gyroscope and Infrared	0.53	0.06
Color, Infrared and Gyroscope	0.56	0.09	Color, Infrared and Gyroscope	0.55	0.09	Color, Infrared and Gyroscope	0.51	0.06

Figure 10: Classifier performance scores for small window size



Low mental workload								
AdaBoost			RandomFo			KNeighbo		
Color	0.54	0.05	Color and Gyroscope	0.54	0.04	Color	0.54	0.03
Color and Gyroscope	0.55	0.06	Color and Infrared	0.55	0.06	Color and Gyroscope	0.51	0.03
Color and Infrared	0.63	0.09	Gyroscope	0.62	0.09	Color and Infrared	0.56	0.04
Gyroscope	0.55	0.05	Infrared	0.55	0.05	Gyroscope	0.53	0.03
Infrared	0.64	0.09	Gyroscope and Infrared	0.62	0.08	Infrared	0.59	0.05
Gyroscope and Infrared	0.6	0.08	Color, Infrared and Gyroscope	0.58	0.07	Gyroscope and Infrared	0.54	0.05
Color, Infrared and Gyroscope	0.6	0.08	rPPG and Gyroscope	0.58	0.08	Color, Infrared and Gyroscope	0.52	0.05
Medium mental workload								
AdaBoost			RandomFo			KNeighbo		
Color	0.51	0.03	Color	0.52	0.03	Color	0.51	0.02
Color and Gyroscope	0.53	0.04	Color and Gyroscope	0.53	0.04	Color and Gyroscope	0.51	0.03
Color and Infrared	0.52	0.06	Color and Infrared	0.53	0.06	Color and Infrared	0.5	0.03
Gyroscope	0.54	0.05	Gyroscope	0.54	0.04	Gyroscope	0.51	0.03
Infrared	0.52	0.07	Infrared	0.54	0.07	Infrared	0.53	0.04
Gyroscope and Infrared	0.53	0.07	Gyroscope and Infrared	0.55	0.06	Gyroscope and Infrared	0.51	0.04
Color, Infrared and Gyroscope	0.53	0.07	Color, Infrared and Gyroscope	0.54	0.07	Color, Infrared and Gyroscope	0.5	0.03
High mental workload								
AdaBoost			RandomFo			KNeighbo		
Color	0.53	0.04	Color	0.53	0.03	Color	0.52	0.02
Color and Gyroscope	0.55	0.05	Color and Gyroscope	0.54	0.05	Color and Gyroscope	0.51	0.04
Color and Infrared	0.58	0.09	Color and Infrared	0.57	0.08	Color and Infrared	0.54	0.04
Gyroscope	0.55	0.05	Gyroscope	0.55	0.05	Gyroscope	0.53	0.05
Infrared	0.58	0.08	Infrared	0.61	0.1	Infrared	0.55	0.04
Gyroscope and Infrared	0.6	0.09	Gyroscope and Infrared	0.57	0.07	Gyroscope and Infrared	0.59	0.08
Color, Infrared and Gyroscope	0.57	0.09	Color, Infrared and Gyroscope	0.58	0.09	Color, Infrared and Gyroscope	0.55	0.08

Figure 11: Classifier performance scores large window size

Appendix 3: An overview of the best performing window size- and features per workload level on the AUC-ROC score.

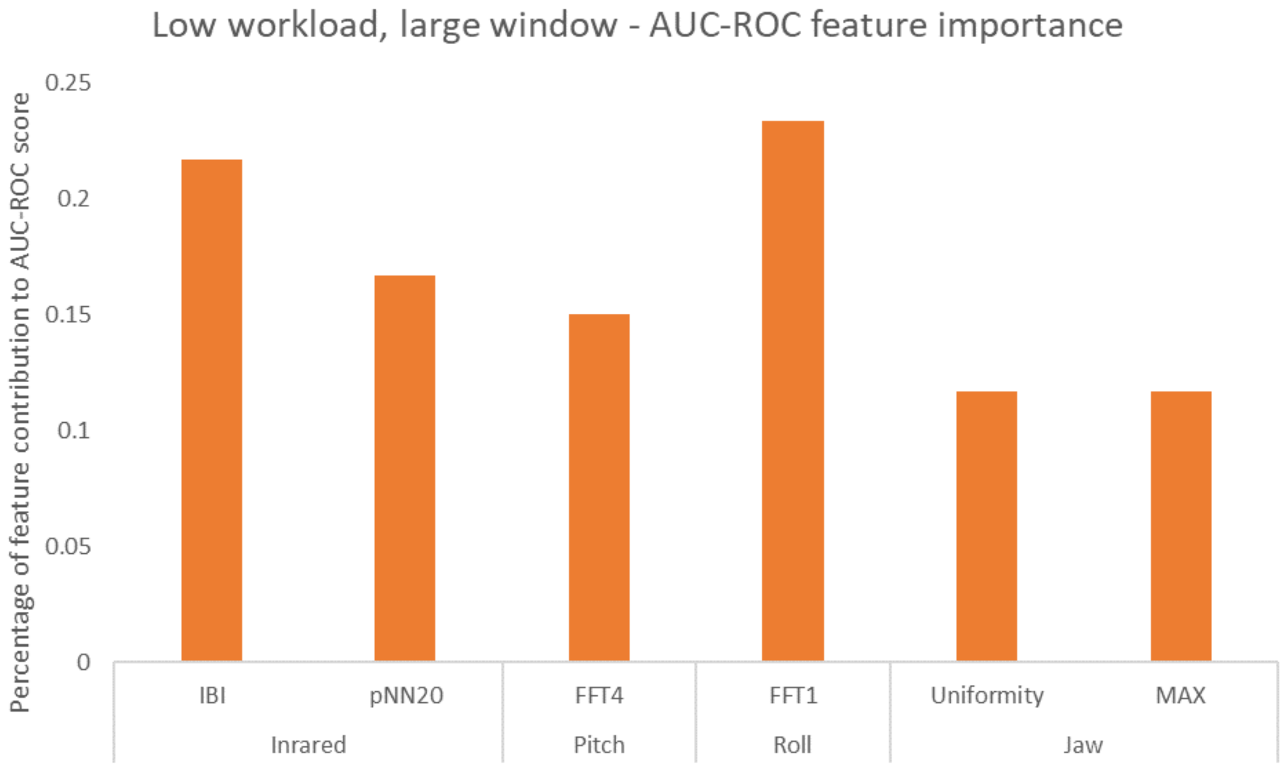


Figure 12: Absolute feature contribution to the AUC-ROC score of low-mental workload classification using a large window

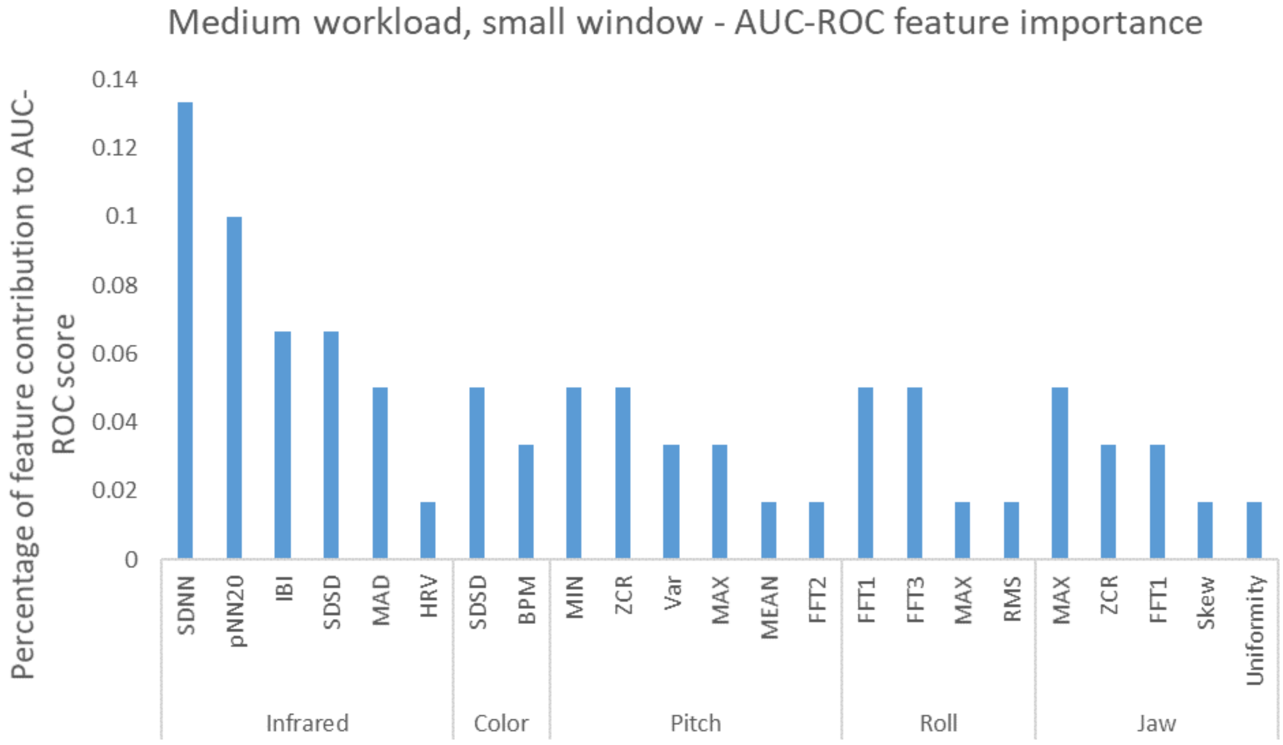


Figure 13: Absolute feature contribution to the AUC-ROC score of medium-mental workload classification using a small window

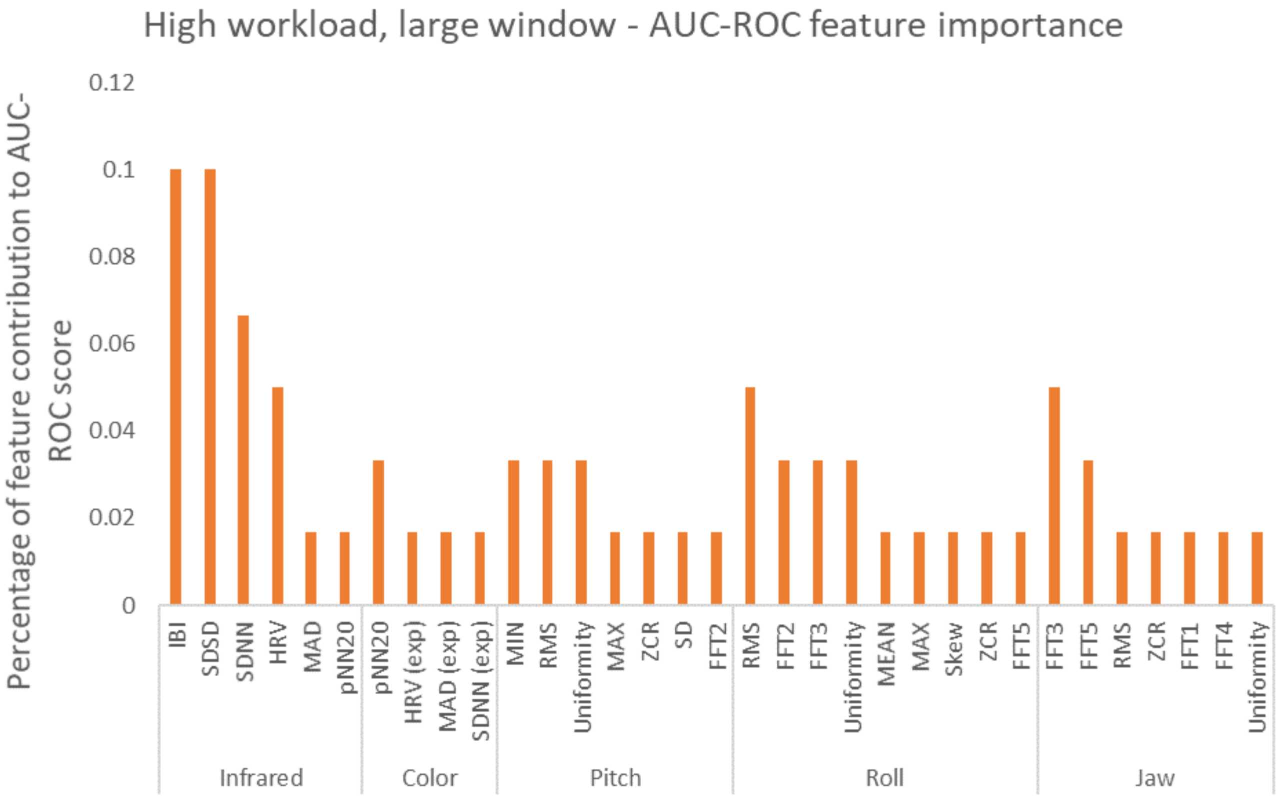


Figure 14: Absolute feature contribution to the AUC-ROC score of high-mental workload classification using a large window

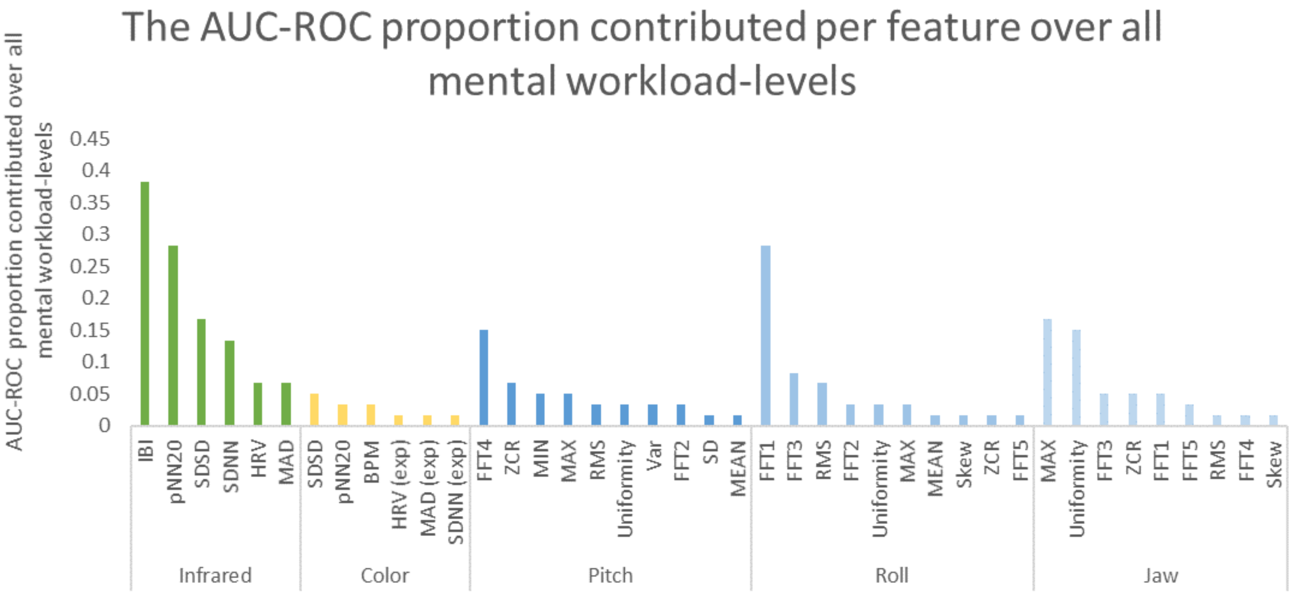


Figure 15: Absolute AUC-ROC score contribution of all best performing features.