Single Trial Classification of fNIRS-based Brain-Computer Interface Mental Arithmetic Data: A Comparison Between Different Classifiers

Günther Bauernfeind[1], David Steyrl[1], Clemens Brunner[1] and Gernot R. Müller-Putz[1]

Abstract— **Functional near infrared spectroscopy (fNIRS) is an emerging technique for the in-vivo assessment of functional activity of the cerebral cortex as well as in the field of brain-computer-interface (BCI) research. A common challenge for the utilization of fNIRS for BCIs is a stable and reliable single trial classification of the recorded spatio-temporal hemodynamic patterns. Many different classification methods are available, but up to now, not more than two different classifiers were evaluated and compared on one data set. In this work, we overcome this issue by comparing five different classification methods on mental arithmetic fNIRS data: linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), support vector machines (SVM), analytic shrinkage regularized LDA (sLDA), and analytic shrinkage regularized QDA (sQDA). Depending on the used method and feature type (oxy-Hb or deoxy-Hb), achieved classification results vary between 56.1 % (deoxy-Hb/QDA) and 86.6 % (oxy-Hb/SVM). We demonstrated that regularized classifiers perform significantly better than non-regularized ones. Considering simplicity and computational effort, we recommend the use of sLDA for fNIRS-based BCIs.**

I. INTRODUCTION

A few years ago, functional near infrared spectroscopy (fNIRS) was proposed as a novel approach in the field of brain-computer communication. Coyle et al. [6] were the first to investigate the suitability of fNIRS for brain-computer interfaces (BCIs), the so-called optical BCI (oBCI). Since that time, a growing number of research groups have investigated different concepts using fNIRS alternatively to, or in combination with, traditional EEG-based BCIs (for an overview see [1]). However, there is still ongoing research needed to investigate the full potential of fNIRS in this field. A common challenge for BCIs is a stable and reliable single trial classification, especially for cognitive (mental) tasks. This challenge applies in particular to oBCIs, because typically only little data is available. Therefore, finding a suitable classifier is mandatory.

Previous studies investigated, among others, the use of linear discriminant analysis (LDA), c. f. [3], [15], [17], [18], quadratic discriminant analysis (QDA), c. f. [15], or support vector machines (SVM), c. f. [16], [18]. However, to the best of our knowledge, up to now not more than two classifiers were evaluated on one data set to investigate their suitability for fNIRS signals.

In the present oBCI simulation, we overcome this issue by evaluating and comparing these previously used classifiers along with their regularized variants: LDA, QDA, SVM, shrinkage regularized LDA (sLDA), and shrinkage regularized QDA (sQDA). Furthermore, these methods are compared to the approach of using antagonistic hemodynamic response patterns for single trial classification, as performed by our group in [3].

*The authors' research has been supported by the "Land Steiermark" research project (A3-22.N-13/2009-8) and FP7 EU research projects Back-Home (No.288566) and BNCI Horizon 2020 (No.609593).

[1]G. Bauernfeind, D. Steyrl, C. Brunner and G. R. Müller-Putz are with Institute for Knowledge Discovery, BCI Lab, Graz University of Technology, 8010 Graz, Austria david.steyrl at tugraz.at

II. METHODS

A. Participants, Experimental Paradigm and Data Recording

For the present investigations, we used the data of eight participants (three male, five female, mean age 26 years, standard deviation SD 2.8 years) which showed antagonistic hemodynamic response patterns (for details see [14]) during the performance of a mental arithmetic (MA) task. These data sets were already used in a first investigation on single trial classification [3] of these responses. Briefly summarized, the participants were instructed to perform cue-guided mental calculations. During this task, they had to sequentially subtract a one-digit number from a two-digit number (e.g., $87 - 4 = 83$, $83 - 4 = 79$, ...; the initial subtraction was presented visually on a monitor) as quickly as possible for 12 s, followed by 28 s of rest. Participants performed 3 or 4 runs with six trials per class and run, resulting in 18 or 24 trials per class, respectively.

For data recording, a continuous-wave fNIRS system (ETG-4000, Hitachi Medical Co., Japan) with a sampling rate of 10 Hz was used. The multi-channel system, consisting of 16 photo-detectors and 17 light emitters (3×11 grid, 52 channels), measured the change of oxy-Hb concentration and deoxy-Hb concentration in millimolar times millimeter ($\mathrm{mM} \cdot \mathrm{mm}$). The distance between source and detector was 3 cm. The lowest line of the grid was arranged along the FP1–FP2 line (international EEG 10-20 system) covering the whole frontal lobe. For further details on data recording, experimental paradigm and channel placement see [3], [14].

B. Preprocessing, Feature Extraction and Evaluation

After removing baseline drifts with a 0.01 Hz high pass filter, the task-related concentration changes referred to a 10 s baseline interval prior to the task were calculated (for further details see [2], [14]). Concentration changes (averaged over 1 s) of the hemodynamic response at seconds 10, 11, 12, 13, and 14 (± 2 s around the end of the MA task, to cover also delayed task-related parts of the response) were labeled as class MA. Changes at seconds 26, 27, 28, 29, and 30 (in between two MA tasks) were labeled as class REST. Features consisted of oxy-Hb or deoxy-Hb concentration values of all channels at the described fixed times. Oxy-Hb

978-1-4244-7929-0/14/$26.00 ©2014 IEEE

and deoxy-Hb features were used separately since combining the features would have worsen the problem of low trials-to-features ratio. For each participant, independent classifiers were calculated using training sets, and subsequent offline BCI simulations were performed on separate evaluation sets. These consisted of the last 8 trials per class of the recording.

C. Linear Discriminant Analysis and Analytic Shrinkage Regularization

The idea behind linear discriminant analysis (LDA) is to use hyperplanes for separating classes [7]. In the case of a binary decision, the classification of a given sample $x \in \mathbb{R}^N$ is performed by

$$\text{sign}(w^T x + b), \tag{1}$$

where b denotes the bias (offset of the hyperplane) and w the projection vector. The projection vector can be calculated by

$$w = \hat{\Sigma}_c^{-1}(\hat{\mu}_1 - \hat{\mu}_2), \tag{2}$$

where $\hat{\mu}_i$ denotes the vector of estimated means per feature dimension of class i and $\hat{\Sigma}_c$ is the estimated common covariance matrix (the average of the individual class covariance matrices). The bias is given by

$$b = -w^T \hat{\mu}, \tag{3}$$

where $\hat{\mu}$ denotes the vector of estimated means per feature dimension over all classes. Although this classification method is very simple, LDA provides the optimal classification if the following assumptions are satisfied [7]:

- True class labels are known
- Features are normally distributed and parameters of the distribution are known
- Features covariance matrices are equal for all classes

Hence, the classification performance of an LDA classifier crucially depends on accurate estimates of the class means and the common covariance matrix. If enough data are available, an accurate estimate is not a problem. However, in cases where insufficient data are available, conventional estimation of the covariance matrix fails. In those cases, the covariance matrix is ill conditioned. One can improve the conditioning of the covariance matrix by regularization.

An efficient regularization method is analytic shrinkage regularization. Analytic shrinkage regularization compensates for systematic errors due to a low amount of data. The systematic error affects the eigenvalue calculation of the covariance matrix: large eigenvalues of the covariance matrix are overestimated, and small eigenvalues are underestimated. One can counteract these systematic errors by decreasing large eigenvalues and increasing small eigenvalues. Analytic shrinkage regularization modifies the estimated covariance matrix by

$$\tilde{\Sigma}(\gamma) = (1-\gamma)\hat{\Sigma} + \gamma \frac{\text{trace}\,\hat{\Sigma}}{d} I, \tag{4}$$

where γ is a tuning parameter ranging from 0 to 1, $\hat{\Sigma}$ is the original estimated covariance matrix, d denotes the dimensionality of the feature space and I denotes the identity matrix. This kind of regularization shrinks the estimated covariance matrix towards the identity matrix.

Essentially, the tuning parameter γ must be selected appropriately; $\gamma = 0$ corresponds to the original covariance matrix estimation, while $\gamma = 1$ discards the original covariance matrix. An optimal value of γ compensates for the systematic error and hence improves the classification accuracy. There exists an analytic method for optimizing γ, introduced by [11]. This so-called analytic shrinkage regularized LDA (sLDA) is computationally very efficient and outperformed ordinary LDA significantly at low trial-to-feature ratios [4]. In our work, we used the analytic shrinkage calculation as implemented in BCILAB [10].

D. Quadratic Discriminant Analysis and Analytic Shrinkage Regularization

Quadratic discriminant analysis (QDA) can be seen as an extension of the LDA classifier [7]. QDA considers quadratic terms of features. In the binary decision case, the classification is given by

$$\text{sign}(x^T A x + w^T x + b), \tag{5}$$

where b denotes the bias term, w the linear projection vector, and A the projection matrix of the quadratic terms. The quadratic terms allow for non-linear decision boundaries. Non-linear decision boundaries are a necessity for optimal classification when the class features are normally distributed, but have different covariance matrices. QDA is the optimal classification strategy if the following assumptions are satisfied [7]:

- True class labels are known
- Features are normally distributed and parameters of the distribution are known

Training of QDA is similar to LDA, except that separate covariance matrices must be estimated for each class. When the feature space is large, this can dramatically increase the amount of parameters to fit [7]. A high amount of parameters to fit intensifies the problem of estimating well conditioned covariance matrices. Therefore, regularization of the covariance matrices is highly recommended. One can use the same analytic solution for estimating a shrinkage regularization parameter compensating for a low amount of training data as used for LDA.

E. Linear Support Vector Machines

Like LDA, linear support vector machines (SVMs) use hyperplanes for separating classes. However, SVMs rely on a different method for finding the separating hyperplanes, which is not based on the statistical properties of the training data. Instead, it tries to find a decision boundary with maximum margin between classes [7]. The classification rule in the binary case is again a linear function in x:

$$\text{sign}(w^T x + b), \tag{6}$$

where w is the projection vector and b is the bias. To calculate w and b, the margin is defined as the distance

between the nearest data points of either class measured perpendicular to the hyperplane. Intuitively, one wants to maximize the margin to maximize the "safety distance" between the classes. It can be shown that maximizing the margin is mathematically equivalent to minimizing $\|w\|^2$ in a rescaled data space where $\min_i |w^T\ x_i + b| = 1$ holds. This formulation runs into problems if there is no perfect separating hyperplane. For dealing with overlapping classes, so-called slack variables ζ_i are introduced. ζ_i measures the amount of constraint violation per data point i. Including the slack variable into the formalism described above leads to the typical SVM notation:

$$\min_{w,b} \|w\|^2 + C \sum_{i=1}^{N} \zeta_i \qquad (7)$$
$$\text{subject to: } \min_i |w^T\ x_i + b| = 1 - \zeta_i$$

C is a tuning parameter that defines the costs of constraint violations and can be determined by cross-validation. It controls the weighting between classification error ($\sum_{i=1}^{N} \zeta_i$) and model complexity ($\|w\|^2$), i. e. a small C leads to a simple model, whereas a large C leads to a complex model. SVMs perform an implicit regularization, because the objective is to minimize the classification error and the model complexity simultaneously. Hence, simple models are preferred over complex models.

We performed a grid search to determine the optimal cost parameter by calculating 10-times 10-fold cross-validations on the training data. Here, we varied the values of C from 2^{-5} to 2^{15} by increasing the exponent in steps of 2. We used a modified SVM implementation from BCILAB [10].

F. Antagonistic Activation Patterns Classification

In [3] we used antagonistic hemodynamic response patterns as features for classification with an LDA. In more detail, features consisted of combinations of oxy-Hb or deoxy-Hb concentration values, selected from three regions of interest (ROI): medial area (ROI_1) of the anterior prefrontal cortex, left (ROI_2) and right (ROI_3) dorsolateral prefrontal cortex. For each participant, independent LDAs were trained and validated (leave-one-out cross validation) with individual antagonistic oxy-Hb responses. All possible feature combinations were evaluated to identify the best performing antagonistic feature combination (ROI_1, ROI_2 or ROI_1, ROI_3). The same procedure was applied also to the antagonistic deoxy-Hb signals (for further details see [3]). By using the best antagonistic oxy-Hb features, a mean classification accuracy of 79.7 % (SD 8.0) was computed. Antagonistic deoxy-Hb features performed worse, reaching a mean classification accuracy of 66.4 % (SD 12.5).

G. Statistical Analysis

We used analysis of variance (ANOVA) to assess statistical differences between the six classification methods and the two feature types. More specifically, we ran a two-way repeated measures ANOVA with the factors "method" (six levels: Antagonistic-LDA, LDA, QDA, sLDA, sQDA, SVM) and "feature" (two levels: oxy-Hb and deoxy-Hb). If Mauchly's test indicated significant violations of the sphericity assumption, we adjusted the degrees of freedom using Greenhouse-Geisser correction. For significant ANOVA results, we performed paired t-tests as post-hoc tests with false discovery rate (FDR) correction to account for multiple comparisons.

III. RESULTS

The results of the BCI simulation and the results of the antagonistic features classification are summarized in Table I.

TABLE I
MEAN AND STANDARD DEVIATION (SD) OF CLASSIFICATION ACCURACIES (IN %) FOR OXY-HB AND DEOXY-HB FEATURES. ANTAGONISTIC-LDA DENOTES THE COMPARISON ACCURACIES, PUBLISHED IN [3]. BEST CLASSIFICATION METHODS PER FEATURE ARE HIGHLIGHTED.

	oxy-Hb		deoxy-Hb	
Classifier	Mean	SD	Mean	SD
Antagonistic-LDA	79.7	8.0	66.4	12.5
LDA	69.2	16.1	76.1	13.8
sLDA	86.3	10.1	77.7	11.3
QDA	62.0	9.5	56.1	6.7
sQDA	82.7	12.7	67.0	17.2
SVM	**86.6**	7.3	**79.4**	14.3

The ANOVA revealed significant main effects of "method" ($F(1.9, 13.5) = 14.34$, $p < 0.01$, $\eta_G^2 = 0.33$) and "feature" ($F(1, 7) = 6.99$, $p = 0.03$, $\eta_G^2 = 0.09$) and a significant interaction between these two factors ($F(2.2, 15.2) = 4.76$, $p < 0.01$, $\eta_G^2 = 0.09$). Pairwise FDR corrected t-tests yielded the following significant differences: both SVM and sLDA were significantly better than sQDA, QDA, LDA, and Antagonistic-LDA; and QDA was significantly worse than Antagonistic-LDA, LDA, and sQDA. The results of the ANOVA are shown in Figure 1.

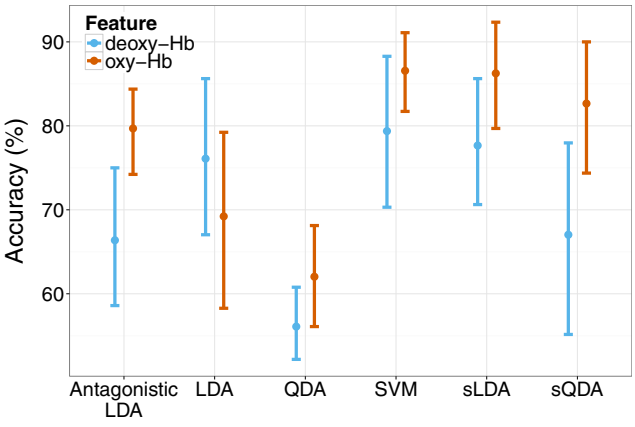

Fig. 1. Mean accuracies and 95 % confidence intervals for all methods and features.

IV. DISCUSSION AND CONCLUSION

Our findings show that the choice of an appropriate features/classifier combination is crucial for single trial classification of fNIRS-based BCI MA data. The results vary significantly and are between 56.1 % for the worst combination (deoxy-Hb/QDA) and 86.6 % (oxy-Hb/SVM) for the best combination.

Concerning features, oxy-Hb features perform significantly better than deoxy-Hb features. Increases in oxy-Hb and only slight decreases in deoxy-Hb are typical for cortical activation [19]. Analogously, increases in deoxy-Hb accompanied by decreases of oxy-Hb indicate cortical deactivation [13]. Although deoxy-Hb responses appear to be more localized and topographically closer to activated areas [5], [8], [9] than the more widespread changes in oxy-Hb [12], [13], deoxy-Hb changes are smaller than oxy-Hb changes and exhibit higher variance. Consequently, oxy-Hb seems to be a more suitable feature for single trial classification. The present investigation also supports this hypothesis. The better performance of oxy-Hb features is consistent over all methods with the exception of LDA (see Figure 1). This exception can be explained by an extremely low classification accuracy in one subject when using oxy-Hb and LDA.

Concerning classification methods, our statistical analysis shows that regularized classifiers (sLDA, sQDA, SVM) perform significantly better than non-regularized classifiers (LDA, QDA). This result was expected, because regularized classifiers have demonstrated their superiority over non-regularized classifiers many times, especially when the trials-to-features ratio is low [4]. This is particularly relevant for fNIRS-based BCIs, where training data are typically scarce due to long measurement time per trial. However, if more training trials are available the performance difference between regularized and non-regularized classifiers should become less [4]. Another result is that regularized linear methods (sLDA, SVM) performed significantly better than the regularized non-linear method (sQDA). This performance difference is mainly caused by two participants, which exhibit low classification accuracies when using sQDA. Although SVM and sLDA performed equally well, training an sLDA model is computationally less demanding than training an SVM model. The former uses an analytic solution for the optimal regularization parameter, whereas the latter relies on time-consuming cross-validation.

Although the comparison revealed significant results, some limitations should be mentioned. First of all, the number of data sets is low. However, considering the increase of the classification accuracy using regularized classifiers, our findings suggest that the use of such methods is recommended, especially if the amount of data is limited. To explore this in more detail, a bigger data set is needed. Another limitation is the type of task. In this investigation, only spatio-temporal hemodynamic patterns induced by performing a MA task (versus rest) were classified. To allow a more general statement on the suitability of the different methods, additional tasks (e. g. motor imagery) have to be investigated. Further it should be mentioned that much more possible classification methods are known. Some of them could be superior to our tested methods.

In conclusion, we strongly recommend to use regularized classifiers with fNIRS-based BCI MA data. Considering simplicity and computational effort, sLDA seems to be the method of choice.

REFERENCES

[1] G. Bauernfeind. *Using Functional Near-Infrared Spectroscopy (fNIRS) for Optical Brain-Computer Interface (oBCI) Applications*. PhD thesis, Graz University of Technology, 2012.

[2] G. Bauernfeind, R. Leeb, S. Wriessnegger, and G. Pfurtscheller. Development, set-up and first results of a one-channel near-infrared spectroscopy system. *Biomed Tech (Berl)*, 53(1):36–43, 2008.

[3] G. Bauernfeind, R. Scherer, G. Pfurtscheller, and C. Neuper. Single trial classification of antagonistic oxyhemoglobin responses during mental arithmetic. *Med Biol Eng Comput*, 49(9):979–984, 2011.

[4] B. Blankertz, S. Lemm, M. Treder, S. Haufe, and K.-R. Mller. Single-trial analysis and classification of erp components - a tutorial. *Neuroimage*, 56:814–825, 2011.

[5] A. F. Cannestra, I. Wartenburger, H. Obrig, A. Villringer, and A. W. Toga. Functional assessment of Broca's area using near infrared spectroscopy in humans. *Neuroreport*, 14(15):1961–1965, 2003.

[6] S. Coyle, T. E. Ward, C. Markham, and G. McDarby. On the suitability of near-infrared (NIR) systems for next-generation brain-computer interfaces. *Physiol Meas*, 25(4):815–822, 2004.

[7] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning, Second Edition, Springer Series in Statistics*. Springer, 2009.

[8] V. Kaiser, G. Bauernfeind, A. Kreilinger, T. Kaufmann, A. Kübler, C. Neuper, and G.R. Müller-Putz. Cortical effects of user training in a motor imagery based brain-computer interface measured by fNIRS and EEG. *Neuroimage*, 85:432–444, 2014.

[9] A. Kleinschmidt, H. Obrig, M. Requardt, K. D. Merboldt, U. Dirnagl, A. Villringer, and J. Frahm. Simultaneous recording of cerebral blood oxygenation changes during human brain activation by magnetic resonance imaging and near-infrared spectroscopy. *J Cereb Blood Flow Metab*, 16(5):817–826, 1996.

[10] C. A. Kothe and S. Makeig. Bcilab: a platform for braincomputer interface development. *J. Neural Eng.*, 10:1–17, 2013.

[11] O. Ledoit and M. Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *J. Multivar. Anal.*, 88:365–411, 2004.

[12] D. Malonek and A. Grinvald. Interactions between electrical activity and cortical microcirculation revealed by imaging spectroscopy: implications for functional brain mapping. *Science*, 272(5261):551–554, 1996.

[13] H. Obrig and A. Villringer. Beyond the visible: imaging the human brain with light. *J Cereb Blood Flow Metab*, 23(1):1–18, 2003.

[14] G. Pfurtscheller, G. Bauernfeind, S. C. Wriessnegger, and C. Neuper. Focal frontal (de)oxyhemoglobin responses during simple arithmetic. *Int J Psychophysiol*, 76(3):186–192, 2010.

[15] S. D. Power, A. Kushi, and T. Chau. Towards a system-paced near-infrared spectroscopy brain-computer interface: differentiating prefrontal activity due to mental arithmetic and mental singing from the no-control state. *J Neural Eng*, 8(6):066004, 2011.

[16] R. Sitaram, H. Zhang, C. Guan, M. Thulasidas, Y. Hoshi, A. Ishikawa, K. Shimizu, and N. Birbaumer. Temporal classification of multichannel near-infrared spectroscopy signals of motor imagery for developing a brain-computer interface. *Neuroimage*, 34(4):1416–1427, 2007.

[17] M. Stangl, G. Bauernfeind, J. Kurzmann, R. Scherer, and C. Neuper. A haemodynamic braincomputer interface based on real-time classification of near infrared spectroscopy signals during motor imagery and mental arithmetic. *J Near Infrared Spectrosc*, 21:157–171, 2013.

[18] K. Tai and T. Chau. Single-trial classification of NIRS signals during emotional induction tasks: towards a corporeal machine interface. *J Neuroeng Rehabil*, 6(1):39, 2009.

[19] I. Vanzetta and A. Grinvald. Coupling between neuronal activity and microcirculation: implications for functional brain imaging. *HFSP J*, 2(2):79–98, 2008.