2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Classification of fNIRS data with LDA and SVM: a proof-of-concept for application in infant studies

Jessica Gemignani [*]

Abstract— **This study presents the implementation of a within-subject classification method, based on the use of Linear Discriminant Analysis (LDA) and Support Vector Machines (SVM), for the classification of hemodynamic responses. Using a synthetic dataset that closely resembles real experimental infant functional near-infrared spectroscopy (fNIRS) data, the impact of different levels of noise and different HRF amplitudes on the classification performances of the two classifiers are quantitively investigated.**

I. Introduction

Functional Near-Infrared Spectroscopy (fNIRS) data is a neuroimaging technique based on the measurement of the optical absorption properties of cerebral blood [1]. Allowing to measure the relative changes of oxygenation in the human head in response to a specific task or at rest, its use has become quite widespread in developmental neuroscience: since it is fully non-invasive, easy to use and silent, it is very well suited even in the youngest participants, e.g. in newborns.

Nevertheless, the specificities of infant fNIRS data can make the subject-level statistical analysis complicated; in particular, the infant hemodynamic response (HRF) often displays large variability in shape [2] and is characterized by smaller amplitudes [3]. This makes the use, for instance, of general linear models (GLMs) more difficult than with adults data.

In a previous work [4], we introduced a data-driven analysis scheme that, by comparing temporal segments of data within the same subject, would characterize a given fNIRS channel as *active* or *not active* by means of a classification method based on Linear Discriminant Analysis (LDA), achieving good classification accuracy on adult data during a motor task (78.7%).

In this work, we aim at investigating the applicability of that method to the analysis of infant data by employing a synthetic dataset that closely resembles it; in fact, for this method to find efficient application in infant data, a clear description of how accuracies change with increasing levels of noise and low or very low HRF amplitudes is necessary. Therefore, the use of a parametrized dataset is extremely valuable. In addition, we also aim at clarifying whether LDA and Support Vector Machines (SVM) perform differently depending on the amount of noise in the data or on the HRF amplitude.

II. Methods

A. Data Generation

Synthetic data was generated according to the montage and stimulus design employed in Gervain et al. (2012, [5]). In that study, fNIRS data was acquired in newborns using a montage with 24 channels, arranged bilaterally on the fronto-temporal areas. Data was generated using tools available in the *Brain AnalyzIR Toolbox* for Matlab [6].

Ten datasets ("participants") were simulated each with 18s-long 14 stimuli. Synthetic HRFs' amplitudes ranged between 0.05 and 0.25 mM x mm (between -0.025 and -0.125 mM x mm for HbR), across channels and subjects, to mimic the inter- and intra-subject variability naturally present in real experimental settings. Importantly, HRFs were only added to 12 channels: this allowed to have a clear ground truth of true *active* and true *not active* channels. Moreover, no HRFs were included in the initial five minutes of each timetrace: this segment of resting state will be used to create "rest" observations to be employed in the classification.

To simulate the contribution of heart rate, respiration and Mayer waves to the NIRS signal, the signal amplitude was increased by a factor ranging between 0.01 and 0.03 mM x mm, at frequencies typical of the newborn NIRS data, namely

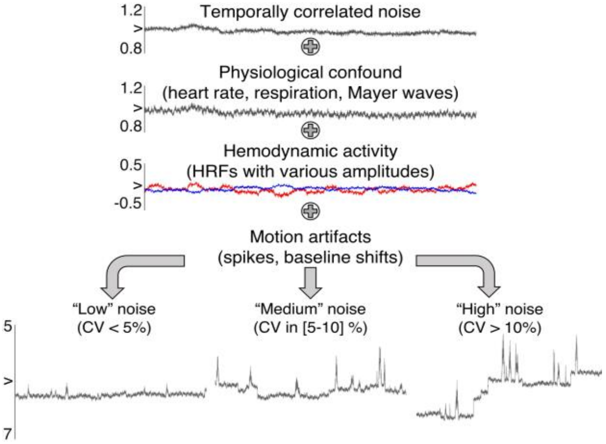

Figure 1: The scheme describes the pipeline employed to generate the synthetic infant fNIRS dataset. For each dataset, three different versions were created with three different levels of noise, based on the amplitude of the motion artifacts [7].

[*] Jessica Gemignani, PhD, is with the University of Padova, Department of Developmental Psychology and Socialization, Via Venezia 8, 35131 Padova, Italy (correspondence: jessica.gemignani@unipd.it).

978-1-7281-1178-0/21/$31.00 ©2021 IEEE

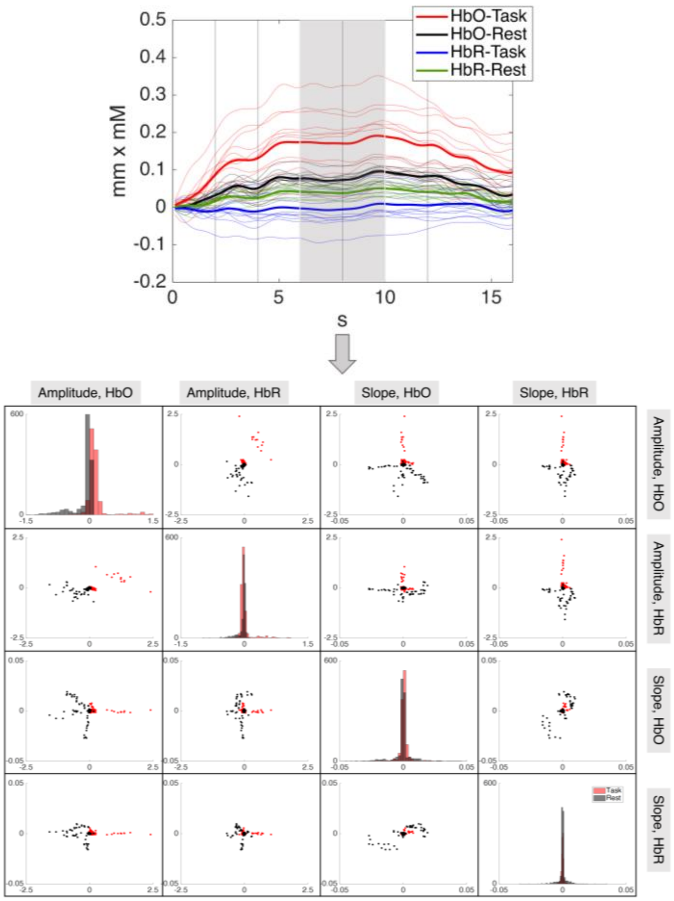

Figure 2: (**Top**) Example of feature extraction. For each channel, from each single-trial epoched signal (soft lines; thicker lines show class averages), features are extracted within seven sliding windows. In particular, for each window average amplitude and slope from both HbO and HbR are calculated. (**Bottom**) Discriminability of task from rest observations based on the chosen features; for visualization purposes, only features extracted from the 4th window are shown (window is highlighted in grey in the top panel). In each *i-j* scatter plot, *i* represents the feature shown on the y-axis and *j* represents that shown on the x-axis.

in the ranges around 1.5 ± 0.2 Hz, 0.25 ± 0.05 Hz and 0.1 ± 0.02 Hz, respectively.

Finally, three different versions of these functional datasets were created, characterized by different levels of noise (Figure 1), measured in terms of coefficient of variation (CV (%), i.e. standard deviation of the timetrace divided by its average).

The scheme in Figure 1 described the simulation steps and shows an example of simulated data. This approach is similar to the one used in our previous work [4], but the characteristics of the data is tailored to better represent typical data acquired on infants [7]. More details on the data generation procedure are provided in [8], along with the code to reproduce it.

B. *Pre-processing*

Data were pre-processed using custom scripts written in Matlab. In particular, raw data were band-pass filtered in the range 0.01-0.7 Hz; the filter was designed as an IIR filter and applied forward-backward with the *filtfilt* function, to achieve zero-phase filtering. Then, filtered raw data were converted into optical densities and concentration changes using the modified Beer-Lambert equation with the following absorption coefficients (μ_a, mm^{-1}-mM^{-1}): μ_a(HbO, 695 nm) = 0.0955, μ_a(HbO, 830 nm) = 0.232, μ_a(HbR, 695 nm) = 0.451 and μ_a(HbR, 830 nm) = 0.179.

C. *Extraction of features*

For the 'task' observations, features were extracted, for each channel, from each of the 14 single-trial epoched signals, using the approach employed in [4] and [9] . In particular, a 4s wide sliding window was moved through the signal within the stimulation period (18s) in 2s steps; for each window, the average signal value and average slope were computed and retained as features. For each channel, the procedure was repeated on both hemoglobin components, and respective features were concatenated; the resulting multivariate feature vector thus included 28 features (2 features x 7 time windows x 2 hemoglobin components).

To create the 'rest' observations, for each channel, 14 markers were randomly positioned throughout the initial five minutes segment of resting state, and features were extracted in the same fashion explained above.

The resulting channel-wise feature matrix included therefore 28 observations with 28 features, and was standardized.

D. *Classification*

Two separate classification procedures were carried out, with regularized Linear Discriminant Analysis (LDA) and Support Vector Machines (SVM); LDA was implemented via tools available in the Berlin Brain-Computer Interfacing (BBCI) toolbox [10], [11]; in this implementation, the optimal shrinkage parameter is obtained analytically [12], [13]. SVMs were employed with linear kernel using functions of the Statistics and Machine Learning Toolbox of Matlab [14].

In both cases, ten repetitions of four-folds cross validation were performed: the 28 observations were separated into four folds, with three folds used for training and one for testing. Classification accuracies were averaged across folds and repetitions.

Additionally, in order to ensure a robust sampling of the rest observations within the five minutes resting state, the whole procedure of sampling observations, extracting features and performing the classification was repeated 100 times. Accuracies were further averaged across these repetitions.

Finally, in order to characterize a channel as *active* or *not active*, channel-wise classification accuracies were compared to their correspondent chance-level classification accuracies: to obtain them, the same classification procedures were repeated on resting state only timetraces; fictitious triggers were placed at random times along the first five minutes, and classification was performed. The process was repeated 100 times.

E. *Statistical Analysis*

Identification of active channels: To classify a channel as *active* or *not active*, the observed classification accuracy was compared to the distribution of chance-level accuracies: a channel was classified as *active* if the ratio between the number of chance-level accuracies equal to or greater than the observed one, divided by the total number (n=100) was 0.05 or less (p<0.05).

Impact of data characteristics on performances: To statistically compare the performances of the algorithms, a mixed-effects linear model was fitted to the classification accuracy as dependent variable. A maximal random-effects structure was planned, with a random intercept for Participant and random uncorrelated slopes for the within-subject factors Noise, Algorithm and Channel. We then added fixed effects for Noise (3 levels) and Algorithm (2 levels) and the continuous variable HRF amplitude, as well as their interaction. As the full model did not converge, we removed first the random slope for Channel, and then for Algorithm. Models were implemented in JASP [15] and subsequent pairwise comparisons were adjusted for multiple comparisons with the Tukey procedure.

III. Results

The goal of the analysis was to compare the performances of LDA and SVM in classifying a given fNIRS channel as *active* or *not active*, in the context of a functional experiment.

On average, LDA and SVM yielded very similar overall classification accuracies (LDA: 75.9 %, SVM: 75.5%), with a standard deviation across channels, subjects and noise levels of 16.2% for LDA and 15.6% for SVM. The true discovery rate, namely the rate of truly active channels classified as such, ranged between 57.5% and 80.8% for LDA and 63.3% and 78.3% for SVM; clearly, it was higher for the less noisy datasets (80.8% LDA, 78.3% SVM) than for the more noisy ones (57.5% LDA, 63.3% SVM).

To investigate quantitively the impact of data characteristics on accuracies and find whether one algorithm is more suited in given conditions, a mixed-effects model was employed. It revealed significant main effects of Noise (F(2, 681)=15.26, p < 0.001) and HRF amplitude (F(1, 689)= 2718.12, p < 0.001) and significant interaction effects of Algorithm x Noise (F(2, 681)=3.09, p < 0.05) and Noise x HRF amplitude (F(2,581)= 15.09, p < 0.001).

Pairwise comparisons on the Algorithm x Noise interaction effect showed that SVM yielded a significantly higher classification accuracy than LDA only at the highest level of noise (mean difference 0.02, SE = 0.008, p = 0.03). No difference between the performance of the algorithms was found at lower levels of noise (CV<10%).

As for the Noise x HRF amplitude interaction, the slope between HRF amplitude and accuracy of the linear prediction for the high level of noise was significantly lower than both the other levels ($\text{slope}_{\text{HIGH}}$= 1.86, SE= 0.07; p < 0.001; $\text{slope}_{\text{MEDIUM}}$= 2.33, SE= 0.07; $\text{slope}_{\text{LOW}}$= 2.36, SE= 0.07).

As for the main effect of Noise, all pairwise contrasts resulted statistically significant (Low-Medium 0.04, t= 5.53, p <0.01; Medium – High 0.08, t= 6.03, p< 0.001; Low-High 0.12, t = 9.81, p <0.001).

Finally, the three-way interaction Algorithm x Noise x HRF did not turn out to be significant; this result suggests that data quality and specifically high levels of noise contribute alone to a lower classification accuracy, regardless of the algorithm employed for discrimination.

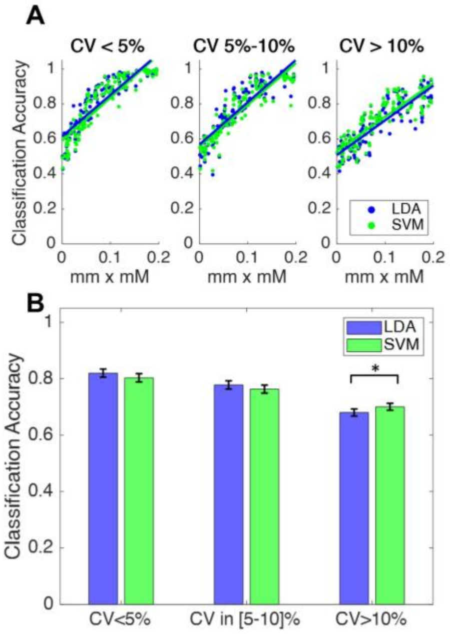

Figure 4: (**A**) Classifications accuracies achieved at different levels of noise, as a function of the HRF amplitude. (**B**) Pairwise comparisons on the Algorithm x Noise interaction effect showed that SVM performs better than LDA at higher levels of noise (70 v 68%, p = 0.03)

IV. Discussion

The statistical analysis of fNIRS data is often complicated by several aspects, like serial autocorrelations, large variability both inter- and intra-subject, systemic oscillations that overlap with the frequencies of the hemodynamic responses and motion artifacts [16].

In this sense, the favorable properties of the classifier we introduced for the analysis of adult fNIRS data [4] would be even more advantageous for the analysis of that acquired on infants: compared to adult data, fNIRS data acquired on infants and toddlers is often of poorer quality, with timetraces characterized by large and frequent motion artifacts. Furthermore, the infant HRF more often displays atypical features (e.g.“inverted responses” [2]) and smaller amplitudes, compared to the adults’. For these reasons, true activations may often remain undetected.

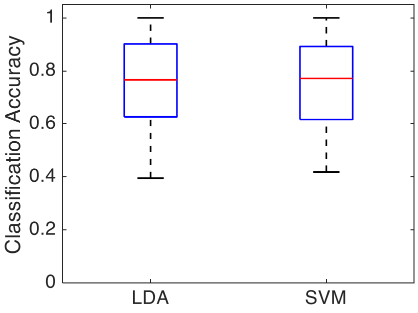

Figure 3: Distribution of classification accuracies pooled from all subjects, channels and noise levels (LDA, M: 75.9%, S: 16.2%; SVM, M: 75.5%, S: 15.6%)

In this work, we employed a synthetic dataset that closely resembles the characteristics of real infant fNIRS data: noisy, artifacted and with hemodynamic activity of very small amplitude (0-0.2 mm x mM). This allowed not only to have a clear ground truth but also to quantitively assess the impact of the data characteristics (HRFs, artifacts) on classification performances.

We classified channels as *active* or *not active* with a self-referencing scheme: within the same subject, classification was performed by comparing time intervals corresponding to resting state and to execution of the task. For the classification, we selected rLDA and SVM; several other works [17], [18] reported that they tend to have similarly good performances, but the question of how these are affected by data characteristics had not yet been explored to our knowledge.

Overall, rLDA and SVM performed very similarly. It is noteworthy that no difference was found under different HRFs amplitudes; moreover, even at the typical small amplitude of 0.1 mm x MM and at the worst level of noise, both yielded performances larger than 60%. This result is very relevant for the applicability of this method to newborn and infant fNIRS.

SVM was found to be slightly superior than rLDA at large noise levels, with a small but significant difference in accuracy (SVM: 70%, rLDA: 68%). But noise and HRFs amplitudes explained the largest share of variability in classification accuracy, regardless of the algorithm employed.

This result highlights the importance of selecting appropriate pre-processing strategies: here, since we specifically aimed at quantifying the impact of noise on performances, we did not either reject bad quality trials neither correct them with semi-automated methods. The application of this method to real experimental data should involve careful consideration of this issue; the selection of the most appropriate strategy depends on a number of factors, in particular the number of available experimental trials. For in-depth discussion on this topic we refer the reader to [7].

Our findings indicate that a linear classifier, such as rLDA or SVM, is suitable for classifying infant NIRS data and that a data-driven analysis based on one of these algorithms can produce results with good accuracy. Future work should involve the application of this method to real experimental data and compare it with other analysis frameworks.

V. Conclusions

- Synthetic infant fNIRS channels, with un-processed artifacts, can be classified as *active* or *not active* with classification accuracy ~ 75%.
- LDA and SVM perform similarly.
- SVM performs slightly better than LDA on the very noisy data.
- Noise and HRF amplitude have the biggest impact on classification accuracies: application to real experimental data should involve careful removal of motion artifacts, by means of trial rejection or correction.

Acknowledgment

This work was supported by the ERC Consolidator Grant "BabyRhythm 773202" awarded to Judit Gervain. The author wishes to thank Dr. Judit Gervain and Dr. Benjamin Blankertz for providing insightful and valuable feedback.

References

[1] M. Ferrari and V. Quaresima, "A brief review on the history of human functional near-infrared spectroscopy (fNIRS) development and fields of application," *Neuroimage*, vol. 63, no. 2, pp. 921–935, 2012.

[2] C. Issard and J. Gervain, "Variability of the hemodynamic response in infants: Influence of experimental design and stimulus complexity," *Dev. Cogn. Neurosci.*, vol. 33, no. January, pp. 182–193, 2018.

[3] R. N. Aslin, M. Shukla, and L. L. Emberson, "Hemodynamic Correlates of Cognition in Human Infants," *Annu. Rev. Psychol.*, vol. 66, no. 1, pp. 349–379, 2014.

[4] J. Gemignani, E. Middell, R. L. Barbour, H. L. Graber, and B. Blankertz, "Improving the analysis of near-infrared spectroscopy data with multivariate classification of hemodynamic patterns : a theoretical formulation and validation," *J. Neural Eng.*, vol. 15, no. 4, p. 045001 (15pp), 2018.

[5] J. Gervain, I. Berent, and J. F. Werker, "Binding at birth: The newborn brain detects identity relations and sequential position in speech," *J. Cogn. Neurosci.*, vol. 24, no. 3, pp. 564–574, 2012.

[6] H. Santosa, X. Zhai, F. Fishburn, and T. Huppert, "The NIRS Brain AnalyzIR Toolbox," *Algorithms*, vol. 11, no. 5, p. 73, May 2018.

[7] J. Gemignani and J. Gervain, "Comparing different pre-processing routines for infant fNIRS data," *Dev. Cogn. Neurosci.*, vol. 48, no. July 2020, p. 100943, 2021.

[8] J. Gemignani and J. Gervain, "A practical guide for synthetic fNIRS data generation," under review *in 43rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2021.

[9] J. Shin, K.-R. Müller, and H.-J. Hwang, "Near-infrared spectroscopy (NIRS)-based eyes-closed brain-computer interface (BCI) using prefrontal cortex activation due to mental arithmetic," *Sci. Rep.*, vol. 6, no. 1, p. 36203, Dec. 2016.

[10] "BBCI toolbox." [Online]. Available: https://github.com/bbci/bbci_public. [Accessed: 03-Nov-2016].

[11] B. Blankertz *et al.*, "The Berlin Brain-Computer Interface: Progress Beyond Communication and Control," *Front. Neurosci.*, vol. 10, p. 530, 2016.

[12] O. Ledoit, M. Wolf, O. Ledoit, and M. Wolf, "A well-conditioned estimator for large-dimensional covariance matrices," *J. Multivar. Anal.*, vol. 88, no. 2, pp. 365–411, 2004.

[13] J. Schäfer and K. Strimmer, "A Shrinkage Approach to Large-Scale Covariance Matrix Estimation and Implications for Functional Genomics," *Stat. Appl. Genet. Mol. Biol.*, vol. 4, no. 1, Jan. 2005.

[14] . The Mathworks, "MATLAB and Statistics and Machine Learning Toolbox Release 2020a." Natick, Massachusetts, United States, 2020.

[15] JASP Team, "JASP (Version 0.14.1)[Computer software]." 2020.

[16] T. J. Huppert, "Commentary on the statistical properties of noise and its implication on general linear models in functional near-infrared spectroscopy," *Neurophotonics*, vol. 3, no. 1, p. 010401, 2016.

[17] M. Misaki, Y. Kim, P. A. Bandettini, and N. Kriegeskorte, "Comparison of multivariate classifiers and response normalizations for pattern-information fMRI," *Neuroimage*, vol. 53, no. 1, pp. 103–118, 2010.

[18] K. Tai and T. Chau, "Single-trial classification of NIRS signals during emotional induction tasks: Towards a corporeal machine interface," *J. Neuroeng. Rehabil.*, vol. 6, no. 1, pp. 1–14, 2009.