Neuroergonomics and Cognitive Engineering, Vol. 102, 2023, 141–151
https://doi.org/10.54941/ahfe1003014

Investigating Feature Set Decisions for Mental State Decoding in Virtual Reality Based Learning Environments

Katharina Lingelbach[1,2], Daniel Diers[3], Michael Bui[1], and Mathias Vukelić[1]

[1]Fraunhofer Institute for Industrial Engineering IAO, 70569 Stuttgart, Germany
[2]Department of Psychology, Carl von Ossietzky University, 26129 Oldenburg, Germany
[3]Institute of Human Factors and Technology Management, University of Stuttgart, 70569 Stuttgart, Germany

ABSTRACT

Brain-Computer Interfaces (BCIs) combined with Virtual Reality (VR) enable the development of user-aware systems for individualized learning that monitor the learner's current mental states and adapt content to their individual skills and needs. We investigate feature set decisions extracted from functional near-infrared spectroscopy (fNIRS) signals and its use in conventional machine learning (ML)-based decoding of working memory. Eleven volunteers participated in a VR study using a visuo-spatial n-back paradigm with simultaneous fNIRS measurements. Single subject and overall decoding performance were compared for different feature sets including exploration of single feature contribution and their localisation within the prefrontal cortex. Our results prove that feature sets combining oxygenated (HbO) and deoxygenated hemoglobin (HbR) features using a sequential feature forward selection have the highest performance. More specifically, HbR peak-to-peak features from premotor regions and right and mid-dorsolateral prefrontal cortex contributed most to the decoding performance. Our results emphasise the need of analysing ML features in mental state decoding and aim to provide empirically supported decision recommendations to reach the next step towards future online decoding pipelines in real-world VR-based learning applications.

Keywords: Brain-computer interfaces (BCI), Functional near-infrared spectroscopy (FNIRS), Visuo-spatial working memory, Learning, Machine learning

INTRODUCTION

Awareness of a user's current mental state and customized adaptability of a system are highly important in the development of efficient and effective learning environments. To identify and continuously monitor individual mental states, Brain-Computer Interfaces (BCIs) measuring neurophysiological signals, e.g., with a functional near-infrared spectroscopy (fNIRS) can be used (Fairclough, 2009; Ayaz et al. 2012). fNIRS is a mobile optical brain imaging technique that records metabolic changes in the concentrations of local oxygenated (HbO) and deoxygenated hemoglobin (HbR). In recent years, it has become a popular brain-imaging technique for real-world applications

due to its good usability and mobility (Peck et al. 2014; Strait and Scheutz, 2014). Given the rapidly changing skill requirements, suitable training and learning environments are of great importance for industrial stakeholders to remain competitive, effective and ensure job satisfaction. In such scenarios, Virtual Reality (VR) has emerged as a revolutionary technology to provide an immersive, interactive, and engaging learning experience (Philippe et al. 2020). Especially when erroneous behaviour is associated with severe consequences or great resources, VR offers the opportunity to explore actions and visualizations of consequences in a safe environment and at affordable costs. In addition, it provides an easy way to personalize educational content, learning speed, and/or format to the individual (Philippe et al. 2020). A good fit between learning environment and the user's skills and needs is decisive to promote self-motivation and, consequently, learning performance (Ryan and Deci, 2000). By adequately integrating BCI-based mental state monitoring in an adaptive VR learning environment, an optimal fit between the user and system can be achieved (Lotte et al. 2012).

Related Work

In most real-world related learning applications, visuo-spatial working memory (WM), describing the ability to process, memorize, and update an object including its visual properties and current location (McAfoose and Baune, 2009), is of great importance. Previous fNIRS studies investigating brain areas associated with (visuo-spatial) WM highlighted the role of the prefrontal cortex (PFC), especially dorsolateral (dlPFC) and ventrolateral (vlPFC) parts (Ayaz et al. 2012; Llana et al. 2022). When recruiting the dlPFC and vlPFC during visuo-spatial WM, local HbO concentration increases and HbR concentration decreases which can be measured with fNIRS (Ferrari and Quaresima, 2012; Llana et al. 2022).

Recently, von Lühmann (2018) investigated different visuo-spatial WM load levels using a colour-based n-back task with fNIRS. The author reported higher discriminability between levels in the mean activation within lateral regions of the PFC. Another research group, investigating single trial WM load decoding with fNIRS, presented an n-back task with three levels in VR (Herff et al. 2014; Putze et al. 2019). In their first decoding approach (Herff et al. 2014), they used the slope of the local HbO and HbR concentrations from time windows of 25 sec and 8 channels located over the prefrontal cortex as feature set. They were able to discriminate 1 from 3-back trials with an average of 78% classification accuracy (Herff et al. 2014). In their second approach (Putze et al. 2019), the authors modified their decoding approach by using smaller time windows of 12 sec and combined the mean HbO and HbR concentration as well as the slope and coefficient of determination of a linear regression as features for each channel. For the classification of 1 and 3-back trials, they achieved averaged accuracies of 49% in a participant-wise classification and 66% when pooling the data of all participants together (Putze et al. 2019).

In single trial decoding, several approaches to extract informative features from HbO and HbR concentrations are proposed, e.g., statistical features

(e.g., average, peak, peak to peak, slope; see Herff et al. 2014; Putze et al. 2019) or coefficients of a general linear model (GLM; von Lühmann et al., 2020). Hence, it is crucial to get an understanding of the selected feature sets and how each feature contributes to the decoding (von Lühmann et al. 2021). This is especially relevant for complex environments, such as VR-supported learning scenarios, where the underlying cognitive processes and associated neuronal activation patterns are still the subject of research. Thus, our goal in the current study was to investigate different statistical feature sets extracted from fNIRS signals, their performance as well as spatial distribution of informative channels when decoding different level of WM load during a VR-based visuo-spatial n-back paradigm.

METHODS

Participants

11 volunteers (four female, right-handed, mean age of 23.73 ± 1.42, range = 21-26 years) participated in the study. One participant was excluded in the analysis due to a different optode montage. Prior to the study, they were checked for sufficient German language knowledge, intact colour vision, no self-reported drug habit, and mental, neurological, or cardiovascular disease using an online questionnaire. All participants received monetary compensation and signed an informed consent according to the recommendations of the Declaration of Helsinki. The study was approved by the ethics committee of the Medical Faculty of the University of Tuebingen, Germany (ID: 827/2020BO1).

Procedure

The experimental task was a colour-based visuo-spatial n-back paradigm in VR adapted from von Lühmann (2018) with a low (1-back) and high WM load condition (3-back) as well as an active baseline using the 0-back condition.

In VR, participants faced a grey wall with eight coloured buttons (red, magenta, blue, light blue, green, yellow, orange, dark grey) arranged evenly in a circle and a coloured target number in the centre of the circle (see Figure 1A). The number indicated the n-back level (i.e., 0, 1, or 3) as well as target colour (i.e., grey as a default for 0-back and any other of the remaining seven colours for 1 and 3-back). It remained the same over a whole block and was presented at the beginning of the block 6 sec before the start of the task. The colour of each button changed trial-wise within the block and participant had to perform the n-back in each trial of 6 sec using the target colour and n-back level. Hence, they compared each trial whether the position of their target colour was the same as n trials before (true n-back) or not (no n-back). In case of a true n-back, they had to press the target colour button with their right hand. In the other case (no n-back), they had to press the grey coloured default button. The experiment consisted of 12 rounds (see Figure 1C). Each round comprised of 4 blocks and 48 trials. It

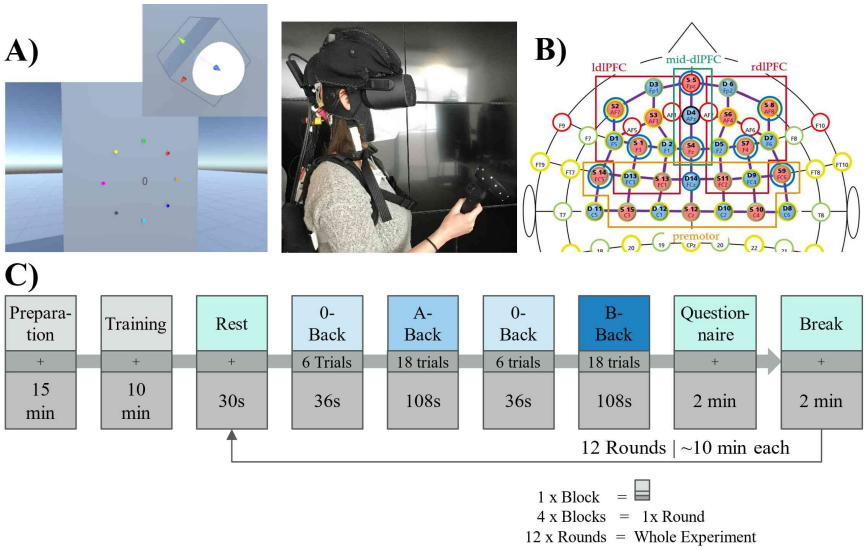

Figure 1: A) Task environment in VR (here: 0-back with grey as target colour). B) Optode montage with 15 sources (red circles), 14 detectors (blue circles), and 8 short channels (blue ring around red sources) covering the PFC. Regions-of-interests, that are the left and right dlPFC, mid-dlPFC and premotor cortex, are defined via coloured squares. C) Block design with 12 rounds á 4 blocks of the visuo-spatial n-back (see von Lühmann et al., 2018).

started with a 30 sec resting state recording followed by an alternating pattern of active baseline using a 0-back (6 trials with target colour grey) and randomly selected either 1- or 3-back block (18 trials). At the end of a round, participants, rated their perceived stress, frustration, and concentration using adapted subscales of the NASA TLX (Hart and Staveland, 1988).

Data Collection

A HP Reverb G1with a per-eye resolution of 2048 × 2048 pixels and refresh rate of 90 Hz was selected as VR hardware. It was preferred over others because of 1) less infra-red-light interference with the fNIRS compared to systems using first and second-generation Lighthouse base stations, 2) visible instead of infrared light positional tracking of controllers, and 3) a smaller head strap facilitating fNIRS optode montage over the PFC (see Figure 1B). A NIRx NIRSport2 system and the Aurora recording software[1] with a sampling rate of 5.8 Hz were used to acquire fNIRS signals.

We utilized three programs synchronized via Lab Streaming Layer (LSL; Kothe et al. 2019) to run the experiment: (1) a Unity program presenting the n-back paradigm and sending triggers to the other programs, (2) the Aurora recording software, and (3) a Python script saving experimental meta-information and behavioural data recorded by the controllers.

[1] https://nirx.net/software

Signal Processing

All analyses were performed in Python 3.8 using MNE-Python (Gramfort et al. 2013, version 1.2.), MNE-NIRS (Luke et al. 2021, version 0.4.), and scikit-learn (Pedregosa et al. 2011, version 1.0.2). The recorded fNIRS raw signals were first converted to optical density followed by channel pruning using the scalp-coupling index and a threshold of below 0.5 (Gramfort et al. 2013; Yücel et al. 2021). To account for baseline shifts and spike artifacts, a temporal derivative distribution repair was applied (Fishburn et al. 2019). In the next step, optical density was transformed into HbO and HbR concentration changes via the modified Beer-Lambert law (partial path-length factor: 6; Luke et al. 2021) and signals filtered with a 4^{th} order zero-phase infinite impulse response (IIR) Butterworth band-pass filter (cut-off: 0.05-0.7 Hz; transition bandwidth: 0.02-0.2 Hz). To decode WM load, we extracted short epochs of 4 sec duration. An epoch rejection threshold of a peak-to-peak amplitude above 80 μM was applied to account for further artefacts. To avoid an imbalanced distribution of classes in the decoding, epochs of the two conditions were equalized resulting in an average number of 208.8 $\pm$ 12.6 (range: 180 to 216) epochs per participant.

Decoding of Working Memory Load

From each epoch we extracted statistical features of HbO and HbR concentration per channel. Features included: peak, minimum, average, slope, peak-to-peak (peak2peak), and time-to-peak (time2peak). A Linear Discriminant Analysis (LDA), Support Vector Machine (SVM) and Gradient Boosting classifier (XGBoost, xgboost, version 1.7.2) were applied participant-wise and feature-set-wise. The analysis pipeline comprised of a nested 5-fold cross-validation (CV) with 20 repetitions, z-standardization of features, and a grid search to optimize hyperparameters (LDA: *solver*, SVM: *C,* xgBoost: *learning rate, max_depth, subsample, n_estimator)*. Decoding performance was statistically evaluated using the area under the receiver operating characteristic curve (AUC ROC) and non-parametric bootstrapping (5000 iterations) over CV folds of a participant. Average performance and its 2.5th and 97.5th confidence interval (CV) were obtained from the bootstrapped values and compared to an empirical chance level (i.e., bootstrapped mean performance of dummy classifiers trained participant-wise). Results were visualized in boxplots with notches within the boxes representing the upper and lower boundaries of the 95% CI of the mean (orange line), boxes comprising 50% of the distribution and whiskers indicating the 5^{th} and 95^{th} quantile of the distribution.

To investigate which cortical regions contributed to the decoding for each feature set, we extracted the coefficients of a typically chosen linear classifier for fNIRS decoding, that is the LDA. Coefficients of each feature set were averaged across participants and visualized on a 3D brain surface. This analysis was limited to the HbO statistical feature because it is assumed that a) effects in the HbO and HbR concentration changes are correlated (Ferrari and Quaresima, 2012; Yücel et al. 2021) and b) HbO effects are better to interpret due to their positive relationship with neuronal activation. In a last step, we

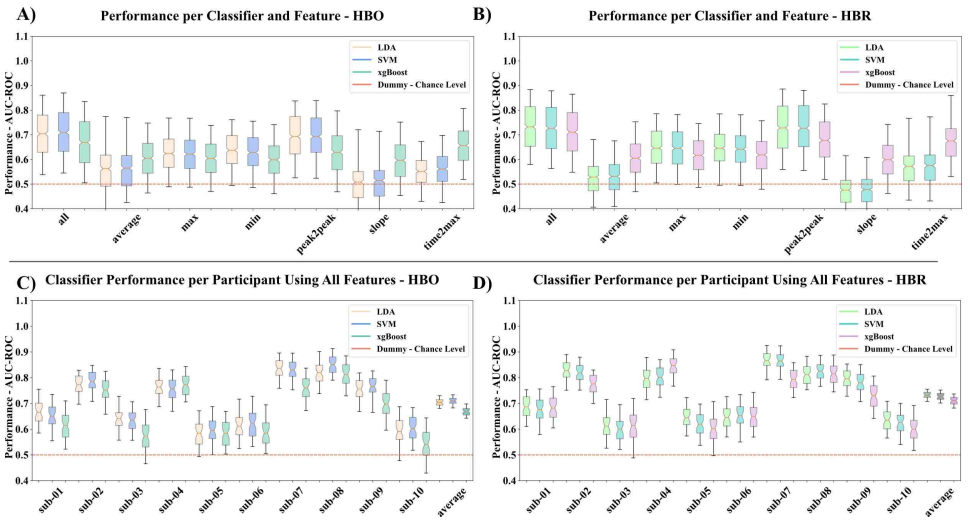

Figure 2: Decoding evaluation using AUC ROC. Average performance across participants for the A) HbO and B) HbR statistical feature sets. Single subject performance using all C) HbO and D) HbR feature sets.

explored optimal feature set size and combination by performing a sequential feature forward selection (mlxtend, version 0.21.0) with a maximum of $k = 20$ features using the combined statistical HbO and HbR feature sets.

RESULTS

Classification performance of all feature sets except of slope in combination with a LDA (HbO and HbR) and SVM (HBR), was significantly above chance level decoding (see Figure 2). The empirical chance level was estimated by a dummy classifier at 50.04% (95% CI [49.9; 50.17]). When comparing the classifiers, peak2peak difference of the HbO and HbR amplitude as well as the combined feature sets yielded in the high accuracies for all three classifiers. For the other feature sets, we observed minor differences between the linear classifiers LDA and SVM. There was a noticeable difference between SVM and LDA to xgBoost when using the HbO and HbR average, slope and time2max as feature sets. For the slope and time2max, xgBoost accuracy was the highest indicating that a set reduction via feature selection performed by the tree-based algorithm is beneficial. When using all HbO feature sets, we observed significant above chance level in all participants with average accuracy of 70.41% [69.92; 70.86] for the LDA (SVM: 70.41[70.49; 71.39]; xgBoost: 66.9 [66.42; 67.47]). For all HbR feature sets, average accuracy was 73.22 % [72.72; 73.66] for the LDA (SVM: 72.65 [72.14; 73.13]; xgBoost: 71.09 [70.53; 71.63]).

When examining the LDA coefficients (Figure 3), the average and max HbO feature sets revealed a similar spatial pattern of informative cortical regions. Higher amplitudes in left frontopolar, dorsolateral, premotor regions as well as in regions of the left motor cortex related to right hand and

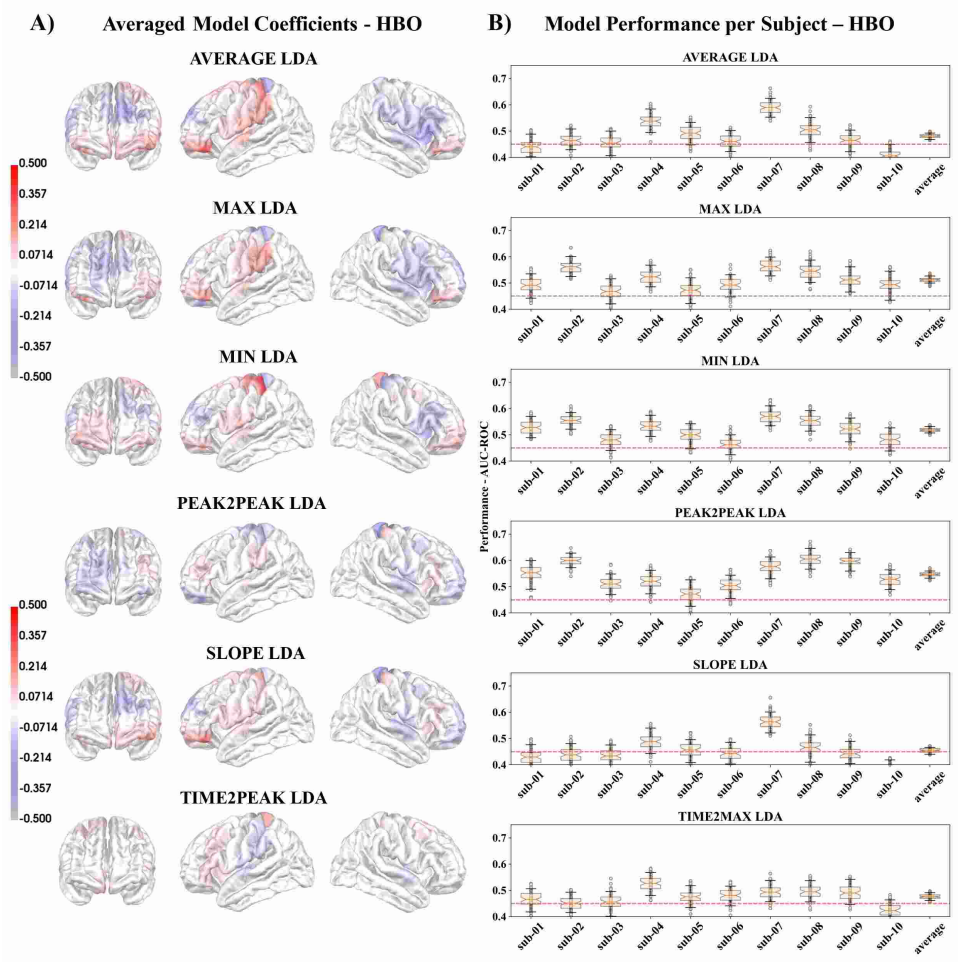

Figure 3: A) Projections of the LDA coefficients per feature set on 3D brain images. Class 0: Low WM Load; Class 1: High WM Load. B) Single subject performance per feature set.

arm movements were associated with the high WM load class (1). Interestingly, coefficients of channels over the left motor region were particularly large when using the min HbO feature set. Epochs with high average and maximum HbO concentration as well as a small peak-to-peak difference in right dorsolateral regions were allocated to the low WM load class (0). For the slope and average feature set, we observed below chance level classification performance in seven (slope) and three (time2peak) out of ten subjects rendering the classifier coefficients as difficult to interpret and less meaningful.

The SFS approach using all HbO and HbR feature sets with a maximum of 20 features revealed that the maximal set size of k = 20 led to the highest mean classification accuracy of 85.49 % [84.82; 86.1] (Figure 4). Especially HbR peak-to-peak features from premotor regions as well as the right and mid dlPFC were selected (Figure 5), followed by HbO peak-to-peak features from premotor regions, average HbO from the left dlPFC and HbR time-to-peak and max features from premotor regions (Figure 5).

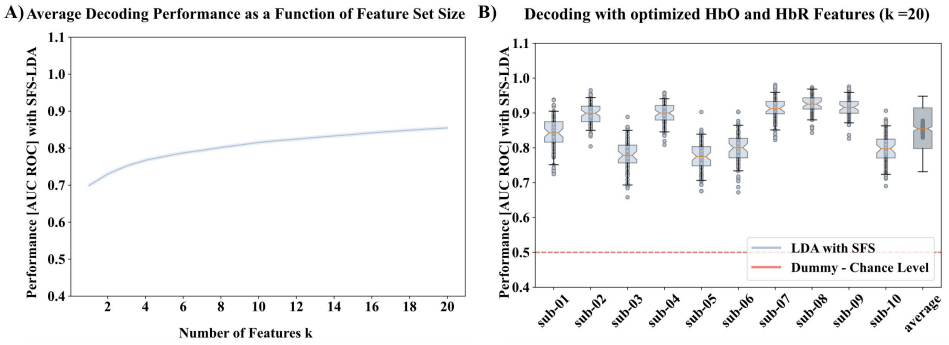

Figure 4: A) Relationship between LDA performance and feature set size with k features. B) Single subject classification performance using the SFS optimized feature set and k = 20.

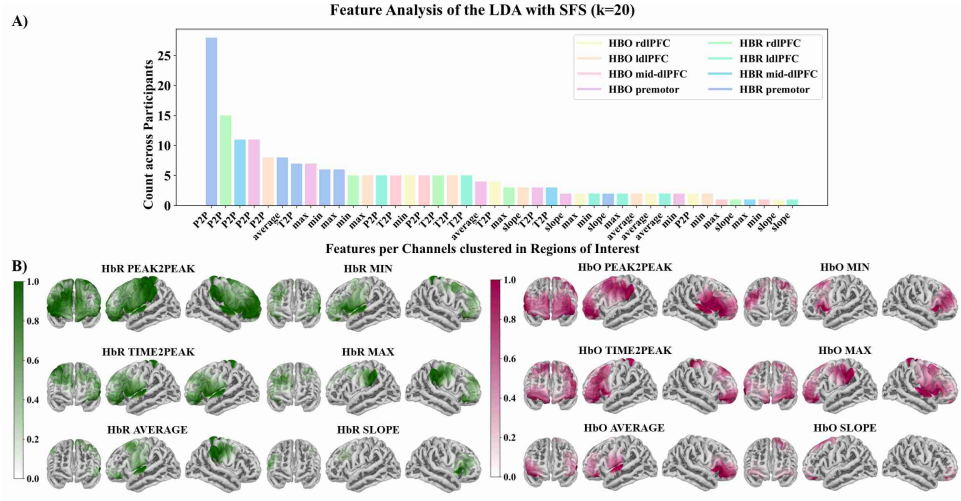

Figure 5: A) Count per selected features across participants divided into region of interest, as defined in Figure 1B), and statistical feature. B) 3D brain projection of counts per selected feature, i.e., channel per chromophore (HbR and HbO) and statistical feature set, across subjects.

DISCUSSION

For the development of robust decoding, knowledge about relevant features underlying cognitive processes and their selection can profoundly affect performance for everyday BCI applications and, thus, need to be systematically investigated (von Lühmann, 2021). We contributed to this knowledge by examining effects of HbO and HbR feature set selections in WM load decoding within a VR-based visuo-spatial n-back paradigm. We could obtain the highest average decoding performance for a combined HbO and HbR feature set with optimized feature selection via SFS (85.49 % [84.82; 86.1]; Figure 4B). The performance even exceeded previously reported WM decoding results (Herff et al. 2014; Putze et al. 2019). Single statistical feature sets, e.g., choosing only the slope per channel, yielded in rather low and below-chance level

decoding performance. Interestingly, HbR-related features contributed the most to our SFS-based decoding (Figure 5A). These chromophore-dependent differences in channel contribution need to be further investigated to identify potential effects of systemic physiological artifacts or non-stationarities (Unni et al. 2017). We did not detect a systematic contribution of any particular region or hemisphere to the decoding performance. Informative channels were rather spatially distributed throughout the PFC. We observed rather strong contribution from channels positioned over the premotor cortex that can be explained by the nature of chosen task requiring a motoric response. Further analyses are needed to systematically study the relation between task-related motoric behaviour and the HbO and HbR concentrations in channels over the premotor cortex as well as their contribution to the WM load decoding. Motoric differences between conditions should be excluded to guarantee sound WM load decoding. Our dataset allows to relate the premotor cortex activation to the time point of first button press as well as head acceleration within each trial. However, acceleration of the controllers during the experiment was not continuously recorded which we would highly recommend for future studies. Beside the feature set, further parameters such as window length strongly influence classification. Contrary to previous studies (Herff et al. 2014; Putze et al. 2019), we selected a short time window for each trial similar to non-overlapping windows within a block. Although rather larger time windows are suggested for fNIRS decoding (Putze et al. 2019), we chose rather shorter window size to avoid possible signal distortion due to motoric behaviour at the end of each trial and to obtain many data samples per class. Rationales for our decision were that, even in shorter windows, activation patterns of the induced WM load should be established after n-trials without any recovery to the baseline between trials of the same block so that statistical features should significantly differ between low and high load. However, we plan to analyse and compare block-wise decoding performance. To proceed towards future everyday world BCI applications, to extend the decoding to more than two WM load levels. One major challenge when using multiple levels or even a continuous scale is to acquire enough data in each class or level. Research on data augmentation is promising to address this challenge (Eastmond et al. 2022; Rommel et al. 2022).

In summary, insights of our study highlight the importance of feature set exploration for the development of everyday world BCI applications in neuro-adaptive VR learning environments.

ACKNOWLEDGMENT AND FUNDING

The authors would like to thank NIRx for providing the fNIRS hardware and technical support during data collection. The research was supported by the Ministry of Economic Affairs, Labour, and Tourism Baden-Wuerttemberg within the project ≫KI-Fortschrittszentrum Lernende Systeme und Kognitive Robotik≪ and Federal Ministry of Economic Affairs and Climate Action [BMWi, NIRcademy, ZF4509902BA9].

REFERENCES

Ayaz, H. et al. (2012) 'Optical brain monitoring for operator training and mental workload assessment', NeuroImage, 59(1), pp. 36–47.

Eastmond, C. et al. (2022) 'Deep learning in fNIRS: a review', Neurophotonics, 9(4), p.41411.

Fairclough, S. H. (2009) "Fundamentals of physiological computing", Interacting with Computers, 21(1-2), pp. 133–145.

Ferrari, M. and Quaresima, V. (2012) 'A brief review on the history of human functional near-infrared spectroscopy (fNIRS) development and fields of application', NeuroImage, 63(2), pp. 921–935.

Fishburn, F. A. et al. (2019) 'Temporal derivative distribution repair (TDDR): a motion correction method for fNIRS', NeuroImage, 184, pp. 171–179.

Gramfort, A. et al. (2013) 'MEG and EEG Data Analysis with MNE-Python', Frontiers in neuroscience, 7(267), pp. 1–13.

Hart, S. G. and Staveland, L. E. (1988) 'Development of NASA-TLX: Results of Empirical and Theoretical Research', in Hancock, P. A. and Meshkati, N. (eds.) Human mental workload. (Advances in Psychology, 52). Amsterdam: North-Holland, pp. 139–183.

Herff, C. et al. (2013) 'Mental workload during n-back task-quantified in the prefrontal cortex using fNIRS', Frontiers in Human Neuroscience, 7, p. 935.

Kothe, C. et al. (2019) Labstreaminglayer. Available at: https://labstreaminglayer.readthedocs.io/index.html (Accessed: 04.01.2023).

Llana, T. et al. (2022) 'Functional near-infrared spectroscopy in the neuropsychological assessment of spatial memory: A systematic review', Acta Psychologica, 224, p.103525.

Lotte, F. et al. (2012) 'Combining BCI with Virtual Reality: Towards New Applications and Improved BCI', in Towards Practical Brain-Computer Interfaces: Springer, Berlin, Heidelberg, pp. 197–220.

Luke, R. et al. (2021) 'Analysis methods for measuring passive auditory fNIRS responses generated by a block-design paradigm', Neurophotonics, 8(2), p.25008.

McAfoose, J. and Baune, B. T. (2009) 'Exploring visual-spatial working memory: a critical review of concepts and models', Neuropsychology Review, 19(1), pp. 130–142.

Peck, E. M. et al. (2014) 'Using fNIRS to Measure Mental Workload in the Real World', in Fairclough, S. H. and Gilleade, K. (eds.) Advances in Physiological Computing. (Human–Computer Interaction Series). London: Springer London, pp. 117–139.

Pedregosa, F. et al. (2011) 'Scikit-learn: Machine learning in Python', the Journal of machine Learning research, 12, pp. 2825–2830.

Philippe, S. et al. (2020) 'Multimodal teaching, learning and training in virtual reality: a review and case study', Virtual Reality & Intelligent Hardware, 2(5), pp. 421–442.

Rommel, C. et al. (2022) 'Data augmentation for learning predictive models on EEG: a systematic comparison', Journal of Neural Engineering, 19(6).

Putze, F. et al. (2019) 'Decoding Mental Workload in Virtual Environments: A fNIRS Study using an Immersive n-back Task', Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference, 2019, pp. 3103–3106.

Ryan, R. M. and Deci, E. L. (2000) 'Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being', The American Psychologist, 55(1), pp. 68–78.

Strait, M. and Scheutz, M. (2014) 'What we can and cannot (yet) do with functional near infrared spectroscopy', Frontiers in Neuroscience, 8, p. 117.
Unni, A. et al. (2017) 'Assessing the Driver's Current Level of Working Memory Load with High Density Functional Near-infrared Spectroscopy: A Realistic Driving Simulator Study', Frontiers in Human Neuroscience, 11, p. 167.
von Lühmann, A. (2018) Multimodal Instrumentation and Methods for Neurotechnology Out of the Lab. Doctoral dissertation. Technischen Universität Berlin. Available at: https://depositonce.tu-berlin.de/items/da8ba104-5dfd-4875-a0a2-871a5890e288/full.
von Lühmann, A. et al. (2021) 'Towards Neuroscience of the Everyday World (NEW) using functional Near-Infrared Spectroscopy', Current Opinion in Biomedical Engineering, 18. doi: 10.1016/j.cobme.2021.100272
von Lühmann, A. et al. (2020) 'Using the General Linear Model to Improve Performance in fNIRS Single Trial Analysis and Classification: A Perspective', Frontiers in Human Neuroscience, 14, p. 30.
Yücel, M. A. et al. (2021) 'Best practices for fNIRS publications', Neurophotonics, 8(1), p. 12101.