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

EEG representation approach based on Kernel Canonical Correlation Analysis highlighting discriminative patterns for BCI applications

Viviana Gómez-Orozco[1] Cristian Blanco Martínez[1] David Augusto Cárdenas-Peña[1] Paula Marcela Herrera[2] and Álvaro Ángel Orozco-Guitiérrez[1]

***Abstract*— Brain-Computer Interface (BCI) is applied in the study of different cognitive processes or clinical conditions as enhancing cognitive skills, motor rehabilitation, and control. However, many approaches focus on using a robust classifier instead of providing a better feature space. This work develops a feature representation methodology through the kernel canonical correlation analysis to reveal nonlinear relations between filter-banked common spatial patterns (CSP) extracted. Our approach reveals nonlinear relations between ranked filter-banked multi-class CSP features and the labels in a finite-dimensional canonical space. We tested the performance of our methodology on the BCI Competition IV dataset 2a. The introduced feature representation using a classic linear SVM achieves accuracy rates competitive with the state-of-the-art BCI strategies. Besides, the processing pipeline allows identifying the spatial and spectral features driven by the underlying brain activity and best modeling the motor imagery intentions.**

***Clinical relevance*— This BCI strategy assesses the nonlinear relationships between time series to improve the interpretation of brain electrical activity, taking into account the spatial and spectral features driven by the underlying brain dynamic.**

I. INTRODUCTION

Brain-computer interface systems allow users to control applications and devices from neuroimaging data. Among all kinds of neuroimages, electroencephalographic (EEG) signals non-invasively record brain electrical activity driven by stimuli or intentions. BCI applications for motor-disabled people include physical therapy, rehabilitation, and motion assistance. One of the most significant open issues in BCI relies on the extraction and selection of features relevant for performing an action and explaining the underlying brain dynamics [1]. On the one hand, feature engineering approaches demand knowledge about brain physiology, which sometimes is either unknown or suffers from insufficient accuracy. On the contrary, the deep learning models outperform thanks to the hierarchically gained complexity. However, the reasons behind the outcomes become inscrutable, and the predictions result disbelieved in critical scenarios.

The literature considers many multi-class motor imagery BCI strategies to enhance the discrimination, either based on the development of more robust and complex classifiers or the relevant information extraction. The formers include neighborhood rough set classifiers with a type-2 fuzzy logic system using the extracted features [2], [3], and a classifier fusion method based on Dempster–Shafer theory to combine the binary classifiers resulting from the one-versus-rest training approach [4]. However, these approaches mainly focus on the multi-class classifier performance improvement to achieve high kappa values. The other kind of approach only analyzes statistical dependencies between activation areas and MI tasks, so that skipping the relevant features associated with neuronal activity [5].

This work proposes a feature representation methodology for BCI that decodes nonlinear relationships from EEG data through the Kernel Canonical Correlation Analysis (KCCA) [6]. The proposal highlights spatial and temporal patterns by analyzing EEG trials belonging to one of two classes. We consider a BCI processing pipeline as follows: First, a filter bank decomposes EEG signals into subbands covering the frequency range between 4 to 40 Hz. Secondly, the common spatial patterns technique extracts spatial filters on each frequency band to maximize the differences between the evaluated tasks. Then, the extracted features concatenate to form a single feature vector with spectral and spatial information. The proposed KCCA-based feature representation maps the characterized trials and task labels into a new joint space. We test the proposed methodology in the well-known BCI Competition IV dataset 2a containing EEG records from nine healthy subjects while performing four motor imagery (MI) tasks, which provided training and test subsets. We fit the multi-class model through a linear support vector machine (SVM) classifier using all training data for selected features and the parameters found by a 10-fold nested cross-validation scheme. Then, we infer the MI condition associated with the testing data employed the model trained from training data. We quantify the performance in terms of Cohen's kappa coefficient κ. Our results show that the proposed feature representation. Besides, we perform a statistical test for comparison purposes between our results and state-of-the-art, resulting in significant improvements with p-values lower than 1%.

II. MATERIALS AND METHODS

A. Filter-banked Multi-class Common Spatial Patterns

Let a labeled EEG signal dataset $\left\{\boldsymbol{X}_n \in \mathbb{R}^{C\times T}, l_n \in \{1,L\}\right\}_{n=1}^{N}$ holding N time-series trials with C channels at T time instants, where each trial is related to one of L classes through its label l_n.

[1]Automatic Research Group, Faculty of Engineerings, Universidad Tecnológica de Pereira, Pereira, Colombia {vigomez, cristian.blanco, dcardenasp, aaog}@utp.edu.co
[2]Psychiatry, Neuroscience, and Community Group, School of Medicine, Universidad Tecnológica de Pereira, Pereira, Colombia p.herrrera@utp.edu.co

For a binary classification task, the Common spatial pattern technique finds the spatial projection maximizing the variance of trials belonging to one class while minimizing the variance of the other [7]. This work considers a filter-banked one-versus-rest (OVR) strategy as the multi-class CSP extension that independently maximizes the separability of studied conditions at each of B frequency bands. The OVR CSP maps each band b into a linearly uncorrelated space through the matrix of column-wise spatial filters $\boldsymbol{W}_b^l \in \mathbb{R}^{C\times M}$, being M the number of spatial filters and l the class index. The CSP problem is stated as a generalized eigenvalue problem $\boldsymbol{\Sigma}_b^l \boldsymbol{W}_b^l = \boldsymbol{\Lambda}_b^l(\boldsymbol{\Sigma}_b^l + \boldsymbol{\Sigma}_b^{-l})\boldsymbol{W}_b^l$, where the superscript $-l$ indicates all classes but l, the diagonal matrix $\boldsymbol{\Lambda}_b^l$ holds the generalized eigenvalues. $\boldsymbol{\Sigma}_b^l \in \mathbb{R}^{C\times C}$ and $\boldsymbol{\Sigma}_b^{-l} \in \mathbb{R}^{C\times C}$ are the covariance matrices in the channel space for EEG trials belonging to class l and for the remaining trials, respectively. Then, the log-relative power in the CSP space compose the extracted feature set:

$$\boldsymbol{y}_{nb}^l = \log\left(\frac{\operatorname{diag}\left(\left(\boldsymbol{W}_b^l\right)^\top \boldsymbol{X}_{nb}\boldsymbol{X}_{nb}^\top \boldsymbol{W}_b^l\right)}{\operatorname{tr}\left(\left(\boldsymbol{W}_b^l\right)^\top \boldsymbol{X}_{nb}\boldsymbol{X}_{nb}^\top \boldsymbol{W}_b^l\right)}\right) \tag{1}$$

where $\operatorname{diag}(\cdot)$ and $\operatorname{tr}(\cdot)$ stand for the diagonal and trace operators, respectively. $\boldsymbol{y}_{nb}^l \in \mathbb{R}^M$ denotes the M spatial features. As a result, the matrix $\boldsymbol{Y} \in \mathbb{R}^{N\times D}$ ($D = B\cdot L\cdot M$) holds the M spatial features for the L classes and B frequency bands extracted from the input EEG dataset.

B. Mutual Information-based feature selection

Since the number of extracted features D grows with the number of frequency bands, spatial filters, and classes; the feature space becomes highly dimensional. Aiming at dealing with such an issue, the Mutual Information of Best Individual Feature (MIBIF) algorithm ranks features according the information shared with the provided labels, as follows:

$$I_d = -\sum_l \pi_l \log(\pi_l) + \sum_l \int p(l|y^d)\log\Big(p(l|y^d)\Big)\,dy^d \tag{2}$$

where y^d corresponds to the d-th feature in $\boldsymbol{Y}$, and I_d denotes the mutual information between the d-th feature and the labels. π_l and $p(y^d|l)$ denote the class prior and posterior distributions. The former is computed from the class histogram. The latter can be approximated using the Bayes theorem and the non-parametric Parzen window estimator [8],[9]. Consequently, MIBIF provides a matrix of selected features $\boldsymbol{Y}_\varepsilon = \left[\boldsymbol{y}_d \in \mathbb{R}^N : \forall d : I_d > \varepsilon\right] \in \mathbb{R}^{N\times D'}$ holding the D' features sharing the most information with the labels.

C. Feature Embedding using Regularized Kernel Canonical Correlation Analysis

The Canonical Correlation Analysis (CCA) linearly embeds features from two spaces into a new single space with the maximum linear correlation. The kernel extension of CCA, termed Kernel CCA (KCCA), decodes nonlinear relationships from the input spaces by linearly combining the nonlinear kernel data embedding $\boldsymbol{z}_Y = \boldsymbol{\alpha}^\top \boldsymbol{K}_Y$ and $\boldsymbol{z}_l = \boldsymbol{\beta}^\top \boldsymbol{K}_l$, where the kernel matrices $\boldsymbol{K}_Y \in \mathbb{R}^{N\times N}$ and $\boldsymbol{K}_l \in \{0,1\}^{N\times N}$ contain the inner products in the Reproduced Hilbert Spaces $k_Y(n,m) = \exp(-\gamma\|\boldsymbol{y}_n - \boldsymbol{y}_m\|^2)$ and $k_l(n,m) = \delta(l_n - l_m)$, respectively. Vectors $\boldsymbol{\alpha} \in \mathbb{R}^N$ and $\boldsymbol{\beta} \in \mathbb{R}^N$ are the canonical basis of $\boldsymbol{K}_Y$ and $\boldsymbol{K}_l$ that maximize the correlation of embedded data [6]:

$$\begin{aligned} \max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \boldsymbol{z}_Y^\top \boldsymbol{z}_l &= \max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \boldsymbol{\alpha}^\top \bar{\boldsymbol{K}}_Y^\top \bar{\boldsymbol{K}}_l \boldsymbol{\beta} \\ s.t.\ \sqrt{\boldsymbol{\alpha}^\top \left(\bar{\boldsymbol{K}}_Y + \rho_Y \boldsymbol{I}\right)^2 \boldsymbol{\alpha}} &= 1 \\ \sqrt{\boldsymbol{\beta}^\top \left(\bar{\boldsymbol{K}}_l + \rho_l \boldsymbol{I}\right)^2 \boldsymbol{\beta}} &= 1 \end{aligned} \tag{3}$$

being $\bar{\boldsymbol{K}} = \left(\boldsymbol{I}_N - \frac{\mathbf{1}_N \mathbf{1}_N^\top}{N}\right)\boldsymbol{K}\left(\boldsymbol{I}_N - \frac{\mathbf{1}_N \mathbf{1}_N^\top}{N}\right)$ the centered kernel matrix, and $\rho_Y \in \mathbb{R}^+$ and $\rho_l \in \mathbb{R}^+$ regularization parameters. Using Lagrange multipliers, the problem in Equation (3) becomes the following generalized eigenvalue problem:

$$\begin{aligned} \boldsymbol{A}\begin{pmatrix}\boldsymbol{\alpha}\\ \boldsymbol{\beta}\end{pmatrix} &= \lambda \boldsymbol{B}\begin{pmatrix}\boldsymbol{\alpha}\\ \boldsymbol{\beta}\end{pmatrix} \\ \boldsymbol{A} &= \begin{pmatrix}\mathbf{0} & \bar{\boldsymbol{K}}_Y^\top \bar{\boldsymbol{K}}_l \\ \bar{\boldsymbol{K}}_l^\top \bar{\boldsymbol{K}}_Y & \mathbf{0}\end{pmatrix} \\ \boldsymbol{B} &= \begin{pmatrix}\left(\bar{\boldsymbol{K}}_Y + \rho_Y \boldsymbol{I}_N\right)^2 & \mathbf{0} \\ \mathbf{0} & \left(\bar{\boldsymbol{K}}_l + \rho_l \boldsymbol{I}_N\right)^2\end{pmatrix} \end{aligned} \tag{4}$$

with $\lambda \in \mathbb{R}^+$ as the Lagrange multiplier for both constraints as well as an eigenvalue, and the vector $[\boldsymbol{\alpha}^\top \boldsymbol{\beta}^\top]^\top$ its corresponding eigenvector. Therefore, the first N resulting eigenvectors map a multichannel trial into an N-dimensional vector space maximally correlated with EEG data labels, so favoring the class separability.

III. Experimental set-up and Results

A. Dataset and preprocessing

This work evaluates the proposed representation methodology using the publicly available BCI Competition IV dataset 2a, released by the Institute for Knowledge Discovery (Laboratory of Brain-Computer Interfaces) at the Graz University of Technology. The dataset records EEG signals from nine healthy subjects performing an instructed MI task per trial. Each trial lasts six seconds, starting with a fixed cross on the computer screen accompanied by a beep. From 2 to 3.25 seconds, an arrow instructs the movement to imagine, namely, move the left hand, right hand, both feet, or tongue. From trial to trial, a one-second blank screen allows a short break. Twenty-two Ag/AgCl electrodes record EEG signals distributed according to the international 10-20 montage system. Also, the dataset holds two subsets: A training one with 288 trials and a testing one with 288 trials. The former is considered for the learning and parameter tuning stages, while the latter for validating and reporting results.

Given that the MI segment lies within 2.5 to 4.5 seconds and that the recordings were downsampled to 250Hz, each trial becomes a time-series lasting $T = 500$ time instants over $C = 22$ channels. For the preprocessing, a bank of $B = 9$ five-ordered Butterworth bandpass filters without phase shift

divide the time-series into non-overlapped subbands of 4 Hz from 4 to 40 Hz [9]. For the feature extraction, the OVR-CSP extracts $M = 4$ spatial filters from each of the $L = 4$ classes, resulting in 16 values per band. Consequently, all values build a single feature vector of $D = 144$ features including spectral and spatial information [10].

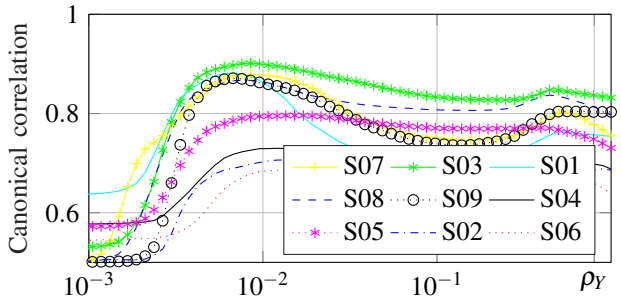
B. Hyperparameter tuning

Note that the proposed approach comprises four hyperparameters, namely, the scale for the RBF kernel γ, the number of selected features D', and the regularization parameters ρ_Y and ρ_l. The kernel scale is fixed to the inverse median pairwise distance among training samples $\boldsymbol{y}_n$. For the remaining, an exhaustive cross-validated grid search fixes the optimal parameters from the grid of $F' \in [5, 144]$, $\rho_Y \in [1 \times 10^{-5}, 1]$, and $\rho_l \in [1 \times 10^{-5}, 1]$. The grid search quantifies the performance of the hyperparameter set according to the mean Cohen's kappa coefficient along ten testing folds, considering a linear SVM classifier.

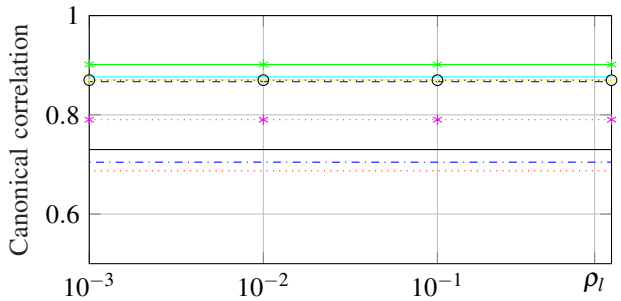
Figure 1 illustrates the ten fold average hyperparameter tuning curves for each subject in the dataset, that is, The canonical correlation versus the regularization parameters, and the Cohen's kappa score along the number of selected features. In general, the tuning curves evidence two subject groups, namely, S07, S03, S01, S08, S09 and S04, S05, S02, S06. The former achieve not only larger canonical correlations, but also large kappa scores. Particularly for the feature kernel regularization ρ_Y, all subjects reach the optimal parameter near the same value before 1×10^{-2}. Note that ρ_Y larger than the optimal decreases the correlation, since highly regularized matrices become diagonal. Regarding ρ_l, the target regularization lacks any effect on the correlation. Such a fact is due to the ill-conditioning of the target kernel matrix, as it has as many nonzero eigenvalues as classes exist. Further, the nonzero eigenvalues are the same because the target kernel is a delta function. Hence, the feature kernel regularization may be fixed for new subjects, while the target regularization can be any small value, so avoiding the exhaustive search for tuning them.

The tuning curve in Figure 1(c) proves that there exists a feature subset with the maximum kappa score. In general, less than 80 features reach the optimal performance, so that increasing the dimension D' hampers the performance. Besides, the performance of the first subjects group rapidly grows, demanding a few predictors. On the contrary, subjects in the second group, with lower performance, present slowly-growing and flatter curves, and usually requiring more features than the former subjects. Therefore, introducing the MIBIF stage not only reduces the redundant information but also correlates with the subject performance.

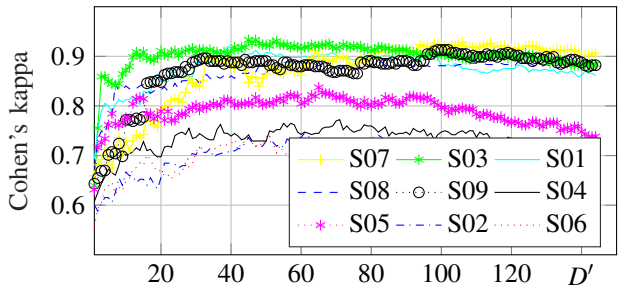
Figure 2 presents the relative number of features selected at each frequency band for each subject, where a value of one means that all features are chosen. Subjects are sorted in descending order according to their kappa score at the optimal hyperparameters. On the left side, subjects with the highest scores demand features that are quite located in a few bands, known as relevant for the MI paradigm



(a) Regularization of $\boldsymbol{K}_Y$



(b) Regularization of $\boldsymbol{K}_l$



(c) Number of selected features.

Fig. 1. Hyperparameter tuning curves. All curves are averaged along ten test folds.

(e.g. 8-12Hz, 20-24Hz, or 28-32Hz). On the right side, the low-performing subjects require information from the whole spectrum to better discriminate classes. And, on the other one subjects with high kappa values need fewer features for each band by choosing the relevant frequency information associated with the paradigm under study. Consequently, the proposed framework yields a feature subset with spectral interpretability related to the expected subject performance in the instructed BCI task.

C. Performance analysis

To graphically analyze the performance of the proposal, Table I color codes the confusion matrices for each subject. Note that green cells in columns agreeing predicted and target labels imply a high sensitivity. Also, red cells in columns mismatching the estimated and target labels identify low false-positive rates. The results evidence that left and hand movements are the best classified among the first subject group (the best-performing ones), which can be related to the handedness preference. On the contrary, the

TABLE I

CONFUSION MATRICES FOR EACH SUBJECT AVERAGED OVER TEN TEST FOLDS

Subject	Target Left (L)				Target Right (R)				Target Feet (F)				Target Tongue (T)			
	L	R	F	T	L	R	F	T	L	R	F	T	L	R	F	T
7	84,3	14,5	1,2	0	1,8	98,2	0	0	0	0	81,3	18,7	0	3,2	14,3	82,5
3	84,2	0	10,5	5,3	3,8	88,5	2,6	5,1	0	1,8	85,5	12,7	0	0	17,2	82,8
1	87,3	7,6	1,3	3,8	1,5	95,5	1,5	1,5	0	0	74	26	1,7	0	17,2	81
8	78,8	7,5	12,5	1,3	2,8	81,7	9,9	5,6	0	0	95,5	4,5	1,3	5,3	13,2	80,3
9	84,7	12,5	2,8	0	6,3	62,5	28,1	3,1	0	14,6	81,3	4,2	0	11,3	12,5	76,3
4	77,8	15,6	2,2	4,4	25	63,2	0	11,8	10,9	10,9	98,8	9,4	0	0	27,5	72,5
5	66,7	5,8	10,1	17,4	13,8	56,4	22,3	7,4	0	20	73,3	6,7	11,2	5,1	33,7	50
2	39,2	33,3	3,9	23,5	31,1	45,9	9,5	13,5	12,6	12,6	55,3	19,4	27,8	11,1	5,6	55,6
6	42	24	12	22	25	48,1	7,7	19,2	20	17,8	62,2	0	14,7	14,7	23,5	47,1
Average	71,7	**13,4**	6,3	8,6	**12,3**	71,1	9,1	7,5	4,8	**8,6**	78,6	**11,3**	6,3	5,6	**18,3**	69,8

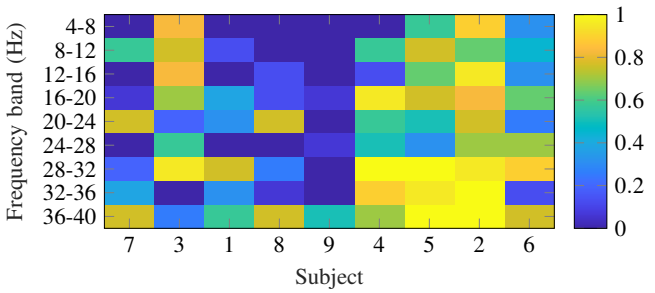


Fig. 2. Feature distribution along the frequency bands for each subject.

TABLE II

KAPPA SCORES ATTAINED BY COMPARED APPROACHES

Subj	Approach				
	He[5]	Kumar[2]	Nguyen[3]	Razi[4]	Ours
S07	54	80	82	81	91
S03	87	90	90	85	90
S01	69	87	74	78	90
S08	97	82	86	86	85
S09	45	76	80	88	84
S04	85	77	52	72	80
S05	78	62	35	67	72
S02	51	62	54	59	67
S06	42	53	37	57	66
Ave.	68±19	74±12	66±20	75±11	81±09
p-val	6.08%	0.17%	0.94%	0.50%	-

worst-performing subjects tend to mislabel the left hand (left arrow) and the right hand (right arrow). Overall, the approach confuses the feet (down) and the tongue (up arrow) no more than 33% of the time. Hence, the proposed KCCA-based representation suitably identifies the target class while exhibiting the most confusion between opposite MI tasks, which agrees with the MI physiology.

Table II compares the attained results against approaches in the state-of-the-art in the provided test subset, proving that the proposal outperforms in four out of nine subjects, two of them being usually low performing. In the remaining ones, the difference between the best result and the proposal is shorter than five percent, except for S08 that considerably improves within a Bayesian framework[5]. Overall, the KCCA reaches the highest grand average kappa score with the shortest deviation, becoming the most balanced approach among the compared ones. Besides, the paired t-test, quantifying the difference between our results and literature, proves a statistically significant improvement with p-values under 1% in three out the four compared works. Consequently, the introduced KCCA-based representation approach highlights the discriminant information among the studied MI conditions for the best and worst-performing subjects.

ACKNOWLEDGMENT

This work was carried out under the grants of the call for "Doctorado Nacional en Empresa - Convocatoria 758 de 2016" funded by Minciencias, the research project 111080763051 funded by minciencias, and the project 6-20-11 funded by the call 850 from Minciencias and the Vicerrectory for Research from Universidad Tecnológica de Pereira.

REFERENCES

[1] F. Lotte, L. Bougrain, A. Cichocki, M. Clerc, M. Congedo, A. Rakotomamonjy, and F. Yger, "A review of classification algorithms for EEG-based brain–computer interfaces: a 10 year update," *Journal of neural engineering*, vol. 15, no. 3, p. 031005, 2018.

[2] S. U. Kumar and H. H. Inbarani, "PSO-based feature selection and neighborhood rough set-based classification for BCI multiclass motor imagery task," *Neural Computing and Applications*, vol. 28, no. 11, pp. 3239–3258, 2017.

[3] T. Nguyen, I. Hettiarachchi, A. Khosravi, S. M. Salaken, A. Bhatti, and S. Nahavandi, "Multiclass EEG data classification using fuzzy systems," in *2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. IEEE, 2017, pp. 1–6.

[4] S. Razi, M. R. K. Mollaei, and J. Ghasemi, "A novel method for classification of BCI multi-class motor imagery task based on Dempster–Shafer theory," *Information Sciences*, vol. 484, pp. 14–26, 2019.

[5] L. He, D. Hu, M. Wan, Y. Wen, K. M. Von Deneen, and M. Zhou, "Common Bayesian network for classification of EEG-based multiclass motor imagery BCI," *IEEE Transactions on Systems, man, and cybernetics: systems*, vol. 46, no. 6, pp. 843–854, 2015.

[6] X. Zhuang, Z. Yang, and D. Cordes, "A technical review of canonical correlation analysis for neuroscience applications," *Human Brain Mapping*, vol. 41, no. 13, pp. 3807–3833, 2020.

[7] B. Blankertz, R. Tomioka, S. Lemm, M. Kawanabe, and K.-R. Muller, "Optimizing spatial filters for robust EEG single-trial analysis," *IEEE Signal processing magazine*, vol. 25, no. 1, pp. 41–56, 2007.

[8] J. C. Principe, *Information theoretic learning: Renyi's entropy and kernel perspectives*. Springer Science & Business Media, 2010.

[9] K. K. Ang and et al, "Filter bank common spatial pattern algorithm on BCI competition IV Datasets 2a and 2b," *Frontiers in neuroscience*, vol. 6, p. 39, 2012.

[10] L. F. Nicolas-Alonso and et al, "Adaptive Stacked Generalization for Multiclass Motor Imagery-based Brain Computer Interfaces," *IEEE Trans Neural Syst Rehabil Eng*, vol. 23, no. 4, pp. 702–712, 2015.