Exploring Machine Learning Approaches for Classifying Mental Workload using fNIRS Data from HCI Tasks

Johann Benerradi*
IDMC
Université de Lorraine
Nancy, France
johann.benerradi@gmail.com

Horia A. Maior
School of Computer Science
University of Nottingham
Nottingham, UK
horia.maior@nottingham.ac.uk

Adrian Marinescu
Faculty of Engineering
University of Nottingham
Nottingham, UK
adrian.marinescu@nottingham.ac.uk

Jeremie Clos
School of Computer Science
University of Nottingham
Nottingham, UK
jeremie.clos@nottingham.ac.uk

Max L. Wilson
School of Computer Science
University of Nottingham
Nottingham, UK
max.wilson@nottingham.ac.uk

ABSTRACT

Functional Near-Infrared Spectroscopy (fNIRS) has shown promise for being potentially *more suitable* (than e.g. EEG) for brain-based Human Computer Interaction (HCI). While some machine learning approaches have been used in prior HCI work, this paper explores different approaches and configurations for classifying Mental Workload (MWL) from a continuous HCI task, to identify and understand potential limitations and data processing decisions. In particular, we investigate three overall approaches: a logistic regression method, a supervised shallow method (SVM), and a supervised deep learning method (CNN). We examine personalised and generalised models, as well as consider different features and ways of labelling the data. Our initial explorations show that generalised models can perform as well as personalised ones and that deep learning can be a suitable approach for medium size datasets. To provide additional practical advice for future brain-computer interaction systems, we conclude by discussing the limitations and data-preparation needs of different machine learning approaches. We also make recommendations for avenues of future work that are most promising for the machine learning of fNIRS data.

CCS CONCEPTS

• **Human-centered computing** → **Interaction paradigms**; • **Computing methodologies** → **Machine learning**.

KEYWORDS

fNIRS, Mental Workload, Machine Learning, Deep Learning

ACM Reference Format:
Johann Benerradi, Horia A. Maior, Adrian Marinescu, Jeremie Clos, and Max L. Wilson. 2019. Exploring Machine Learning Approaches for Classifying Mental Workload using fNIRS Data from HCI Tasks. In *Proceedings of the Halfway to the Future Symposium 2019 (HTTF 2019), November 19–20, 2019, Nottingham, United Kingdom.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3363384.3363392

*University of Nottingham, School of Computer Science, Nottingham, UK.

HTTF 2019, November 19–20, 2019, Nottingham, United Kingdom

ACM ISBN 978-1-4503-7203-9/19/11.
https://doi.org/10.1145/3363384.3363392

1 INTRODUCTION

Assessing mental workload in users is a long established concern and well evaluated concept in HCI and human factors, especially in safety critical domains like air traffic control [38]. Past work developed and relied on self-reporting methods like NASA-TLX [18], which can retrospectively judge the workload involved in a task. One of the longstanding goals for the future is that technology will be able to reliably identify people that are at risk of becoming overloaded, and automatically adjust their task demand accordingly. So far, existing research into workload estimation has focused on more established physiological sensors such as eye-tracking devices or EEG. More recently, functional Near-Infrared Spectroscopy (fNIRS) has shown promise as an alternative technique for measuring brain activity in HCI [1, 24, 27, 41, 48], because measures of blood oxygenation are more tolerant of physical movement than electrical activity in the brain measured by EEG, whilst still being just as portable [25, 42]. fNIRS, however, has received less attention in terms of MWL classifications using machine learning and it is not clear that established approaches for other brain data will work for fNIRS data.

While some examples of prior work use machine learning to classify mental workload levels from fNIRS data [1, 48], they typically provide very little information about the machine learning models, and do not compare different approaches for generating them. Here, we specifically explore three different approaches for classifying mental workload from fNIRS data:

- Approach 1: a logistic regression model
- Approach 2: Support Vector Machines (SVM), a standard supervised learning approach
- Approach 3: Convolutional Neural Networks (CNN), a deep learning approach

The first approach is a simple linear model, while the SVM and the CNN are standard shallow and deep approaches in the literature.

To support the use of fNIRS data in the future of brain-computer interaction, we provide code samples for all of our processing pipeline stages and machine learning techniques.

Our investigation is guided by these Research Questions:

RQ1 How should data be prepared and how should features be selected for these approaches?
RQ2 How well do these approaches perform at classifying mental workload?
RQ3 Do personalised models outperform generalised models for fNIRS data?

2 RELATED WORK

Mental workload is a well established concept based upon the multiple resources model from human factors [47], where mental workload levels increase significantly when a user has to cognitively process large amounts of information within one modality (spatial or verbal), and within the same stage of cognitive processing. A user, for example, will struggle to hold two number sequences in their head, but can scan a piece of text for a particular keyword while rehearsing a single number sequence, or can do so while processing spatial information. More broadly for HCI, Sharples & Magaw [38] describe mental workload as *"the relationship between primary task performance and the resources demanded by the primary task"*, where task performance drops if a user has too little to do to remain cognitively engaged in the task, or where task demand is too high for the user to perform it at a suitable level of performance. Similar concepts are captured within the cognitive load literature [33], and this term is often used synonymously in publications (e.g. [14, 17]).

Well established approaches to evaluating mental workload have traditionally depended on subjective reporting. NASA TLX [18] is perhaps the most established one for retrospectively assessing an entire task for both mental and physical workload, where papers vary on whether they report overall differences, or differences in individual subscales like mental demand and effort. With more desire for understanding the *current* mental workload that a user is experiencing during a task, like when the workload is becoming too much for an air traffic controller, the Instantaneous Self Assessment (ISA) [9, 23] scale was developed to allow participants to quickly report mental workload on a simple Likert scale. A recognised consequence of this technique is that self-reporting mental workload during a task can act as a secondary task that itself impedes the performance of the primary task [45]. Consequently, much work has focused on physiological measurements to estimate mental workload.

2.1 Mental workload and physiological measures

Many psycho-physiological changes have been observed to correlate with mental workload changes. An observable change, which is sometimes built into eye-tracking products, is pupil dilation, where dilation in a consistently lit environment is an indication of increased mental workload [5, 21, 29]. Skin temperature changes are also observable from a thermal camera, where Marinescu et al. [28, 29] have shown that nose temperature often decreases with increased mental workload. On the body, galvanic skin response [11, 39, 40] and fluctuations in cardiac activity [7, 17, 31, 44] (measured from e.g. the wrist), have often been correlated with mental workload changes.

A more direct approach, often used to estimate mental workload, is to take measurements of brain activity. Electroencephalography (EEG) is now a consumer-grade technology for estimating mental workload [3], where changes in EEG data have been shown to correlate highly with working memory load, integration of information and analytical reasoning [6]. The commercialisation of EEG has also meant that very cheap EEG devices (<$200) can be easily integrated into brain-computer interaction responsive systems [34, 36].

In the last decade, however, an increasing amount of research has investigated the use of fNIRS in the field of HCI [25, 35, 42] due to its better spatial resolution and tolerance to movements than EEG, even though it has a slightly lower temporal resolution [32]. fNIRS measures blood oxygenation levels, and is typically applied to the prefrontal cortex due to the involvement of this brain area in working memory [20]. Blood oxygenation change is a reliable indicator of the prefrontal cortex activation which reflects an increase in the amount of oxygenated haemoglobin (HbO) and changes in the de-oxygenated haemoglobin (Hb). These changes are affected by both a) the individuals underlying bodily blood oxygenation levels (which may be higher for a healthier person, or indeed for someone that is currently more alert), and b) the Blood Oxygen Level-Dependent (BOLD) delay, where the body can take 2-6 seconds (varying across individuals) to fulfil oxygen demands from the brain. This type of brain activation (oxygen in regions of the brain) correlates to the activation observed in fMRI studies [12]. While not yet as commercialised as EEG, fNIRS devices can be fully portable (via e.g. Bluetooth), and are worn in a similar way to non invasive EEG sensors. This portability in addition to its tolerance to movements thus makes fNIRS well suited for the evaluation of real-world HCI tasks such as computer usage [26, 35, 42].

2.2 Machine learning of mental workload

Supervised learning is a subcategory of machine learning where data is labelled with some measure of interest that we are trying to estimate, and classification is a subcategory of supervised learning where that label is a category. Typical approaches to workload classification involve either two (low and high) or three classes (low, medium, and high). We start by reviewing machine learning models of physiological data, and then more specific examples as applied to fNIRS data.

2.2.1 Machine learning with physiological data. Because it can be computed using a camera and is thus less intrusive than most other sensors, the most common set of features for mental workload estimation is the position and dilation of the pupils. Zhang et al. [49] used a decision tree classifier, with 2 classes (low and high), on a vehicle driving task. They used summary statistics (mean and standard deviation) on gaze data (pupil diameter, detection of the direction of gaze) as well as driving data, e.g. velocity, lane position, steering angle and acceleration and achieved significant results using all features.

Marshall [30] compared neural networks with discriminant function models, and found that neural network models performed as good or better in all cases on a binary classification task where

the classes are relaxed/engaged. Haapalainen et al. [17] used a naive Bayes classifier on a binary classification problem, mixing a variety of sensors in order to determine their relative usefulness: eye-tracker (eye movement and change in pupil size), ECG armband (used to collect galvanic skin response (GSR), heat flux (rate of heat transfer) and median absolute deviation (MAD - measure of variability) of the ECG), EEG headset (EEG signal converted into two mental state outputs, attention and meditation), HR monitor (HR and HRV). Chen and Epps [10] used eye-tracking data (pupil size, blink number) to detect the level of mental workload during a mental arithmetic task. They labelled the amount of work on a 5 point scale, and then grouped those in either 2 classes (1 and 2 versus 4 and 5) or 3 classes (1 versus 4 versus 5) to generate different classification tasks. They then used a Gaussian mixture model classifier to perform the classification. Solovey et al. [43] performed an evaluation of multiple learning algorithms: decision trees, logistic regression, 1-nearest neighbour, multilayer perceptron, and naive Bayes using heart rate and heart rate variability as physiological data as well as data extracted from the vehicle that was being driven. Fridman et al. [14] compared a hidden Markov model with their 3D CNN model on a 3 class classification problem using a working memory task.

2.2.2 Machine learning with fNRIS data. Comparatively little work has been done on classifying mental workload using fNIRS data. Early work used the task of counting the coloured faces of a cube [15, 37] as a way to generate low, medium and high mental workload, and then trained machine learning algorithms to perform multi-class classification. Firstly using a 3-nearest neighbours approach [37], classification accuracy was then much improved using a multilayer perceptron with a sliding window [15] which brought the performance to the 41.15% - 69.7% range.

Because most of the existing research has focused on a batch processing method, which is flawed for realistic applications of mental workload estimation, later work then focused on bringing this performance to a real-time setting. It was first done by Girouard et al. [16] using an unspecified sequence classification algorithm to categorise tasks in a 3-class classification problem. Afergan et al. [1] used this approach to adapt the difficulty of a UAV piloting task by estimating the mental workload associated with the current reaction, and Yuksel et al. [48] used these estimations for a brain-computer system enhancing piano learning, which enabled learners to play faster and with a higher accuracy.

Our investigation below builds on this kind of prior work to specifically investigate the value of *different* machine learning approaches with fNIRS data for the use in brain-computer interaction applications.

3 MENTAL WORKLOAD DATASET

In this paper we sought to investigate, compare, and release the software for three alternative ways of analysing and classifying various levels of mental workload based on the measurements coming from fNIRS. We generated a dataset consisting of performance data, subjective workload information, and physiological responses during a controlled experiment. The study design, described below, closely follows the study performed by Marinescu et al. [29].

3.1 Dataset task

A specific computer-based task was designed to impose different levels of mental demands on participants. As shown in Figure 1, the task consists of aiming at the target balls using a joystick, and shooting them using a button on the joystick, before the balls reach the yellow line; reaching the yellow line drags it down. The yellow line moved down the screen with the lowest missed target, or moved up the screen if all targets were destroyed. The position of the joystick is indicated by a red circular cursor that turns green once it is within range of the target. We preferred this task in favour of a n-back task because it is a more naturalistic and continuous task that allows us to easily model and understand the task demands imposed on the individuals.

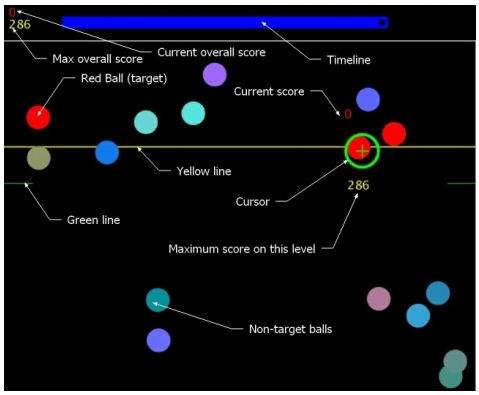

Figure 1: Study task of shooting red balloons

Participants played this task three times, each lasting approximately 10 minutes. As presented in Figure 2, demand increased and decreased within each task. This demand was set by incrementally increasing the number of targets from 3 to a maximum of 13 at the mid-point of each round, then reducing the number of targets back to 3. During Type 1, the participant had to shoot all red balls. To increase Mental Workload, Type 2 involved shooting only the balls with odd numbers on them regardless of the colour. Sample screen recordings of the task are available online[1].

3.2 Data collection protocol

Eleven students and staff from the University of Nottingham took part in the study (6 men and 5 women; mean age = 29 years; SD = 6.8; range = 19-42). Each participant was invited to read the information sheet and provide consent. They then played a training version of the stimulus task until they became familiar with the rules and the controls. After the training was finished, the physiological sensors were placed upon the participants. When ready, participants performed each condition with the corresponding stimulus task. Every 45 seconds during the tasks, the participant was asked to verbally rate their level of mental workload using the Instantaneous Self-Assessment (ISA) technique [22]. Participants were compensated with a £20 voucher as remuneration for their time. This protocol was approved by the Ethics Committee.

[1] T1 sample: https://goo.gl/uiimKg; T2 sample: https://goo.gl/2FVxA2

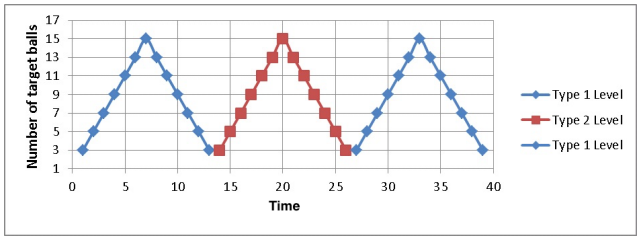

Figure 2: Variation in demand during study tasks

3.2.1 fNIRS measurements. Measures of brain activity were recorded using an fNIRS300 device and the associated Cognitive Optical Brain Imaging (COBI) Studio hardware integrated software platform provided Biopac Systems Inc [4]. The headband shaped device is a sixteen-channel transducer for continuous Near-Infrared Spectroscopy (NIRS). The headband consists of four infrared (IR) emitters operating on a range between 700 to 900 nm, and ten IR detectors. See Figure 3 for how the headband is positioned. The acquisition rate of the device is 2 Hz.

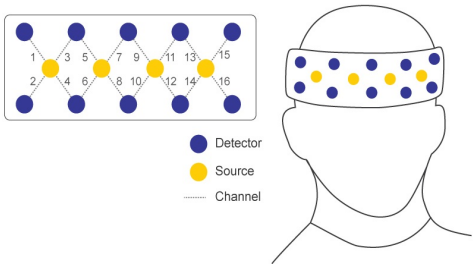

Figure 3: Sensor layout for the Biopac fNIRS used [2]

3.2.2 ISA scores: category labels. To capture subjective workload information, participants were surveyed during the tasks on a regular interval of 45 seconds using the 5-point ISA scale. The mean ISA score levels were then split within a number of classes in order to label the fNIRS data for: 2 classes (high and low), and 3 classes (high, medium and low). In order to translate this information from a 5-point score into a 2, respectively 3 levels of workload we had to split the data such that we keep a balanced number of labels in each class. Figure 4 illustrates how scores were split.

3.2.3 Data exclusions. Due to the limitations of the equipment (one headband does not fit all) some pieces of data are missing or is heavily corrupted with noise. Therefore, the data from two participants (p01 and p10) was excluded for certain analyses.

4 PRE-PROCESSING PIPELINE

In the following two sections, we present a software pipeline developed to pre-process, process, analyse and classify mental workload from fNIRS data (code is made available online[3]). This section describes our pre-processing pipeline necessary for preparing the data for classification.

[2]Image by Hyosun Kwon
[3]https://gitlab.com/HanBnrd/fnirs-learning

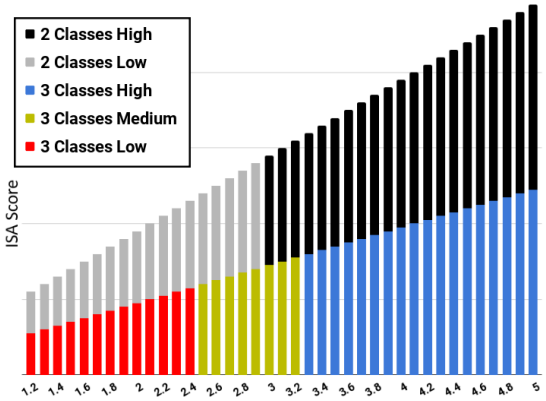

Figure 4: 5-point ISA score split into a 2 (high and low), respectively 3 classes (high, medium and low) of workload

1 - Modified Beer-Lambert Law (MBLL). Using a typical fNIRS sensor, an important pre-processing step is needed in order to transform raw data from the device into oxygenated (oxy-Hb) and deoxygenated (deoxy-Hb) haemoglobin levels using the Modified Beer-Lambert Law (MBLL) [46]. Thereafter, filtering algorithms remove high-frequency noise, and physiological artefacts such as heartbeats and other motion related artefacts. These steps are usually performed by the recording software that comes with the sensor, and the two resulting values are provided to use for real-time and offline monitoring and analysis.

2 - Correlation Based Signal Improvement (CBSI). Cui et al. addresses the challenge of improving the signal quality in fNIRS data and propose the CBSI Filter [13]. Designed for fNIRS in particular, this technique filters the signal from movement artefacts, even those induced by head motion. Carefully studying how such artefacts affect the fNIRS measurements of oxy-Hb and deoxy-Hb, the two which are typically strongly negatively correlated, were found to become more positively correlated in the presence of movement artefacts. Therefore, the proposed method for filtering fNIRS signal reduces noise based on the principle that the concentration changes in oxy-Hb and deoxy-Hb should be negatively correlated [13]. In practice, the filtering function takes as input the oxy-Hb and deoxy-Hb measurements and provides a resulting measure (that we simply call CBSI) that indicates changes in activity over the targeted region of the brain. This filter is useful for both real-time and offline use.

3 - Resulting CBSI data. Based on the CBSI filtering technique, Figure 6 shows the strong link between the resulting fNIRS data and the workload experienced by participants, where stronger correlations were observed on different channels. We correlated each channel of the resulting CBSI filtered data with the Mean ISA scores of participants subjective workload reports. Further, we normalised the ISA scores and normalised and averaged the fNIRS data such that they are comparable, and Figure 5 shows the strong connection between the two.

4 - Normalising the resulting data. Because fNIRS is a relative rather than absolute measure, it is typically used in a within-participants study design, therefore, ideally comparing different study conditions on a single-continuous recorded session. That

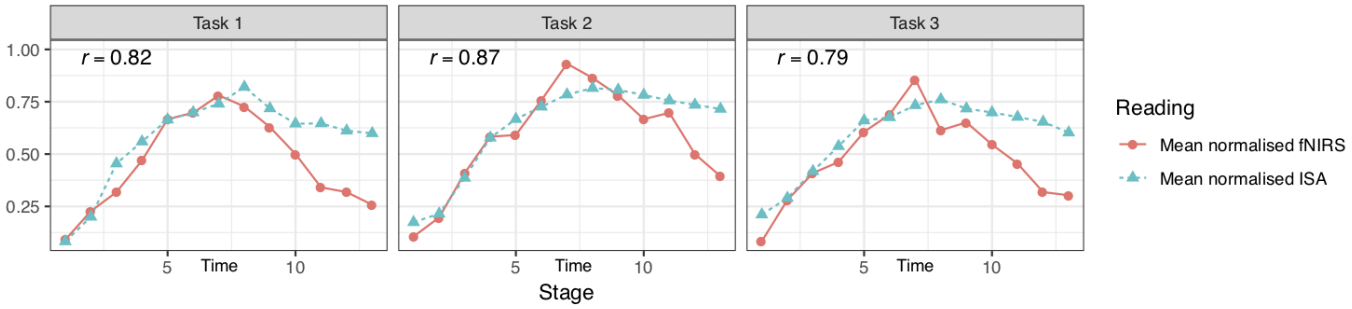

Figure 5: fNIRS refined workload measure vs ISA subjective workload technique with correlation coefficient

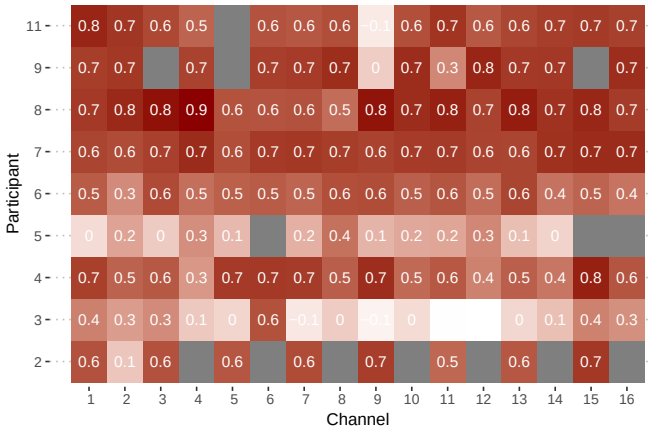

Figure 6: Correlation table of ISA subjective workload and fNIRS workload measure per channel and participant

means there are certain limitations when comes to using fNIRS to compare between participants, or when comes to multiple recordings lets say over multiple days.

One straight forward way to overcome this limitation is to normalise the fNIRS data, such that at any point the fNIRS measurements are relative to the previous measurements and states, and will always vary in the range of 0 and 1. This technique can be useful in both, offline and real time scenarios, and we have implemented a version using Python, available in the links provided with this paper.

5 LEARNING APPROACHES AND RESULTS

Three approaches to detect user workload levels are presented. An approach to classify workload levels using a logistic regression model, and two machine learning techniques are detailed, each being representative of the state of the art in their specific category - Support Vector Machines (SVM), for shallow classifiers, and Convolutional Neural Networks (CNN), which are a category of deep neural networks. Across our three approaches, we also consider two techniques based on: personalised and generalised learning.

Personalised learning. These techniques build a model that is specific to one person. Their main advantage is that personalised models are usually better able to perform predictions on the person they were learned on. Their main drawback, however, is that they need to be trained for every new participant, requiring to gather enough data before being able to classify mental workload. Our analysis achieves this by learning on the two first tasks and tests are made on the remaining one, for each participant.

Generalised learning. Generalised learning refers to machine learning techniques used to build a model over a population. Its main advantage is to be able to generalise from multiple users in order to be usable for a new user without any new data to train on. Its main drawback is that unless given enough data, it tends to perform generally worse than a personalised model for a user. In our analysis, the training of generalised models was done by holding out data from one participant for testing and the remaining for learning. We repeated the process through to test the data from all the participants, this way performing k-fold cross validation.

5.1 Approach 1: logistic regression

The first proposed method classifies mental workload levels based on a logistic regression for ordinal responses model that is trained with the labelled normalised CBSI filtered fNIRS data. Logistic regression is a type of classification algorithm used when the response variable is categorical. Logistic regression uses a maximum likelihood estimation to determine the regression coefficients of the linear model. The sigmoid function is used to output the probability of a predictor variable belonging to one of the classes, in our case one of the levels of subjective workload. This approach was chosen to demonstrate the classification results that can be obtained by using a simple approach.

5.1.1 Data and feature selection. Both, the personalised and generalised logistic regression models use the same input features. They consisted of the mean CBSI values during each of the task blocks.

5.1.2 Results. A personalised model was trained for each participant on the first two tasks and tested on the third. Table 1 shows the accuracy of the prediction made by the model on new data from the third task. In the same way, a generalised model was trained on the data from all participants except for one that it was tested on. This process was repeated for each participant and table 2 shows the resulting accuracy.

Table 1: Personalised logistic regression using normalised fNIRS data: classification accuracy

Participant	2 classes	3 classes
p02	69.23 %	61.53 %
p03	46.15 %	30.76 %
p04	84.61 %	46.15 %
p05	69.23 %	38.46 %
p06	76.92 %	61.53 %
p07	76.92 %	53.84 %
p08	92.30 %	46.15 %
p09	76.92 %	46.15 %
p11	84.61%	30.76 %
Average	75.21 %	46.15 %

Table 2: Generalised logistic regression using normalised fNIRS data: classification accuracy. Training on 8 participants testing on 1.

Test on	2 classes	3 classes
p02	69.23 %	53.85 %
p03	61.54 %	43.59 %
p04	71.79 %	56.41 %
p05	66.67 %	33.33 %
p06	74.36 %	51.28 %
p07	74.36 %	66.67 %
p08	71.79 %	58.97 %
p09	69.23 %	56.41 %
p11	53.85 %	38.46 %
Average	68.09 %	50.99 %

5.2 Approach 2: support vector machines

Support Vector Machines [19] (SVMs) are maximal margin classifiers. They work by finding a hyperplane that can accurately separate the data while simultaneously maximising the distance of this hyperplane from each of the data points which are closest to it. SVMs then progressed with the introduction of the kernel trick [8], which consists in replacing the dot product part of the optimisation process by simple functions defined on pairs of input patterns.

They achieve a high generalisation power by introducing a slack variable in the optimisation process which allows the SVM to tolerate some misclassification if it results in a significantly smoother hyperplane. That trade-off is controlled by a regularisation parameter which can be manually tuned to each problem. In the context of our experiments, this parameter was fixed and not optimised.

5.2.1 Data and feature selection. Both, the personalised and generalised SVMs use the following features:

- **CBSI mean**, by averaging the values for each set of four connected channels as per figure 3. This results in a vector of four means, where the first value corresponds to the mean of channels 1, 2, 3 and 4, the second value to channels 5, 6, 7 and 8, and so on. This correspond for the whole headset to 4 values per time sample.
- **CBSI standard deviation**, by taking the standard deviation calculated similarly to the previous feature, corresponding again to 4 values.
- **CBSI gradient**, computed as the slope of the linear regression on the 5 previous seconds of each CBSI mean, as describe above, which yet again corresponds to 4 values.

The mean is used to lessen noise that can be present in some channels and the standard deviation enables to still keep information about variability between the averaged channels. Those features enable to create a training set of size 2160x12, respectively 25920x12 for the personalised, respectively generalised learning and a testing set of size 1080x12, respectively 3240x12 for the personalised, respectively generalised learning. The labels used for classification are those described in figure 4. The learning dataset was shuffled before training each model.

Table 3: Personalised SVM (linear kernel): classification accuracy. Training on 2 tasks testing on 1.

Participant	2 classes	3 classes
p02	58.70 %	46.57 %
p03	66.67 %	41.20 %
p04	81.94 %	47.50 %
p05	81.57 %	45.46 %
p06	66.02 %	45.65 %
p07	81.57 %	43.61 %
p08	72.04 %	55.65 %
p09	73.33 %	64.81 %
p11	73.43 %	46.57 %
Average	72.81 %	48.56 %

5.2.2 Results. Table 3 shows the SVM accuracy for two classes (high, low workload) and three classes (high, medium, low workload) using a linear kernel with personalised learning on each participant. The k-fold cross-validation average accuracy is 72.81 % for 2 classes and 48.56 % for 3 classes.

Table 4 on the other hand shows the SVM accuracy for two classes and three classes using a linear kernel with generalised learning. The k-fold cross-validation average accuracy is here 71.27 % for 2 classes and 53.90 % for 3 classes.

5.3 Approach 3: convolutional neural network

Neural networks are a function approximator built from the succession of layers of computational units (neurons) where each unit is connected to every unit from the previous layer, and produces an output from the non-linear transformation to the weighted sum of the outputs of the units in the previous layer, weighted by the strength of their respective connection to the unit. This layering produces an incremental transformation of the data until a linear classifier (typically a logistic regression) is run on the output of the last layer, producing a prediction. A neural network is fully differentiable, and therefore the training process occurs by modifying the weights of the connections between units using the gradient of

Table 4: Generalised SVM (linear kernel): classification accuracy. Training on 8 participants testing on 1.

Test on	2 classes	3 classes
p02	75.15 %	54.38 %
p03	51.57 %	29.07 %
p04	77.78 %	59.32 %
p05	66.60 %	44.57 %
p06	67.59 %	52.41 %
p07	76.36 %	63.73 %
p08	76.70 %	61.76 %
p09	68.18 %	59.26 %
p11	81.54 %	60.68 %
Average	71.27 %	53.90 %

the error produced from the prediction with respect to the current weight of the connection. The process of computing the gradient of the error with respect to each weight, starting from the last layer and going backwards toward the first layer, is called backpropagation.

Convolutional Neural Networks (CNNs) are deep neural networks containing one or more convolutional layers. Convolutional layers are layers with specific space-invariant property which makes them useful in analysing raw data such as sensor data and images.

As this type of neural network typically requires large amounts of data, we here describe only generalised learning, enabling to perform training on multiple participants.

5.3.1 Data and feature selection. The CNN uses the following features:

- **CBSI mean**, by averaging the values for each set of four connected channels as per figure 3. This results in a vector of four means per time sample as described for the SVM approach.
- **CBSI standard deviation**, by taking the standard deviation calculated similarly to the previous feature, corresponding again to 4 values per time sample.

Computing the CBSI mean and standard deviation enables us to filter through and remove data channels that are too noisy while keeping the same data matrix for every input (see Figure 6). Inputs of 10 sec (20 time points) of those features were used for the model with a 9 sec overlapping. No overlap was made between inputs from different classes. This creates training inputs of size 2x20x4 corresponding to [average and standard deviation] x [10sec of 2Hz data] x [spatial locations of averages and standard deviations]. This shape enables to perform convolutions more easily across time and space. In order to show the evolution of the performance with data increase we used respectively 4, 6 and 8 participants to train the model, corresponding to respectively 51844, 7776 and 10368 inputs. Testing was made on one participant corresponding to 1296 inputs. Again, the labels used for classification are those described in figure 4.

Table 5: CNN k-fold cross-validation average accuracy with increasing number of data. Training on n-1 participants testing on 1 (with n number of participants in the dataset).

Dataset size	Number of training data	Accuracy 2 classes	Accuracy 3 classes
5 participants	5184	67.52 %	42.61 %
7 participants	7776	71.76 %	42.68 %
9 participants	10368	72.77 %	49.53 %

5.3.2 CNN architecture. The model architecture is described in figure 7. The CNN is composed of two convolutions (one across time and channels, the other across time only), each followed by max-pooling down sampling. Those convolutions are then followed by a fully connected layer with a ReLU activation function, which is then fed into another smaller fully connected layer. The output of that final layer is passed through a Log-Softmax normalisation in order to produce a vector of class probabilities, which is used to compute the cross-entropy error and perform the training by backpropagation. The learning rate was set to 0.001 and the momentum to 0.8.

5.3.3 Results. In the same training configuration as the other approaches for generalised learning (train on 8 participants test on 1), the CNN performed quite well with a k-fold cross-validation average accuracy of 72.77 % for 2 classes and 49.53 % for 3 classes.

In Table 5 we present the results from the CNN with an increasing size of the training dataset. Even though it appears that an increasing the train set size may improve the model performance, no statistical test could highlight any correlation between the number of training samples and the accuracy with a confidence level of 5 %.

5.4 Approach comparison

The models were compared using paired-sample Student t-tests and the thresholds for significance levels were set at 5 %.

5.4.1 Models based on personalised learning. When it comes to the models based on personalised learning we could only compare between the logistic regression and the SVM approaches as CNN could only be trained using generalised learning. Table 6 shows how the logistic regression and the SVM performed quite similarly with respectively 75.21 % and 75.81 % of k-fold cross-validation average accuracy for 2 classes. Those accuracies were respectively 46.15 % and 48.56 % for 3 classes. No statistically significant differences were shown between those two approaches either for 2 or 3 classes.

5.4.2 Models based on generalised learning. Models based on generalised learning allowed all of the investigated techniques to be compared. As shown in table 6 the best performances were achieved by the CNN and the SVM for 2 classes with k-fold cross-validation average accuracies of respectively 72.77 % and 71.27 %. Even though no significance is shown, it appears that the CNN outperforms the logistic regression (68.09 % accuracy) with a p-value of 0.0967. For 3 classes the highest accuracy is achieved by the SVM with 53.90 % which appears to be better than the CNN with a p-value of 0.0758.

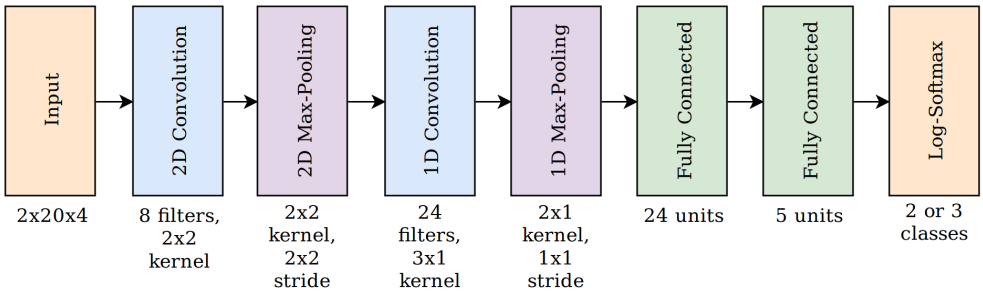

Figure 7: CNN architecture for 2 and 3 classes classification

Table 6: k-fold cross-validation average accuracies for each approaches to classify mental workload levels based on generalised (training on 8 participants, testing on one) and personalised learning

Classes	Approach	Personalised	Generalised
2 classes	Logistic regression	75.21 %	68.09 %
	SVM	72.81 %	71.27 %
	CNN	N/A	72.77 %
3 classes	Logistic regression	46.15 %	50.99 %
	SVM	48.56 %	53.90 %
	CNN	N/A	49.53 %

5.4.3 Personalised vs. generalised learning. Wilcoxon tests were performed to evaluate the difference between personalised and generalised learning for logistic regression and SVM. No statistical differences were found for 2 and 3 classes, which means that generalised models perform similarly to personalised ones.

6 DISCUSSIONS AND PRACTICAL ADVICE

We begin our discussion by explaining feature selection and addressing the research questions, before addressing practical considerations for using these approaches for different situations.

6.1 Data preparation and feature selection

RQ1 aimed at finding how the data should be prepared and features selected for each approach. The first step of the pre-processing was to apply a CBSI filtering in order to remove head motion artefacts. However, some channels may be affected more severely by noise for various reasons. For example, the brain scanner did not fit every participant perfectly. Those noisy channels were removed by visual inspection and this is why we decided to use the mean signal between each 4 nearby channels, in order to have the same input size for every participant and task. The standard deviation of those same 4 nearby channels was also computed as it could reflect differences between those channels that can be a good insight on oxygenation differences across space. Those are the two features that we decided to feed to the CNN as those kind of models have the specificity to learn patterns by themselves. The SVM on the other hand is more dependant on features that can lead to split data into the different classes. We decided to introduce a third feature which was the slope of the linear regression on 5 sec worth of CBSI means. This feature is a good indicator of the evolution of brain oxygenation across time which can give insight on the mental workload. Indeed, an increase in mental workload will lead to a an increase of oxygenated blood which can be highlighted by an increase in the slope of the linear regression of CBSI means.

6.2 Model specificities

RQ2 was concerned the performance, here reflected by the accuracy, of our approaches at classifying correctly mental workload into different classes. Besides accuracy, several important factors have to be kept in mind about the convenience of each model in real-world use which links this second research question to RQ3 about differences between personalized and generalized models. From one perspective, an ideal candidate would be a model based on generalised learning, such that it could be trained on a large dataset, and then freely applied to new participants. This would mean that no training period would be needed for each new participant which is more convenient, and the results points toward the fact that those generalised approaches can be suitable for mental workload classification.

Below we discuss some of the practical advantages and disadvantages of each approach with further detail.

6.2.1 Logistic regression. The model based on a logistic regression was a more simplistic approach to classifying workload levels. Table 6 shows how this approach proved to perform in a close range to the SVM for both personalised and generalised models. By order of simplicity and speed to train, however, the logistic regression model is the fastest of the three, making it a realistic candidate for situations where quick starting, without much training data, is desired.

6.2.2 SVM. The SVM approach especially stands out on 3 classes classification with generalised learning as shown in table 6. The dataset is substantial but not too large which enables the SVM to be trained relatively quickly with a linear kernel, which makes it also usable for personalised learning. In comparison to a CNN model, SVM would work faster and with a smaller dataset. This model is also more reliable and the training is less affected by randomness than the CNN. One limiting factor of the SVM however is that it can less easily learn temporal patterns if specific features enabling to

relate them are not used. We tried to reflect this temporal evolution by including the linear regression slope on the 5 previous seconds but this doesn't reflect more complex patterns that can be observed with physiological data.

6.2.3 CNN. The CNN approach really stands out for 2 modalities classification with generalised learning as shown in table 6. A benefit of deep learning approaches like CNNs, is that they don't need a lot of specific features to develop their own understanding of the data. There is significant scope, however, developing the complexity of the CNN through the number of layers. So while CNNs continue to show promise for eventually better models, they require both complex development and large amounts of training data. The choice of this specific deep learning model was oriented by the fact that it could allow to make both spatial and temporal convolution to find features from those two modalities that are crucial for mental workload assessment with fNIRS data.

The main issue of this kind of deep learning approach is that it is *data hungry* and gets better over time as it learns from thousands of samples. Indeed, we are convinced that the CNN has a good potential for classifying mental workload from fNIRS but would benefit from larger datasets. In this study, we decided to use 10 sec input samples with a 9 sec overlap for two reasons. The first is that overlapping enables to have more data and the second is that it then makes the model predict mental workload every second, which would be useful for real-time classification. The training set size might also be the explanation for lower performance compared to other models with 3 classes. Indeed, in this configuration, the training is made on approximately a third of the training set for each class (because of the way labels were made) instead of a half for 2 classes. The fact that the dataset size is at the low end for CNN requirements could also explain that no significant improvement was found with the dataset size increase that we performed which was at maximum with 8 training participants. Further analysis with more participants would help justifying this assumption.

6.3 Other choices

6.3.1 ISA scores. Subjective techniques for assessing users' mental workload are useful ways to capture the subjective experiences of participants experiencing various levels of work demands. In this experiment we used the real-time, continuous ISA technique to survey participants verbally, on a regular interval of 45 seconds, about their perceived mental workload changes during the tasks.

We collected this information in order to be able to correctly label participants fNIRS data with the corresponding low, medium or high workload state. As subjective measures such as ISA rely on the user's ability to self-judge and report the state throughout the task (which requires not only extra effort, but also skill and potential training), we averaged and used all participants' ISA scores as labels for each participant fNIRS data. This was only possible as all participants experienced the exact same level of task demands.

6.3.2 Normalising data. Normalising data was a stage of our preprocessing pipeline, but its less practical to do this in real-time. In real-time, normalisation can only occur by using max and min values within a sliding window, rather than retrospectively with the whole data sample.

6.4 Future work

There is a significant amount of scope for developing the complexity and increasing the accuracy of machine learning classification approaches for fNIRS, which might warrant a significant amount of future research.

One, perhaps obvious, starting point would be to investigate further the potential accuracy that a CNN could reach with larger data samples. On the one hand, more work can be done to gather larger fNIRS datasets in order to take full advantage of deep learning models such as CNN. On the other hand, the model type as well as the model structure can be further investigated in order to be less data hungry and perform on-shot learning or few-shot learning.

Another consideration would be to investigate a universal background models approach, in which a long term generalised CNN is developed and used as a starting point to reduce the training time needed to produce a personalised model for each person. Similarly, a transfer learning approach could be explored, in which different archetype models are created to then be selected to best match each user.

It is also important to note that our investigation was based upon a primarily spatial task that invoked a certain kind of mental workload changes. Future research would benefit from investigating data that is created from different forms of cognitive activity, that might manifest in different ways in the prefrontal-cortex. Indeed, much research into the use of fNIRS considers full-scalp measurements that might benefit from observing concurrent changes in other regions of the brain.

More broadly, both shallow and deep learning models typically benefit from multiple data comparison points, and a large opportunity exists to build stronger models that augment fNIRS data with e.g. facial thermography or galvanic skin response data. Indeed, previous work by Ahn et al. [2] integrate fNIRS and EEG in their models to classify state of restfulness, and find that the multimodal input significantly improve their accuracy.

Finally, in this paper we performed offline analysis which enabled to benefit from CBSI filtering as well as normalisation. Future work will aim at implementing and testing those models for real time analysis. This will require to perform pre-processing on a sliding window which duration will need to be investigated in order to still have good performance while making predictions often enough to be suitable for a real-time neuro-feedback.

7 CONCLUSIONS

While some sensor data solutions, such as step identification from gyroscope data, are now relatively mature, the classification of mental workload from brain data is still largely an unsolved problem. Examples of SVMs have been used in related work, but little work has evaluated the different machine learning approaches that will work best for the task, especially adapting the features used to fit the best specificities of each model. We considered three types of models, including a) a logistic regression, b) a Support Vector Machine (SVM), and c) a Convolutional Neural Network (CNN). While a CNN would typically be expected to work better with large numbers of training samples, we accounted for this factor by restricting its depth. We also considered personalised and generalised models within these three approaches, given that fNIRS produces a relative

measure of blood oxygenation that is widely reported as being subject to individual differences. Generalised models are practically beneficial for removing the need to train personalised models for each user, and our results show that such approach can achieve good performance, especially when simply classifying between low and high workload. There is vast opportunity, however, for future research to investigate more advanced deep learning techniques that generate better and more accurate generalisable models.

ACKNOWLEDGMENTS

This work was supported by the EPSRC [grant numbers EP/G037574/1, EP/N50970X/1, EP/M000877/1]. We would also like to thank Siyang Song, Dr. Enrique Sánchez-Lozano and Dr. Michel Valstar for their advice on the application of machine learning and more specifically deep learning for fNIRS data.

Data Access Statement: Consent was not gained from participants for this dataset to be made available to other researchers.

REFERENCES

[1] Daniel Afergan, Evan M Peck, Erin T Solovey, Andrew Jenkins, Samuel W Hincks, Eli T Brown, Remco Chang, and Robert JK Jacob. 2014. Dynamic difficulty using brain metrics of workload. In *Proc. SIGCHI*. ACM, 3797–3806.

[2] Sangtae Ahn, Thien Nguyen, Hyojung Jang, Jae G Kim, and Sung C Jun. 2016. Exploring neuro-physiological correlates of drivers' mental fatigue caused by sleep deprivation using simultaneous EEG, ECG, and fNIRS data. *Frontiers in human neuroscience* 10 (2016), 219.

[3] Aurélien Appriou, Andrzej Cichocki, and Fabien Lotte. 2018. Towards Robust Neuroadaptive HCI: Exploring Modern Machine Learning Methods to Estimate Mental Workload From EEG Signals. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, LBW615.

[4] Hasan Ayaz and Banu Onaral. 2005. *Analytical software and stimulus-presentation platform to utilize, visualize and analyze near-infrared spectroscopy measures*. Drexel University.

[5] Jackson Beatty. 1982. Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological bulletin* 91, 2 (1982), 276.

[6] Chris Berka, Daniel J Levendowski, Michelle N Lumicao, Alan Yau, Gene Davis, Vladimir T Zivkovic, Richard E Olmstead, Patrice D Tremoulet, and Patrick L Craven. 2007. EEG correlates of task engagement and mental workload in vigilance, learning, and memory tasks. *Aviation, space, and environmental medicine* 78, 5 (2007), B231–B244.

[7] George E Billman. 2011. Heart rate variability–a historical perspective. *Frontiers in physiology* 2 (2011).

[8] Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. 1992. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*. ACM, 144–152.

[9] SD Brennan. 1992. An experimental report on rating scale descriptor sets for the instantaneous self assessment (ISA) recorder. *Portsmouth: DRA Maritime Command and Control Division. DRA Technical Memorandum (CAD5)* 92017 (1992).

[10] Siyuan Chen and Julien Epps. 2013. Automatic classification of eye activity for cognitive load measurement with emotion interference. *Computer methods and programs in biomedicine* 110, 2 (2013), 111–124.

[11] C Collet, E Salvia, and C Petit-Boulanger. 2014. Measuring workload with electrodermal activity during common braking actions. *Ergonomics* 57, 6 (2014), 886–896.

[12] Xu Cui, Signe Bray, Daniel M Bryant, Gary H Glover, and Allan L Reiss. 2011. A quantitative comparison of NIRS and fMRI across multiple cognitive tasks. *Neuroimage* 54, 4 (2011), 2808–2821.

[13] Xu Cui, Signe Bray, and Allan L Reiss. 2010. Functional near infrared spectroscopy (NIRS) signal improvement based on negative correlation between oxygenated and deoxygenated hemoglobin dynamics. *Neuroimage* 49, 4 (2010), 3039–3046.

[14] Lex Fridman, Bryan Reimer, Bruce Mehler, and William T. Freeman. 2018. Cognitive Load Estimation in the Wild. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 652, 9 pages. https://doi.org/10.1145/3173574.3174226

[15] Audrey Girouard, Erin Treacy Solovey, Leanne M Hirshfield, Evan M Peck, Krysta Chauncey, Angelo Sassaroli, Sergio Fantini, and Robert JK Jacob. 2010. From brain signals to adaptive interfaces: using fNIRS in HCI. In *Brain-Computer Interfaces*. Springer, 221–237.

[16] Audrey Girouard, Erin Treacy Solovey, and Robert JK Jacob. 2013. Designing a passive brain computer interface using real time classification of functional near-infrared spectroscopy. *International Journal of Autonomous and Adaptive Communications Systems* 6, 1 (2013), 26–44.

[17] Eija Haapalainen, SeungJun Kim, Jodi F Forlizzi, and Anind K Dey. 2010. Psychophysiological measures for assessing cognitive load. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*. ACM, 301–310.

[18] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Human mental workload* (1988).

[19] Marti A. Hearst, Susan T Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. 1998. Support vector machines. *IEEE Intelligent Systems and their applications* 13, 4 (1998), 18–28.

[20] Yoko Hoshi, Brian H Tsou, Vincent A Billock, Masato Tanosaki, Yoshinobu Iguchi, Miho Shimada, Toshikazu Shinba, Yoshifumi Yamada, and Ichiro Oda. 2003. Spatiotemporal characteristics of hemodynamic changes in the human lateral prefrontal cortex during working memory tasks. *Neuroimage* 20, 3 (2003), 1493–1504.

[21] Shamsi T Iqbal, Xianjun Sam Zheng, and Brian P Bailey. 2004. Task-evoked pupillary response to mental workload in human-computer interaction. In *CHI'04 extended abstracts on Human factors in computing systems*. ACM, 1477–1480.

[22] Jordan and Brennen. 1992. Instantaneous self-assessment of workload technique (ISA). Retrieved from http://www.skybrary.aero/bookshelf/books/1963.pdf. (1992).

[23] CS Jordan. 1992. Experimental study of the effects of an instantaneous self assessment workload recorder on task performance. *Report No. DRA/TM (CAD5)/92011. Farnborough: Defence Evaluation & Research Agency* (1992).

[24] Kristiyan Lukanov, Horia A Maior, and Max L Wilson. 2016. Using fNIRS in usability testing: understanding the effect of web form layout on mental workload. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 4011–4016.

[25] Horia A Maior, Matthew Pike, Sarah Sharples, and Max L Wilson. 2015. Examining the reliability of using fNIRS in realistic hci settings for spatial and verbal tasks. In *Proceedings of CHI*, Vol. 15. 3807–3816.

[26] Horia A Maior, Matthew Pike, Max L Wilson, and Sarah Sharples. 2014. Continuous detection of workload overload: an FNIRS approach. In *Contemporary Ergonomics and Human Factors 2014: Proceedings of the international conference on Ergonomics & Human Factors 2014, Southampton, UK, 7-10 April 2014*. CRC Press, 450.

[27] Horia A Maior, Max L Wilson, and Sarah Sharples. 2018. Workload Alerts - Using Physiological Measures of Mental Workload to Provide Feedback During Tasks. *ACM Transactions on Computer-Human Interaction (TOCHI)* 25, 2 (2018), 9.

[28] Adrian Cornelius Marinescu. 2018. *Facial thermography for assessment of workload in safety critical environments*. Ph.D. Dissertation. University of Nottingham.

[29] Adrian Cornelius Marinescu, Sarah Sharples, Alastair Campbell Ritchie, Tomas Sánchez López, Michael McDowell, and Hervé P Morvan. 2017. Physiological parameter response to variation of mental workload. *Human factors* (2017), 0018720817733101.

[30] Sandra P Marshall. 2007. Identifying cognitive state from eye metrics. *Aviation, space, and environmental medicine* 78, 5 (2007), B165–B175.

[31] Peter Nickel and Friedhelm Nachreiner. 2003. Sensitivity and diagnosticity of the 0.1-Hz component of heart rate variability as an indicator of mental workload. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 45, 4 (2003), 575–590.

[32] Ryota Nishiyori. 2016. fNIRS: An emergent method to document functional cortical activity during infant movements. *Frontiers in psychology* 7 (2016), 533.

[33] Fred Paas, Alexander Renkl, and John Sweller. 2003. Cognitive load theory and instructional design: Recent developments. *Educational psychologist* 38, 1 (2003), 1–4.

[34] Matthew Pike, Max L Wilson, Steve Benford, and Richard Ramchurn. 2016. # Scanners: A BCI Enhanced Cinematic Experience. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, 293–296.

[35] Matthew F Pike, Horia A Maior, Martin Porcheron, Sarah C Sharples, and Max L Wilson. 2014. Measuring the effect of think aloud protocols on workload using fNIRS. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*. ACM, 3807–3816.

[36] Richard Ramchurn, Max L Wilson, Sarah Martindale, and Steve Benford. 2018. # Scanners 2-The MOMENT: A New Brain-Controlled Movie. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, D210.

[37] Angelo Sassaroli, Feng Zheng, Leanne M Hirshfield, Audrey Girouard, Erin Treacy Solovey, Robert JK Jacob, and Sergio Fantini. 2008. Discrimination of mental workload levels in human subjects with functional near-infrared spectroscopy. *Journal of Innovative Optical Health Sciences* 1, 02 (2008), 227–237.

[38] Sarah Sharples and Ted Megaw. 2015. Definition and Mesurement of Human Workload. In *Evaluation of human work*, John R Wilson and Sarah Sharples (Eds.). CRC Press.

[39] Yu Shi, Natalie Ruiz, Ronnie Taib, Eric Choi, and Fang Chen. 2007. Galvanic skin response (GSR) as an index of cognitive load. In *CHI'07 extended abstracts on Human factors in computing systems*. ACM, 2651–2656.

[40] Yoshihiro Shimomura, Takumi Yoda, Koji Sugiura, Akinori Horiguchi, Koichi Iwanaga, and Tetsuo Katsuura. 2008. Use of frequency domain analysis of skin conductance for evaluation of mental workload. *Journal of physiological anthropology* 27, 4 (2008), 173–177.

[41] Erin Solovey, Paul Schermerhorn, Matthias Scheutz, Angelo Sassaroli, Sergio Fantini, and Robert Jacob. 2012. Brainput: enhancing interactive systems with streaming fNIRs brain input. In *CHI*. ACM, 2193–2202.

[42] Erin Treacy Solovey, Audrey Girouard, Krysta Chauncey, Leanne M Hirshfield, Angelo Sassaroli, Feng Zheng, Sergio Fantini, and Robert JK Jacob. 2009. Using fNIRS brain sensing in realistic HCI settings: experiments and guidelines. In *Proc. UIST*. ACM, 157–166.

[43] Erin T Solovey, Marin Zec, Enrique Abdon Garcia Perez, Bryan Reimer, and Bruce Mehler. 2014. Classifying driver workload using physiological and driving performance data: two field studies. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 4057–4066.

[44] Kazushi Taoda, Masanori Kawamura, Kinzou Wakara, Yasuma Fukuchi, and Katsuo Nishiyama. 2001. Heart rate variability during long truck driving work. *Journal of human ergology* 30, 1-2 (2001), 235–240.

[45] Andrew J Tattersall and Penelope S Foord. 1996. An experimental evaluation of instantaneous self-assessment as a measure of workload. *Ergonomics* 39, 5 (1996), 740–748.

[46] Arno Villringer and Britton Chance. 1997. Non-invasive optical spectroscopy and imaging of human brain function. *Trends in neurosciences* 20, 10 (1997), 435–442.

[47] Christopher D Wickens. 2008. Multiple resources and mental workload. *The Journal of the Human Factors and Ergonomics Society* 50, 3 (2008), 449–455.

[48] Beste F Yuksel, Kurt B Oleson, Lane Harrison, Evan M Peck, Daniel Afergan, Remco Chang, and Robert JK Jacob. 2016. Learn piano with BACh: An adaptive learning interface that adjusts task difficulty based on brain state. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 5372–5384.

[49] Yilu Zhang, Yuri Owechko, and Jing Zhang. 2004. Driver cognitive workload estimation: A data-driven perspective. In *Intelligent Transportation Systems*. Citeseer, 642–647.