In-Ear SpO_2 for Classification of Cognitive Workload

Harry J. Davies, Ian Williams, Ghena Hammour, Metin Yarici,
Michael J. Stacey, Barry M. Seemungal and Danilo P. Mandic

***Abstract*—The brain is the most metabolically active organ in the body, which increases its metabolic activity, and thus oxygen consumption, with increasing cognitive demand. This motivates us to question whether increased cognitive workload may be measurable through changes in blood oxygen saturation. To this end, we explore the feasibility of cognitive workload tracking based on in-ear SpO_2 measurements, which are known to be both robust and exhibit minimal delay. We consider cognitive workload assessment based on an N-back task with randomised order. It is shown that the 2-back and 3-back tasks (high cognitive workload) yield either the lowest median absolute SpO_2 or largest median decrease in SpO_2 in all of the subjects, indicating a measurable and statistically significant decrease in blood oxygen in response to increased cognitive workload. This makes it possible to classify the four N-back task categories, over 5 second epochs, with a mean accuracy of 90.6%, using features derived from in-ear pulse oximetry, including SpO_2, pulse rate and respiration rate. These findings suggest that in-ear SpO_2 measurements provide sufficient information for the reliable classification of cognitive workload over short time windows, which promises a new avenue for real time cognitive workload tracking.**

I. Introduction

COGNITIVE workload is defined as the level of mental effort undertaken by an individual in response to a task. The mental effort is usually related to working memory, and thus corresponds to the utilisation of brain resources [1]. Cognitive workload affects almost every task-related aspect of our daily lives, from general learning to driving to internet browsing. The ability to accurately measure cognitive load would yield manifold benefits, as too little cognitive workload leaves us vulnerable to distraction, whereas too much cognitive workload makes us prone to making mistakes. Depending on the task a person is engaged in, these mistakes can be more benign, such as less efficient studying, through to life threatening as is the case with driving and the possibility of fatal accidents. The ability to accurately measure and predict cognitive workload would therefore make possible personalised task adaptation, together with the associated benefits on an individual and a societal level, from increasing productivity to decreasing the likelihood of mistakes. Classification of cognitive workload therefore promises immense benefit in diverse areas ranging from driver safety to augmenting human capability with closed loop brain computer interface.

A. Cognitive workload tracking

It is natural to attempt to track cognitive workload based on scalp electroencephalography (EEG), with examples including the classification of skilled vs bad driver performance [2] and the prediction of performance in working memory tasks [3]. Whilst scalp EEG has proven effective at discerning the relevant brain activity changes that arise from changes in cognitive workload, it is generally obtrusive and thus impractical for daily life applications. Discrete and non-stigmatising wearable solutions are still being developed, such as Hearables [4] and ear-EEG [5] [6].

In recent years, eye gaze tracking has become a useful tool for estimating cognitive workload, such as in classification of cognitive workload as well as predicting correctness in an N-back task whilst in a driving simulator [7] [8]. However, the ways to measure gaze and pupil dilation inevitably involve cameras; these are generally fixed and positioned to track the face and eyes and can be embedded into glasses for wearable gaze tracking.

Other sensing modalities relevant for the estimation of cognitive load include electrocardiography (ECG) and photoplethsmography (PPG) and the use of the corresponding heart rate metrics to classify cognitive workload in a range of scenarios, including driving whilst performing an N-back memory task [9], taking maths tests of varying difficulty [10] and when engaging in a partially automated task with a machine based component [11]. Whilst ECG and PPG are both less obtrusive in daily life than scalp EEG, and offer a wearable solution to cognitive workload tracking, it remains unclear whether the documented increases in heart rate are associated with the stress of performing well during higher cognitive workload tasks [12] [13], or indeed the increased cognitive workload itself. Namely, heart rate is known to correlate strongly with stress level whilst driving, as well as skin conductivity (sweat level) [14]. For the purpose of rigorous cognitive workload tracking, it is therefore important to consider tasks whereby the aspect of stress that a maths test or driving may cause is reduced, whilst still maintaining the ability to vary cognitive workload.

B. The brain, oxygen and cognitive workload

The brain is the most metabolically active organ in the human body. At rest, the brain consumes 20% of the body's oxygen [15] and this percentage increases with increased cognitive demand. Oxygen restriction has significant effects on cognitive function; for example, less oxygen delivery to the brain has been observed in those with memory impairments [16] [17]. Moreover, the administration of oxygen, through the breathing of supplemental oxygen and the associated increase in blood oxygen, has been shown to result in a significantly better memory performance and faster reaction times [18] [19] [20] [21].

Functional near infrared spectroscopy (fNIRS), a tool for measuring oxygenation of tissue and thus oxygen consumption, has shown increases in oxygen consumption of the brain with an increase in cognitive workload in drivers [22]. Furthermore, fNIRS has helped to detect increased oxygenation

of specific brain regions (such as the left inferior frontal gyrus, involved in language processing) with an increase in the difficulty of a letter based N-back memory task in pilots [1]. This motivates us to investigate whether these changes in oxygen consumption are also observable in spectral analysis of blood, or if they manifest themselves through changes in breathing rate or breathing magnitude.

C. In-ear SpO_2

The notion of blood oxygen saturation refers to the proportion of haemoglobin binding sites (four for each molecule) that are occupied with oxygen, out of the total number of haemoglobin binding sites available in the blood, and is formulated as

$$\text{Oxygen Saturation} = \frac{HbO_2}{HbO_2 + Hb} \tag{1}$$

where the symbol Hb refers to haemoglobin not bound with oxygen and HbO_2 to haemoglobin bound to oxygen.

Photoplethysmography (PPG) refers to the non-invasive measurement of light absorption through the blood. With each heart beat, there is a pulsatile increase in blood volume and, when more blood is present, more light is absorbed resulting in less light reflected back to the sensor. Through this mechanism, PPG effectively measures the pulse. Arterial blood oxygen saturation is typically estimated using pulse oximetry to yield a percentage value (SpO_2), with subjects that have a healthy respiratory system typically exhibiting SpO_2 values of 96-98% at sea level [23]. Pulse oximetry uses PPG simultaneously at different wavelengths of light (red and infrared) to estimate blood oxygen levels. The level of blood oxygen saturation is mirrored in a change in the ratio of light absorbance between the red and infrared light. The extinction coefficient of oxygenated haemoglobin with red light ($\approx 660nm$) is lower than that of deoxygenated haemoglobin, and the reverse holds true for infrared light ($\approx 880-940nm$) [24]. The simultaneous measurement of the absorbance/reflectance of both infrared and red light therefore allows for an estimation of blood oxygen saturation.

Pulse oximetry can be measured at most skin sites, but for convenience it is usually measured from the finger. On the other hand, enthusiasm for wearable eHealth technology and Hearables [4] has promoted the ear canal as a preferred site for the measurement of vital signs, given its proximity to the brain (ear-EEG [5]) and the comparatively stable position of the head with respect to vital signs in daily life. This is in stark contrast with currently used sites such as the wrist and finger.

In-ear pulse oximetry offers significant benefits over conventional finger pulse oximetry. Firstly, the in-ear location is robust to changes in blood volume which occur owing to vasoconstriction and hypothermia, giving a stable and accurate photoplethysmogram during cold exposure [25]. Next, it has been documented that in-ear PPG exhibits both larger respiratory induced intensity variations and larger amplitude variations with respiration [26], allowing for a improved measurement of respiration rate. Also, a significant delay has been evidenced between ear pulse oximetry and pulse oximetry on the right index finger [27] and on the hand/foot [28] for detection of hypoxemia (low levels of blood oxygen). Indeed, SpO_2 from the right index finger was shown to take an average of 12.4 seconds longer to respond when compared with SpO_2 from the right ear canal during breath holds across different subjects [27]. This is because the ear is supplied by the common carotid artery, in close proximity to the heart, making ear-SpO_2 an effective non-invasive proxy for priority core blood oxygen.

Given that the brain is the most metabolically active organ in the body and that it increases oxygen consumption with cognitive workload, we here hypothesise that increased cognitive workload may be measurable through blood oxygen saturation. Considering that wearable in-ear pulse oximetry provides a robust SpO_2 signal with minimal delay, we set out to answer whether in-ear pulse oximetry can be used to accurately classify different levels of cognitive workload, and furthermore can this classification be performed in an almost real-time fashion?

II. Methods

A. Hardware

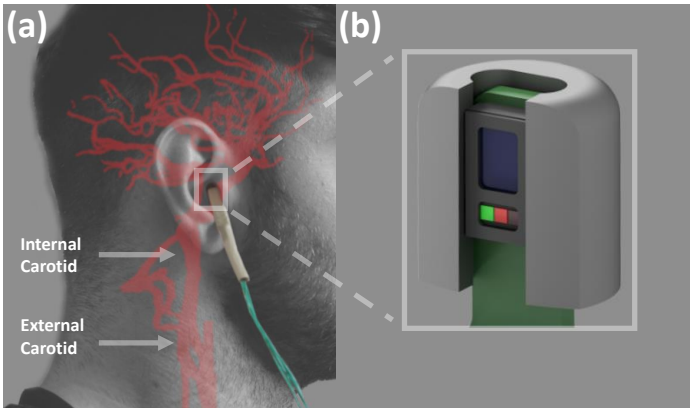

Fig. 1. The in-ear photoplethysmography sensor used in our study. (a) The sensor placement within the ear canal, with the major arteries supplying the brain and the ear highlighted. (b) Zoom-in of the pulse oximetry sensor, with a form factor of a viscoelastic memory foam earbud.

The MAX30101 digital PPG chip by Maxim Integrated (San Jose, CA, USA) was used in our in-ear sensor, consisting of red (660nm) and infrared (880nm) light emitting diodes as well as a photo-diode to measure the reflected light. The PPG sensor was embedded in a cut-out rectangular section of a viscoelastic foam earbud [29], allowing for comfortable insertion. The two light wavelengths (at 660nm and 880nm) give two measures of pulse, and also the signals required to estimate blood oxygen saturation and respiratory waveforms. The connected PPG circuitry, including decoupling capacitors and level shifting circuitry that enable digital communication between the 1.8V and 3V domains, was neatly covered with heat shrink and positioned just outside of the ear, as shown in Fig. 1. The sensor apparatus was connected to a circuit board which stores the data for each trial on an SD card.

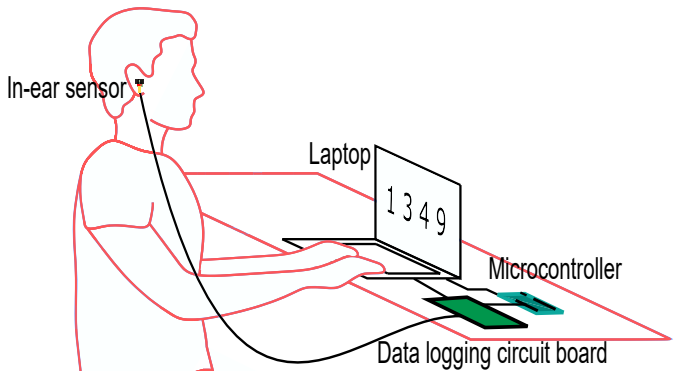

Fig. 2. Illustration of the recording of in-ear SpO_2 during an N-back task. The SpO_2 sensor links to a circuit board which logs the data stream and also accepts input from the microcontroller. The four single digit numbers displayed on the laptop refresh every 5 seconds, and communicate this refresh time to the microcontroller, which in turn sends an electrical pulse to the circuit board to align the task with the physiological data.

MATLAB 2018a by MathWorks (Natick, MA, USA) was used to create a graphical user interface which refreshed four single digit numbers every 5 seconds on a screen in front of the subject. The MATLAB program also communicated with an Arduino Uno by Arduino (Somerville, MA, USA) with each refresh, which in turn communicated with the data logging circuit board with an electrical pulse to align the PPG data to each 5 second window.

B. Experimental Protocol

The participants in the recordings were 10 healthy subjects (5 male, 5 female) aged 22 - 29 years. A single PPG sensor was used per subject and was inserted within the right ear canal. The subjects were seated in front of a monitor during the recording where a MATLAB graphical user interface updated with 4 randomly generated single digit numbers every 5 seconds, as shown in Fig. 2. Subjects were asked to count the number of odd numbers and, depending on the N-back trial, they were tasked with entering the current number of odd numbers using the keyboard (0-back), the previous number (1-back), the number 2 steps back (2-back), or the number of odd numbers 3 steps back (3-back). Each trial lasted for 5 minutes and 40 seconds (68 5 second epochs), with 6 epochs used for calibration, leaving 62 epochs for analysis. Four trials were performed by each subject, corresponding to the four levels of N-back task that were presented in a quasi-randomised order. Each subject was given between 5 and 10 minutes rest between trials, and allowed to practice until they were confident with the tasks before the recordings started.

The recordings were performed under the Imperial College London ethics committee approval JRCO 20IC6414, and all subjects gave full informed consent.

C. Signal Processing

The ratio of absorbance of infra-red to red light within the PPG sensor changes depending on the proportion of haemoglobin that is oxygenated in the blood. This change can

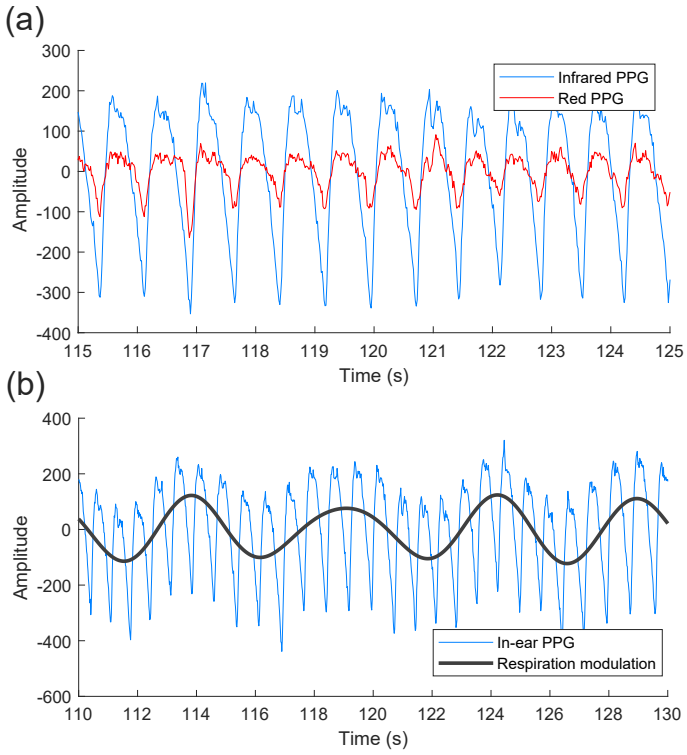

Fig. 3. Overview of the signals recorded from in-ear photoplethysmography. (a) Both the red and infrared AC photoplethysmography signals, bandpass filtered between 1Hz and 30Hz for the calculation of heart rate and SpO_2. (b) The infrared in-ear PPG signal bandpass filtered between 0.2 and 30Hz with the respiration modulation superimposed in black.

be quantified through the so called *ratio of ratios* metric [30], given by

$$R = \frac{\frac{AC_{red}}{DC_{red}}}{\frac{AC_{infrared}}{DC_{infrared}}} \tag{2}$$

An empirically derived linear approximation can then be used to calculate an SpO_2 value as a proxy to oxygen saturation. Using the manufacturers suggested calibration [31], the SpO_2 value was calculated as

$$SpO_2 = 104 - 17R \tag{3}$$

To obtain the alternating current (AC) components within the PPG measurements, the raw signals were firstly bandpass filtered between 1Hz and 30Hz. Peak detection was performed using the MATLAB function *findpeaks*, with a minimum peak prominence that varied between 80 and 150 arbitrary units for the infrared signal, while for the red signal it was set at 30 arbitrary units. In general, the values chosen for minimum peak prominence translated to approximately half of the maximum peak value. This was to account for decreases in pulse amplitude that occur due to the effects of respiration. The same procedure was repeated for the inverted signals to find the troughs. Next, the peak values and trough values were separated and interpolated, before their absolute values were added together to give a continuous estimate of

TABLE I
SUMMARY OF FEATURES USED FOR CLASSIFICATION

Category	Features
SpO_2	SpO_2 mean, relative change in SpO_2, red amplitude mean, infrared (IR) amplitude mean, relative change in red amplitude, relative change in IR amplitude, red AC/DC ratio mean, IR AC/DC ratio mean, red peak prominence mean, IR peak prominence mean, red/IR AC ratio mean, red amplitude variance, IR amplitude variance.
Pulse	Heart rate mean, relative change in heart rate, pulse full-width-half-maximum (FWHM) mean, pulse width ratio† mean, pulse width ratio variance.
Breathing	Breathing rate mean, relative change in breathing rate breathing amplitude mean.

† Pulse width ratio is the ratio between the FWHM of the peak and the FWHM of the trough, giving a systolic to diastolic duration ratio.

the AC amplitude. The direct current (DC) components were obtained by low-pass filtering the raw signals at 0.01Hz.

The peak detection procedure of the AC infrared troughs was also used to calculate pulse rate, given that the PPG peak from the ear canal is broader than the peak from the finger (a characteristic of the pressure wave found in the carotid artery) and would thus give a noisy pulse rate estimate. An example of the photoplethysmography pulse signal from the ear is shown in Fig. 3(a).

Fluctuations in the baseline of ear-PPG due to inspiration and expiration have been evidenced as 8-fold stronger from the ear-canal than from the finger [26]. For the calculation of respiration rate, the raw PPG signal was first band-pass filtered between 0.2Hz and 30Hz, followed by a moving average filter with a 3-second window. Peak detection was performed using the MATLAB function *findpeaks* with a minimum peak prominence of 10, to give respiration peaks. The difference of the timings of these peaks was then used to give a breathing interval, shown in Fig. 3(b). The inverse of the interval signal was then multiplied by 60 to give breathing rate (in breaths per minute). The amplitude values of the respiration peaks were also used as an estimate of breathing amplitude. No epochs of data were discarded, even in the presence of motion artefacts.

D. Feature extraction

For each 5-second epoch, 21 time domain features (13 SpO_2 based features, 5 pulse based features and 3 breathing based features) were extracted. Frequency-based features were not used as the 5-second window is too short for reliable heart rate variability metrics from PPG. Five features were calculated using both the 5-second epoch and the calibration data from the start of the task. This was particularly important in the case of SpO_2, as although healthy SpO_2 levels generally fall within a small range of 94-100%, the changes we detected due to cognitive load were less than 1%. Whilst absolute values are adequate for testing and training on the same subject, features that are relative to a calibration period are more useful for generalising across subjects. The 21 features used are summarised in Table I.

1) SpO_2 features. The SpO_2 mean was calculated based on the ratio of ratios defined in equations (2) and (3). Infrared amplitude mean and variance were defined as the mean and variance of the infrared light peak amplitudes when the infrared signal has been band-pass filtered between 1Hz and 30Hz, and the red amplitude mean and variance were defined as the mean and variance of the red light peak amplitudes when the red signal has been filtered in the same way. Alternating current to direct current (AC/DC) ratios were defined as the mean peak amplitudes after band-pass filtering between 1Hz and 30Hz, divided by the mean of the signal low-pass filtered at 0.01Hz. Peak prominence was defined as the peak value minus the minimum of the signal, either between two peaks that had larger peak values than itself, or across the whole signal if it was the highest peak. All relative features were calculated as the feature minus the same feature calculated from the 6 calibration epochs at the start of the task.

2) Pulse features. Pulse based features were calculated from the infrared light signal band-pass filtered between 1Hz and 30Hz, shown in Fig. 3(a). Heart rate is defined as 60 divided by the peak to peak time interval in seconds. The mean heart rate across the 5-second window and mean relative heart rate compared with the initial calibration period were used as features. The pulse width was implemented using the full-width-half-maximum (FWHM), defined has the width of the peak at half of the peak height relative to the rest of the signal [32]. Pulse width ratio is the ratio between the FWHM of the peak and the FWHM of the trough, giving a systolic (heart beating) to diastolic (heart resting) duration ratio given by

$$\frac{Width_{systolic}}{Width_{diastolic}} \simeq \frac{FWHM_{peak}}{FWHM_{trough}} \tag{4}$$

where FWHM is the full-width-half-maximum which was used as pulse width [32]. Both the mean and variance of the pulse width ratio were used as features.

3) Breathing features. Breathing related features were calculated from the infrared light signal band pass filtered between 0.2Hz and 30Hz and moving average filtered over a window of 3-seconds, resulting in a similar respiratory modulation signal to the example shown in Fig. 3(b). The breathing rate was calculated as 60 divided by the interval in seconds between breathing modulation peaks. The mean of this across the 5-second interval was used as mean breathing rate, and relative breathing rate was calculated as the mean breathing rate of the current segment minus the breathing rate of the calibration period. Breathing amplitude mean was calculated as the mean peak amplitude of the breathing modulation signal.

E. Classification and evaluation

A random forest classifier with AdaBoost was employed from the publicly available scikit-learn Python toolbox [33]. For the random forest base, the number of trees was set to 50, the class weight was set to 'balanced subsample'. For the AdaBoost framework, the random forest was set as the base classifier, the maximum number of estimators was set to 50, the learning rate was set to 1.0 and the real boosting algorithm

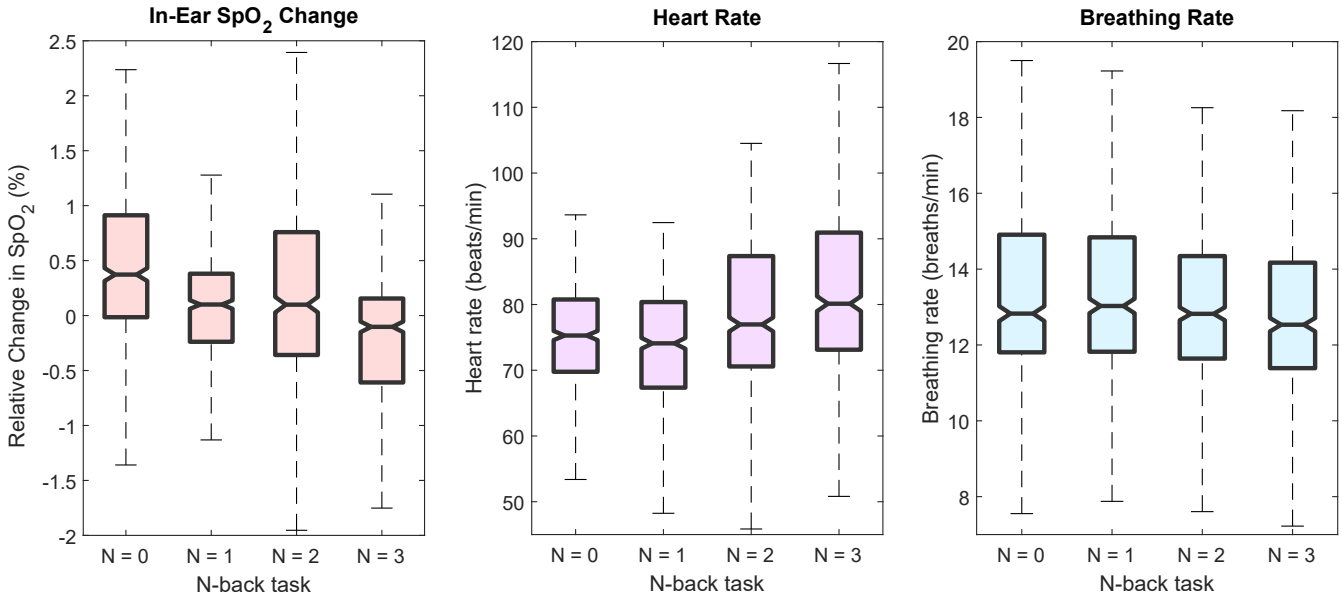

Fig. 4. Box plots of the relative change in SpO_2 (left, red), heart rate (middle, purple) and breathing rate (right, blue) from the in-ear sensor, split into N-back categories and including each 5 second epoch. The top and bottom of each box represent respectively the upper and lower quartiles, the center notches of each box designate the median, and the whisker lines extending out of the box the range.

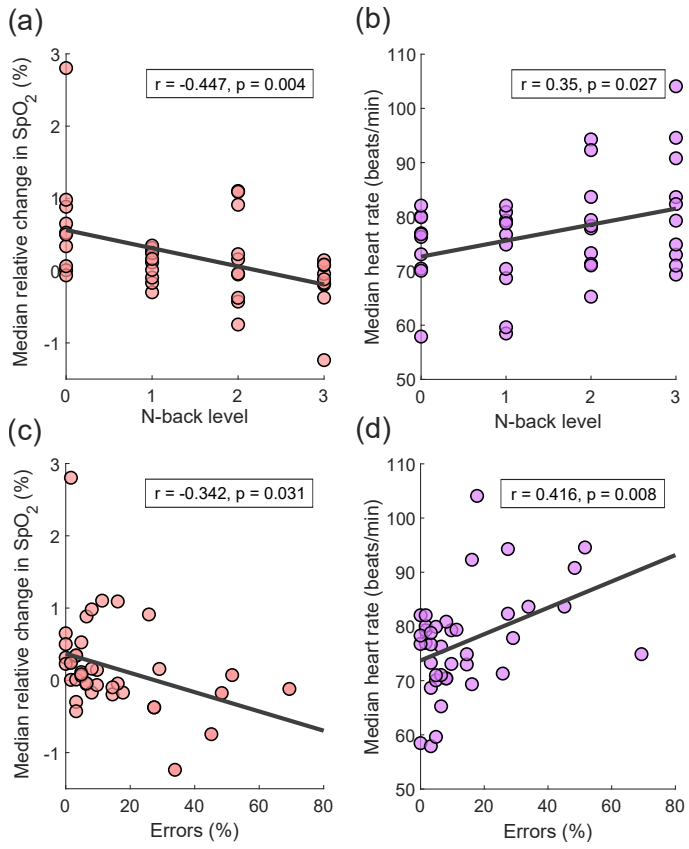

Fig. 5. Scatter plots of median relative change in SpO_2 (red) and median heart rate (purple) against N-back level and proportion of errors, with trend lines (black) and correlation coefficients and p values superimposed. (a) Median relative change in SpO_2 against N-back level. (b) Median heart rate against N-back level. (c) Median relative change in SpO_2 against proportion of errors. (d) Median heart rate against proportion of errors.

"SAMME-R" [34] was used. Whilst it is unconventional to use random forest as a weak learner for AdaBoost, it was found when training on a separate data set that the random forest alone still exhibited notable training bias, highlighting the need for boosting. A possible reason for this bias is that with the small pool of 10 subjects, the physiological data was relatively sparse.

Ten-fold cross-validation was employed on the fully shuffled data for the case of four-category classification (0-back, 1-back, 2-back, 3-back). Leave-one-subject-out cross-validation was employed on two-category classification (0-back and 3-back). All 21 features were used in ten-fold cross-validation, but only the mean SpO_2, the relative change in SpO_2 and the mean heart rate were used in leave-one-subject-out cross-validation. In the case of 10-fold cross-validation, the maximum number of features was set to 10 while for leave one subject out cross-validation the maximum number of features was set to 3, as only 3 features were used. Class-specific accuracy and overall accuracy were used as metrics to evaluate classification performance.

Given the three categories of features used (pulse, SpO_2 and breathing), feature importance by means of a reduction in tree impurity was calculated for each feature. This was used to ascertain the relative contribution of SpO_2 derived features compared with conventionally used features such as heart rate and breathing rate.

III. Results

The mean mistake percentages across subjects for each N-back stage were $4.0\% \pm 3.3\%$, $4.8\% \pm 4.3\%$, $17.1\% \pm 14.4\%$ and $29.4\% \pm 21.1\%$ for 0-back, 1-back, 2-back and 3-back tasks, respectively. The substantial increase in mistakes between 1-back, 2-back and 3-back tasks indicates that the

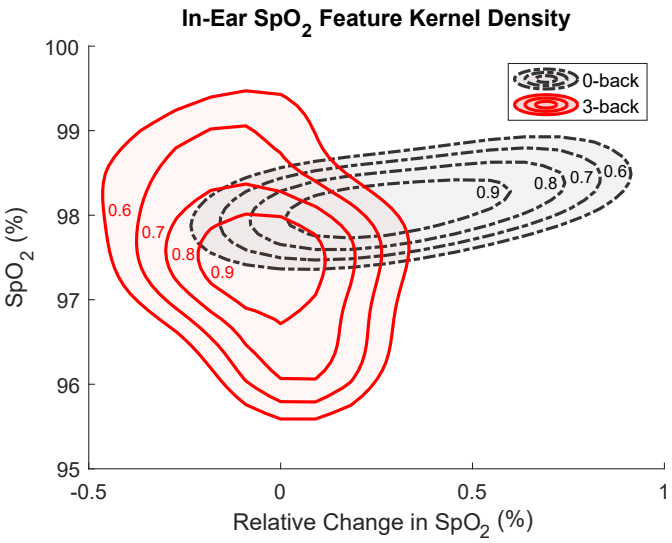

Fig. 6. Two-dimensional contour plots of the kernel density estimates for mean SpO_2, and relative change in SpO_2, derived from all of the 5 second epochs of data. The kernel density estimates are plotted independently for two categories: 0-back (black dotted line) and 3-back (red solid line). Kernel density was normalised between 0 and 1 for each category, and the corresponding values for each contour line shown are marked within the contour lines themselves.

(a) 10-fold Cross Validation

True N-back \ Predicted N-back	N = 0	N = 1	N = 2	N = 3
N = 0	**93.4**	3.4	1.6	1.6
N = 1	2.4	**89.2**	2.7	5.6
N = 2	3.2	4.2	**89.5**	3.1
N = 3	1.3	5.5	3.1	**90.2**

(b) Leave One Subject Out: Binary

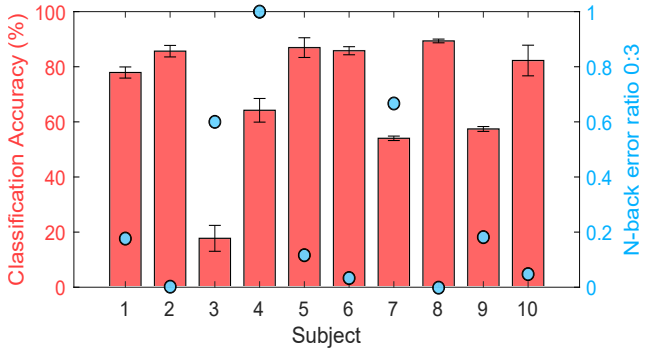

Fig. 7. Classification accuracy of SpO_2 based cognitive workload estimation. (a) Mean confusion matrix for the results of ten-fold shuffled cross-validation in four-category prediction (0-back, 1-back, 2-back and 3-back). The rows correspond to the true N-back category, and the columns to the category predicted by the classifier. (b) Classification accuracy for testing on each subject, with the classifier trained exclusively on the other 9 subjects (red bars) and the N-back error ratio between the 0-back and 3-back tasks for each subject (blue circles). An error ratio of 1 means the same number of errors were made on the 3-back task as on the 0-back.

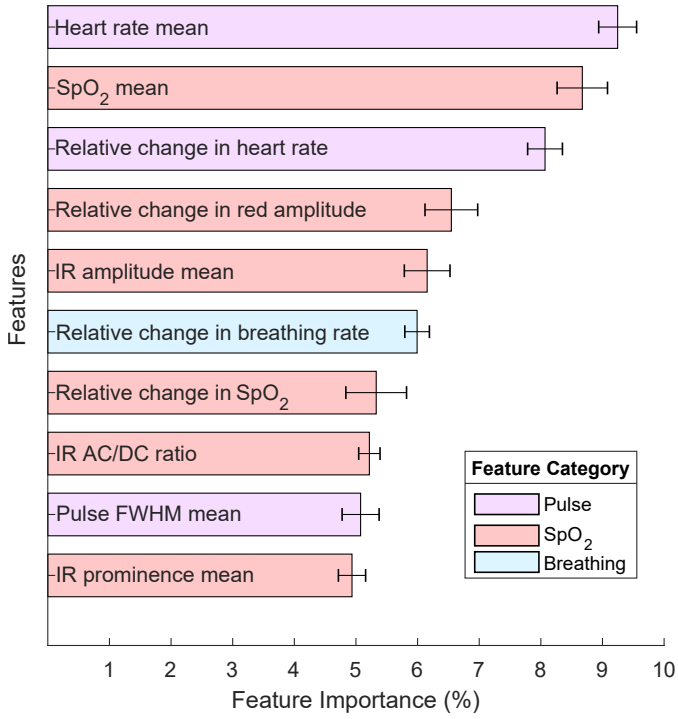

Fig. 8. Feature importance for the 10 most significant features out of the 21 features considered during 10-fold cross validation. Feature importance was derived from the reduction in tree impurity based on the contribution of each feature to the random forest classifier. The features were split into three categories, depending on the physiological metric from which they were derived. The SpO_2 features are shown in red, pulse features are in purple, and breathing features are in blue.

3-back and 2-back tasks were difficult enough to create a meaningful increase in cognitive workload at a group level, although not necessarily being the case for some specific subjects.

A. Change in blood oxygen, heart rate, breathing rate

The mean recorded SpO_2 across all subjects and trials was $97.0\% \pm 1.7\%$, the mean heart rate was 76.7 ± 11.4 beats per minute, and the mean estimated breathing rate was 13.5 ± 3.3 breaths per minute. All results therefore fell into the physiologically expected range. Recorded values of SpO_2, heart rate and breathing rate, across all participants and all 5 second windows are presented in Fig.4 for each N-back task. Each physiological metric had statistically significant changes across N-back tasks, with one-way ANOVA yielding p-values of $P < .001$.

We observed a decrease in median SpO_2, relative to the start of the task, with every increase in N-back difficulty. The median SpO_2 relative to the start of the N-back task was +0.373%, +0.101%, +0.099% and -0.102%, for 0-back, 1-back, 2-back and 3-back respectively, as shown in Fig. 4. Moreover, for the most difficult 3-back task either the median SpO_2 or median SpO_2 relative to the start of the task was the lowest out of all tasks in 8 out of the 10 subjects, while in the remaining two subjects this occurred for the 2-back task. Similarly, the overall median heart rate was highest in the 2-back and 3-back tasks, but on an individual subject basis the

highest median heart rate only occurred in the 3-back task in 5 out of 10 subjects. A slight decrease in median breathing rate was also observed with the 3-back task.

The Pearson's correlation coefficients between the task median relative change in SpO_2 and the task median heart rate, and both N-back difficulty and proportion of errors, were also examined. Fig. 5 highlights the prominence of a medium negative correlation between the median relative change in SpO_2 and N-back difficulty (r = -0.45, p = 0.004) and between the median relative change in SpO_2 and the proportion of errors (r = -0.34, p = 0.031). Similarly, there was a medium positive correlation between the median heart rate and N-back difficulty (r = 0.35, p = 0.027) and between the median heart rate and proportion of errors (r = 0.416, p=0.008). The strongest correlation was therefore seen between a decrease in relative SpO_2 and increase in workload. Importantly, relative SpO_2 was more correlated with workload than with errors and the opposite was true of heart rate. A negative correlation was seen between median breathing rate and workload, but this correlation was not significant (r = -0.28, p = 0.086).

Fig. 6 further demonstrates the high separability of 0-back and 3-back tasks with SpO_2 features, through a two-dimensional kernel density plot of mean SpO_2 and relative change in SpO_2.

B. Classification

1) Shuffled ten-fold cross-validation. With ten-fold cross-validation, shuffled across all participants, we were able to classify the 5-second 0-back epochs with an average accuracy of 93.4%, 1-back epochs with an accuracy of 89.2%, 2-back with 89.5% and 3-back with 90.2%, giving a total average classification accuracy of 90.6%. The largest errors occurred with the miss-classification of 1-back as 3-back and vice versa, with larger errors also occurring between 0 and 1, 1 and 2, and 2 and 3. Classification accuracy was notably better for 0-back and 3-back tasks, as evidenced by the full confusion matrix averaged across 10-fold cross validation in Fig. 7(a).

Averaged feature importance (according to reduction in tree impurity in the random forest) across each fold for ten-fold cross-validation was calculated with the top 10 features presented in Fig. 8. The two most important features for classification in the case of shuffled ten-fold cross-validation were the mean heart rate and the mean SpO_2. Moreover, 6 of 10 most important features were derived from SpO_2.

2) Leave-one-subject-out cross-validation. Binary leave-one-subject-out cross validation with 3 features had varied performance, but performed well across the majority of subjects, with 6 subjects having an mean accuracy greater than 77.9%, and 4 of those subjects having an accuracy greater than 85%. The 6 subjects with highest classification accuracy made an average of 15 times more errors in the 3-back task, reflected in a low 0-back to 3-back error ratio, and the 4 subjects with the lowest classification accuracy made an average of 2 times more errors in the 3-back task, reflected in a high 0-back to 3-back error ratio. The accuracy percentages for testing on each subject, along with the 0-back to 3-back error ratios are shown in Fig. 7(b). Notably mean SpO_2 and the relative change in SpO_2 were the most valuable features in terms of reducing tree impurity in binary leave-one-subject-out cross validation.

IV. Discussion

In general, an increase in cognitive workload led to a decrease in the measured in-ear SpO_2 levels. The decrease in the measured in-ear SpO_2 with increased cognitive workload was consistent across all 10 subjects, with the lowest median relative change in SpO_2 or the lowest median absolute SpO_2 occurring in the 3-back or 2-back tasks in all subjects. This demonstrates the robustness of the in-ear SpO_2 response to changes in cognitive workload, compared with the commonly used metric of heart rate, where an increase in heart rate did not necessarily correspond to increased cognitive workload. As expected, errors were highly correlated with increased cognitive workload due to the increased task difficulty. Importantly, the relative change in SpO_2 was more correlated with the level of N-back task than it was with the proportion of errors made, and the opposite was true for the heart rate. A possible reason for this is that some subjects became stressed when making errors, thus triggering an increase in heart rate [12]. Whilst it is important to note that we did not provide live feedback to participants when they made errors, during an N-back memory task it is feasible that subjects were aware of when they have forgotten a number. In this particular experiment, the memory aspect of the task contributed more to errors than the counting of the odd numbers, as evidenced by an increase in the mean error rate from 4% to 29% between the 0-back and 3-back tasks. Given the small sample size, conclusions cannot be drawn on whether in-ear SpO_2 tracks cognitive workload independent of stress. Future work will address this issue further with the addition of a control task that purposefully induces stress and questionnaires to assess subjective stress of the participants. Importantly, the correlation between the relative change in SpO_2 and an increase in cognitive workload was the strongest and most significant correlation found.

The robustness of the measured SpO_2 changes were further reflected in the high classification accuracy. With ten-fold cross-validation, the 5-second epochs of in-ear data achieved an average classification accuracy of 90.6% across the four N-back task categories, with the two most important features for classification being the mean heart rate and the mean SpO_2. The performance of leave-one-subject-out cross-validation using the two categories of 0-back and 3-back was less consistent, but was reasonably good in the majority of subjects, with an accuracy of 77.9% and above in 6 subjects, 4 of which achieved a classification accuracy of over 85%. Notably, the leave-one-subject-out evaluation was implemented with just 3 features, the mean SpO_2, the mean heart rate and the relative change in SpO_2. In this case, the most important features were the SpO_2-derived features. Our average classification accuracy across unseen subjects with a 5-second window (70.1%) is comparable to that of gaze and pupil derived features (70.4%) [35] and the reported accuracy achieved by fNIRs when training and testing on the same subject (63.5% and 78% with 15-second to 25-second windows respectively) [36].

In general, cognitive workload tasks induce different levels of cognitive workload in different people, which is evidenced

through the large standard deviation in mistakes, suggesting highly subjective levels of difficulty. The physiological response to increased cognitive workload also varies widely between different people. The ability of leave-one-subject-out training to perform well when testing on a majority of the 10 subjects conclusively demonstrates a robustness in the SpO_2 response to changes in cognitive workload that becomes visible even across a few subjects. Importantly, in the subjects where the classifier performed poorly, there were comparable errors between the 3-back and 0-back tasks and the absolute errors for the 3-back task were low. This could indicate that both tasks were found to be comparatively easy and thus classification performance was reduced because the experiment failed to induce large changes in cognitive workload, but we cannot know this for certain in the absence of questionnaires that assess subjective cognitive load [37]. An argument can also be made that subjects may have been less stressed when making fewer mistakes and that a lack of stress may also have contributed to worse classification accuracy. Whilst it is clear from this proof-of-concept study that in-ear SpO_2 generally allows for high accuracy classification of cognitive load, understanding why specific breakdowns in classification accuracy happen and if they are related to experimental flaws or the metric itself should be investigated more comprehensively in the future. The overlap in task performance across subjects was further exaggerated when comparing 0-back to 1-back, or 2-back to 3-back, making full four category leave-one-subject-out classification unfeasible.

In our data, the in-ear SpO_2 decrease in response to cognitive load is visible within the first two 5-second segments of increased cognitive load and this decrease tends to accumulate gradually across the trial. This is comparable to galvanic skin response which has a response time to emotion evoking stimuli in the range of 1 to 5 seconds [38] and a tonic response in the range of 10 to 100 seconds [39]. Notably, in-ear SpO_2 is slower than more instantaneous measures such as EEG which has a response time on the order of hundreds of milliseconds [40] and therefore it is recommended that in-ear SpO_2 be used to measure sustained periods of cognitive load in the period of tens of seconds and longer for maximal effectiveness, rather than to explore the cognitive load induced instantaneously by a single stimulus.

Whilst it has been demonstrated that in-ear SpO_2 is an effective measure for distinguishing aggregate levels of cognitive load, a limitation when compared with more sophisticated measures such as fMRI and fNIRs is that in-ear SpO_2 cannot distinguish the type of cognitive load, such as whether cognitive load is induced by increased memory demands (as is the case with an N-back task) or induced by audio/visual feedback or motor control. In uncontrolled environments where external stimuli are a factor, it would therefore be difficult to relate in-ear SpO_2 measured cognitive load changes solely to a single task.

It is also important to note that the external carotid artery supplies the ear canal with oxygen, whereas the internal carotid artery supplies the brain with oxygen. More experimentation is needed to ascertain the extent to which the observed robust decrease in in-ear SpO_2 is caused by the increased oxygen consumption of the brain, as opposed to other physiological factors. The impact of sympathetic tone was investigated but a change in heart rate variability metrics was not found to be predictive of an increasing cognitive load in this study. Further investigation is also needed to determine whether this SpO_2 response is specific to the ear canal.

V. Conclusion

A proof of concept for cognitive workload estimation using a novel wearable in-ear pulse oximetry sensor has been introduced. Pulse oximetry from the ear canal has been shown to be capable of discriminating between 4 categories of cognitive workload based on an N-back task over 5-second epochs, with a mean accuracy of 90.6%. High cognitive workload in the 2-back and 3-back tasks has led to either the lowest median absolute SpO_2 or largest median decrease in SpO_2 in all of the subjects, therefore demonstrating a robust decrease in measured blood oxygen in response to increased cognitive workload. We conjecture that the decrease in measured SpO_2 with increased cognitive load could be related to the increased oxygen consumption of the brain under increased cognitive demands, and to this end we have examined the predictability of the change in in-ear SpO_2 in response to changes in cognitive workload. The consistency of the SpO_2 response has been further evidenced by an ability to generalise across subjects, even in a relatively small subject pool. In combination with the previously documented rapid reaction speed of in-ear SpO_2 measurements [27], this indicates the promise of in-ear SpO_2 as a tool for close to real-time cognitive workload classification. Overall, this pilot study has established in-ear SpO_2 as an effective tool for the classification of cognitive workload, which can be used alone or in combination with current gold standard workload tracking equipment such as EEG and ECG, or within the emerging multi-modal Hearables.

Acknowledgment

This work was supported by the Racing Foundation grant 285/2018, MURI/EPSRC grant EP/P008461, and the Dementia Research Institute at Imperial College London.

References

[1] H. Ayaz, B. Willems, B. Bunce, P. Shewokis, K. Izzetoglu, S. Hah, A. Deshmukh, and B. Onaral, "Cognitive workload assessment of air traffic controllers using optical brain imaging sensors," *Advances in Understanding Human Performance: Neuroergonomics, Human Factors Design, and Special Populations*, pp. 21–31, 2010.

[2] M. Hajinoroozi, Z. Mao, T. P. Jung, C. T. Lin, and Y. Huang, "EEG-based prediction of driver's cognitive performance by deep convolutional neural network," *Signal Processing: Image Communication*, vol. 47, pp. 549–555, Sep 2016.

[3] J. K. Johannesen, J. Bi, R. Jiang, J. G. Kenney, and C.-M. A. Chen, "Machine learning identification of EEG features predicting working memory performance in schizophrenia and healthy adults," *Neuropsychiatric Electrophysiology*, vol. 2, no. 1, p. 3, Dec 2016.

[4] V. Goverdovsky, W. Von Rosenberg, T. Nakamura, D. Looney, D. J. Sharp, C. Papavassiliou, M. J. Morrell, and D. P. Mandic, "Hearables: Multimodal physiological in-ear sensing," *Scientific Reports*, vol. 7, no. 1, pp. 1–10, Dec 2017.

[5] D. Looney, P. Kidmose, C. Park, M. Ungstrup, M. Rank, K. Rosenkranz, and D. Mandic, "The in-the-ear recording concept: User-centered and wearable brain monitoring," *IEEE Pulse*, vol. 3, no. 6, pp. 32–42, 2012.

[6] T. Nakamura, Y. D. Alqurashi, M. J. Morrell, and D. P. Mandic, "Hearables: Automatic Overnight Sleep Monitoring With Standardized In-Ear EEG Sensor," *IEEE Transactions on Biomedical Engineering*, vol. 67, no. 1, pp. 203–212, 2020.

[7] R. Wang, P. V. Amadori, and Y. Demiris, "Real-time workload classification during driving using HyperNetworks," in *IEEE International Conference on Intelligent Robots and Systems*. Institute of Electrical and Electronics Engineers Inc., Dec 2018, pp. 3060–3065.

[8] P. Vito Amadori, T. Fischer, R. Wang, and Y. Demiris, "Decision anticipation for driving assistance systems," in *IEEE International Conference on Intelligent Transportation Systems (ITSC 2020)*, 2020.

[9] A. Tjolleng, K. Jung, W. Hong, W. Lee, B. Lee, H. You, J. Son, and S. Park, "Classification of a driver's cognitive workload levels using artificial neural network on ECG signals," *Applied Ergonomics*, vol. 59, pp. 326–332, Mar 2017.

[10] C. Wang and J. Guo, "A data-driven framework for learners' cognitive load detection using ECG-PPG physiological feature fusion and XGBoost classification," in *Procedia Computer Science*, vol. 147. Elsevier B.V., Jan 2019, pp. 338–348.

[11] J. Zhang, S. Li, and R. Wang, "Pattern recognition of momentary mental workload based on multi-channel electrophysiological data and ensemble convolutional neural networks," *Frontiers in Neuroscience*, vol. 11, no. MAY, p. 310, May 2017.

[12] W. von Rosenberg, T. Chanwimalueang, T. Adjei, U. Jaffer, V. Goverdovsky, and D. P. Mandic, "Resolving Ambiguities in the LF/HF Ratio: LF-HF Scatter Plots for the Categorization of Mental and Physical Stress from HRV," *Frontiers in Physiology*, vol. 8, p. 360, Jun 2017.

[13] T. Chanwimalueang, L. Aufegger, T. Adjei, D. Wasley, C. Cruder, D. P. Mandic, and A. Williamon, "Stage call: Cardiovascular reactivity to audition stress in musicians," *PLOS ONE*, vol. 12, no. 4, p. e0176023, Apr 2017.

[14] J. A. Healey and R. W. Picard, "Detecting stress during real-world driving tasks using physiological sensors," *IEEE Transactions on Intelligent Transportation Systems*, vol. 6, no. 2, pp. 156–166, Jun 2005.

[15] R. F. Butterworth, *Hypoxic Encephalopathy*, 6th ed., A. R. Siegel GJ, Agranoff BW, Ed. Lippincott-Raven, 1999.

[16] L. Owen and S. I. Sunram-Lea, "Metabolic agents that enhance ATP can improve cognitive functioning: A review of the evidence for glucose, oxygen, pyruvate, creatine, and l-carnitine," *Nutrients*, vol. 3, no. 8, pp. 735–755, Aug 2011.

[17] F. Eustache, P. Rioux, B. Desgranges, G. Marchal, M. C. Petit-Taboué, M. Dary, B. Lechevalier, and J. C. Baron, "Healthy aging, memory subsystems and regional cerebral oxygen consumption," *Neuropsychologia*, vol. 33, no. 7, pp. 867–887, Jul 1995.

[18] A. B. Scholey, M. C. Moss, N. Neave, and K. Wesnes, "Cognitive performance, hyperoxia, and heart rate following oxygen administration in healthy young adults," *Physiology and Behavior*, vol. 67, no. 5, pp. 783–789, Nov 1999.

[19] M. C. Moss, A. B. Scholey, and K. Wesnes, "Oxygen administration selectively enhances cognitive performance in healthy young adults: A placebo-controlled double-blind crossover study," *Psychopharmacology*, vol. 138, no. 1, pp. 27–33, 1998.

[20] S. P. Kim, M. H. Choi, J. H. Kim, H. W. Yeon, H. J. Yoon, H. S. Kim, J. Y. Park, J. H. Yi, G. R. Tack, and S. C. Chung, "Changes of 2-back task performance and physiological signals in ADHD children due to transient increase in oxygen level," *Neuroscience Letters*, vol. 511, no. 2, pp. 70–73, Mar 2012.

[21] S. C. Chung, J. H. Sohn, B. Lee, G. R. Tack, J. H. Yi, J. H. You, J. H. Jun, and R. Sparacio, "The effect of transient increase in oxygen level on brain activation and verbal performance," *International Journal of Psychophysiology*, vol. 62, no. 1, pp. 103–108, Oct 2006.

[22] H. Tsunashima and K. Yanagisawa, "Measurement of brain function of car driver using functional near-infrared spectroscopy (fNIRS)," *Computational Intelligence and Neuroscience*, 2009.

[23] G. B. Smith, D. R. Prytherch, D. Watson, V. Forde, A. Windsor, P. E. Schmidt, P. I. Featherstone, B. Higgins, and P. Meredith, "SpO2 values in acute medical admissions breathing air-Implications for the British Thoracic Society guideline for emergency oxygen use in adult patients?" *Resuscitation*, vol. 83, no. 10, pp. 1201–1205, Oct 2012.

[24] M. Nitzan, A. Romem, and R. Koppel, "Pulse oximetry: Fundamentals and technology update," pp. 231–239, Jul 2014.

[25] K. Budidha and P. A. Kyriacou, "Investigation of photoplethysmography and arterial blood oxygen saturation from the ear-canal and the finger under conditions of artificially induced hypothermia," in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, vol. 2015-Novem. Institute of Electrical and Electronics Engineers Inc., Nov 2015, pp. 7954–7957.

[26] H. J. Davies, P. Bachtiger, I. Williams, P. L. Molyneaux, N. S. Peters, and D. P. Mandic, "Wearable In-Ear PPG: Detailed Respiratory Variations Enable Classification of COPD," *IEEE Transactions on Biomedical Engineering*, vol. 69, no. 7, pp. 2390–2400, 2022.

[27] H. J. Davies, I. Williams, N. S. Peters, and D. P. Mandic, "In-ear SpO2: A tool for wearable, unobtrusive monitoring of core blood oxygen saturation," *Sensors*, vol. 20, no. 17, p. 4879, Aug 2020.

[28] E. A. Hamber, P. L. Bailey, S. W. James, D. T. Wells, J. K. Lu, and N. L. Pace, "Delays in the detection of hypoxemia due to site of pulse oximetry probe placement," *Journal of Clinical Anesthesia*, vol. 11, no. 2, pp. 113–118, Mar 1999.

[29] V. Goverdovsky, D. Looney, P. Kidmose, and D. P. Mandic, "In-ear EEG from viscoelastic generic earpieces: Robust and unobtrusive 24/7 monitoring," *IEEE Sensors Journal*, vol. 16, no. 1, pp. 271–277, Jan 2016.

[30] T. L. Rusch, R. Sankar, and J. E. Scharf, "Signal processing methods for pulse oximetry," *Computers in Biology and Medicine*, vol. 26, no. 2, pp. 143–159, 1996.

[31] MaximIntegrated, "Recommended configurations and operating profiles for MAX30101/MAX30102 EV Kits," Tech. Rep., 2018.

[32] M. Elgendi, "On the analysis of fingertip photoplethysmogram signals," *Current Cardiology Reviews*, vol. 8, no. 1, pp. 14–25, 2012.

[33] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, no. Oct, pp. 2825–2830, 2011.

[34] J. Zhu, S. Rosset, H. Zou, and T. Hastie, "Multi-class AdaBoost," *Ann Arbor*, vol. 1001, p. 48109, 2006.

[35] T. Appel, C. Scharinger, P. Gerjets, and E. Kasneci, "Cross-subject workload classification using pupil-related measures," in *Proceedings of the 2018 ACM Symposium on Eye Tracking Research Applications*, ser. ETRA '18. New York, NY, USA: Association for Computing Machinery, 2018.

[36] C. Herff, D. Heger, O. Fortmann, J. Hennrich, F. Putze, and T. Schultz, "Mental workload during n-back task—quantified in the prefrontal cortex using fnirs," *Frontiers in Human Neuroscience*, vol. 7, p. 935, 2014.

[37] K. Ouwehand, A. van der Kroef, J. Wong, and F. Paas, "Measuring Cognitive Load: Are There More Valid Alternatives to Likert Rating Scales?" *Frontiers in Education*, vol. 6, 2021.

[38] G. I. Christopoulos, M. A. Uy, and W. J. Yap, "The Body and the Brain: Measuring Skin Conductance Responses to Understand the Emotional Experience:," *Organizational Research Methods*, vol. 22, no. 1, pp. 394–420, Dec 2016.

[39] A. Jimenez-Molina, C. Retamal, and H. Lira, "Using Psychophysiological Sensors to Assess Mental Workload During Web Browsing," *Sensors 2018, Vol. 18, Page 458*, vol. 18, no. 2, p. 458, Feb 2018.

[40] S. Dong, L. M. Reder, Y. Yao, Y. Liu, and F. Chen, "Individual differences in working memory capacity are reflected in different ERP and EEG patterns to task difficulty," *Brain Research*, vol. 1616, pp. 146–156, Aug 2015.