A Simple Subject Independent Channel Selection in EEG for Motor Imagery Task

Raghav Dev, Sandeep Kumar, and Tapan Kumar Gandhi, *Member, IEEE*

***Abstract*— The classification of Motor Imagery (MI) tasks through EEG is valuable in brain-computer interfacing. Selecting EEG channels for MI task classification is well discussed problem and is challenging due to its combinatorial nature. Therefore to narrow down the search field certain assumption is generally made. However such assumptions bring disadvantages if they are not reasonably backed by established literature. This paper introduces a subject-independent EEG channel selection approach consisting of two stages. First, we employ channel selection by minimizing divergence between channels. We rank channels based on their divergence from a reference channel Cz. We hypothesize that channels less divergent from Cz are more relevant for MI task classification. In the second stage, we employ a three-step feature selection and classification model to evaluate the selected channels for MI task classification accuracy. This stage consists of a bandpass Type II Chebyshev filter of 4-40 Hz cutoff frequency, followed by a 2-component CSP filter for dimensionality reduction followed by three classifiers viz. SVM, 1-NN and 5-NN. We assess this methodology on two publicly available datasets viz. PhysioNet MI EEG dataset and BCI Competition III IVa dataset. Empirical results support our hypothesis that divergence between channels and a reference channel Cz can be used as a ranking measure for channel selection. Empirical comparison implies that the subject-independent approach performs better than classical models such as CSP Rank, fishers rank, and normalized mutual information significantly.**

***Index Terms*— Brain-computer interface, Channel Selection, electroencephalography, KL-Divergence.**

I. Introduction

ELECTROENCEPHALOGRAPHY (EEG) is a cost-effective, easy to use, and rich with temporal dynamics modality to study the human brain [1]. EEG channel selection for various cases such as motor imagery (MI) tasks, epileptic seizure detection, source localization, and emotion recognition is a well-researched topic and crucial for brain-computer interface applications [2]–[5]. Channel selection is essentially an NP-hard problem since it is combinatorial [6]. Therefore we need priors and assumptions to narrow down the search.

One common assumption in channel selection is that if a particular channel contributes to increasing classification accuracy more than another, it is better and more relevant [7]–[10]. However, there is no theoretical reason to believe that just because a channel doesn't contribute to classification accuracy, it is redundant or noisy. It might be useful for generalizability. Contrary to this, the inclusion of noise in signals has the potential to enhance the accuracy of classification [11]. Another commonly made assumption is that an effective set of electrodes should be in close proximity to certain crucial electrodes [12], [13]. However, Helmholtz theorem [14] and analysis on electrode sensitivity [15], [16] based on it, state otherwise because of the shunting effect. Therefore, we think, the priors and assumptions to narrow down the search are at the heart of this problem and must be based on established notions.

Various other (than the distance between the electrodes) distance-based measures have been widely used for channel selection [17]. Examples such as normalized mutual information (NMI) among channels [18] and mutual information between channels and labels [19]–[22] have been used as the distance parameter to rank and select channels for the classification of the tasks. However, selected channels using mutual information based models are dependent on labels and therefore lack generalizability. Similarly, in [18], researchers used NMI to find the divergence between the channels, however, their channel selection algorithm is highly dependent on tasks. The other issue with existing NMI based models [18] is that it has been evaluated between every couple of channels. But if two channels are noisy, the NMI between those will be high too, and hence NMI between all couple of channels is not the best way to deal with channel selection. Similar is the issue with correlation based channel selection proposed in [23]. In this work, we have attempted to address it.

Classical methods such as recursive feature elimination [24] are largely subject and task specific, since, the channel selection depends on the classifier's loss function (margin of SVM). Similarly, spatial filters based channel selection models [25]–[29] are task dependent since filters have to be derived for each label. Although very few, there are studies that have attempted the subject-independent approach for channel selection [30], [31]. However, experiments in [30] are limited, and [31] is the deep learning model hence the channel selection is dependent on the labels for training the model. Similar is the case with the other deep learning based channel selection models that they are either subject dependent or the task dependent [10], [32]. The critical challenge faced by these channel selection models that beg for a solution is that their selection criteria are dominantly dependent on the specific subject and task at hand. This raises concerns regarding the broader applicability of these models, prompting questions about their ability to

All the authors are with the Department of Electrical Engineering, Indian Institute of Technology Delhi, New Delhi, 110016, India. Corresponding Author: Raghav Dev, rdevm23@gmail.com.

generalize effectively to larger and more diverse datasets. We hypothesize that the mutual information between channels can however be strong scaffolds for the subject-independent channel selection. To the best of our knowledge, there has not been any attempt to devise such a subject-independent model.

An additional unsettled issue that got our attention during this study is that several works are showing that a subset of the electrodes of EEG can outperform 'all channels' in classifying the stimulus [33], [34]. On the other hand researchers in [9], [32], empirically show that a subset of the channels is not significantly overperforming the set of all the channels for classifying the MI task. However, the dataset used in [9] has only 22 EEG channels which is smaller than other studies showing it does outperform. In this study, we find that there is no evidence to believe that a subset of channels can outperform all the channels.

In this work, we hypothesize that a subset of channels is good if a *measure of divergence* among them is as small as possible. Furthermore, by minimizing the divergence the selected channels shall be fairly subject and task-independent. To accomplish that, we find the KL divergence between a reference channel Cz and all other channels at all time incidents. Afterwards, the expectation of the KL divergence is considered as the rank of the channel. The smaller the rank, the greater the importance of the channels. In addition, we find there is no significant gap in decoding accuracy to believe that a subset of channels can outperform all the channels.

The rest of the paper is organized in three sections. In Section II, methodology has been described. We have set the notations, described the channel selection algorithm, and finally feature selection and classifier pool have been explained in this section. In Section III, we have drawn the results and discussions on the proposed method. Section IV concludes the paper.

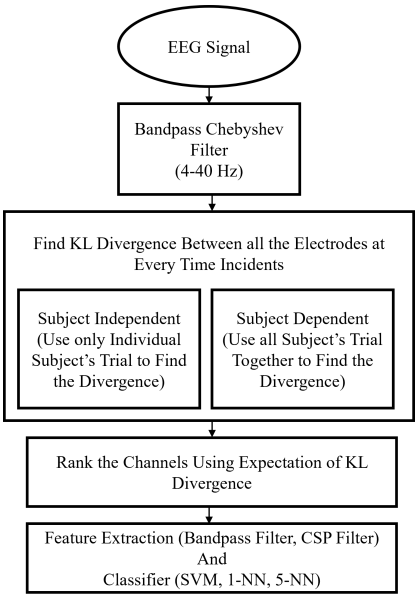

Fig. 1. This figure shows the flow of the methodology. The EEG signal will be first filtered using a bandpass filter. Afterward, channel selection is done followed by the classification stage. In channel selection methods we use KL divergence over all the trials to rank the channels. Depending on how this divergence score is evaluated, the proposed method is subject-dependent or independent to it. Finally, we used minimal level of feature reduction using CSP filters, before evaluating our method based on classification accuracy.

II. Methodology

A. Notations and Conventions

The general convention is that the bold faced capital letters are matrices and small letters are vectors, while italic capital letters are mostly probabilities and random variables with a few exceptions that shall be clear with context. The italic small letters are scalar variables. The indices are represented with small letters that run from 1 to the capital letters. For example, $t \in \{1, 2, \cdots, T\}$ and $s \in \{1, 2, \cdots, S\}$.

B. Background and Formulation

The basic idea is to rank each channel based on how minimally it is divergent from other channels in the group of selected channels. However, if we find the divergence between every possible channel pair, the noisy channel could be selected. One idea to narrow down the search is to find the divergence of channels from one reference channel and we can fairly assume that if two channels are less divergent from the reference channel, those two channels are themselves less divergent from each other.

The issue is which channel should be considered as a reference channel. A good reference channel would be that which has the least impedance and noise, but, there is little to no literature dedicated to channel-wise noise and impedance analysis. However, in almost all existing channel selection algorithms across all kinds of tasks, Cz has been consistently common as one of the first few selected channels [8], [13], [32]. Another reason for choosing it as a reference is that it is often considered as reference channel in bipolar EEG recordings. Since impedance is measured using this reference channel during the setup, usually people place such electrodes with extreme caution and make sure its contact with the scalp well [35]. Therefore, Cz will be a good choice as the reference channel. Henceforth, a sense of divergence will be evaluated from this reference for each channel and then channels will be ranked based on that. Once the channels are selected, filtering is done followed by classification. Fig. 1 summarises this flow.

To formulate the problem of finding the divergence among channels, let us set the notations. Let $\mathbf{X}_y^{(t,s)} \in \mathbb{R}^{N \times M}$ be the EEG signal of t^{th} trial of s^{th} subject recorded during y^{th} stimulus, where N is number of channels and M is the number of samples of a trial. We can ignore y for the channel selection part because the proposed model is independent of it. Hence, it can be described as $\mathbf{X}^{(t,s)} = [\mathbf{x}_1^{(t,s)}, \mathbf{x}_2^{(t,s)}, \mathbf{x}_3^{(t,s)} \cdots \mathbf{x}_m^{(t,s)} \cdots \mathbf{x}_M^{(t,s)}]$, where, $\mathbf{x}_m^{(t,s)} = [x_{1m}^{(t,s)}, x_{2m}^{(t,s)}, x_{3m}^{(t,s)} \cdots x_{nm}^{(t,s)} \cdots x_{Nm}^{(t,s)}]'$, where, $x_{nm}^{(t,s)}$ is EEG signal of n^{th} channel and m^{th} sample. Let the underlined probability space of the EEG signal is $(\Omega, \mathcal{F}, P)$. and $A_{nm}^s(\omega_{nm}^s)\ s.t.\ \omega_{nm}^s \in \Omega$ is the random variable representing the EEG signal of n^{th} channel and m^{th} sample of s^{th} subject such that,

$$x_{nm}^{(t,s)} = A_{nm}^s(\omega_{nm}^s = \omega_{nm}^{(t,s)}), \tag{1}$$

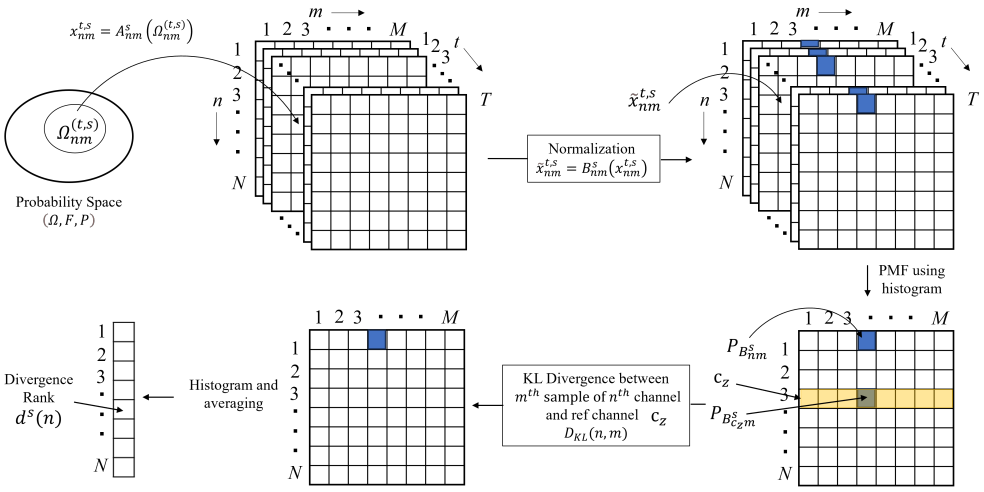

Fig. 2. In this figure, the methodology for channel selection has been elaborated pictorially. First, we normalized the data, followed by evaluating the PMFs of each sample and channel using the histogram of all the trail data. Thereafter, we take the KL divergence between a channel and the reference channel and finally average the KL divergence over all the samples to find the divergence rank for the channel.

where, $\omega_{nm}^{(t,s)} \subset \Omega$ is the probabilistic event that generate the EEG signal at t^{th} trial.

EEG signals often contain spurious signals. Such spurious signals can affect the divergence between the channels, significantly. Therefore to attenuate them we used a logarithmic function to normalize the signal. The normalized signal $\tilde{x}_{nm}^{(t,s)} \in \mathbb{R}^+$ can be evaluated as,

$$\tilde{x}_{nm}^{(t,s)} = \log\left\{1 + \mathcal{N}\left(x_{nm}^{(t,s)}\right)\right\}, \tag{2}$$

where. $\mathcal{N}(.)$ is $0-1$ normalization function along all the samples M. The $0-1$ normalization would help us to prevent the saturation of the signal $\tilde{x}$ because of the log function. The random variable $B_{nm}^s : x_{nm}^s \to \mathbb{R}^+$ representing the normalized signal shall therefore be defined such that $\tilde{x}_{nm}^{(t,s)} = B_{nm}^s(x_{nm}^s = x_{nm}^{(t,s)})$.

Next, to find the divergence among the channels we have to estimate the probability mass functions (PMFs) over the random variables B_{nm}^s corresponding to the normalized EEG signal $\tilde{x}_{nm}^{(t,s)}$ for all possible n, m and s. Therefore we will be estimating a total of $N \times M \times S$ number of PMFs. Let $P_{B_{nm}^s}$ be the PMF of the random variable B_{nm}^s. To estimate the PMF $P_{B_{nm}^s}$ we have T number of the trail of EEG signal for each possible n, m, and s. The number of trials T is not very large, hence, a non-parametric model will be more appropriate to estimate the probabilities. It will not be a significant computational burden. Hence, we define PMF $P_{B_{nm}^s}(x_{nm}^s)$ over random variable B_{nm}^s using histogram with H number of bins. Therefore the half bin size (centers of histogram) $c = \log(2)/2H$, since maximum value B_{nm}^s can take is $\log(2)$. Hence, the PMF $P_{B_{nm}^s}(x_{nm}^s)$ is defined as,

$$P_{B_{nm}^s}(B_{nm}^s = \tilde{x}_{nm}^{(t,s)}) = \frac{1}{2T}\left(1 + \sum_{i=1}^{T} \phi\left(h \cdot c - \tilde{x}_{nm}^{(i,s)}\right)\right), \tag{3}$$

where,

$$\phi(x) = \begin{cases} 1 & |x| \leq c, \\ 0 & \text{otherwise}, \end{cases} \tag{4}$$

and,

$$h = \underset{h \in \{1,\cdots,H\}}{\operatorname{argmin}} \|\tilde{x}_{nm}^{(t,s)}) - h \cdot c\|. \tag{5}$$

Eqs. 3-5 defines the histogram over the EEG signal $\tilde{\mathbf{x}}_{nm}^s = [\tilde{x}_{nm}^{(1,s)}, \tilde{x}_{nm}^{(2,s)}, \tilde{x}_{nm}^{(3,s)}, \cdots, \tilde{x}_{nm}^{(t,s)} \cdots, \tilde{x}_{nm}^{(T,s)}]$ to estimate the PMF $P_{B_{nm}^s}$ at for each channels n, samples m of each subject s. Eq. 5 find the closest bin index of the signal and eq. 3 count the number of EEG trials in that bin. This enables us to estimate all the PMFs. We will use these PMFs to find the divergence between a channel and the reference channel. The divergence can be subject-dependent and independent and hence the channel selection algorithm based on it has been described in the following subsections as subject-dependent and subject-independent.

C. Subject dependent Channel Selection

Mutual information, divergence, and several other derived measures have been used in literature to rank the EEG channels for channel selection or feature selection [17]–[19]. However, they face limitations as discussed in the first section. In this work, we shall use the expectation of KL divergence between the channel and reference channel to find its rank. Let $D_{KL}^s(n,m)$ be the KL divergence between the n^{th} channel and Cz for the m^{th} sample and s^{th} subject. It is evaluated as,

$$D_{KL}^s(n,m) = \sum_{t=1}^{T} P_{B_n^s m}(\tilde{x}_{nm}^{(t,s)}) \log\left(\frac{P_{B_{nm}^s}(\tilde{x}_{nm}^{(t,s)})}{P_{B_{c_z m}^s}(\tilde{x}_{c_z m}^{(t,s)})}\right), \tag{6}$$

where, $P_{B_{nm}^s}$ is probability mass function of n^{th} channel, m^{th} sample and s^{th} subject as defined in eq. 3. The ranking of the channel will be done based on an average of the KL divergence over all the M samples. To avoid the spurious value of $D_{KL}^s(n,m)$ affecting the average, we find the histogram

and evaluate the sum of the centers weighted with their counts as the divergence metric between the channel and reference channel. To find the Let $d^s(n)$ be the divergence between n^{th} channel and the reference channel Cz for the s^{th} subject. It is defined as,

$$d^s(n) = \sum_{h=1}^{H} w(h) \cdot c(h), \tag{7}$$

where $w(h)$ and $c(h)$ are histogram counts and centers of M KL divergences between n^{th} channel and Cz. The histogram has been evaluated over samples (time incidents).

Using eqs. 6–7, we can evaluate a divergence score for every channel of every subject. Finally, the channels will be selected based on how small this divergence score $d^s(n)$ of the channel is. The channel set $C_N = \{C_1, C_2, C_3, \cdots, C_n, \cdots, C_N\}$ is sorted such that $d^s(n) \leq d^s(n+1)\ \forall n$. If we need to select best K channels then the selected channel set $C_K = \{C_k |\ k \leq K \leq N\}$. The method has been graphically explained in fig 2 for more clarity over the dimensions of the data at every step.

D. Subject Independent Channel Selection

There are two ways to make the channel selection method subject-independent. One is to take the average over the divergence score $d^s(n)$ across the subjects (say, average ranks). Let d^{AR} is divergence scores of the average rank method, then it can be evaluated as

$$d^{\text{AR}}(n) = \frac{1}{S} \sum_{s=1}^{S} d^s(n). \tag{8}$$

Another way to find the divergence over all the data of all the subjects (say, subject-independent). That can be done by treating all subjects as a single subject and taking the KL divergence over all the trials of all the subjects. Basically, in formulation the number of trials will be $T \times S$ lets call it $\bar{T}$ and its index $\bar{t}$. A random variable $\bar{A}$ represent the EEG signal such that, $x^{\bar{t}}_{nm} = \bar{A}_{nm}(\omega_{nm} = \omega^{\bar{t}}_{nm})$, where, $\omega^{\bar{t}}_{nm} \in \Omega$, is the event that generate the EEG signal at $\bar{t}^{\text{th}}$ trial. Similarly, let $\bar{B}_{nm}$ be the random variable representing normalized signal $\tilde{x}^{\bar{t}}_{nm}$. Let PMF over the random variable $\bar{B}_{nm}$ is $P_{\bar{B}_{nm}}(\bar{B}_{nm}$ evaluated similarly, except for all the trials $\bar{T}$, then, subject independent KL divergence $D_{KL}(n, m)$ between a channel and reference channel is evaluated as,

$$D_{KL}(n,m) = \sum_{\bar{t}=1}^{\bar{T}} P_{\bar{B}_{nm}}(\tilde{x}^{\bar{t}}_{nm}) \log \left(\frac{P_{\bar{B}_{nm}}(\tilde{x}^{\bar{t}}_{nm})}{P_{\bar{B}_{c_z m}}(\tilde{x}^{\bar{t}}_{c_z m})} \right), \tag{9}$$

therefore, the divergence rank $d^{\text{INDP}}(n)$ for subject independent method for n^{th} channel is,

$$d^{\text{INDP}}(n) = \sum_{h=1}^{H} w(h) \cdot c(h), \tag{10}$$

where, $w(h)$ and $c(h)$ are histogram counts and centers of KL divergence $D_{KL}(n, m)$ between n^{th} channel and Cz, similar to as in eq. 7 for the subject dependent version. Finally, we can rank the channels as per the value of $d^{\text{INDP}}(n)$ and $d^{\text{AR}}(n)$ and select the channels with minimum divergence score.

E. Inclusion of C3, C4 and Cz

The existing literature shows that 3C (C3, C4, and Cz) channels are the best choice if you want to choose just three channels and are most frequently selected [8], [13], [24], [36]. However, we did not find any convincing argument supporting that except for empirical evidence that suggests strongly as such. We compared 3C with randomly selecting channels and it outperforms all by a great margin (can be seen in fig. 9 & 10). The probable explanation could be lying in the impedance distribution of the channels and scalp, however, we find little to no literature on impedance distribution on electrodes. Nevertheless, 3C is at least empirically the best 3 channel that can be selected and hence, we included it as the first three channels for all the proposed methods and from 4^{th} to the number of selected channels, we used divergence measure.

F. Feature Extraction

The goal of the work was to develop a channel selection algorithm and test it using the most basic feature extraction methods for EEG and finally classification. In literature [2], the most common and basic feature extraction pipeline has three stages viz. a) filterbank with overlapping frequency bands, b) common spatial pattern (CSP) Filtering, c) feature Reduction (selection) using mutual information (between features and labels). However, through experiments, we did not find any significant effect of filterbank over a single filter. Moreover, with a minimal number of CSP components, we do not require the feature reduction stage at all. In our view, the feature reduction stage is relevant only when there is filterbank rather than just one filter.

Therefore we have used two-stage simple feature extraction stages. The most commonly used frequency band used in MI classification using EEG is $4-40$ Hz [38] therefore we used 10^{th} order Type II Chebyshev filter with cutoff frequencies of $4-40$ Hz. Secondly, we used the CSP filter with 2 components to keep the model overall computationally efficient. For the implementation of the Type II Chebyshev filter and CSP filter stages, we replicated the part of the work presented in [9] with changes for our suitability.

G. Classifier Pool

We used three classifiers to perform the experiments viz. a) support vector machines (SVM), b) 1-nearest neighbor (1-NN), and 3) 5-NN. The reason to chose 1-NN is that it is proven that it has a theoretical bound on its error, which is twice as much of Bay's classifier [39]. Therefore, to have a simple check on SVM and 5-NN to be not overfit, we have 1-NN.

The training protocol is that, first, the datasets are divided into 80:20 training-testing sets. Afterward, the training dataset is split into 10 folds. Then SVM, 1-NN, and 5-NN are trained and validated in the 10-fold cross-validation manner for each subject. The best model is selected that has maximum accuracy over the validation set.

TABLE I

THIS TABLE SHOWS THE EMPIRICAL RESULTS OF THE SUBJECT-INDEPENDENT METHODS ON THE BCICIVA DATATSET [37]. RESULTS SHOW THAT THE PROPOSED METHOD OUTPERFORM CLASSICAL METHODS CSP RANK [26] AND NMI [18] SIGNIFICANTLY. IT PERFORMS **15.21%** MORE THAN 3CS AND JUST **2.91%** LESS THAN *All Channels* ACCURACY WITH AS FEW AS **20** CHANNELS OUT OF **118**. RESULTS IN BOLD ARE THE BEST AMONG THE 3CS, CSP RANK, NMI, AND SUB INDP. ALL CHANNEL ACCURACY HAS BEEN PROVIDED FOR REFERENCE.

Subjects	C3, C4 & Cz			CSP Rank [26]			NMI [18]			Sub Indp (20 Channels)			All Channels		
	SVM	1-NN	5-NN	SVM	1-NN	5-NN	SVM	1-NN	5-NN	SVM	1-NN	5-NN	SVM	1-NN	5-NN
aa	64.53	64.14	64.14	83.85	74.89	74.89	**92.04**	**82.07**	**80.68**	88.64	83.46	83.46	89.64	88.25	86.86
al	85.75	85.75	88.02	85.75	84.50	83.68	90.29	84.30	88.02	**93.19**	**87.20**	**87.20**	96.28	96.28	97.73
av	66.31	51.04	56.83	66.53	59.92	63.64	68.61	64.26	66.53	**76.24**	**71.08**	**67.36**	77.89	70.45	68.18
aw	65.51	57.53	57.53	85.99	87.50	82.33	91.82	84.69	90.51	**96.99**	**93.97**	**93.97**	98.49	95.68	93.53
ay	89.59	78.70	76.67	85.53	78.22	80.09	83.50	85.37	85.37	**92.69**	**96.42**	**92.69**	100.00	96.58	94.71
Average	74.34	67.43	68.64	81.53	77.01	76.92	85.25	80.14	82.22	**89.55**	**86.43**	**84.94**	92.46	89.45	88.20
STD	10.97	12.96	12.03	7.54	9.64	7.29	8.88	8.02	8.49	**7.16**	**8.97**	**9.57**	8.10	9.99	10.62

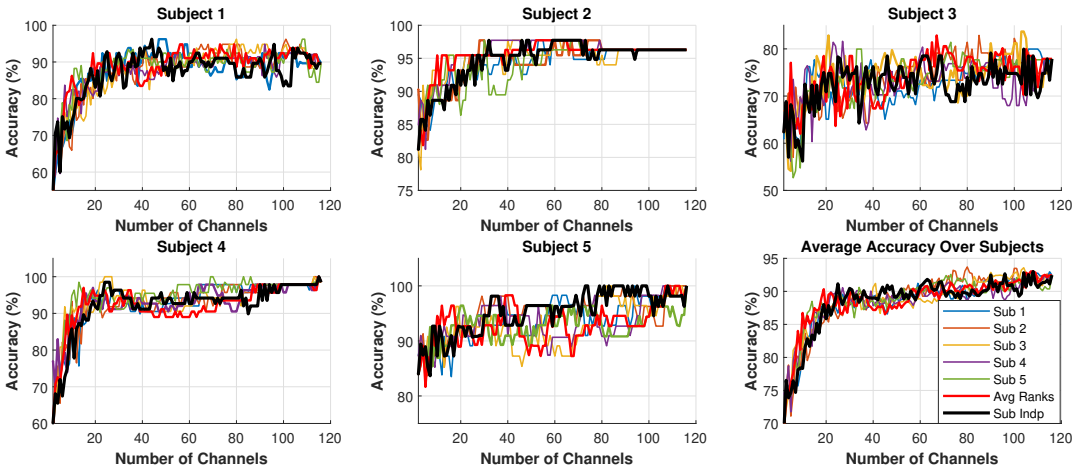

Fig. 3. This plot shows how classification accuracy varies with different numbers of channels on the BCICIVa dataset [37] without using bandpass filter stage for channel selection. Seven channel selection methods have been compared in this figure. The subject's (eg. Sub 1) methods are methods where channel ranks are evaluated using divergence of the dataset of the individual subjects, while average rank (Avg Rank) is just the average of the individual subject's ranks. In Subject independent (Sub Indp) method, we used all the trials together of all the subjects to find the ranks of the channels. It can be seen that *Avg Rank* and *Sub Indp* perform better than subject-dependent models most of the time with the smaller number of channels, however, subject-dependent methods perform better with $\mathbf{80-100}$ channels.

III. EXPERIMENTS AND RESULTS

A. Dataset Description

We used two publicly available datasets viz BCI competition III, IVa Motor Imagery dataset (BCICIVa) [37], [40] and PhysioNet EEG Motor Moverment/Imagery dataset (PhysioNet) [41], [42]. The former is a rich dataset because it was recorded on 118 electrodes and there are a total of 280 trials per subject. It is useful to judge motor imagery classification models because of its high-density electrodes and a suitable number of trials for training classification models. On the other hand, the latter one is rich because the number of subjects is 109, although electrode density is average (64), hence it will be useful for judging a model that claims to be subject-independent.

1) BCI competition III dataset IVa: The dataset [37], [40] is a very rich dataset for channel selection since it has a large number of channels 118. The dataset was recorded at 1000 Hz. The data was recorded for five healthy subjects while visual cues of two classes viz. right hand and right foot (originally for three cues) were presented for 3.5 seconds, with random intervals for a total of 280 trials per subject. For our purposes, we clipped the data at the onset sample till 3500 samples from the onset.

2) PhysioNet: In this dataset [41], [42], 64 channel EEG was recorded for 109 subjects performing various MI tasks. Originally, there were 14 runs, out of which we used 3 runs 4^{th}, 8^{th}, and 12^{th} corresponding to *Task 2 (imagine opening and closing left or right fist)*. The publicly available version of the dataset is at 160Hz with each trial duration of 4.1 seconds. The number of trials per run is 30 out of which 15 is of MI stimulus and 15 is of resting. We excluded 15 subjects because of the data incompleteness. The excluding criteria are - a) the number of events in any of the three runs is less than 30, b) if any of the trials is recorded duration is less than 4.1s then the whole subject data is excluded, and c) if the number of samples in any run is less than 19200 (160 Hz $\times$4 seconds $\times$30 runs) the subject has been discarded. Based on these criteria data of subjects 34, 37, 41, 51, 64, 72, 73, 74, 76, 88, 89, 92, 100, 102, and 104 was discarded. So, we used data of 94 subjects, 3 runs each with 15 trials of 640 samples (4 seconds from onset of the stimulus).

B. Parameters Tuning and other Details for Replicability

There are several parameters to be tuned. To start with, we decided $H = 10$ to devise the histogram in eq 8. Therefore

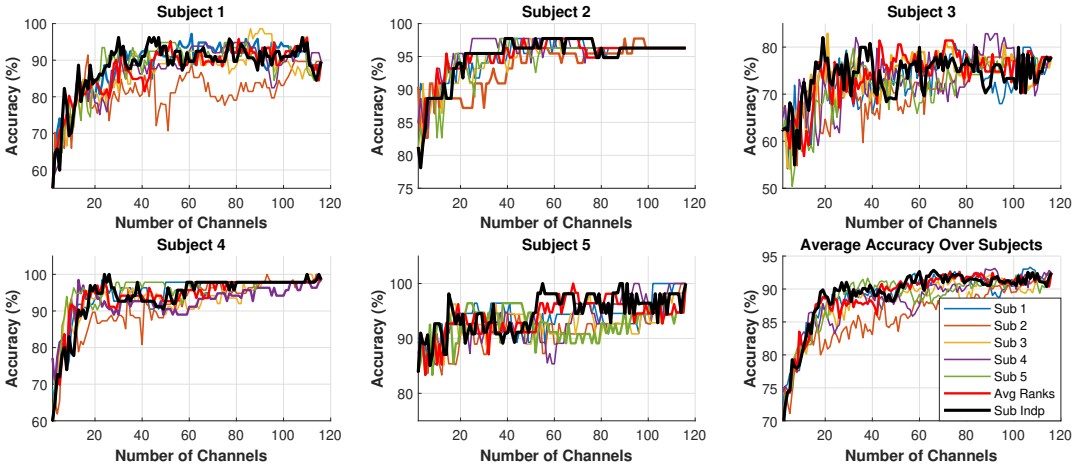

Fig. 4. In this plot, how classification accuracy varies with different numbers of channels on the BCICIVa dataset [37] using bandpass filter stage for channel selection have been shown. The most important point to notice in this plot is that the accuracy plot is more stable and consistent than that of *without filters* in fig. 3. Most clearly it can be seen with plots of *Sub Indp* plots of Subject 2 & 3. However, overall with or without filters, *Sub Indp* method performs the best among all for the smaller to average number of channels.

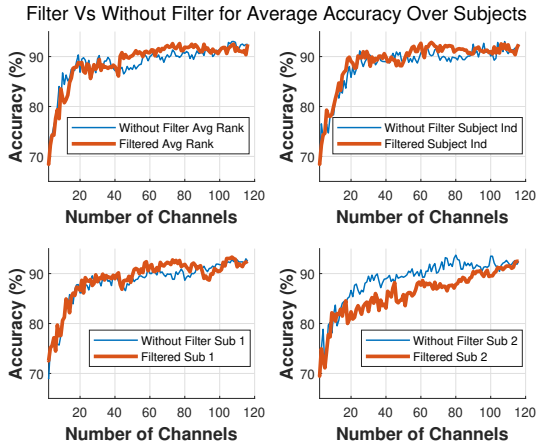

Fig. 5. The impact of bandpass filtering on the selection of channels on the assessment of classification accuracy on the BCICIVa dataset [37]. The plots reveal that filtering has a discernible impact on classification accuracy, influencing a smoother and more consistent change in accuracy with the number of channels compared to cases without filters. However, the difference between the accuracy is not visibly significant. There are up to **20** channel differences in the selected channels contributing to variations in the accuracy.

the size of $w(h)$ will be $H \times 1$. We find the PMF using MATLAB `hist()` with 10 bins. We added 1 to all the counts while finding the PMF so that in case if count in any of the bins is zero, it does not create `NaN` or `inf` in KL divergence since it has the log function. In bandpass filtering, we used $4 - 40$ Hz. The order of the filter was 10 since we found it computationally efficient enough for our purposes. For CSP filtering, we used two components of the CSP filter corresponding to the highest and lowest eigenvalues both corresponding to both classes, hence in total 8 components, similar as in [9]. Hence at the input end of the classifiers, the feature dimension is 8.

C. Empirical Results

Assessment of proposed methods – subject-dependent (Sub Dep), average ranks (Avg Ranks), and subject-independent (Sub Indp) based on MI classification accuracy on both datasets have been drawn in TABLEs I & II and figs. 3, 4, & 6. We compared our results with three existing methods (its subject-independent version) CSP Rank [26], Fishers's score (FS) [7], [43]–[45] and NMI [18]. The comparison results have been shown in figs. 9 & 10 for both datasets. Comparison of the results has been drawn in detail in Section III-D.5.

TABLEs I & II shows that the proposed subject-independent method outperforms the existing models. It can be seen that from the TABLE I, on the BCICIVa dataset, as few as 20 channels (approx. 17% of all channels) perform almost 15.21% more than 3C and just 2.91% less than *all channels*. Similarly, for the PhysioNet dataset, the proposed subject-independent outperforms the existing channel selection models as can be seen in the TABLE II.

How the classification accuracy varies with various numbers of channels of all the proposed methods has been shown in figs. 3, 4, & 6. Fig. 3 shows how channel selection worked in the absence of bandpass filtering in the channel selection stage. The observation reveals that channel selection is effective even without filters; however, the method demonstrates increased robustness when filters are incorporated, as can be seen in figs. 4 & 5. Figs. 4 & 6 shows that the subject-independent method performs most consistently for most of the subjects. Another important conclusion that can be drawn from the results in figs. 4, & 6 is that the classification accuracy almost monotonically increases with increasing number of channels. This is in contrast with the studies in several state-of-the-art works on this topic that a set of fewer channels can outperform all the channels in classification [13].

D. Discussion

In this section, we have discussed how the various aspects of channel selection affect the proposed methods, how the

TABLE II

IN THIS TABLE WE SHOW EMPIRICAL RESULTS OF THE PROPOSED METHOD ON THE PHYSIONET DATASET [41]. RESULTS ARE COMPARED WITH RANDOMLY SELECTED CHANNELS AVERAGED OVER **10** TIMES (SAY, RAND AVG), CSP RANK [26], AND NMI [18] FOR DIFFERENT NUMBERS OF CHANNELS. THE SHOWN RESULTS ARE AVERAGED OVER ALL THE **94** SUBJECTS. RESULTS IN BOLD ARE THE BEST SVM CLASSIFICATION ACCURACY AMONG ALL THE METHODS. 1-NN AND 5-NN ARE FOR REFERENCE. THE RESULTS SHOW THAT THE PROPOSED MODEL OUTPERFORMS THE EXISTING MODELS MOST OF THE TIME.

Methods	10 Channels			25 Channels			40 Channels			55 Channels		
	SVM	1-NN	5-NN	SVM	1-NN	5-NN	SVM	1-NN	5-NN	SVM	1-NN	5-NN
Random Avg	66.81	69.16	69.12	79.01	81.58	82.20	84.46	87.42	87.48	87.48	89.94	89.90
CSP Rank [26]	66.93	69.42	70.16	80.83	83.14	83.33	**87.36**	87.53	88.14	87.58	89.58	89.53
FS [7]	65.20	70.26	70.95	80.76	84.20	85.02	86.97	89.93	90.51	89.02	91.67	91.92
NMI [18]	67.44	70.66	70.79	81.09	83.83	84.28	86.72	90.78	91.00	88.67	91.59	92.09
Sub Indp	**67.61**	68.73	68.54	**81.53**	81.80	82.18	87.03	89.95	89.81	**90.12**	91.30	91.35

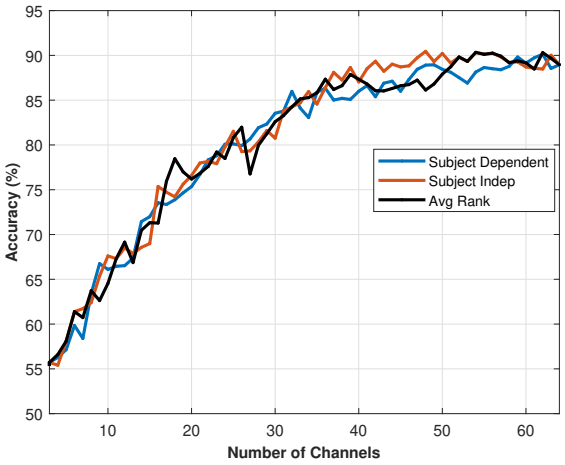

Fig. 6. This graph illustrates how the accuracy of classification changes with varying numbers of channels on the PhysioNet dataset [41]. The classification accuracy is averaged over **94** subjects. The shown three methods are subject-dependent, average rank, and subject-independent. Again, it can be seen that on average subject-independent methods perform better (at least as good as) than subject-dependent methods, with the additional advantage of it is more generalizable.

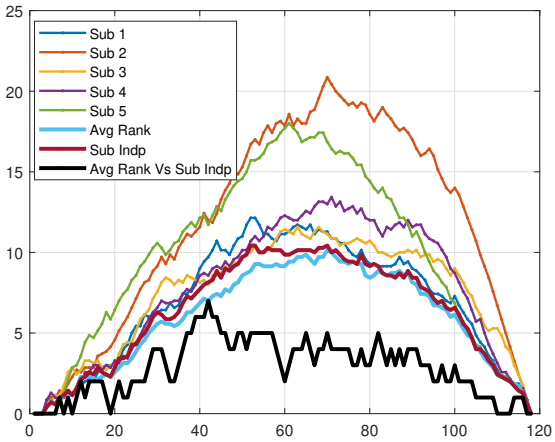

Fig. 7. In this plot, we examined the divergence among various proposed channel selection models based on the number of distinct channels selected on the BCICIVa dataset [37]. The associated plot for the method illustrates the average disparity in selected channels compared to all other methods. For instance, *Sub 1* represents the average difference from other methods such as *Sub 2, 3, 4,...* based on the selected channels. It shows that both subject-independent methods are close to each other up to a great extent.

proposed model is subject-independent, and compared the results with the existing models.

1) Effect of Bandpass Filter on Channel Selection: We analyzed how channel selection is affected by whether we use signal filtering or not. How much KL divergence based ranking of the channel affected by the filtering? In most EEG processing, the bandpass filter(s) are used which is generally computationally costly, to suppress the artifacts and power line interference. However, we find that the classification accuracy is not heavily affected because of filtering. Fig. 5, shows that difference between accuracy because of filtering is marginal however crucial. Most of the time accuracy is marginally better than that of without filters, however, an important use we can see from the plot is that, with filters, the accuracy curve seems more consistent and hence seems generalizable.

2) Subject Independence Analysis From Selected Channel Point of View: The main goal of the study was to devise a truly subject-independent channel selection method. We ranked the channels based on a divergence measure. We analyzed how much the subject-independent channel selection method shares the selected channels with the subject-dependent version of it. Figs. 7 & 8 graphically shows that the subject-independent methods share most of the channels with subject-dependent methods. Fig. 7 refers to the dataset BCICIVa [37], where we can see that, at its peak the number of different channels on average is approximately 15% for the subject independent method. Furthermore, this difference is almost half of it with the average rank method (black plot). Similar inference can be made from the fig. 8 for the PhisioNet dataset [41], where the subject-independent method is different from other methods maximum up to 15% averaged over 94 subjects. It shows that the proposed subject-independent method is generalizable up to a great extent.

3) Subject Independence Analysis From Accuracy Point of View: Inferences about how the subject-independent method is independent of subjects can be made from the figs. 4 & 6 for the datasets BCICIVa [37] and PhisioNet [41] respectively. In the first dataset since the number of subjects is just 5 it is easy to visualize that for all the subjects, *Sub Indp* method performs as good as subject-dependent methods for most cases and better than them at the smaller number of

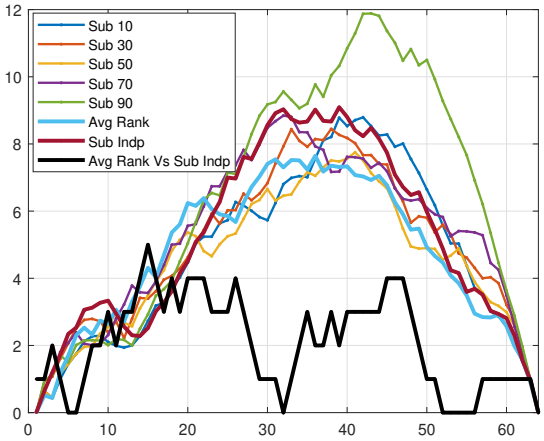

Fig. 8. Plot on how all methods are different from each number based on the difference in selected channels for the classification of MI on the PhysioNet dataset [41]. It shows the average of the number of different channels of the method from all other methods. The average if over $\mathbf{96}$ ranking models ($\mathbf{94}$ subjects, Avg Rank and Sub Indp). Again, it can be seen that the subject-independent method is almost similar to the other models, have peak average difference of less than $\mathbf{15\%}$ of channels, hence it is generalizable up to a great extent.

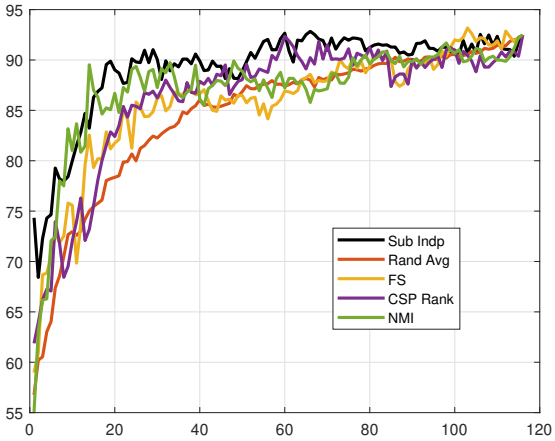

Fig. 9. This figure illustrates the comparison between the proposed method and classical models on the BCICIVa dataset [37]. The proposed model is evaluated against three classical models (CSP Rank [26], FS [7], and NMI [18]) as well as randomly selected channels (Rand Avg). The plots distinctly demonstrate the superior performance of the proposed method over all other approaches. Another notable point is that the accuracy monotonically (approximately) increases with the number of channels.

channels. Fig. 6 shows that for the smaller number of channels *Sub Indp* method performs adequately well, and performs better for $40 - 64$ channels. Overall we can infer that the proposed method performs adequately without sacrificing on classification accuracy.

4) Relevancy of Channels and Divergence: In fig. 4 for the 2^{nd} and 4^{th} subject, we can see that the *Sub Indp* has streigh line while channels varies between $80 - 118$. This shows that divergence information has been captured by already selected channels and no new addition of enough divergence is brought up by the incoming channel to affect the classification accuracy. This also shows that divergence is an excellent measure for channel selection.

5) Comparison of Methodology: The novelty of the proposed model lies in its subject-independent and generalizable nature. Most of the recent models, for one, are subject and task-dependent, and secondly, it is almost impossible to have a subject-independent version of it [13], [32], [36] and yet be fair while comparing with their work. We can classify the channel selection methods into two classes, a) where the channel selection algorithm and classification stage are modular and separate from each other [7], [13], [26], and, b) where both models are entangled and depend on each other [32], [36]. By the nature of the method itself, it is not possible to compare with the latter class. There is another concern about the comparison study in the existing works that those are not compared with randomly selected channels. It should be noted that the comparison of randomly selected channels is not equivalent to the comparison of random classification.

We have, therefore, compared our work with classical methods that are steadily extendable to its subject-independent version. We compared our work with three channel selection models FS [7], CSP Rank [26], and NMI [18] in addition to comparing it with the average of randomly selected channels

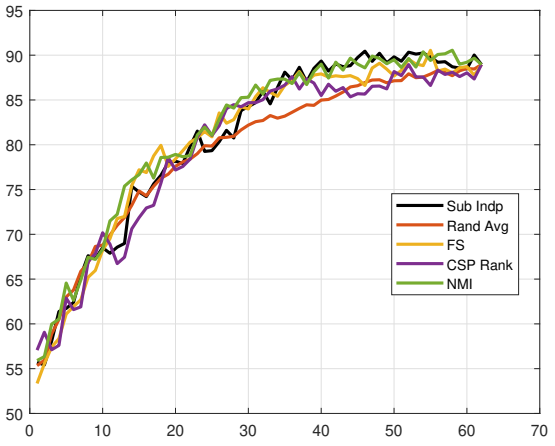

Fig. 10. Plot illustrating the comparison of the proposed method with existing models in terms of classification accuracy on the PhisioNet dataset [41]. The results have been compared with three classical models viz. CSP Rank [26], FS [7], and NMI [18] and randomly selected channels. The results shows that the proposed method outperforms CSP Rank, FS, and Rand Avg methods significantly. However, the proposed method underperforms the NMI for channels less than $\mathbf{35}$ and significantly outperforms afterward.

10 times (Rand Avg).

The comparison results are shown in fig. 9 & 10. Fig. 9 shows how the performance of the proposed model *Sub Indp* varies with the number of channels in comparison with the classical models like FS, CSP Rank, and NMI, along with Rand Avg on the BCICIVa dataset [37]. It can be seen that the proposed method has outperformed all the methods significantly. However, this is not so apparent in the case of the PhysioNet dataset as it can be seen in fig. 10. For the $35 - 50$ number of channels, the proposed model outperforms significantly over all other methods. However, for the smaller

number of channels $(3-35)$ NMI and FS dominate the proposed method. However, for channels $3-35$ the accuracy is steeply increasing and yet to be saturated, hence it may not be the best point of comparison. It can be said that it is 'yet to be the saturated area' for classification accuracy for all the methods. To support this reasoning, it can be seen that even with BCICIVa dataset in fig. 9 the proposed method is not a winner in the 'yet to be saturated area' (in this case it is much smaller between $3-15$). Overall, the proposed method performs better than the existing methods.

IV. Conclusions

In this work, we presented a subject-independent EEG channel selection algorithm and analyzed various aspects of it. We hypothesized that a good way to pursue that would be to use a metric that has a sense of divergence to find a rank for all the channels. We used KL divergence of trials between channels and a reference channel and further their expectation as rank to the channels. We proposed the subject-dependent and independent versions of it and analyzed the differences. We validated our method on two publicly available datasets. We find that the set of channels with smaller divergence among them is a better choice for the motor imagery classification. Experimental results show that the proposed approach performs better than existing models and is generalizable. Future work includes extending the experiments on other tasks than motor imagery to test how it works with other tasks.

Acknowledgement

We thank Dr. Pasquale Arpaia and Dr. Antonio Esposito for their help in replicating their work, especially with CSP implementation. We also thank Mr. Rohit Misra (Ph.D. Candidate, Otto Von Guericke University, Germany) for the discussion and debate during the development of our model,

Data Availability

The datasets used in this work are publicly available. The codes used in this manuscript are available at `https://github.com/rdevm23/eeg_ch_sel_kl_div` for reproducibility.

References

[1] C. M. Michel and M. M. Murray, "Towards the utilization of eeg as a brain imaging tool," *Neuroimage*, vol. 61, no. 2, pp. 371–385, 2012.

[2] Abdullah, I. Faye, and M. R. Islam, "EEG channel selection techniques in motor imagery applications: A review and new perspectives," *Bioengineering*, vol. 9, no. 12, p. 726, 2022.

[3] K. F. Razi and A. Schmid, "Epileptic seizure detection with patient-specific feature and channel selection for low-power applications," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 16, no. 4, pp. 626–635, 2022.

[4] E. Guttmann-Flury, X. Sheng, and X. Zhu, "Channel selection from source localization: A review of four EEG-based brain–computer interfaces paradigms," *Behavior Research Methods*, vol. 55, no. 4, pp. 1980–2003, 2023.

[5] L. Yang, S. Chao, Q. Zhang, P. Ni, and D. Liu, "A grouped dynamic eeg channel selection method for emotion recognition," in *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 3689–3696, IEEE, 2021.

[6] V. Tzoumas, Y. Xue, S. Pequito, P. Bogdan, and G. J. Pappas, "Selecting sensors in biological fractional-order systems," *IEEE Transactions on Control of Network Systems*, vol. 5, no. 2, pp. 709–721, 2018.

[7] T. N. Lal, M. Schröder, T. Hinterberger, J. Weston, M. Bogdan, N. Birbaumer, and B. Schölkopf, "Support vector channel selection in BCI," *IEEE Transactions on Biomedical Engineering*, vol. 51, pp. 1003–1010, 6 2004.

[8] M. Arvaneh, C. Guan, K. K. Ang, and C. Quek, "Optimizing the channel selection and classification accuracy in EEG-based BCI," *IEEE Transactions on Biomedical Engineering*, vol. 58, pp. 1865–1873, 6 2011.

[9] P. Arpaia, F. Donnarumma, A. Esposito, and M. Parvis, "Channel Selection for Optimal EEG Measurement in Motor Imagery-Based Brain-Computer Interfaces," *International Journal of Neural Systems*, vol. 31, 3 2021.

[10] G. Ghorbanzadeh, Z. Nabizadeh, N. Karimi, P. Khadivi, A. Emami, and S. Samavi, "Dgaff: Deep genetic algorithm fitness formation for eeg biosignal channel selection," *Biomedical Signal Processing and Control*, vol. 79, p. 104119, 2023.

[11] W.-W. Fan and C.-H. Lee, "Classification of imbalanced data using deep learning with adding noise," *Journal of Sensors*, vol. 2021, pp. 1–18, 2021.

[12] H. Vikram Shenoy and A. P. Vinod, "An iterative optimization technique for robust channel selection in motor imagery based brain computer interface," in *2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 1858–1863, 2014.

[13] V. S. Handiru and V. A. Prasad, "Optimized Bi-Objective EEG Channel Selection and Cross-Subject Generalization with Brain-Computer Interfaces," *IEEE Transactions on Human-Machine Systems*, vol. 46, pp. 777–786, 12 2016.

[14] H. Helmholtz, "Ueber einige gesetze der vertheilung elektrischer ströme in körperlichen leitern mit anwendung auf die thierisch-elektrischen versuche," *Annalen der Physik*, vol. 165, no. 6, pp. 211–233, 1853.

[15] S. Rush and D. A. Driscoll, "EEG electrode sensitivity-an application of reciprocity," *IEEE Transactions on Biomedical Engineering*, no. 1, pp. 15–22, 1969.

[16] S. Rush and D. A. Driscoll, "Current distribution in the brain from surface electrodes.," *Anesthesia & Analgesia*, vol. 47, no. 6, pp. 717–723, 1968.

[17] T. Alotaiby, F. E. A. El-Samie, S. A. Alshebeili, and I. Ahmad, "A review of channel selection algorithms for EEG signal processing," *EURASIP Journal on Advances in Signal Processing*, vol. 2015, pp. 1–21, 2015.

[18] Z.-M. Wang, S.-Y. Hu, and H. Song, "Channel selection method for EEG emotion recognition using normalized mutual information," *IEEE Access*, vol. 7, pp. 143303–143311, 2019.

[19] J. Meng, L. Yao, X. Sheng, D. Zhang, and X. Zhu, "Simultaneously optimizing spatial spectral features based on mutual information for eeg classification," *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 1, pp. 227–240, 2014.

[20] K. M. Hassan, M. R. Islam, T. T. Nguyen, and M. K. I. Molla, "Epileptic seizure detection in eeg using mutual information-based best individual feature selection," *Expert Systems with Applications*, vol. 193, p. 116414, 2022.

[21] O. Özdenizci and D. Erdoğmuş, "Information theoretic feature transformation learning for brain interfaces," *IEEE Transactions on Biomedical Engineering*, vol. 67, no. 1, pp. 69–78, 2019.

[22] A. Tiwari and A. Chaturvedi, "A novel channel selection method for bci classification using dynamic channel relevance," *IEEE Access*, vol. 9, pp. 126698–126716, 2021.

[23] J. Jin, Y. Miao, I. Daly, C. Zuo, D. Hu, and A. Cichocki, "Correlation-based channel selection and regularized feature optimization for mi-based bci," *Neural Networks*, vol. 118, pp. 262–270, 2019.

[24] T. N. Lal, M. Schroder, T. Hinterberger, J. Weston, M. Bogdan, N. Birbaumer, and B. Scholkopf, "Support vector channel selection in bci," *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 6, pp. 1003–1010, 2004.

[25] X. Yong, R. K. Ward, and G. E. Birch, "Sparse spatial filter optimization for eeg channel reduction in brain-computer interface," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 417–420, IEEE, 2008.

[26] J. Meng, G. Liu, G. Huang, and X. Zhu, "Automated selecting subset of channels based on CSP in motor imagery brain-computer interface system," in *2009 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pp. 2290–2294, IEEE, 2009.

[27] H. Lu, H.-L. Eng, C. Guan, K. N. Plataniotis, and A. N. Venetsanopoulos, "Regularized common spatial pattern with aggregation for EEG classification in small-sample setting," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 12, pp. 2936–2946, 2010.

[28] H.-I. Suk and S.-W. Lee, "A novel bayesian framework for discriminative feature extraction in brain-computer interfaces," *IEEE Transactions on*

Pattern Analysis and Machine Intelligence, vol. 35, no. 2, pp. 286–299, 2012.

[29] A. Das and S. Suresh, "An effect-size based channel selection algorithm for mental task classification in brain computer interface," in *2015 IEEE International Conference on Systems, Man, and Cybernetics*, pp. 3140–3145, IEEE, 2015.

[30] M. Schröder, T. N. Lal, T. Hinterberger, M. Bogdan, N. J. Hill, N. Birbaumer, W. Rosenstiel, and B. Schölkopf, "Robust eeg channel selection across subjects for brain-computer interfaces," *EURASIP Journal on Advances in Signal Processing*, vol. 2005, pp. 1–10, 2005.

[31] A. Nagarajan, N. Robinson, and C. Guan, "Relevance-based channel selection in motor imagery brain–computer interface," *Journal of Neural Engineering*, vol. 20, no. 1, p. 016024, 2023.

[32] B. Sun, Z. Liu, Z. Wu, C. Mu, and T. Li, "Graph Convolution Neural Network based End-to-end Channel Selection and Classification for Motor Imagery Brain-computer Interfaces," *IEEE Transactions on Industrial Informatics*, 2022.

[33] T. Alotaiby, F. E. El-Samie, S. A. Alshebeili, and I. Ahmad, "A review of channel selection algorithms for EEG signal processing," *Eurasip Journal on Advances in Signal Processing*, vol. 2015, pp. 1–21, 12 2015.

[34] Z. Qiu, J. Jin, H.-K. Lam, Y. Zhang, X. Wang, and A. Cichocki, "Improved sffs method for channel selection in motor imagery based bci," *Neurocomputing*, vol. 207, pp. 519–527, 2016.

[35] G. A. Light, L. E. Williams, F. Minow, J. Sprock, A. Rissling, R. Sharp, N. R. Swerdlow, and D. L. Braff, "Electroencephalography (EEG) and event-related potentials (erps) with human participants," *Current Protocols in Neuroscience*, vol. 52, no. 1, pp. 6–25, 2010.

[36] J. Jin, C. Liu, I. Daly, Y. Miao, S. Li, X. Wang, and A. Cichocki, "Bispectrum-based channel selection for motor imagery based brain-computer interfacing," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 28, no. 10, pp. 2153–2163, 2020.

[37] G. Dornhege, B. Blankertz, G. Curio, and K.-R. Muller, "Boosting bit rates in noninvasive eeg single-trial classifications by feature combination and multiclass paradigms," *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 6, pp. 993–1002, 2004.

[38] W. Ko, E. Jeon, S. Jeong, and H.-I. Suk, "Multi-scale neural network for eeg representation learning in bci," *IEEE Computational Intelligence Magazine*, vol. 16, no. 2, pp. 31–45, 2021.

[39] R. O. Duda, P. E. Hart, *et al.*, *Pattern classification*. John Wiley & Sons, 2006.

[40] B. Blankertz, K.-R. Muller, D. J. Krusienski, G. Schalk, J. R. Wolpaw, A. Schlogl, G. Pfurtscheller, J. R. Millan, M. Schroder, and N. Birbaumer, "The bci competition iii: Validating alternative approaches to actual bci problems," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 14, no. 2, pp. 153–159, 2006.

[41] G. Schalk, D. J. McFarland, T. Hinterberger, N. Birbaumer, and J. R. Wolpaw, "BCI2000: a general-purpose brain-computer interface (BCI) system," *IEEE Transactions on Biomedical Engineering*, vol. 51, pp. 1034–1043, 6 2004.

[42] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C. K. Peng, and H. E. Stanley, "PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, 2000.

[43] Y.-W. Chen and C.-J. Lin, "Combining SVMs with various feature selection strategies," *Feature Extraction: Foundations and Applications*, pp. 315–324, 2006.

[44] T. Markiewicz and S. Osowski, "Data mining techniques for feature selection in blood cell recognition.," in *European Symposium on Artificial Neural Networks (ESANN)*, pp. 407–412, 2006.

[45] V. Muralidharan, "feature_rank (input, labels, numindices, method) `https://www.mathworks.com/matlabcentral/fileexchange/54906-feature_rank-input-labels-numindices-method`, MATLAB Central File Exchange,," 2023.