IOP Publishing Journal of Neural Engineering
J. Neural Eng. **15** (2018) 045001 (15pp) https://doi.org/10.1088/1741-2552/aabb7c

# Improving the analysis of near-spectroscopy data with multivariate classification of hemodynamic patterns: a theoretical formulation and validation

**Jessica Gemignani[1,2], Eike Middell[1], Randall L Barbour[3,4], Harry L Graber[5] and Benjamin Blankertz[2]**

[1] NIRx Medizintechnik GmbH, Gustav-Meyer-Allee 25, 13355 Berlin, Germany
[2] Neurotechnology Group, Technische Universität Berlin, Marchstraße 23, 10587 Berlin, Germany
[3] Department of Pathology, SUNY Downstate Medical Center, 450 Clarkson Ave, Brooklyn, NY 11203, United States of America
[4] NIRx Medical Technologies LLC, 15 Cherry Lane, Glen Head, NY 11545, United States of America
[5] Photon Migration Technologies Corp, 15 Cherry Lane, Glen Head, NY 11545, United States of America

E-mail: jessicagemignani@gmail.com

Received 1 November 2017, revised 20 March 2018
Accepted for publication 4 April 2018
Published 9 May 2018

**Abstract**

*Objective.* The statistical analysis of functional near infrared spectroscopy (fNIRS) data based on the general linear model (GLM) is often made difficult by serial correlations, high inter-subject variability of the hemodynamic response, and the presence of motion artifacts. In this work we propose to extract information on the pattern of hemodynamic activations without using any *a priori* model for the data, by classifying the channels as '*active*' or '*not active*' with a multivariate classifier based on linear discriminant analysis (LDA). *Approach.* This work is developed in two steps. First we compared the performance of the two analyses, using a synthetic approach in which simulated hemodynamic activations were combined with either simulated or real resting-state fNIRS data. This procedure allowed for exact quantification of the classification accuracies of GLM and LDA. In the case of real resting-state data, the correlations between classification accuracy and demographic characteristics were investigated by means of a Linear Mixed Model. In the second step, to further characterize the reliability of the newly proposed analysis method, we conducted an experiment in which participants had to perform a simple motor task and data were analyzed with the LDA-based classifier as well as with the standard GLM analysis. *Main results.* The results of the simulation study show that the LDA-based method achieves higher classification accuracies than the GLM analysis, and that the LDA results are more uniform across different subjects and, in contrast to the accuracies achieved by the GLM analysis, have no significant correlations with any of the demographic characteristics. Findings from the real-data experiment are consistent with the results of the real-plus-simulation study, in that the GLM-analysis results show greater inter-subject variability than do the corresponding LDA results. *Significance*. The results obtained suggest that the outcome of GLM analysis is highly vulnerable to violations of theoretical assumptions, and that therefore a data-driven approach such as that provided by the proposed LDA-based method is to be favored.

Keywords: fNIRS, GLM, LDA, hemodynamic, HRF

Supplementary material for this article is available online

(Some figures may appear in colour only in the online journal)

## 1. Introduction

Functional near infrared spectroscopy (fNIRS) is a non-invasive neuroimaging technique based on the measurement of the optical absorption of cerebral blood. Thanks to the different absorption spectra of oxygenated and deoxygenated hemoglobin (HbO and HbR, respectively) in the near-infrared region of the electromagnetic spectrum (650–900 nm), it is possible to estimate the relative changes of oxygenation and blood perfusion in the human head, and therefore the level of oxygenation in the area of interest in response to a specific task [1, 2].

Although it is a relatively young technique, fNIRS is used in a wide range of fields, including (among many others) language studies, social interaction, and motor studies. Some features, such as portability, relative inexpensiveness, make fNIRS particularly advantageous over other functional techniques like functional magnetic resonance (fMRI) for certain populations of subjects, for example infants and children [3–5].

In a standard task-related fNIRS experiment, the subject usually performs several trials of one or more experimental conditions. After acquisition, data need to be pre-processed to remove cardiac and respiratory-related oscillations and possibly artifacts, and then the raw light-intensity data are converted into hemoglobin concentration changes through a modified Beer–Lambert law. To assess if a task induced a significant increase in the local neuronal activity, typically a general linear model (GLM) is employed to model the hemoglobin data $\mathbf{Y}$ (HbO, HbR, or Hb total) as $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon$, where $\mathbf{X}$ is the design matrix obtained by convolving the stimulus design with the expected hemodynamic response [6], $\boldsymbol{\beta}$ are the regressors representing the effect of each condition on the responses, and $\varepsilon$ is the measurement error [7].

An issue that has received substantial attention is that valid estimation of $\boldsymbol{\beta}$ requires that $\varepsilon$ have zero mean and be spherical (i.e. it must be 'white noise') [8]; these assumptions usually are greatly violated by fNIRS data, due to physiological noise, temporal and spatial correlations in the measurement data, and presence of artifacts. For these reasons, the GLM method is susceptible to yielding high false discovery rates. One strategy to overcome the problem is to remove structured noise from the residual term by filtering the data with a whitening filter based on the autoregressive model of the data [8, 9]. In contrast, largely unaddressed is the issue of inter- and intra-subject variability of the hemodynamic response; if the time course of the 'expected' hemodynamic response used to generate $\mathbf{X}$ is not a good approximation to the one that actually underlies the data $\mathbf{Y}$, then a true condition-induced change in neural activity could remain undetected. This is an especially relevant concern when data from very young subjects, or from a particular clinical population that under certain circumstances show atypical hemodynamic responses [10].

A strategy for addressing the variability in shape of the real hemodynamic response is incorporation of temporal and dispersion derivatives into the model [6]. The rationale for this procedure is that the additional regressors can capture the variance arising from small differences in the duration of the response and regress it out of the data. However, the method is time consuming and it complicates the interpretation of results, especially in group-level analyses [11–13].

As an alternative to the model-based approach, we propose to use a multivariate classifier based on linear discriminant analysis (LDA) [14] to distinguish two classes of NIRS signals that we will call '*active*' and '*not active*'. LDA has several features (low computational requirements, good performances, easy to use) that make it suitable for brain–computer interface (BCI) applications, the field where at present it is most frequently used [15]. Here we want to assess if its characteristics make it also a convenient tool for offline statistical analysis in quest of interpreting hemodynamic patterns with respect to the experimental conditions that elicited them.

The advantage of using a classifier for this purpose is that no assumptions on the structure of the noise are necessary, and that no prior knowledge of the shape of the expected hemodynamic response is needed. Furthermore, while GLM is a univariate approach to data analysis, in that time series of HbO, HbR or HbTot are considered independently of each other, in LDA information regarding the simultaneous variations of two or more hemoglobin components can be combined in a multivariate strategy. In fact, the use of combined features from HbO and HbR has been already reported to achieve higher accuracy than the use of separate features [16]. In this way, the analysis would yield a single metric for 'activation' for each channel, and this would be easier to test than separately testing $\beta$ coefficient from HbO and HbR, especially at group-level. In addition, comparisons between the results yielded by the (data-driven) classifier and (model-based) GLM may be informative in the sense that the classifier might identify unpredictable effects that elude the model-based analysis. For example, in a case where GLM analysis reports a channel as '*not active*', the availability of LDA results could facilitate the process of deciding whether activation truly was absent (i.e. because LDA also classified the channel as '*not active*') or if the hemodynamic model used for GLM was not optimal (i.e. because LDA classified it as '*active*').

The present work comprises two steps: the first is a comparison of the proposed LDA-based method with canonical GLM analysis. In order to do this, an extensive volume of simulated data is used to characterize the two algorithms in terms of receiver operator curves (ROC) when no systemic oscillation is present (i.e. simulated hemodynamic responses were added to simulated resting-state data) or when a considerable amount of systemic oscillation is present (i.e. simulated hemodynamic responses were added to experimental resting-state data); the real-data

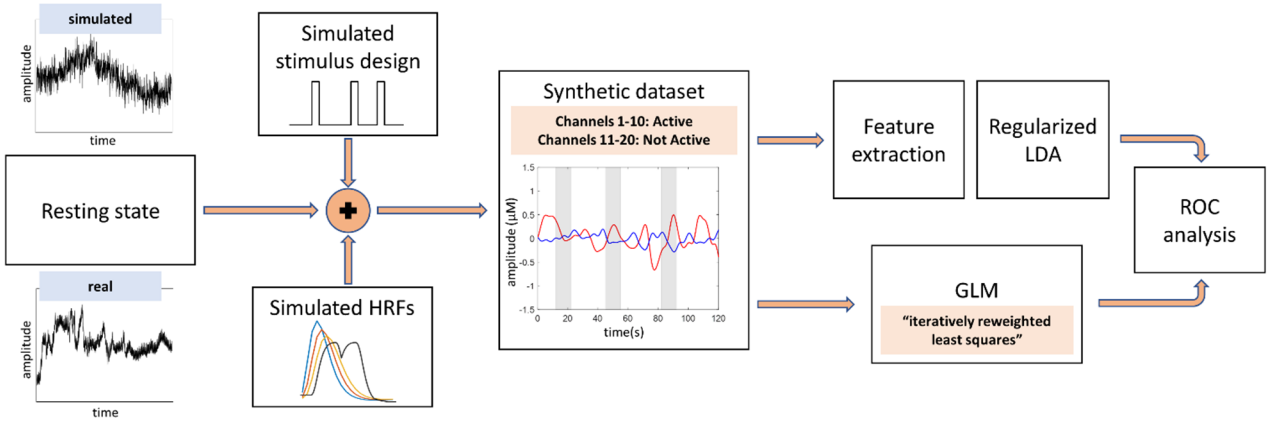



**Figure 1.** In each iteration, data are simulated, based on either synthetic or real resting-state data; HbO and HbR (red and blue time traces in the 'Synthetic dataset' panel, respectively) were analyzed with GLM or LDA, and ROC analysis was performed to compare the classification accuracies. The simulated HRFs vary in shape and size, and 30% of them are characterized by a 'double bump' as a simplified model of stimulus-locked Mayer waves.

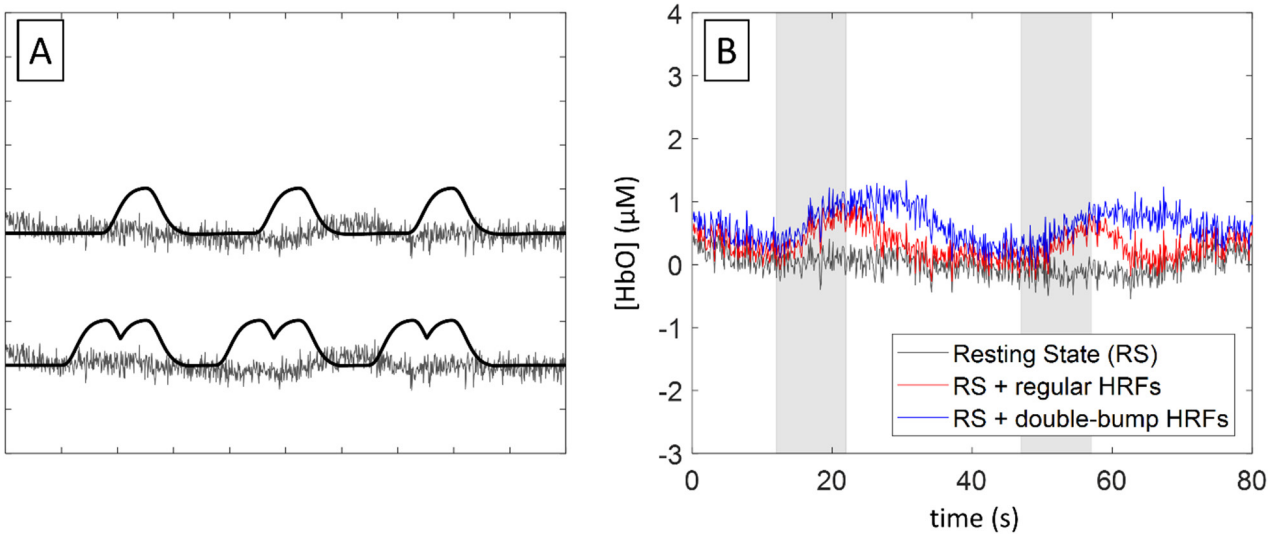



**Figure 2.** (A) Example of how simulated HRFs are created (black line) and added to a real resting-state time trace (dark grey line). The top trace is a simple HRF while the bottom trace contains a double bump. (B) The red time trace represents the HbO signal before the hemodynamic activations are added; the grey and blue ones are, respectively, the time traces after a simple HRF or a double-bump HRF have been added.

results also were used to characterize the impact of inter-subject variability on the outcomes of the classification analyses. Second, the two algorithms were used to analyze and classify the task-induced activations in a set of experimental data.

## 2. Methods

In order to compare the performances of the LDA-based and GLM-based methods under controlled conditions, sensitivity and specificity were quantified by recovering a known synthetic hemodynamic response added to either synthetic or real resting-state data. This approach has been used in several reports [8, 9, 17, 18] and its particularly suited for studies that make use of ROC analysis, because it permits an exact quantification of true and false discovery rates. In a second step, the two analysis pipelines were applied to real experimental data and channel-wise statistical assessments of each subject were compared.

Figure 1 reports a summary description of the whole procedure followed in this work. Figure 2 shows how known hemodynamic responses were added over resting state time traces.

### *2.1. Theoretical formulation*

*2.1.1. Generation of the synthetic dataset.* 5000 datasets of NIRS data were iteratively simulated by combining temporally correlated ('colored') noise and synthetic hemodynamic response functions (HRFs).

Baseline noise was produced by first generating white noise, then imposing temporal correlation on it by employing an autoregressive model of order 30, via tools in the fNIRS toolbox [8].

Each dataset contained 20 channels, and synthetic HRFs were added to the resting state for half of them (i.e. Channels 1–10). Channels that include synthetic HRFs in their HbO and HbR data called '*active*' and the others are '*not active*'. For

the '*active*' channels, 'Start' and 'Stop' markers were created according to an experimental paradigm with three episodes, each of 10 s duration. The position of the 'Start' markers was randomized, with the constraint that successive ones were separated in time by at least 35 s. The total duration of each time series was 4 min, at a sampling frequency of 7.81 Hz.

To model the inter- and intra-subject variability of real hemodynamic responses, synthetic evoked HRFs had variable size and shape across subjects and channels. While each HRF had the mathematical form of a canonical HRF [6], their peak amplitudes ranged from 0.01 to 0.1 $\mu$M [18], while, based on experience and existing literature on the variability of the hemodynamic response [19], the onset-to-peak times ranged from 2 to 8 s and the onset-to-undershoot times ranged from 14 to 18 s.

Positive-going synthetic HRFs (see figure 1(A)) were added to the resting-state data for the HbO time series. The synthetic HRFs added to the corresponding HbR resting-state data had the same form as those for HbO, but were 50% reduced in magnitude and reversed in algebraic sign (i.e. they were negative-going). In addition, for 30% of the '*active*' channels in each dataset the synthetic HRF included a 'double bump' (see figure 1(A)), as an elementary model of systemic hemodynamic activity time-locked with the experimental condition. An example of this sort of additional activity that frequently is present in experimental data is so-called 'Mayer waves', which are systemic oscillations originating in the superficial tissue layers [20], and which occur, more or less prominently, at ~0.1 Hz frequencies. Such oscillations are particularly difficult to treat in a GLM-analysis framework, owing to their extensive spectral overlap with typical event-related activity (i.e. they cannot be eliminated via straightforward frequency filtering) [21].

The use of synthetic baseline noise has the clear benefit that a large volume of data can be created, and it allows us to benchmark our methodology against recent literature on the topic [8, 9]. However, synthetic data might not capture all the properties of real physiological data. Therefore we complemented the synthetic-data analysis by using experimental resting-state data as baseline noise. For this purpose, 15 young adults (mean age $\pm$ SD: 28.1 $\pm$ 4.0 years old; age range: 23–38; 11 women, four men) participated in the collection of 4 min of resting-state data. For a subset of the participants, this measurement was followed by the motor-task study that was used in a later stage of this analysis (see section 2.2.1 for descriptions of the experimental setup and data collection).

Experimental resting-state recordings were used as a source of real physiological and correlated data, and were employed in the same manner as described above for the synthetic resting-state data, namely by performing 5000 randomizations of the positions of the 'Start' markers and adding simulated HRFs of variable shapes and amplitudes to only Channels 1–10 (left hemisphere), and labeling those channels as '*active*'.

#### 2.1.2. Data analysis.

*2.1.2.1. Pre-processing.* Both simulated and real data underwent the same pre-processing steps. Hemoglobin concentration changes were calculated using the modified Beer–Lambert law (differential pathlength factor (DPF): 6, absorption coefficients ($\mu_a$, $cm^{-1}$ $M^{-1}$) for HbO: $\mu_a(760\,nm) = 1349$ and $\mu_a(850\,nm) = 2436$, for HbR: $\mu_a(760\,nm) = 3565$ and $\mu_a(850\,nm) = 1592$).

Data was bandpass filtered in the range [0.01–0.2] Hz, with a zero-phase distortion digital FIR filter designed and implemented, respectively, with the Matlab commands *firls* and *filtfilt*.

For the subsequent statistical analysis, filtered data was used for the LDA analysis, in accordance with most fNIRS-based BCI literature [22], while unfiltered data was used for the GLM computations because it has been reported that frequency filtering can produce biased estimates of the regressors [8]. In this way, both methods were used at their optimal settings.

*2.1.2.2. Analysis with GLM.* GLM was applied using the *autoregressive iteratively reweighted least squares* algorithm available in the fNIRS toolbox. This algorithm is reported to efficiently remove serial correlations from data, thereby achieving an acceptable false discovery rate [9]. HbO and HbR time traces were analyzed independently.

After the regressors $\boldsymbol{\beta}$ are estimated, the null hypothesis that there was no hemodynamic response (H0: $\beta = 0$) is tested by defining a contrast vector ($\mathbf{c}$) and calculating the channel-wise t-statistic via the formula [7]:

$$t = \frac{\mathbf{c}^T \beta}{\sqrt{\mathbf{c}^T \ \mathrm{cov}\,(\beta)\,\mathbf{c}}}. \qquad (1)$$

In this case, with only one experimental condition to be tested, the contrast vector would be [1 0], with the second column referring to the constant column added to the GLM design matrix. The p-values corresponding to the t-statistics from equation (1) were computed via two-tailed t-tests.

*2.1.2.3. Analysis with LDA.* For each HbO and HbR time series, trials were defined as the signal in the 15 s time interval following each 'Start' marker. Each trial was baseline-corrected by removing the mean value of the signal over the 3 s interval prior to stimulus onset. Channel-wise block averages were obtained by averaging across all trials within each channel.

Features were extracted from the channel-wise block averages (figure 3(A)). To do this, a 3 s wide window was moved through the block-average time series in 1 s steps and the mean value and mean slope (computed as the change in signal amplitude over the time window divided by its size in number of samples) were computed within each window, yielding a 30-features vector (2 features $\times$ 15 windows) for each of HbO and HbR. Each feature vector was normalized to zero mean value and unit variance. Then the HbO and HbR feature vectors were concatenated, resulting in a 60-features vector that was used for the classification (figure 3(B)).

Channels were classified as '*active*' or '*not active*' with regularized linear discriminant analysis, via tools available in the Berlin brain–computer interfacing (BBCI) toolbox [23, 24]. Ten repetitions of four-fold cross validation was performed: 20 trials (ten '*active*' channels and ten '*not active*' channels) were separated into four folds, with three folds used

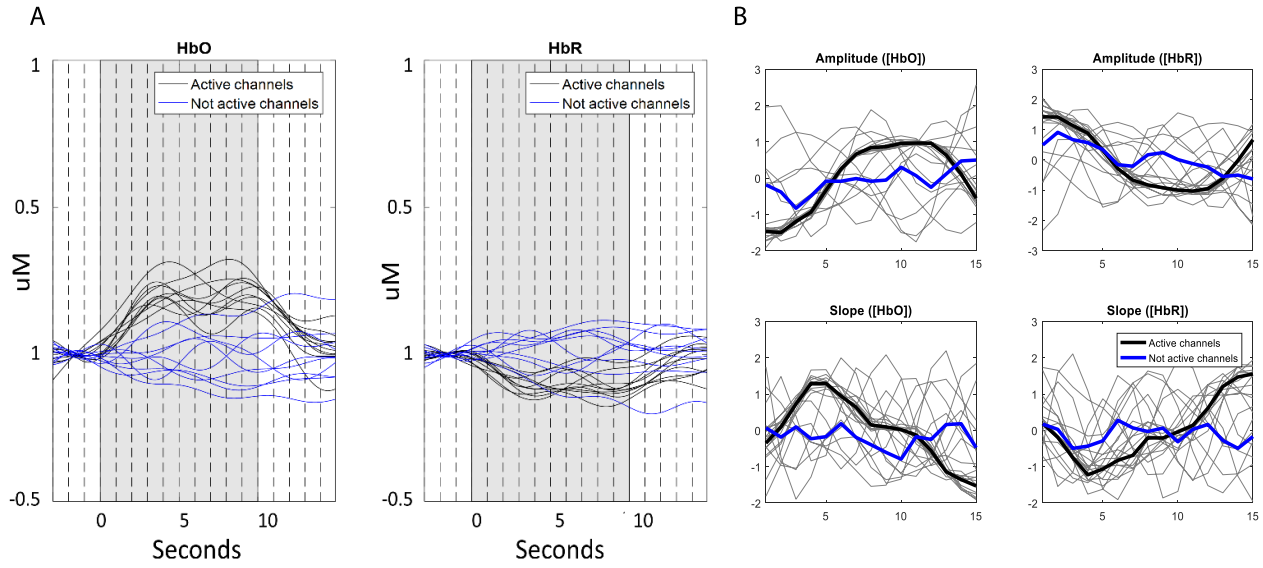



**Figure 3.** (A) Block averages of HbO (left) and HbR (right) signal used for feature extraction; dashed lines represent the 1 s steps used for the moving-window computation of amplitude and slope. (B) Features vectors are obtained from the block averages by computing mean and slope of the signal over a sliding window of 3 s duration with 1 s steps, resulting in a 30-features vectors that were then normalized and concatenated to produce the 60-features multivariate (HbO + HbR) classifier. Grey lines represent individual trials, black lines highlight the mean value of the feature vectors corresponding to '*active*' channels, and blue lines highlight the mean value of the feature vectors corresponding to the '*not active*' channels. Values of the *y* axis are normalized values.

for training and the remaining fold as the test dataset. The procedure was repeated ten times. Each feature vector $\boldsymbol{x} \in \mathbb{R}^N$ is assigned an output by the application of the formula of the separating hyperplane characterizing the LDA classifier [23]:

$$w^T\boldsymbol{x} + b = 0 \tag{2}$$

where **w** is the projection vector characterizing the classifier and $b$ is a bias term. The projection vector **w** is calculated based on the difference between the estimated mean values of the two classes and the common covariance matrix (for further details about binary linear classifiers, see [14, 25]). The bias term $b$ is chosen such that the separating threshold between the two classes is 0; therefore the classification function assigns each vector **x** a class label according to the algebraic sign of the output, sign($\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b$). In this implementation, the class '*not active*' was assigned to negative or zero outputs and class '*active*' was assigned to positive outputs.

*2.1.2.4. Evaluation of performance.* The performances of the GLM analysis and the LDA-based method were evaluated by computing receiver operating characteristic (ROC) curves. ROC curves for the GLM results were computed by varying the significance threshold for the *t*-test p-values, from 0 to 1 in increments of 0.001, and computing the corresponding false positive rate and true positive rate for each threshold. ROC curves for the LDA results were computed by comparing the distributions of output values of '*active*' and '*not active*' channels. We defined a significance threshold, varying from 0 to 1 in increments of 0.001, in the following manner: on the distribution of '*not activ*e' outputs, we defined a reference value as the percentile corresponding to the considered threshold. We defined as true negatives (TN) the samples of the '*not active*' distribution that were smaller than the reference value, false positive (FP) the samples of the '*not active*' distribution that were equal or greater than reference value, true positives (TP) the samples of the '*active*' distribution that were equal or greater than the reference value and false negative (FN) the samples of the '*active*' distribution that were smaller than the reference value. We repeated the procedure by sliding the reference value until 100% of the '*not active*' distribution was covered (i.e. significance threshold = 1). For example, by setting the threshold at 0.05, we computed the reference value on the distribution of '*not active*' outputs corresponding to its 5% percentile, and based on this reference value we computed TP, TN, FP, FN at p = 0.05.

For both GLM and LDA, classification accuracy was computed as the rate of correct classifications, (TP + TN)/(TP + TN + FP + FN), at p = 0.05. In order to investigate the impact of double-bump HRFs on classification accuracy, we conducted two separate analyses on the two subsets of data characterized by, respectively, only HRFs with no-double bumps and only HRFs coupled with double-bumps.

*2.1.2.5. Analysis of the correlation between subject demographics and classification accuracy.* Subject demographics such as gender, age and chronobiology have been reported to play a role in the cerebral metabolism [26–29], and therefore we tested for correlations between the individual-subject classification accuracies and each subject's characteristics. To do so, a linear mixed effects (LME) model was fitted in Matlab 2017, with a random intercept for each participant and fixed effects for age, gender, hair color (two levels: blond, brown), and time-of-day of the measurement (three levels: 10 AM–1 PM, 1 PM–3 PM, 3 PM–6 PM):

$$\text{Accuracy} \sim \text{Age} + \text{HairColor} + \text{Gender} + \text{TimeMeasurement} + (1|\text{Participant}).$$

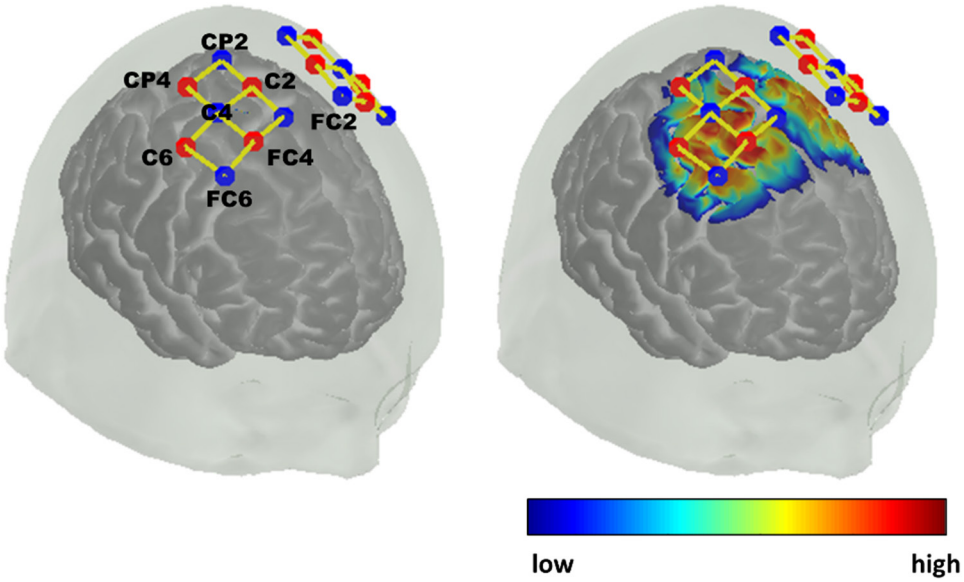



**Figure 4.** (A) Probe setup. The arrangement of optodes follows the 10–20 standard and the placement is analogous in the other hemisphere. Red dots indicate the sources, blue dots indicate the detectors, and yellow lines indicate the formed channels. (B) Sensitivity profile of the probe setup.

This analysis was carried out for LDA (HbO + HbR), GLM (HbO) and GLM (HbR) separately. Analysis of variance (ANOVA) was performed on each model to test the significance of the effects (error DF = 10 (15 observations minus five modelled effects)).

### 2.2. Application of the algorithms to experimental data

To provide a practical example of use of the proposed algorithm and compare it with the GLM analysis in the framework of a real experiment, a paradigm was chosen—finger tapping—that has a well-known effect on the motor cortex. In particular, it is expected to elicit a recognizable and significant response in the primary motor cortex (M1, Brodmann area 4, likely to underlie the C3/C4 positions of the EEG 10–20 system) and the premotor cortex (PMC, Brodmann area 6, likely to underlie the FC3/FC4 positions) [30].

*2.2.1. Experimental setup and data collection.* Seven healthy young adults (a subset of the 15 participants in the preceding study; mean ± SD age 26.0 ± 2.3 years, age range 23–30 years; five female, two male) participated in this study. The experiment consisted of 16 episodes of finger tapping (eight left, eight right, alternating), each of 10 s duration, with 20 s rest periods between successive episodes. Before the experiment began, the subject was required to sit quietly for the collection of 4 min of resting-state data.

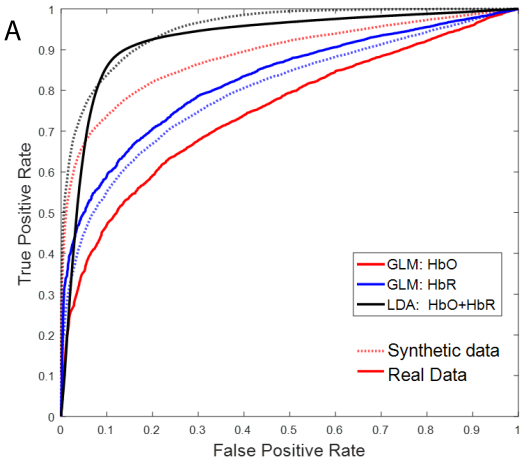
NIRS recordings were conducted with a NIRSport system (NIRx GmbH, Berlin, Germany), with sampling frequency 7.81 Hz, at wavelengths 760 nm and 850 nm, with eight sources and eight detectors. Sources and detectors were placed into a cap (EASYCAP, Herrsching, Germany), arranged according to the International 10–20 system. The source-detector distance was 2.5–3 cm, to form 20 channels evenly distributed between the hemispheres. A spatial sensitivity profile was calculated based on the Monte Carlo photon migration modeling available in the AtlasViewer software [31], to prove that the probe design was selective for the regions relevant to the finger tapping task (underlying the 10–20 positions FC3/FC4 and C3/C4). The Monte Carlo modeling was performed with $10^6$ photons. Figure 4 shows the probe arrangement and the resulting sensitivity profile. Additional details about the probe arrangement can be found in the supplementary material available at stacks.iop.org/JNE/15/045001/mmedia.

*2.2.2. Data analysis.* The data analysis aims at identifying which channels are significantly activated by the motor task and can therefore be labeled '*active*', as opposed to the '*not active*' channels that are not significantly activated by the task. For this reason, no distinction was made between left and right-hand finger tapping. The data was analyzed with the GLM analysis and the LDA-based method described in the previous section.

*2.2.2.1. Analysis with GLM.* For the GLM analysis, the stimulus times of the task were convolved with a canonical hemodynamic response function (peakTime = 6 s) to produce the single column ('Task' condition) of the design matrix. A GLM was applied using the *autoregressive iteratively reweighted least squares* algorithm available in the fNIRS toolbox [8].

*2.2.2.2. Analysis with LDA.* For the LDA analysis, amplitudes and slopes were computed for each episode of finger tapping. For the 'Rest' condition, an equal number (n = 16) of time intervals were produced by randomly sampling the initial 4 min of resting-state data of each measurement, and features were extracted. The sampling of 'Rest' trials and the classification Task versus Rest was iterated 2000 times for each channel and for each subject, to ensure robustness of the analysis. Ten repetitions of four-fold cross validation were conducted and each of the 32 trials (16 task and 16 rest) was assigned a classifier output via equation (2).

A




B

| Overall mean classification accuracy at p = 0.05 | | |
|---|---|---|
| | Synthetic resting state | Real resting state |
| LDA(HbO+HbR) | 85.25 % | 78.76% |
| GLM(HbO) | 79.08 % | 65.76% |
| GLM(HbR) | 67.75 % | 70.29% |

**Figure 5.** (A) ROC curves for GLM and LDA, using HbO, HbR and HbO + HbR features (only for LDA). The curves for the real resting-state data (solid lines) are obtained by averaging the individual curves across subjects, while dotted lines refer to the completely synthetic dataset. (B) The table reports the mean classification accuracies, over all iterations, of the three algorithms applied to synthetic and real resting-state data. The classification accuracy is computed from the ROC curves at the false positive rate of 0.05.

*2.2.3. Statistical analysis.* The results of the GLM analysis were statistically assessed by computing the channel-wise *t*-statistics (equation (1)) from the resulting $\beta$ values, then testing them via two-tailed *t*-tests. The outputs of the LDA analysis were divided into 'Task' and 'Rest', then averaged over folds and over repetitions, and tested by comparing the two distributions (Task and Rest). The rationale of this procedure is that, if the task elicited a hemodynamic response *and* the classifier had a good discrimination between 'Task' and 'Rest', then the distributions of the outputs should be well separated and the channel will be labeled as 'active'. If, on the contrary, the two distributions are not well separated, it means that for that channel the execution of the task did not elicit a response substantially different from the resting state, and the channel will be labeled as 'not active'. As explained in section 2.1.2, the class label '*active*' is assigned to positive outputs, while '*not active*' to the negative outputs. Therefore, the channel-wise p-value in this case was computed as the fraction of 'Rest' outputs equal or greater than the mean value of the distribution of the 'Task' outputs [32].

## 3. Results

### *3.1. Theoretical formulation*

*3.1.1. Performance of the algorithms: overall classification accuracies.* Our first goal was to theoretically compare the two algorithms in terms of overall classification accuracy, both on synthetic and on real data. The other important objective was to evaluate whether, with real data, the achieved results are consistent across subjects, and to evaluate the impact of inter-subject variability on the performance of each algorithm.

Figure 5(A) shows the ROC curves obtained using synthetic and real resting-state data. In both cases the LDA classifier based on HbO + HbR features outperforms GLM applied to either HbO or HbR, with results tabulated in figure 5(B). A difference between GLM results for synthetic and real resting-state data is also seen, in that GLM(HbO) is more accurate than GLM(HbR) in the former case, while GLM(HbR) is more accurate than GLM(HBO) in the latter. We speculate that this difference indicates that the synthetic data do not entirely represent the properties of the real physiological data. For example, it certainly does not reflect the frequency structure of the resting-state signal, or its spatial dependence across the different channel positions. In addition, the temporal correlation in the synthetic data was imposed by using an autoregressive model of fixed order (N = 30) [8], which does not account for the variability that can be found in real data from different subjects.

To further investigate the performances of the three methods, we computed the classification accuracies for each subject individually (figure 6(A)). The barplots indicate the classification accuracy as computed from the individual subjects' ROC curves at p = 0.05, and the red line indicates the mean accuracy over all subjects, respectively (mean ± SD) 78.76 ± 5.1% for LDA(HbO + HbR), 65.76 ± 10.2% for GLM(HbO) and 70.29 ± 8.9% for GLM(HbR), the standard deviation being computed across the 15 subjects. The individual errorbars represent the standard error of the mean across the 5000 repetitions. Finally figure 6(B) shows the classification accuracies computed on two separate sets of data: data for all the channels that did not have Mayer waves modeled (i.e. no 'double bumps' (figure 1(A))) and data for all the channels that did have them. In this case, for the data without Mayer waves we found that LDA achieves an accuracy of 79.1 ± 6%, GLM(HbO) 77.8 ± 9.3% and GLM(HbR) 82.4 ± 8.2%, while for data with Mayer waves, the accuracy decreases to 77.0 ± 11% for LDA, 62.4 ± 7.5% for GLM(HbO) and 64.8 ± 6.8% for GLM(HbR).

*3.1.2. Correlation between classification accuracies and individual measures.* The goal of this analysis was to quantitatively assess the impact of individual characteristics (hair color, gender, age), and of the measurement time of day, on the individual classification accuracy. Table 1 shows the results of the LME analysis. The model shows a significant correlation between Hair Color and individual accuracies for GLM(HbR), but not for any of the other fixed effects in the

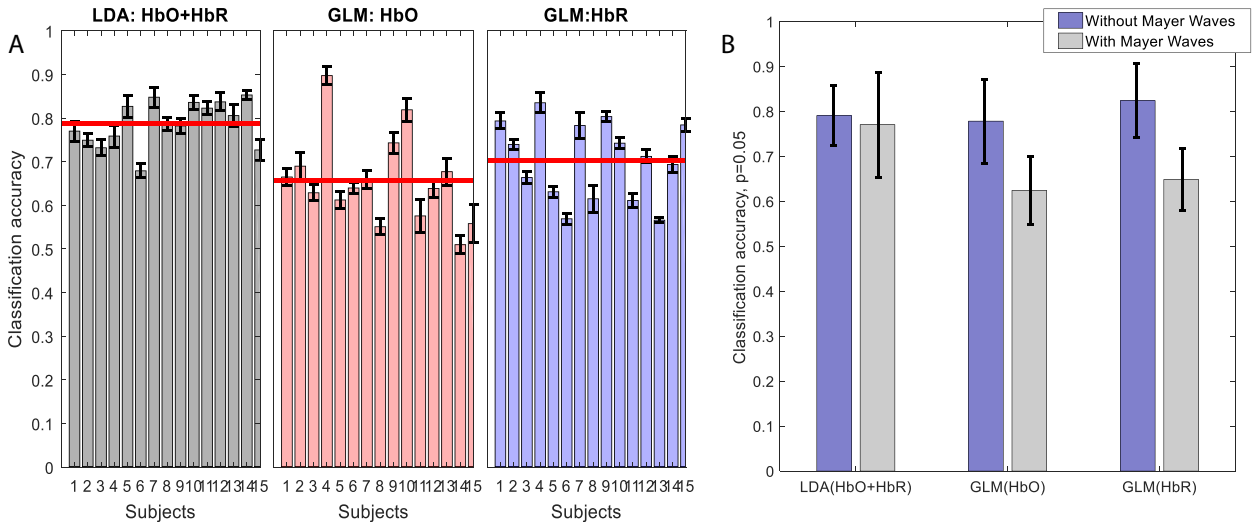



**Figure 6.** (A) Individual classification accuracies for the real-resting-state datasets, for LDA(HbO + HbR), GLM(HbO), and GLM(HbR) (left, middle, right). The red line indicates the mean accuracy reached by each algorithm over all the subjects. The errorbars represent the standard error of the mean for each individual subject, over all the iterations performed. The individual mean accuracies achieved by the LDA method is significantly higher than those achieved by the GLM(HbO) ($p = 0.0002$) and GLM(HbR) ($p = 0.01$), but no significant difference was found between GLM(HbO) and GLM(HbR) ($p = 0.24$, Repeated Measures ANOVA 1-way with Fixed Effect: 'Analysis Method'). Also, the individual standard errors of the mean yielded by the LDA are significantly lower than those achieved by the GLM(HbO) and GLM(HbR) ($p = 0.021$ and $p = 0.022$, respectively), but no difference was found between those yielded by GLM(HbO) and GLM(HbR) ($p = 0.97$). B) Classification accuracies computed on two subsets of the real-resting-state datasets, one completely free from Mayer-wave oscillations and the other one with all the HRFs tainted by double-bumps. For the LDA, there is no significant difference between the accuracies reached in presence and absence of Mayer waves (paired *t*-test, $p = 0.44$), while for the GLM the difference was statistically significant (GLM(HbO), $p < 0.001$, GLM(HbR), $p < 0.001$).

**Table 1.** Results of the linear model fitted to the individual classification accuracies, with fixed effects: age, hair color, gender and time of measurement.

| | LDA: HbO + HbR | | GLM: HbO | | GLM: HbR | |
|---|---|---|---|---|---|---|
| | $\beta$ | p-value | $\beta$ | p-value | $\beta$ | p-value |
| Age | 0.0042 | 0.1762 | 0.0010 | 0.8734 | 0.0068 | 0.1120 |
| Hair color | 0.0361 | 0.1875 | −0.0936 | 0.1020 | *−0.1408* | *0.0052* |
| Gender | 0.0038 | 0.1476 | −0.0077 | 0.1552 | −0.0026 | 0.4904 |
| Time of measurement | −0.0187 | 0.5109 | 0.0573 | 0.3308 | −0.0072 | 0.7540 |

model, and there are no significant correlations for either LDA or GLM(HbO). Figure 7 reports distributions of individual accuracies grouped by hair color. More plots of accuracy distributions grouped by the other effects used in the model can be found in the supplementary material.

### 3.2. Experimental results

Results from the experimental-data study are reported in figure 8, as t-statistic values for GLM(HbO) and GLM(HbR), and classifier outputs for LDA, for $p \leqslant 0.05$ threshold. White cells indicate that the channel did not reach statistical significance.

To better understand the source of the variability in the results, plots of the block-averaged trials for those channels, and topographic images of the channel-wise output values (LDA output values, and $\beta$s for GLM HbO/HbR), were produced for each subject. The images were produced via the visualization tool available in nirsLAB v2017.06 [33]. Plots for all subjects and corresponding output values are available in the supplementary material, while here only the plots for subject 1 and subject 2 are reported.

*3.2.1. Subject 1.* Subject 1 results are non-significant at $p = 0.05$ for every channel according to the GLM(HbO) analysis, and significant for channels 1, 4 and 16 in the GLM(HbR) analysis, while the great majority of channels are classified as '*active*' ($p < 0.05$) by the LDA classifier.

In the plots of table 2, we observe that the HbO time traces are greatly affected by the double-bump typical of the 0.1 Hz systemic oscillation, and also that the first peak after stimulus occurs earlier than the onset-to-peak time of the theoretical model (6 s). An enlarged depiction of block-average behavior for Channel 16 is shown in figure 9, together with the plots of the HRF model used in the GLM analysis and the block averages of the resting-state trials used by the LDA classifier. The resting-state HbO trace also includes a feature that is qualitatively similar to a hemodynamic response, but the task response is correctly discriminated from the resting-state time series nevertheless ($p < 0.001$).

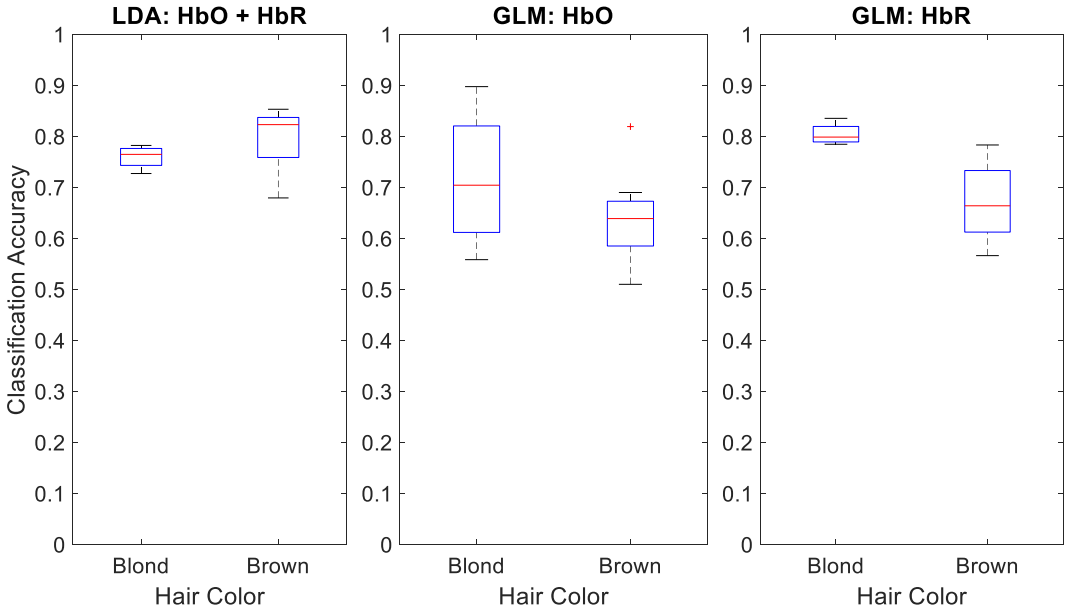



**Figure 7.** Distribution of individual classification accuracies within the two hair color (six blond and nine brown) subject classes. The accuracies reached by the GLM(HbR) are significantly higher for blond-haired subjects than brown-haired. More distributions for the other modeled effects can be found in the supplementary material. The central red marks represents the median values, the blue boxes extend from the 25th to the 75th percentiles, and the black whiskers extend to the most extreme data points not considered outliers (which are marked with red crosses).

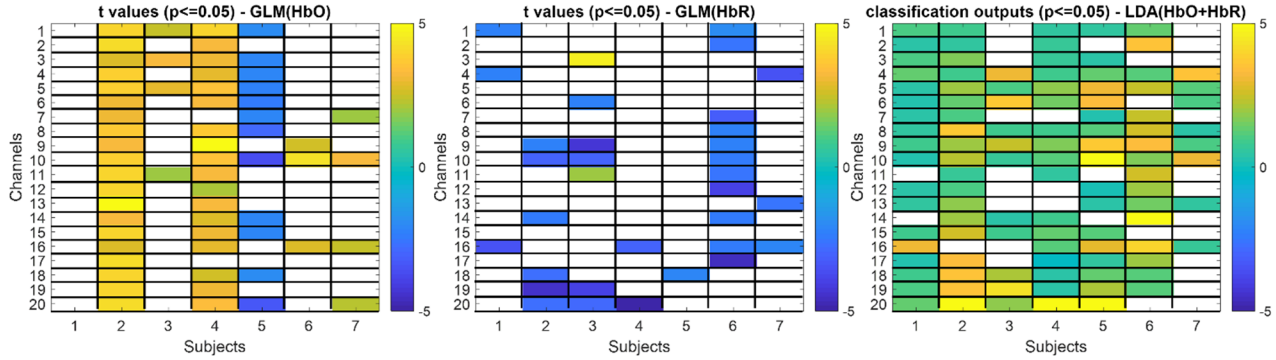


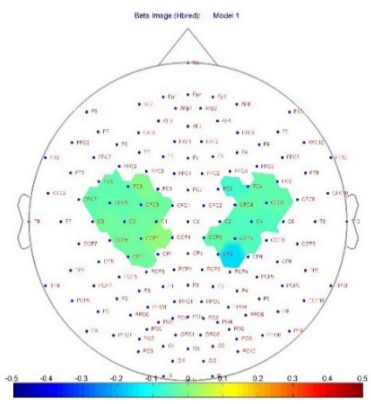
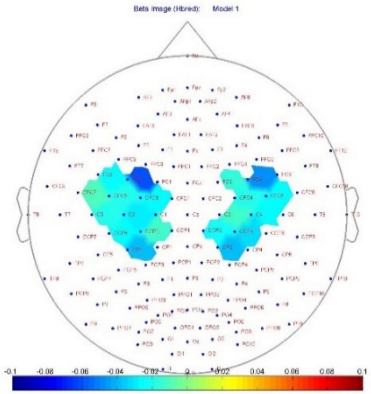
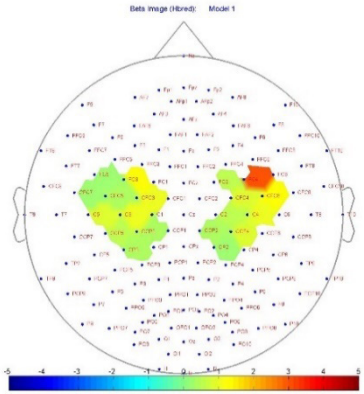
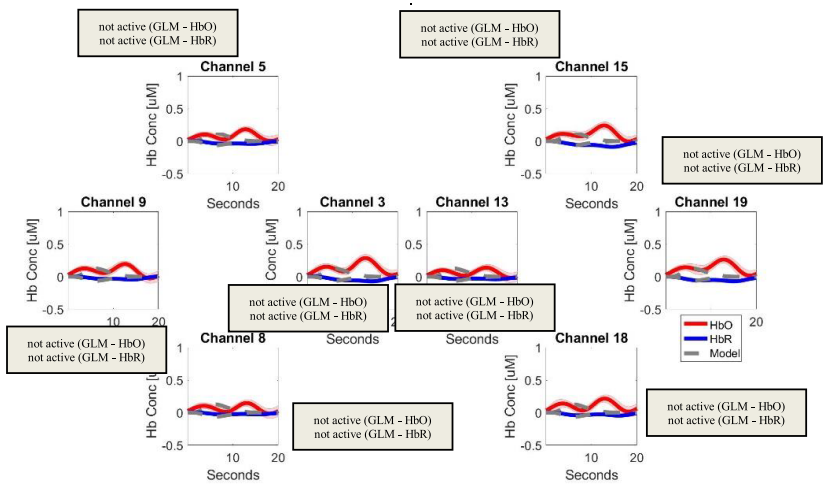
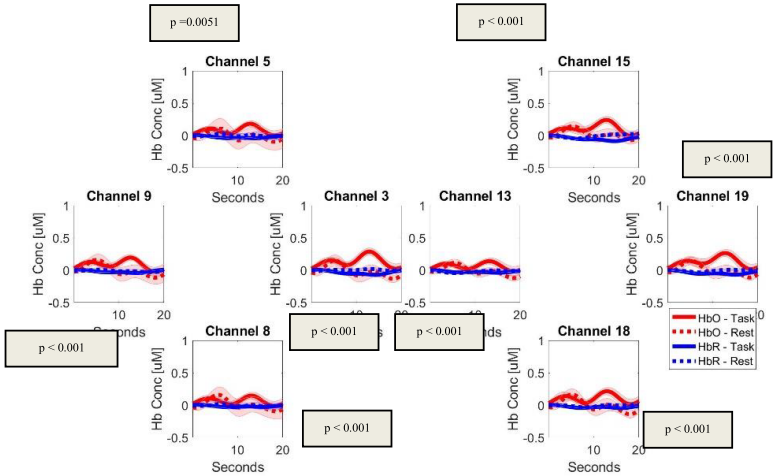

**Figure 8.** Classifier results for the finger-tapping experimental data, for the three different analyses. GLM t-statistic values and LDA classifier outputs (the latter derived from application of the separating hyperplane formula) are thresholded at $p \leqslant 0.05$. Blank cells indicate non-significant values (i.e. that the corresponding channel was classified as '*not active*'). The individual minimum value for statistical significance for the results of the LDA classifier varied across channels, ranging from 0.12 to 1.18. The numbers of channels classified as '*active*' by the three analyses are significantly different ($p = 0.01$, 1-way repeated measures ANOVA).

*3.2.2. Subject 2.* All channels of Subject 2 are classified as '*active*' ($p = 0.05$) in the GLM(HbO) analysis, while six of 20 are classified as '*active*' by the GLM(HbR) analysis, and 19 of 20 by the LDA classifier (table 3, figure 10). A depiction of Channel 5 is presented in figure 10. While the activation is correctly classified by the GLM(HbO) analysis, the same does not happen for the HbR. We speculate that the reason may be that the peak of the response is quite delayed (around 14 s post-stimulus) with respect to the 6 s assumed by the model. Nevertheless, the response is quite different from the resting state and the LDA picks up this difference, classifying the channel as '*active*'.

Corresponding results for all subjects can be found in the supplementary material. For each subject, a table is reported with:

– Topographic images of channel-wise GLM $\beta$ values and LDA classifier outputs. Large positive(negative) $\beta$ values indicate a good fit of the GLM model to the HbO(HbR) data, and a correspondingly better chance of that channel having a statistically significant hemodynamic response. LDA outputs are negative if the channel is classified as '*not active*' and positive is the channel is classified as '*active*'. Therefore, a large positive classifier output value indicates a good chance that the channel is labeled as '*active*'.
– Block averages of the signal in response to the stimulus (read and blue curves for HbO and HbR, respectively). The shaded error bars indicate the standard error computed over the experimental trials. The GLM plots are accompanied by the canonical basis function used by the model; the LDA plots are accompanied by the block averages of the resting-state trials. The block averages are shown only for the channels covering the motor cortex.

**Table 2.** Topographic images and block averages for all the analyses on Subject 1.

| | SUBJECT 1 | | |
|---|---|---|---|
| | GLM-HbO | GLM-HbR | LDA |
| $\beta$-value images for GLM; Classifier outputs for LDA |  |  |  |
| Block averages for the 'motor' channels: 3: s1-d3 5: s2-d3 8: s3-d3 9: s4-d3 13: s5-d7 15: s6-d7 18: s7-d7 19: s8-d7 |  | |  |

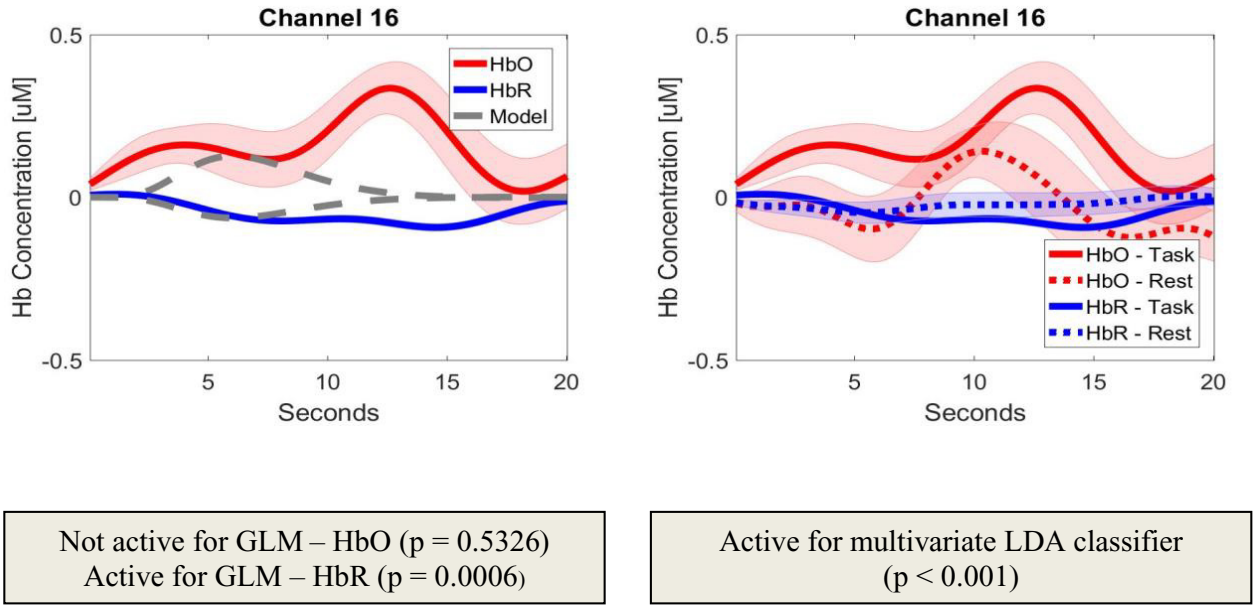



**Figure 9.** Block-average data for Subject 1, Channel 16. On the left the plot of the averaged signal is accompanied by the plot of the model used by the GLM analysis, namely a canonical HRF with peak time = 6 s. On the right, the same plot is accompanied by the plot of an example of average of resting state trials against which the task trials are classified.

## 4. Discussion

Statistical analysis of fNIRS data is often complicated by serial correlations, inter-subject variability of the hemodynamic response, and the presence of systemic oscillations and possibly motion artifacts. The study presented in this paper demonstrates that a data-driven approach (linear discriminant analysis, LDA) to data analysis is more robust than the most commonly employed model-based approach (general linear model, GLM) to many of these issues, and can therefore improve the detection of the hemodynamic activation.

Advantages of the proposed LDA approach are that no assumptions on the structure of the noise are necessary, and that no prior knowledge of the shape of the expected hemodynamic response is assumed. The LDA method compares data from different temporal segments of the same recording; namely, it compares, within the same subject, time intervals corresponding to the resting state and to execution of the task. Thus it constitutes a self-referencing approach, and in other fNIRS imaging contexts it has been shown that this data-analysis strategy can enhance detectability of effects that are small in comparison to other sources of intra- and inter-subject variance [34]. As such, LDA can generate information potentially superior, or at least complementary, to the information yielded by a model-based approach. For example, if LDA recognizes activation where the GLM does not, it could mean that the GLM model does not accurately represent the real HRF, and it might be worth investigating why this is so.

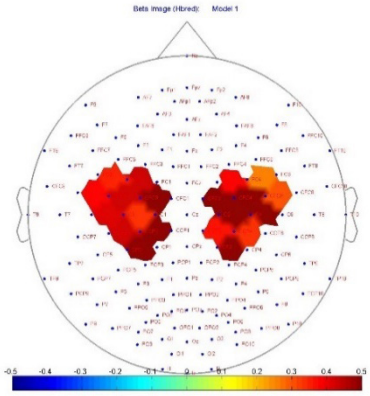
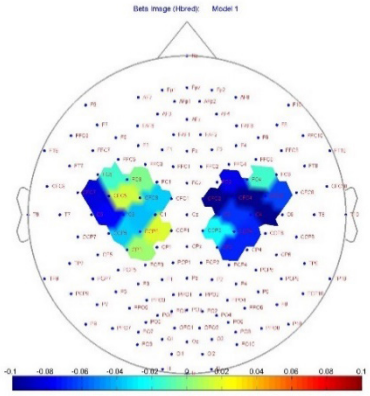
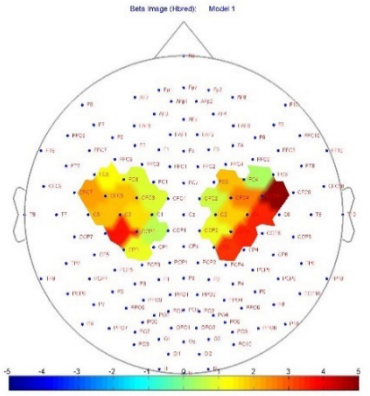
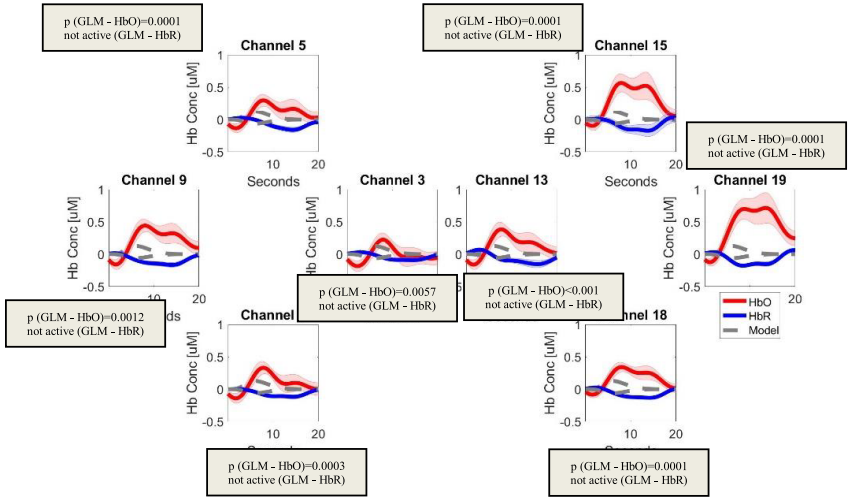
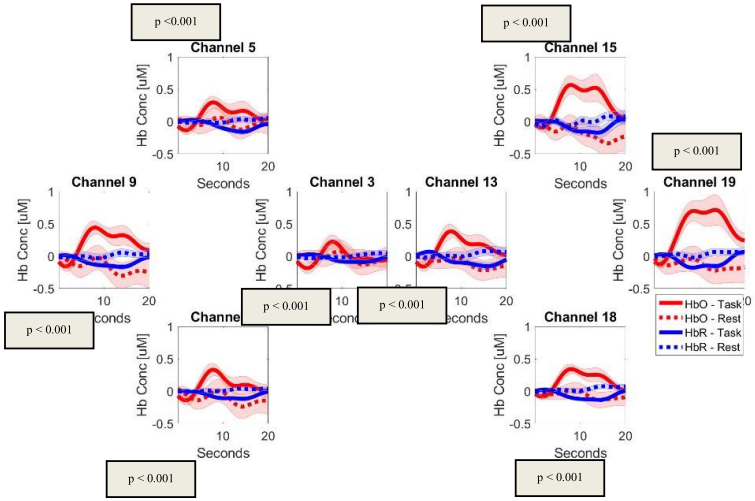
An additional strength of the multivariate LDA classifier proposed in this study is that it combines features from the simultaneous variations in HbO and HbR time series, while the GLM approach analyzes them independently. This results in the former yielding a univariate channel-wise metric for 'activation,' while the latter yields separate beta coefficients for HbO and HbR. Performing statistical tests on a single metric is highly desirable, especially for group-level studies.

To quantify and compare the classification performances of the three methods, we made use of both synthetic and real resting-state data. The use of synthetic data, for which the ground truth is known with certainty, also allowed us to benchmark our methodology against recent literature regarding GLM classification accuracy [9].

The multivariate LDA classifier yielded greater classification accuracy than GLM, for both the synthetic and real resting-state data (78.7% for LDA, 65.76 for GLM(HbO) and 70.3% for GLM (HbR), in the real resting-state data case (figure 5(A))). Moreover, we demonstrated that the LDA had less inter-subject variability, as illustrated in figure 6(A), where the standard deviation of individual results was 5.1% about the mean for LDA, as opposed to 10.2% for GLM(HbO) and 8.9% for GLM(HbR). In addition, the linear mixed model fit of individual-subject accuracies to predictors Age, Hair Color, Gender, and Time of Measurement revealed a significant effect only for Hair Color and only on the accuracy achieved with GLM(HbR) (blond-hair accuracy > brown-hair accuracy). The latter findings show that the observed differences between accuracies of the model-based and data-driven approaches is not simply accounted for by obvious (and easily absorbed into the classification model) demographic or physical characteristics of either the subject (e.g. gender) or the measurement (e.g. time of day).

The LDA results also are less sensitive than those for GLM to the 'double bumps' that were used to approximate Mayer waves synchronized with hemodynamic task responses (in the presence of double bumps, classification accuracy falls from 79.1 ± 6% to 77.0 ± 11% for LDA, from 77.8 ± 9.3% to 62.4 ± 7.5% for GLM(HbO), and from 82.4 ± 8.2% to 64.8 ± 6.8% for GLM(HbR) (figure 6(B))). These results

**Table 3.** Topographic images and block averages for all the analyses on Subject 2.

| | SUBJECT 2 | | |
|---|---|---|---|
| | GLM-HbO | GLM-HbR | LDA |
| $\beta$-value images for GLM; Classifier outputs for LDA |  |  |  |
| Block averages of the motor channels: 3: s1-d3 5: s2-d3 8: s3-d3 9: s4-d3 13: s5-d7 15: s6-d7 18: s7-d7 19: s8-d7 |  | |  |

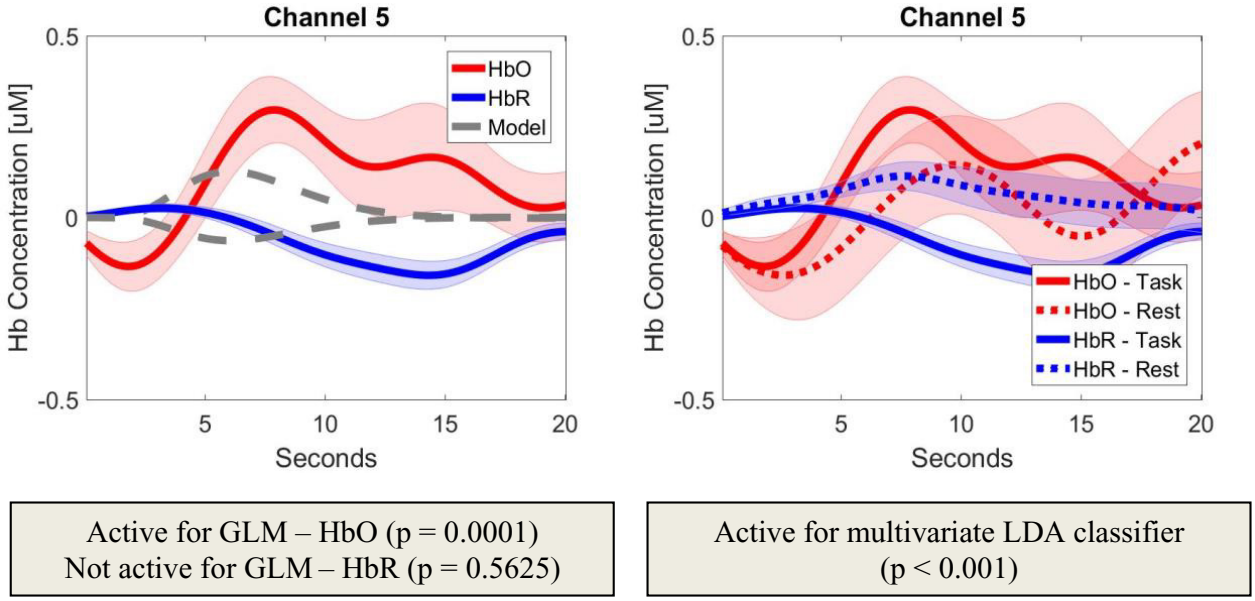



**Figure 10.** Block-average data for Subject 2, Channel 5.

confirm that the GLM, at least when used with a fixed basis function for all subjects, as was the case here, is less successful than LDA at picking up individual variability and atypical activation patterns, and is at risk of false negatives. These results suggest the possibility that the model used does not accurately represent the real hemodynamic responses and that therefore a different model would need to be designed. In this respect, the results of the one analysis can be used in support of interpreting the results of the other.

As a control study, additional simulations were performed to identify the classification performance of the LDA-based method when applied to data that did not actually contain any task-induced responses (either real or simulated) in the 'Task' time intervals. The classifier performed at chance level in these cases (results not shown), suggesting that the possibility of false-positive results in the analyses of simulated and real task-response data is not an important concern.

To further understand and characterize the performance of the two pipelines in a real application, we used the two methods to analyze data from a motor experiment. The optode array covered the motor cortex and its vicinity on both hemispheres. Eight out of twenty channels were placed over the scalp positions most likely to cover the motor area. In this scenario we could verify that the LDA-based classifier is less susceptible to than GLM to 0.1 Hz systemic oscillations. This is illustrated for the subject considered in table 2: due to the systemic oscillations, the block-average HbO and HbR traces in the eight channels over the motor cortex differ from the hemodynamic response modeled in the GLM computations. Consequently, the GLM recognizes none of these channels as '*active*'. Conversely, by contrasting 'Task' temporal segments with 'Rest' temporal segments, the LDA classifier finds significant differences in all of these channels, regardless of the presence of Mayer waves.

On the other hand, the LDA-based method generally classified more channels as '*active*' in response to the motor task than the canonical GLM analysis did. Because no established ground truth exists in the real data, this classification result warrants cautious interpretation. Especially for channels that extend beyond the center of the motor cortex, the activations found cannot be unambiguously attributed to neural activation caused by the motor task. However, the resting-state data classification results, and inspection of the experimental hemoglobin time traces, show that these classification results also are not easily dismissed as false positives.

In fact, as shown in recent literature [35], the fNIRS signal is not composed exclusively of cerebral task-evoked signal but also includes cerebral non-evoked signal ('cerebral resting state'), extracerebral task-evoked signal ('extracerebral confound'), and extracerebral non-evoked signal. Quantitative characterization of the three latter components is still an open research question [36], which is why they could not be modeled separately in our simulations. However, they are likely to be present in the motor experiment data and offer a plausible explanation for the classification results: when a difference is found between 'Task' and 'Rest', not only the cerebral response to the task, but also all the systemic hemodynamic changes provoked in the extracerebral compartment by the execution the task (e.g. changes in heart rate, blood pressure, respiration rate), is discriminated from the resting state. These changes involve the whole extracerebral layer, and therefore their effect extends beyond the probes that specifically illuminate the motor cortex [35]. To exclusively associate the found activations with cerebral recruitment, a step that would be necessary, but is beyond the scope of the present work, would be to remove from the data the physiological component measured in the extracerebral layers before the analysis, for example with multi-distance NIRS measurements [37].

Finally, it is worthy of note that the experimental data, in agreement with the results of the theoretical simulations, reveal great inter-subject variability in the comparative sensitivities of GLM(HbO) and GLM(HbR). That is, some subjects' hemodynamic patterns are better interpreted, and hemodynamic task responses more detectable, using HbO data, while others' are better explained using HbR. This is a manifestation of the highly subject-specific hemodynamic fingerprint that has been reported [9]. A classifier, such as the one proposed, that takes into account simultaneous variations of both hemoglobin components has the potential to overcome this limitation and offer a more flexible analysis that adapts to the individual's own hemodynamic characteristics. In this respect, the proposed approach would be of especial application for populations, such as young children, that exhibit 'atypical' patterns of hemodynamic responses, such as uncoupled HbO and HbR or inverted response direction [38, 39].

Nevertheless, in the current form the proposed approach only classifies 'activations' versus 'non-activations'. As a future development, a non-parametric framework can be formulated in order to test more complex hypotheses on the distributions of classifier outputs, such as the comparison of amplitudes of the responses induced by different conditions, within subjects or between groups of subjects.

## Acknowledgments

The authors would like to thank the Marie Skłodowska-Curie ITN 'PREDICTABLE' grant (Grant agreement no: 641858) 'Understanding and predicting developmental language abilities and disorders in multilingual Europe,' funded by the European Commission for providing financial support. The authors would also like to express their gratitude to Dr Christoph H Schmitz for his review and his helpful comments. The authors are also grateful to the anonymous reviewers for their valuable and constructive feedback.

## ORCID iDs

Jessica Gemignani https://orcid.org/0000-0002-7722-4489

## References

[1] Delpy D T and Cope M 1997 Quantification in tissue near-infrared spectroscopy *Phil. Trans. R. Soc.* B **352** 649–59
[2] Scholkmann F *et al* 2014 A review on continuous wave functional near-infrared spectroscopy and imaging instrumentation and methodology *Neuroimage* **85** 6–27
[3] Lloyd-Fox S, Blasi A and Elwell C E 2010 Illuminating the developing brain: the past, present and future of functional near infrared spectroscopy *Neurosci. Biobehav. Rev.* **34** 269–84
[4] Rossi S, Telkemeyer S, Wartenburger I and Obrig H 2012 Shedding light on words and sentences: near-infrared spectroscopy in language research *Brain Lang.* **121** 152–63
[5] Quaresima V and Ferrari M 2016 Medical near infrared spectroscopy: a prestigious history and a bright future *NIR News* **27** 10–3
[6] Penny W D, Friston K J, Ashburner J T, Kiebel S J and Nichols T E 2011 *Statistical Parametric Mapping: the Analysis of Functional Brain Images* (New York: Elsevier)
[7] Tak S and Ye J C 2014 Statistical analysis of fNIRS data: a comprehensive review *Neuroimage* **85** 72–91
[8] Huppert T J 2016 Commentary on the statistical properties of noise and its implication on general linear models in functional near-infrared spectroscopy *Neurophotonics* **3** 10401
[9] Barker J W *et al* 2013 Autoregressive model based algorithm for correcting motion and serially correlated errors in fNIRS *Biomed. Opt. Express* **4** 35–54
[10] Gervain J *et al* 2011 Near-infrared spectroscopy: a report from the McDonnell infant methodology consortium *Dev. Cogn. Neurosci.* **1** 22–46
[11] Calhoun V D, Stevens M C, Pearlson G D and Kiehl K A 2004 fMRI analysis with the general linear model: removal of latency-induced amplitude bias by incorporation of hemodynamic derivative terms *NeuroImage* **22** 252–7
[12] Liao C H, Worsley K J, Poline J B, Aston J A D, Duncan G H and Evans A C 2002 Estimating the delay of the fMRI response *NeuroImage* **16** 593–606
[13] Lindquist M A and Wager T D 2012 Validity and power in hemodynamic response modeling: a comparison study and a new approach *Hum. Brain Mapp.* **28** 764–84
[14] Blankertz B, Lemm S, Treder M, Haufe S and Müller K-R 2011 Single-trial analysis and classification of ERP components—a tutorial *NeuroImage* **56** 814–25
[15] Lotte F, Congedo M, Lécuyer A, Lamarche F and Arnaldi B 2007 A review of classification algorithms for EEG-based brain–computer interfaces *J. Neural Eng.* **4** R1
[16] Shin J, Müller K-R and Hwang H-J 2016 Near-infrared spectroscopy (NIRS)-based eyes-closed brain-computer interface (BCI) using prefrontal cortex activation due to mental arithmetic *Sci. Rep.* **6** 36203
[17] Gagnon L, Perdue K, Greve D N, Goldenholz D, Kaskhedikar G and Boas D A 2011 Improved recovery of the hemodynamic response in diffuse optical imaging using short optode separations and state-space modeling *NeuroImage* **56** 1362–71
[18] Barker J W 2014 Estimation of Cerebral Physiology and Hemodynamics via Near-Infrared Spectroscopy *Doctoral dissertation* University of Pittsburgh
[19] Aguirre G, Zarahn E and D'Esposito M 1998 The variability of human, BOLD hemodynamic responses *NeuroImage* **8** 360–9
[20] Julien C 2006 The enigma of Mayer waves: facts and models *Cardiovasc. Res.* **70** 12–21
[21] Yücel M A *et al* 2016 Mayer waves reduce the accuracy of estimated hemodynamic response functions in functional near-infrared spectroscopy *Biomed. Opt. Express* **7** 3078
[22] Naseer N and Hong K-S 2015 fNIRS-based brain-computer interfaces: a review *Front. Hum. Neurosci.* **9** 1–15
[23] Blankertz B *et al* 2016 The Berlin brain–computer interface: progress beyond communication and control *Front. Neurosci* **10**
[24] *BBCI Toolbox* (https://github.com/bbci/bbci_public) (Accessed: 03 November 2016)
[25] Duda R O, Hart P E and Stork D G 2012 *Pattern Classification* (New York: Wiley)
[26] Kameyama M, Fukuda M, Uehara T and Mikuni M 2004 Sex and age dependencies of cerebral blood volume changes during cognitive activation: a multichannel near-infrared spectroscopy study *NeuroImage* **22** 1715–21

[27] Yang H *et al* 2007 Gender difference in hemodynamic responses of prefrontal area to emotional stress by near-infrared spectroscopy *Behav. Brain Res.* **178** 172–6
[28] Herrmann M J, Walter A, Ehlis A C and Fallgatter A J 2006 Cerebral oxygenation changes in the prefrontal cortex: effects of age and gender *Neurobiol. Aging* **27** 888–94
[29] Anderson J A, Campbell K, Amer T, Grady C and Hasher L 2014 Timing is everything: age differences in the cognitive control network are modulated by time of day *Psychol. Aging* **29** 648–57
[30] Lancaster J L *et al* 2000 Automated Talairach Atlas labels for functional brain mapping *Hum. Brain Mapp.* **10** 120–31
[31] Aasted C M *et al* 2015 Anatomical guidance for functional near-infrared spectroscopy: atlasviewer tutorial *Neurophotonics* **2** 20801
[32] Cohen M X 2014 *Analyzing Neural Time Series Data: Theory and Practice* (Cambridge, MA: MIT Press)
[33] Xu Y, Graber H L and Barbour R L 2014 nirsLAB: a computing environment for fNIRS neuroimaging data analysis *Biomedical Optics 2014* p BM3A.1
[34] Graber H L *et al* 2015 Enhanced resting-state dynamics of the hemoglobin signal as a novel biomarker for detection of breast cancer *Med. Phys.* **42** 6406–24
[35] Tachtsidis I and Scholkmann F 2016 False positives and false negatives in functional near-infrared spectroscopy: issues, challenges, and the way forward *Neurophotonics* **3** 031405
[36] Caldwell M, Scholkmann F, Wolf U, Wolf M, Elwell C and Tachtsidis I 2016 Modelling confounding effects from extracerebral contamination and systemic factors on functional near-infrared spectroscopy *Neuroimage* **143** 91–105
[37] Brigadoi S and Cooper R J 2015 How short is short? Optimum source–detector distance for short-separation channels in functional near-infrared spectroscopy *Neurophotonics* **2** 025005
[38] Chen S, Ning P, Sakatani K, Zuo H, Lichty W and Zhao S 2002 Auditory-evoked cerebral oxygenation changes in hypoxic-ischemic encephalopathy of newborn infants monitored by near infrared spectroscopy *Early Hum. Dev.* **67** 113–21
[39] Sakatani K, Chen S, Lichty W, Zuo H and Wang Y P 1999 Cerebral blood oxygenation changes induced by auditory stimulation in newborn infants measured by near infrared spectroscopy *Early Hum. Dev.* **55** 229–36