Benchmarking Framework for Machine Learning with fNIRS

Johann Benerradi[1], Jeremie Clos[1], Aleksandra Landowska[1], Michel F. Valstar[1], Max L. Wilson[1]

[1]School of Computer Science, University of Nottingham, Nottingham, United Kingdom

Functional near-infrared spectroscopy (fNIRS) (Jobsis 1977) is being increasingly used for cognitive neuroscience and brain-computer interfaces (BCIs) (Naseer and Hong 2015). It is often used with the aim of determining what type of task a subject is doing or assessing a task's level of intensity, and is becoming growingly popular for classifying types and levels of mental activity (Herff et al. 2014, Benerradi et al. 2019). For classification, machine learning is used widely, whether it be standard machine learning with models such as linear discriminant analysis (LDA) or support vector machines (SVM), or more recently deep learning with techniques ranging from standard artificial neural networks (ANNs) to convolutional neural networks (CNNs) and recurrent neural networks (RNNs) (Naseer et al. 2016, Trakoolwilaiwan et al. 2017, Yoo et al. 2018). Unlike other communities that have developed standardised and comparable approaches to machine learning of physiological measures, standards and good practices are still emerging for fNIRS. Consequently, in some cases, these techniques can appear to be effective, but researchers need to be aware of good practices and avoid common pitfalls that undermine the reliability of the end results (Lipton and Steinhardt 2019).

Our work aims to raise awareness of those issues and provide a framework to implement a robust methodology to produce meaningful and reproducible benchmarked results of machine learning classifications with fNIRS data. This framework comes in the form of an open-source Python library with an implementation of signal processing pipelines, feature extraction, and machine learning models applied to fNIRS data with rigorous validation. Some of the most common features used in the fNIRS literature can be extracted, including the mean, the standard deviation, and the slope of the linear regression (Naseer and Hong 2015). Models are validated with nested k-fold cross-validation (the value of k can be set), thus enabling systematic hyperparameter optimisation on the inner loop (Bengio 2012, Schmidt et al. 2020). It also calculates metrics like accuracy, whilst plotting training graphs and confusion matrices. This whole methodology has been systematically applied to 5 open-access datasets of fNIRS tasks from the literature, as shown in Table 1. Those tasks include n-back, word generation, mental arithmetic, and motor execution.

Based on this framework, we produced a benchmarking of the most popular models used in fNIRS BCIs, including LDA, SVM, ANN, CNN and a type of RNN called long short-term memory (LSTM) network (Naseer and Hong 2015). Early results show that the performance of models obtained with this methodology are below the performances reported in literature, highlighting the fact that different methodologies can lead to drastically different results. This emphasises the necessity to have a unified robust methodology for machine learning classification of fNIRS data in order to reliably compare results to each other. Precautions should be taken when claiming better performances of specific models as our results show that this tends to be very dataset dependent, sometimes giving results close to chance level and being unstable with hardly optimisable hyperparameters. It also appears that classification of mental tasks yields poorer performance compared to motor tasks. This is likely explained by the more complex brain processes involved for higher level tasks (Masters and Schulte 2020).

Finally, our work aims to encourage researchers to report more exhaustively in manuscripts the methodology and parameters used in their machine learning models for the sake of reproducibility and comparison.

Our framework presents the machine learning fNIRS community with a challenge platform akin to other benchmarking datasets such as CIFAR-10 and MNIST (Krizhevsky and Hinton 2009, LeCun et al. 1998), for researchers to compare their results. This framework will be published under the form of a repository available online, open to contributions from the community in order to add support for new datasets and techniques. We encourage machine learning researchers in the fNIRS community to use this framework when introducing new techniques, to enable a clearer depiction of performance improvements, especially for others choosing machine learning approaches for fNIRS in the context of cognitive neuroscience experiments or BCIs.

Dataset	Classes	Number of participants	Total number of examples
Herff et al. 2014	1-back; 2-back; 3-back	10	300
Shin et al. 2018	0-back; 2-back; 3-back	26	702
Shin et al. 2018	rest; word generation	26	1,560
Shin et al. 2016	rest; mental arithmetic	29	1,740
Bak et al. 2019	right hand; left hand; foot	30	2,250

Table 1. Summary of the datasets currently supported by the framework.

References:

Jobsis, F. F. (1977). Noninvasive, infrared monitoring of cerebral and myocardial oxygen sufficiency and circulatory parameters. Science, 198(4323), 1264-1267. doi: 10.1126/science.929199

Naseer, N., and Hong, K. S. (2015). fNIRS-based brain-computer interfaces: a review. Frontiers in human neuroscience, 9, 3. doi: 10.3389/fnhum.2015.00003

Herff, C., Heger, D., Fortmann, O., Hennrich, J., Putze, F., and Schultz, T. (2014). Mental workload during n-back task—quantified in the prefrontal cortex using fNIRS. Frontiers in human neuroscience, 7, 935. doi: 10.3389/fnhum.2013.00935

Benerradi, J., A. Maior, H., Marinescu, A., Clos, J., and L. Wilson, M. (2019, November). Exploring machine learning approaches for classifying mental workload using fNIRS data from HCI tasks. In Proceedings of the Halfway to the Future Symposium 2019 (pp. 1-11). doi: 10.1145/3363384.3363392

Naseer, N., Qureshi, N. K., Noori, F. M., and Hong, K. S. (2016). Analysis of different classification techniques for two-class functional near-infrared spectroscopy-based brain-computer interface. Computational intelligence and neuroscience, 2016. doi: 10.1155/2016/5480760

Trakoolwilaiwan, T., Behboodi, B., Lee, J., Kim, K., and Choi, J. W. (2017). Convolutional neural network for high-accuracy functional near-infrared spectroscopy in a brain–computer interface: three-class classification of rest, right-, and left-hand motor execution. Neurophotonics, 5(1), 011008. doi: 10.1117/1.NPh.5.1.011008

Yoo, S. H., Woo, S. W., and Amad, Z. (2018, October). Classification of three categories from prefrontal cortex using LSTM networks: fNIRS study. In 2018 18th International Conference on Control, Automation and Systems (ICCAS) (pp. 1141-1146). IEEE.

Lipton, Z. C., and Steinhardt, J. (2019). Troubling Trends in Machine Learning Scholarship: Some ML papers suffer from flaws that could mislead the public and stymie future research. Queue, 17(1), 45-77. doi: 10.1145/3317287.3328534

Bengio, Y. (2012). Practical recommendations for gradient-based training of deep architectures. In Neural networks: Tricks of the trade (pp. 437-478). Springer, Berlin, Heidelberg. doi: 10.1007/978-3-642-35289-8_26

Schmidt, R. M., Schneider, F., and Hennig, P. (2020). Descending through a Crowded Valley--Benchmarking Deep Learning Optimizers. arXiv preprint arXiv:2007.01547.

Shin, J., Von Lühmann, A., Kim, D. W., Mehnert, J., Hwang, H. J., and Müller, K. R. (2018). Simultaneous acquisition of EEG and NIRS during cognitive tasks for an open access dataset. Scientific data, 5(1), 1-16. doi: 10.1038/sdata.2018.3

Shin, J., von Lühmann, A., Blankertz, B., Kim, D. W., Jeong, J., Hwang, H. J., and Müller, K. R. (2016). Open access dataset for EEG+ NIRS single-trial classification. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 25(10), 1735-1745. doi: 10.1109/TNSRE.2016.2628057

Bak, S., Park, J., Shin, J., and Jeong, J. (2019). Open-access fNIRS dataset for classification of unilateral finger-and foot-tapping. Electronics, 8(12), 1486. doi: 10.3390/electronics8121486

Masters, M., and Schulte, A. (2020, September). Investigating the Utility of fNIRS to Assess Mental Workload in a Simulated Helicopter Environment. In 2020 IEEE International Conference on Human-Machine Systems (ICHMS) (pp. 1-6). IEEE. doi: 10.1109/ICHMS49158.2020.9209549

Krizhevsky, A., and Hinton, G. (2009). Learning multiple layers of features from tiny images.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278-2324. doi: 10.1109/5.726791