Relevance based Channel Selection in Motor Imagery Brain-Computer Interface

Aarthy Nagarajan, Neethu Robinson and Cuntai Guan

Nanyang Technological University, 50 Nanyang Avenue, Singapore

E-mail: aarthy001@e.ntu.edu.sg, nrobinson@ntu.edu.sg, ctguan@ntu.edu.sg

October 2022

Abstract.

Channel selection in electroencephalogram (EEG)-based brain-computer interface (BCI) has been extensively studied for over two decades, with the goal to select optimal subject-specific channels that can enhance the overall decoding efficacy of BCI. With the emergence of deep learning (DL) based BCI models, there arises a need for fresh perspectives and novel techniques to conduct channel selection. In this regard, subject-independent channel selection is relevant, since DL models trained using cross-subject data offer superior performance, and the impact of inherent inter-subject variability of EEG characteristics in subject-independent DL training is not yet fully understood. Here, we propose a novel methodology for implementing subject-independent channel selection in DL based motor imagery (MI)-BCI, using layer-wise relevance propagation (LRP) and neural network pruning. Experiments were conducted using Deep ConvNet and 62-channel MI data from Korea University (KU) EEG dataset. Using our proposed methodology, we achieved a 61% reduction in the number of channels without any significant drop (p=0.09) in subject-independent classification accuracy, due to the selection of highly relevant channels by LRP. LRP relevance based channel selections provide significantly better accuracies compared to conventional weight based selections while using less than 40% of the total number of channels, with differences in accuracies ranging from 5.96% to 1.72%. The performance of the adapted sparse-LRP model using only 16% of the total number of channels is similar to that of the adapted baseline model (p=0.13). Furthermore, the accuracy of the adapted sparse-LRP model using only 35% of the total number of channels exceeded that of the adapted baseline model by 0.53% (p=0.81). Analyses of channels chosen by LRP confirm the neurophysiological plausibility of selection, and emphasize the influence of motor, parietal, and occipital channels in MI-EEG classification.

Keywords: MI-BCI, channel selection, deep learning, explainable AI, layer-wise relevance propagation

1. Introduction

Brain-computer interface (BCI) systems have become an integral part of today's society for healthy and paralyzed users, by offering a means of non-muscular communication with the environment [1, 2]. Specifically, the motor-imagery (MI)-BCI, which involves voluntary modulation of the neural electrical activity to communicate movement intention, has some important clinical applications and is popularly researched [3–5]. While traditionally, MI-BCI operated using machine learning (ML) algorithms such as common spatial patterns (CSP) [6] and the filter-bank CSP (FBCSP) [7], the advantage of learning complex tasks in an end-to-end fashion has recently encouraged BCI researchers to shift to deep learning (DL) based MI models [8–10] for detecting user intentions from the brain signal. This much welcomed change also brings about the need to revisit and address some traditional issues in BCI from the perspective of deep learning, and with more robust solutions [11].

The recording of MI-based electroencephalography (EEG) signal using non-invasive cortical electrodes opened up questions on the importance of channel selection, which has been studied by BCI researchers for over two decades. Current channel selection literature in MI-BCI highlights methodologies specific to subjects and sessions, and are meant to remove noisy channels yet maintain or even improve classification performance. Channel selection, in these studies, is usually performed as an additional step in the ML pipeline of BCI algorithms. Three are three kinds of channel selection techniques discussed in literature: (i) filter methods – that use statistical measures such as correlation coefficients and mutual information to identify optimal channel sets [12–15]. Although these techniques take less computation time, they often lead to suboptimal performance using selected channels (ii) wrapper methods – that follow an iterative process to find the optimal channel sets, while evaluating their performance using a classifier [16, 17]. Thus, they are computationally inefficient and can have varied results based on the classification algorithm applied, and (iii) hybrid or embedded methods – which are a combination of filter and wrapper, and hence, enjoy the benefits of both [18]. Regularization and optimization techniques are applied in the hybrid approach to fine-tune the classifier performance alongside inducing sparsity of channels [19].

Relatively few studies have explored subject-independent [16, 17, 20] and session-independent [19, 21] selection of channels. Arvaneh et al. [19] optimized the CSP algorithm using $\frac{L1}{L2}$ norm to induce sparsity in CSP coefficients pertaining to noisy channels. The sparse CSP (SCSP) channel selection was evaluated using data from a different session, thus illustrating the transferability of channel selection from one session to another. In a subsequent paper [21], Arvaneh et al. evaluated the robustness of SCSP (RCSP) across new sessions using stroke data. The channels selected using RCSP from the first session were found to perform well on subsequent sessions. In [16], Schröder et al. investigated cross-subject transfer of channel rankings based on recursive channel elimination (RCE). A small dataset, containing EEG recorded from 8 subjects using 39 electrodes, were used to derive the findings for this study. In [20], Parashiva et

al. trained a sparse autoencoder neural network with sparsity regularization to select channels for both subject-specific and subject-independent cases. Data from selected channels was then fed to conventional ML based BCI pipeline for classification. They reported results using EEG data from 10 healthy subjects collected with 31 electrodes. In another subject-independent study [17], Arpaia et al. progressively selected channels common to all subjects, using sequential forward selection method. They evaluated their approach using both 2-class as well as 4-class MI data from BCI competition dataset IV-2a. Apparently, these subject-independent and session-independent channel selection studies were conducted using small datasets, by applying ML algorithms and extensive computations, nevertheless, had not obtained notably high performance. In addition, none of the above discussed channel selection methods are relevant to DL models in MI-BCI.

The introduction of DL in BCI has rekindled our interest in channel selection and how it could be addressed using sophisticated DL based methods, more so in a subject-independent setup. The deep neural networks are considered as black-box models which approximate any arbitrary function that cannot be clearly identified from its structure. Even deep learning experts find it challenging to understand how all the neurons work together to arrive at the output. It is an added concern when such black-box models fail to perform if the dataset is even slightly perturbed. This growing concern sparked the introduction of an avenue of research called the explainable AI (XAI) or trustworthy AI [22, 23], that has enhanced our understanding of how neural networks think, which in turn has improved its interpretability and therefore, applicability. In this study, we exploit the benefits of XAI to solve the problem of channel selection in DL based MI classifiers. In addition, a robust channel selection procedure is one that is generalizable towards unseen subjects. A precise evaluation of such a procedure can be done only with the help of a large enough dataset.

This study is an attempt to address the above-mentioned points through five main contributions:

(i) We propose a novel relevance based subject-independent channel selection procedure for MI classifiers based on DL, using layer-wise relevance propagation (LRP) [24], a popular XAI tool, and neural network pruning.

(ii) We evaluate our proposed method using Deep ConvNet [8], a state-of-the-art MI-BCI model, and Korea University (KU) dataset [25] that contains MI data collected from 54 subjects.

(iii) We compare the performance of our proposed relevance based channel selection against weight magnitude based channel selection.

(iv) We validate our channel selection results against those obtained using 20 motor channels and random selection of channels.

(v) We investigate the impact of our channel selection method on the model adaptation performance, using a small amount of data from the unseen subject.

The rest of the paper is structured as follows. We begin with related work in

section 2, and we briefly introduce the Deep ConvNet architecture, discuss our proposed relevance based channel selection, and sparsification (selection) of channel weights in Deep ConvNet using neural network pruning in section 3. We also specify the model training, optimization and adaptation settings, and our paradigm for subject-independent classification. Following this, we describe the KU dataset and the flow of our subject-independent channel selection experiments including some additional analyses in section 4, discuss the results and their analyses in sections 5 and 6, before concluding in section 7.

2. Related Work

XAI methods are traditionally used to enhance the interpretability and trustworthiness of a deep neural network, such that one can explain the output of the network better [26]. However, the application of XAI extends beyond model explainability. As XAI reveals the importance of each neuron or a set of neurons, deep learning researchers and architects can exploit this to introduce useful structural changes to the neural network. This brings about an interesting connection of XAI with neural network pruning. Particularly, the LRP's [24] direct link to the network output and the conservative nature of relevance propagation between layers makes it a suitable criterion for network pruning. In [27], Yeom et al. experimentally highlighted the effectiveness and robustness of LRP based relevance as a pruning criterion. They showed that relevance based pruning is both scalable and efficient, and offers better results in transfer learning scenarios compared to other criteria.

Neural network pruning has typically been used for compressing deep learning models for the purpose of efficiency and speed [28–30]. Pruning can involve removal of weights or/and neurons, and can be performed globally across the network, layer-wise or in a random manner [31]. The framework for pruning is generally consistent across research studies, but what differs is the criterion used to select candidates for pruning. One of the most popularly studied approach is magnitude based weight pruning, where the magnitude of model weight with respect to an individual or a set of neurons is used as a criterion for weight removal. Different types of weight-based criteria have been explored in pruning literature. In the simplest case of unstructured pruning, individual neuronal weights with low magnitude are identified for removal [32]. A similar approach for performing magnitude based weight pruning in MI-BCI was introduced in [33].

Although commonly used for network compression, in this study, we implement pruning as a method for channel selection (or deselection) by using LRP [24] as a criterion. Our objectives, using the proposed methodology, are two-fold: (i) to perform subject-independent channel selection for facilitating user comfort and convenience during data collection, and, (ii) to sparsify DL classifiers for MI for better generalizability, interpretability, and efficiency. We compare the performance of our proposed LRP based channel selection method with weight magnitude based channel selection, which is one of the state-of-the-art criteria used for neural network pruning.

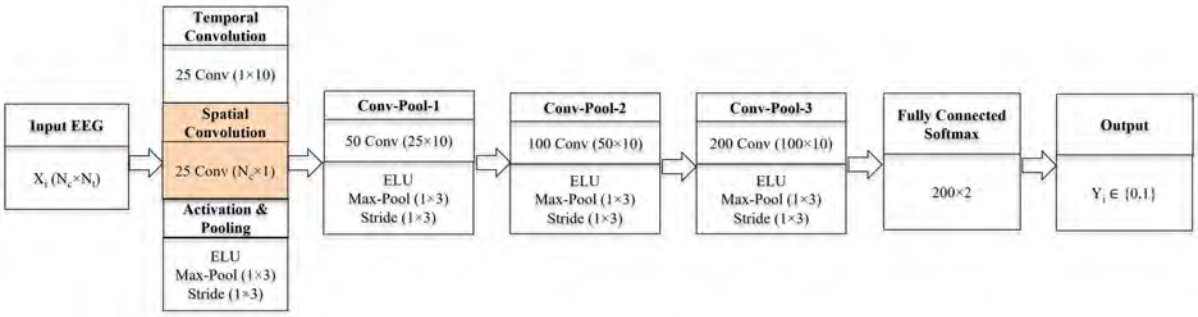

Figure 1: Deep ConvNet Architecture [8] with highlighted spatial convolution filters.

To the best of our knowledge, this is the first study to investigate channel selection, in particular subject-independent channel selection, in DL models for MI detection. Furthermore, this is the first study to apply neural network pruning based on XAI for performing channel selection in BCI.

3. Methodology

In this section, we provide details of the Deep ConvNet architecture and related training parameters. We further discuss our proposed channel selection algorithm after outlining the procedure for LRP based channel relevance estimation and spatial filter sparsification by network pruning.

3.1. Network Architecture

A convolutional neural network or CNN [34] is a deep learning algorithm that applies filters to the input, resulting in feature activation maps. A convolution operation is represented mathematically by equation 1, where, X is the input, F is the filter and the * sign symbolizes convolution. The filter weights are typically two-dimensional and expressed as $f_1 \times f_2$. A one-layer CNN generally consists of a convolution layer, followed by activation and pooling for feature extraction, and a fully connected classification layer. Equation 2 mathematically defines a one-layer CNN. Here, σ represents the activation function and W^T and B represent the weights and biases of the dense layer.

$$Z = X * F \tag{1}$$

$$Output = Softmax(W^T.(\sigma(X * F)) + B) \tag{2}$$

To demonstrate our novel subject-independent channel selection in BCI, we use the Deep ConvNet, a state-of-the-art CNN classifier for MI detection introduced by Schirrmeister et al. in [8]. The Deep ConvNet includes a temporal convolution and a spatial convolution, followed by max-pooling, three convolution-max-pooling blocks, and a fully-connected softmax classification layer as illustrated in Figure 1. Deep ConvNet uses the exponential linear unit (ELU) [35] as the activation function in every layer.

Considering that the input EEG is of dimension $N_c \times N_t$, where N_c is the number of channels and N_t the number of time samples in every channel, the temporal convolution performs convolution over time and the spatial convolution implements convolution over all the channels. The 25 spatial convolution filters (highlighted in orange), each of dimension $N_c \times 1$, consist of weights for every electrode present in the EEG data. This allows us to deselect certain channels in the data by setting their respective spatial filter weights to zero.

3.2. Proposed Channel Selection by LRP Relevance Score

Layer-wise relevance propagation or LRP [24] is a popular and widely used XAI framework that explains a network's decision by decomposing it into relevance scores attributed to the intermediate neurons and inputs. LRP works by the principle of conservation and deep Taylor decomposition (DTD) [36]. LRP highlights the nodes that contribute to the model's decision layer by layer. This overcomes the limitation of other gradient-based methods that focus more on sensitivity analysis of the network rather than contributions from individual neurons to the model's output. The relevance decomposition performed by LRP is expressed mathematically in equation 3, where the total relevance of node j in layer l which is closer to the input, is equal to the sum of relevances of all nodes i from the subsequent layer $l+1$, that are connected with node j.

$$R_j^{(l)} = \sum_i R_{j \leftarrow i}^{(l+1)} \tag{3}$$

The basic LRP attribution rule, usually denoted as LRP_z, performs a proportional decomposition with regard to localized activations, $z = xw$, that are propagated through the network during prediction. The definition of LRP_z is given in equation 4, where $z_{ji} = x_j w_{ji}$ and $z_i = \sum_j z_{ji}$.

$$R_j^{(l)} = \sum_i \frac{z_{ji}}{z_i} R_i^{(l+1)} \tag{4}$$

In order to prevent division by zero while applying the basic LRP_z rule, a small stabilizer term, $\epsilon > 0$, was later included in the denominator of equation 4. This improvised rule is often called as the LRP-epsilon (LRP_ϵ) rule and is defined as follows.

$$R_j^{(l)} = \begin{cases} \sum_i \frac{z_{ji}}{z_i + \epsilon} R_i^{(l+1)} & \mathrm{z}_i \geq 0 \\ \sum_i \frac{z_{ji}}{z_i - \epsilon} R_i^{(l+1)} & \mathrm{z}_i < 0 \end{cases} \tag{5}$$

Several other variations of LRP rule, such as $LRP_{\alpha\beta}$ and LRP_γ, were also implemented to cater to the different attribution requirements. However, even the basic LRP rule has been observed to provide highly interpretable heatmaps for better understanding of EEG classifiers [37]. In this study, we apply the LRP_ϵ rule to perform the subject-independent channel selection in DL based MI-BCI. We used the Pytorch

framework [38] for Deep ConvNet implementation and the Captum XAI library [39] for computing LRP relevance scores.

The computational efficiency, theoretical affirmation, popularity, and trustworthiness make LRP a promising tool for estimating EEG channel relevance scores using MI classifiers based on DL. The fact that LRP backpropagates the network output to identify the contributions of EEG input from specific channels, offers us an opportunity to obtain channel-wise relevance scores that could be exploited for performing channel selection. We implemented the following steps using the Captum's application programming interface (API) for LRP [39].

The first step in calculating the channel-wise relevance scores is to define an instance of the LRP criterion for the pre-trained subject-independent model by applying the chosen rules to the underlying layers. In the next step, the multi-subject training trials are sent individually to the "attribute" function of the LRP criterion. The attributions are obtained class-wise by indicating the target class of the input trial to the attribute function. This is to ensure that the output attributions are relevant to the prediction of that particular class. The attribute function performs backward propagation of the output score sequentially through all layers of the model and returns the corresponding relevance scores for each neuron in the underlying layers. The relevance estimation is defined by the rules chosen for each layer, which in our case is LRP_{ϵ}. The ϵ term was set to 1e-9, which is the default value in Captum's API for LRP. The output attributions are of the same dimension as that of the EEG input trial (N_c, N_t), hence they were averaged across time samples to get a single score for each channel in the input. These relevance scores were class-wise averaged across all trials, and further normalized between 0 and 1, for each training subject. The final set of scores were obtained by averaging the scores across all subjects. The resulting class-wise average channel relevance scores were saved and used for channel selection.

3.3. Spatial Filter Sparsification by Neural Network Pruning

Neural network pruning [40][41], which involves systematic removal or sparsification of weights from a pre-trained model, is popularly applied by deep learning architects for better memory efficiency and faster application of the model during testing time. Such sparse models are also more structurally interpretable and prevent overfitting of parameters. In this study, we proposed a novel application of pruning to perform channel selection, or rather channel deselection, by sparsifying the Deep ConvNet of those channel weights whose LRP relevances are low.

The Deep ConvNet implements spatial convolution using kernel weights whose dimensions match the number of channels present in the input EEG data. The channel relevance scores estimated using LRP could be used to select and retain the weights of highly relevant channels in the spatial convolution filters, and prune the weights of less relevant ones, thus performing channel selection. In doing so, only the channels with non-zero weights contribute to the prediction. Network pruning is followed by

re-training for a few epochs, as is usually done, in order to adapt the remaining weights to the changes caused by pruning and to regain the performance of the model. We used the pruning methods offered by Pytorch utilities [38] to implement the spatial filter sparsification in Deep ConvNet.

3.4. Channel Selection and Model Sparsification using LRP Relevance Score

In our proposed method, we select channels based on LRP computed channel relevance scores. The number of top relevant channels (topN) selected will remain the same for all pre-trained models and is pre-determined. The procedure for implementing this channel selection method is described below in three steps. For better understanding, the method is also outlined as a pseudocode in Algorithm 1. The algorithm makes reference to a separate function that calculates the channel-wise relevance scores using the LRP_ϵ rule.

Step 1: For every target/test subject s_{test}, the corresponding EEG test data x_{test}, training data x_{train}, and pre-trained subject-independent model $m_{s_{test}}$, are loaded.

Algorithm 1: Channel Selection and Model Sparsification using LRP Relevance Score

Input: EEG data X^*, subjects S, pre-trained models m, LRP rule r, number of top channels to select *topn*

Output: Sparse models and accuracies

foreach *test subject* $s_{test} \in S$ **do**
- *Train subjects* $s_{train} = S$ - s_{test};
- *load test data* $x_{test} = X_{s_{test}}$;
- *load training data* $x_{train} = X_{s_{train}}$;
- *load pre-trained subject-independent model* $m_{s_{test}}$;
- *sorted channels based on relevance scores* $C_{sort} =\leftarrow CHANREL(x_{train}, m_{s_{test}}, r)$;
- *selected channels* $C_{sel} = C_{sort}[topn]$;
- $m'_{s_{test}} = m_{s_{test}}$;
- *sparsify* $m'_{s_{test}}$, *retain* C_{sel} *and prune weights of remaining channels in spatial conv filter;*
- *finetune* $m'_{s_{test}}$ *using* x_{train};
- $acc \leftarrow m'_{s_{test}}(x_{test})$;
- *save* $m'_{s_{test}}$ *and acc ;*

end

* $X = R^{N \times C \times T}$, *where N = No. of trials, C = No. of channels, T = No. of time samples*

Step 2: The channel-wise relevance scores are computed using x_{train}, by following the steps described in 3.2. As averaging across two classes neutralizes the channel

Function *CHANREL*(x_{ref}, m, r) :
  A_{c0} = *[];*
  A_{c1} = *[];*
  foreach *training subject* $s_{ref} \in x_{ref}$ **do**
    load training subject data, $x_{s_{ref}}$*;*
    Create an instance of LRP criterion cr, using pre-trained model m and the chosen LRP rules r for each layer;
    Calculate the attributions Attr using cr, $x_{s_{ref}}$ *and the target class;*
    foreach *target class* $c \in c0, c1$ **do**
      Time averaged attributions $(N \times C \times 1)A_{timeavg(c)} \leftarrow mean(Attr_c, dimension = -1)$*;*
      Average attributions $(C \times 1)A_{avg(c)} \leftarrow mean(A_{timeavg(c)}, dimension = 0)$*;*
      Normalized mean attributions $(C \times 1)A_{norm(c)} \leftarrow normalize(A_{avg(c)})$*;*
      $A_c + \leftarrow A_{norm(c)}$*;*
    end
  end
  $(C \times 1)Rel_{c0} \leftarrow mean(A_{c0})$*;*
  $(C \times 1)Rel_{c1} \leftarrow mean(A_{c1})$*;*
  foreach *channel* $ch \in C$ **do**
    $Rel_{ch} \leftarrow max(Rel_{c0-ch}, Rel_{c1-ch})$*;*
  end
  return sort$((C \times 1)Rel)$*;*
end

relevances, we selected the larger out of the two scores (of two classes) as the relevance score for each channel. The channel-wise relevances are sorted, and the channels are rated from 1 to n, where n is the total number of channels. The topN channels, based on rating, are selected and evaluated on x_{test}.

Step 3: The topN channels are retained and the weights of remaining channels are pruned from the spatial convolution kernels of the pre-trained model. The resulting sparse model, $m'_{s_{test}}$, is re-trained using x_{train}. The evaluation on x_{test} is done after re-training. This procedure is repeated for different values of N, where N is varied from 2-60 in steps of 2. The resulting sparse models for different values of N, and their corresponding accuracies, are recorded for further analysis.

We would like to note here that the channel selection is not performed based on the results obtained using x_{test}. We repeat the evaluations for different values of N for the purpose of analysis.

3.5. Channel Selection and Model Sparsification using Magnitude of Weight

For comparison with our proposed method summarized in the previous section, we also selected channels based on the magnitude of weights of respective channels in the spatial convolution filters of the pre-trained model.

Step 1: Remains the same as in section 3.4

Step 2: The channel-wise magnitude of weights are calculated by averaging the respective channel weights across all spatial convolution kernels in the pre-trained model. The channels are sorted by their mean weights and rated from 1 to n, where n is the total number of channels. The topN channels, based on rating, are selected and evaluated using the test set.

Step 3: Remains the same as in section 3.4

3.6. Model Training and Adaptation

The Deep ConvNet was trained using Adam optimizer [42] and negative log-likelihood loss function to update the model weights. We also performed Batch Normalization [43] and Dropout [44] for each convolution-max-pooling block. Model training was performed in 200 epochs, and the epoch with the lowest training loss was selected for evaluation and further analysis. Re-training after spatial filter sparsification was performed for 5 epochs before final evaluation of the sparse model using test data.

In addition, to account for the inter-subject variabilities in the subject-independent scenario, we adapted the pre-trained subject-independent models using some data from the corresponding target subject. This was performed both before and after model sparsification, and the respective performances of adapted models were compared.

3.7. Subject-independent Classification

The subject-independent MI classification models were trained using the leave-one-subject-out cross-validation (LOSO-CV) paradigm. For every test/target subject, the training data consists of all trials from the remaining subjects. The classification performance is measured by the average accuracy across all subjects. Further details regarding the data division will be discussed in section 4.1.

4. Experiments

In this section, we describe the KU MI-EEG dataset and provide details on the data division used in our experiments. We also introduce the additional analyses performed for validating our experimental results. The entire flow of our channel selection experiments is visualized as a block diagram in Figure 2. The computations for this work were performed using multiple GPU and CPU resource clusters available from the School of Computer Science and Engineering at NTU and from the National Supercomputing Centre, Singapore.

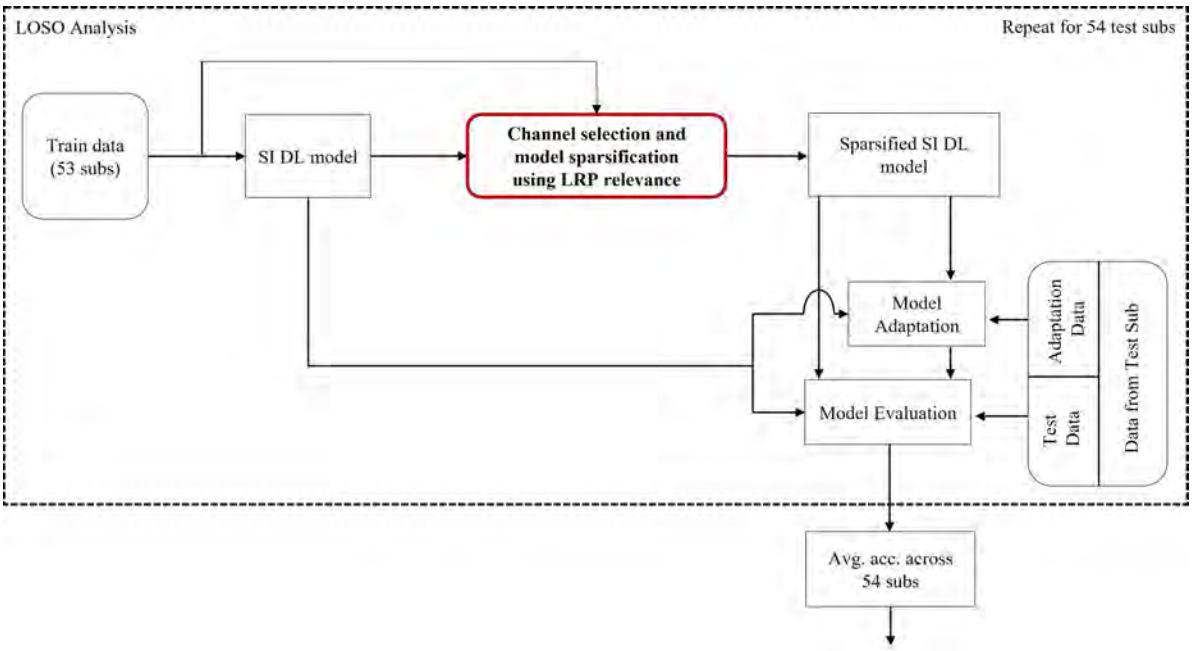

Figure 2: The flow of our channel selection experiments using subject-independent (SI) deep learning (DL) models for MI.

4.1. Dataset and Data Division

Our channel selection experiments were conducted using two-class (left and right hand) MI data from the Korea University EEG dataset [25], that contains data collected from 54 healthy people (aged 24-35). For every subject, EEG data were obtained from two sessions, collected using 62 Ag/AgCl electrodes at 1000 Hz sampling frequency. The data consists of 200 MI trials from each session, of which 100 trials belong to each class. Each session was divided into an offline phase (phase 1), to collect data for constructing the classifier, and an online test phase (phase 2) with visual feedback. At the beginning of each trial, a fixation mark was displayed at the center of a screen in order to prepare the subject for the trial. Subsequently, a visual cue consisting of a left or a right arrow was shown for 4s, during which the subject performed the respective MI task. The screen remained blank for about 6s after every trial. We used 0-4s post-cue data and downsampled it by four for our experiments.

For every target subject, the 100 trials from phase 2 of session 2 was used as the test set in all our channel selection experiments. This test data is not used during model training and estimation of channel relevances for channel selection. The subject-independent MI model, for every target subject, is pre-trained using 400 trials each from the remaining 53 subjects. No validation data was used during training. For model adaptation experiments, the adaptation set for every target subject included session 1 data for training and phase 1 of session 2 data for validation, following the settings used in [45].

Table 1: Average subject-independent accuracies (leave-one-subject-out) of Deep ConvNet using all 62 channels and 20 motor channels. The 20 motor channel accuracy was obtained by pruning the 62-channel subject-independent model.

	Full – 62 channels	Motor – 20 channels
Mean LOSO-CV Accuracy ± SD	84.82±12.07	81.72±12.61**

** ($p<0.001$) indicates that the accuracy using 20 motor channels is significantly lower than the accuracy achieved using all 62 channels.

4.2. Analyses using Motor Channels and Randomly Selected Channels

In order to validate the performance of our method and highlight its significance, we conducted additional experiments by selecting fixed and random channel sets. A fixed set of 20 motor channels that are typically used for MI prediction [10, 25] were chosen. These channels are FC-5/3/1/2/4/6, C-5/3/1/z/2/4/6, and CP-5/3/1/z/2/4/6. We simply sparsified the spatial convolution filter weights of the pre-trained 62-channel subject-independent models such that only the 20 motor channels are retained, and measured the classification accuracies after re-training. Furthermore, to invalidate the randomness of our results using the proposed method, we compared the performance of topN channels selected based on LRP relevance with that of randomly selected N (randomN) channels, where N is varied from 2-60 in steps of 2. The average LOSO-CV classification accuracies of the 20 motor channel set and the different randomN sets were recorded for analysis. In order to prevent selection bias, for each N we repeated classification using randomN for 10 times using 10 different random channel sets, and averaged the accuracies to obtain the mean subject-independent accuracy of randomN.

4.3. Model Adaptation

As part of the subject-independent channel selection experiments, we included an additional analysis of comparing the performances of the adapted baseline versus adapted sparse models. For this purpose, we adapted the original and the sparse models using a small part of the target subject's data, which is not included in the test set and training set, and evaluated the performances of the adapted models on the test set.

5. Results

The results include average LOSO-CV accuracies of Deep ConvNet before and after applying relevance based channel selection. In addition, performance comparisons between relevance based selections, weight magnitude based selections, motor channels and random selections are included. Furthermore, visual illustrations of top selected channels and average channel-wise relevances are provided. Subject-wise classification accuracies are analyzed as well. Finally, yet importantly, we report the performance differences between the adapted baseline model and the adapted sparse models.

5.1. Channel Selection Results using Deep ConvNet

The baseline for our study is the LOSO-CV accuracy of Deep ConvNet using 62-channel data from KU dataset. Following the training settings outlined in section 3.6, we obtained an average accuracy of 84.82% across 54 subjects. This is similar to the LOSO-CV result reported in literature [45] using the same dataset.

In our proposed method, we used the LRP estimated channel-wise relevance scores to select the topN channels. The classification accuracies of sparse models, in which the topN channels were selected and the remaining pruned from their respective spatial filter weights, were estimated. For a given N, the number of selected channels remains the same for all target subject models, however, the channels that get selected may vary. The average accuracies obtained using our proposed method are available in column 2 of Table 2. Those accuracies that are not significantly different from the baseline LOSO-CV accuracy are highlighted in bold, and are achieved using $N = 24$, 32-40, 48, and 52-60. We observe that the accuracies are above 80% for all $N > 10$. $N = 24$ is the minimum number of channels required to acquire a performance not significantly different (p=0.09) from the baseline.

To compare the results of our proposed channel selection method with that of a closely related state-of-the-art technique, we used the channel-wise mean weights to select the topN channels, and the performances of resulting sparse models were computed. The results obtained using weight magnitude based channel selection are available in column 3 of Table 2. None of the accuracies were similar to baseline, with p-values less than 0.05 for all values of N. We observe that the accuracies are above 80% for all $N \geq 16$, with the exception of $N = 18$. Those accuracies that are significantly lower than that of the corresponding topN selection by LRP relevance scores, are indicated with * ($p < 0.05$) and ** ($p < 0.001$). Results show that weight magnitude based channel selection achieves relatively lower performance, compared to relevance based selection, while using fewer channels where $N = 6$ up to 24, and thereafter for $N = 32$, 34, 36, and 44.

5.2. Results using 20 Motor Channels

BCI researchers typically use 20 channels from the motor region to enhance the MI classification performance [10, 25]. The average LOSO-CV accuracy acquired using the 20 motor channels is 81.72%, against the baseline 62-channel accuracy of 84.82%, as presented in Table 1. The performance of motor channels is significantly lower than the baseline, with $p < 0.001$.

5.3. Performances of topN Channels Selected by Relevance and Weight versus Randomly Selected Channels

The performance comparison between randomN and topN is tabulated in Table 2 and visualized in Figure 3. The topN channels selected by LRP relevances are indicated in

Table 2: Average subject-independent accuracy (in %) of Deep ConvNet using relevance based vs weight based vs random channel selection.

No of channels (N)	Accuracy of topN-LRP (%)	Accuracy of topN-weight (%)	Accuracy of randomN (%)
6	77.94±12.91	71.98±12.03**	73.22±10.35**
8	78.41±12.92	74.98±11.30*	74.64±10.71**
10	79.93±13.63	77.11±11.57*	75.99±11.03**
12	81.52±12.25	78.07±12.30*	76.92±11.42**
14	82.48±12.22	79.00±12.13**	78.67±11.59**
16	82.57±12.08	80.43±11.44*	79.46±11.44**
18	81.83±12.29	79.94±12.51*	80.08±11.55*
20	83.19±11.60	80.82±11.98*	81.03±11.74**
22	82.94±12.17	81.22±11.85*	80.81±11.66*
24	**83.59**±12.15	80.67±12.20*	81.35±11.53*
26	83.26±12.50	81.67±12.06	80.96±11.96**
28	82.85±12.76	81.98±11.97	81.59±11.45
30	82.93±13.13	82.80±12.13	82.02±11.56
32	**83.94**±12.62	82.57±11.69*	82.19±11.64*
34	**84.06**±12.83	82.41±11.60*	82.89±11.87*
36	**83.93**±13.07	82.39±12.25*	83.14±11.78
38	**83.83**±12.49	82.32±12.11	83.29±11.64
40	**83.80**±12.97	82.83±12.43	83.56±11.82
42	83.13±12.97	83.02±12.20	83.25±12.02
44	83.67±13.23	82.56±12.07*	83.62±11.91
46	83.20±13.58	83.00±12.78	83.58±11.61
48	**83.85**±13.33	83.74±12.08	83.62±11.57
50	83.63±13.11	83.13±12.65	83.58±11.87
52	**83.98**±12.75	83.06±13.29	83.75±11.84
54	**84.11**±12.81	83.06±12.81	83.97±11.64
56	**84.33**±12.22	83.65±12.56	83.91±11.66
58	**84.17**±12.67	83.63±12.56	**84.00**±11.84
60	**84.00**±12.61	83.57±12.84	83.94±11.94

The $*$ ($p<0.05$), and ** ($p<0.001$) indicate that the accuracies are significantly lower than the accuracy of corresponding topN selected based on LRP relevance. Results that are not significantly different from baseline accuracy using all 62 channels (84.82%), are highlighted in bold.

this section as topN-LRP and the topN channels selected by weight are indicated as topN-weight, for clarity and convenience.

The performances of topN-LRP are significantly higher than the corresponding performances of topN-weight and randomN channels up until $N = 24$. Beyond this point, randomN channels continue to obtain significantly lower accuracies for $N = 26$, 32, and 34, and topN-weight for $N = 32$, 34, 36, and 44. In addition, topN-LRP achieves a subject-independent accuracy not significantly different from the 62-channel baseline by using only 24 channels ($p=0.09$), compared to topN-weight, which never reached a similar performance, and randomN, which requires 58 channels to attain a

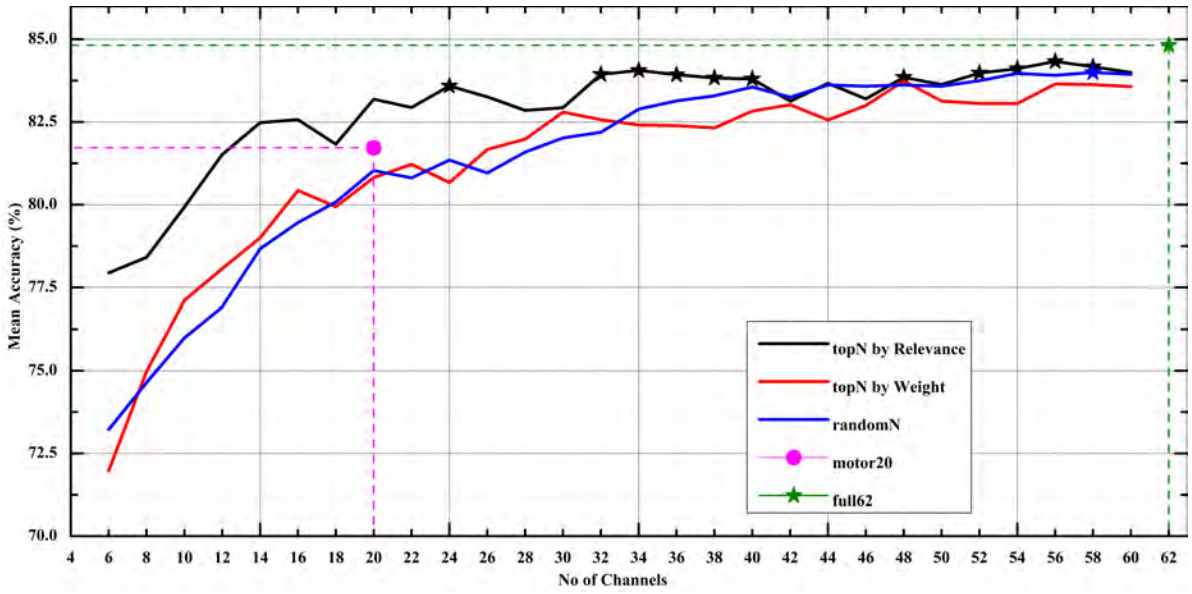

Figure 3: Comparison of average subject-independent classification accuracies of Deep ConvNet, using top channels selected by relevance and weight, randomly selected channels, 20 motor channels and all 62 channels. Those accuracies that are not significantly different from the baseline accuracy (full62) are indicated using $\star$.

similar level of performance. The performances of all three types of selections saturate beyond 40 channels, nevertheless, we notice sustained fluctuations in the performance of topN-weight.

The performance of topN-LRP clearly stands out in comparison with topN-weight and randomN for implementing subject-independent channel selection in DL models such that, good performance can be achieved for an unknown target subject even with fewer channels. This is evident by observing the results of topN-LRP for $N = 6 - 24$. Even for as few as 6 channels selected using LRP relevance, we secured an average LOSO-CV accuracy of 77.94% and with just 10 channels the accuracy reaches close to 80%. After reaching an accuracy level that is not significantly different from baseline at $N = 24$, the topN-LRP maintains this performance level across several channel subsets such as $N = 32 - 40$, 48, and thereafter from 52-60. However, the performance of topN-LRP never exceeded the baseline performance for any value of N.

In Figure 3, we have also indicated the performance using 20 motor channels, which is close to the performance of random20 and top20-weight, and is 1.47% lower than the accuracy of top20-LRP. Although the performance of top20-LRP is not significantly different from that of 20 motor channels, the p-value is close to significance level with p=0.056. These results indicate the need for following an interpretable channel selection protocol, more so in subject-independent scenarios, as simply using motor channels may not necessarily lead to an optimal performance for all datasets. Closely related to this discussion are our results using the top24-LRP, and we note here that there were electrodes in these sets coming from parietal and occipital areas, as they were identified

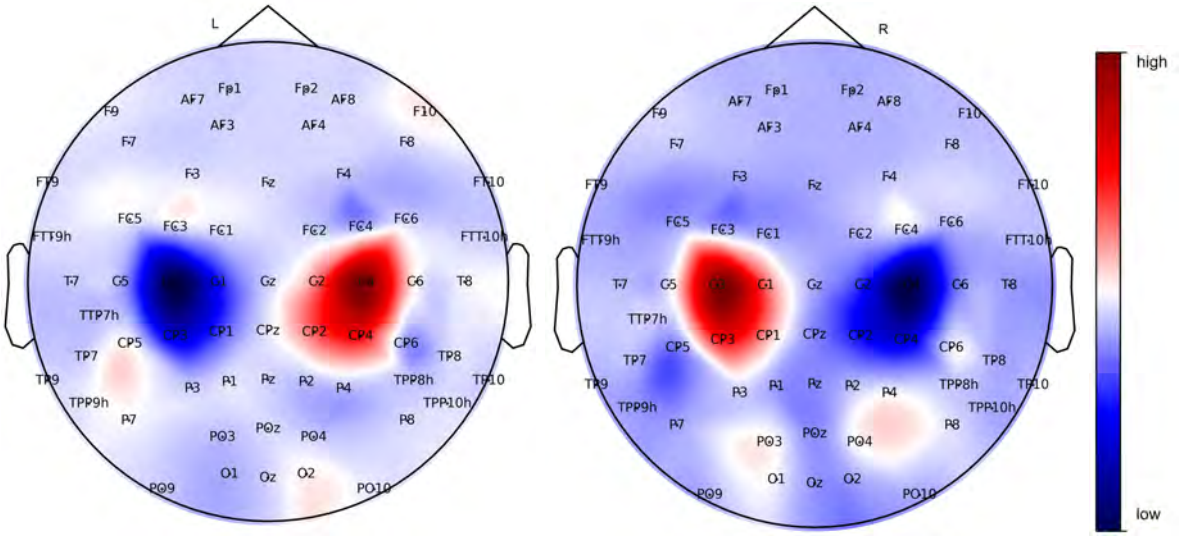

Figure 4: Channel-wise mean LRP relevances across 54 subject-independent models for two-class MI. LRP relevances are class-wise averaged, where "L" indicates left-hand MI, and "R" indicates right-hand MI.

to be relevant for two-class MI classification of KU data. This further justifies the need for relevance based selection of channels in comparison with fixed selection of motor channels for all datasets.

5.4. Channel-wise LRP Relevances

Figure 4 visualizes the channel-wise mean LRP relevances across 54 subject-independent models for two-class MI classification using KU dataset.

In Figure 4, the negatively relevant channels are emphasized in blue, while the positively relevant ones are highlighted in red. The two-class topomaps of channel relevances illustrate event related desynchronization and synchronization (ERD/S) like pattern typically observed during MI, clearly demonstrating the neurophysiological plausibility of the estimated channel relevances. For the left-hand MI, we notice strong positive contributions coming from the right side motor channels such as C2, C4, CP2, and CP4, and negative contributions from C1, C3, CP1, and CP3. We observe a vice versa behaviour for the right-hand MI class. In addition, some of the parietal and occipital channels are also observed to be relevant for MI prediction using KU EEG data. The clear localization of highly relevant channels from the motor region is an indicator of consistency and low variability of channel relevances across different subject-independent models.

5.5. The top24 Channels

For a deeper understanding regarding the neurophysiological plausibility and variability of the selected channels while using LRP relevances versus channel weights, we plotted

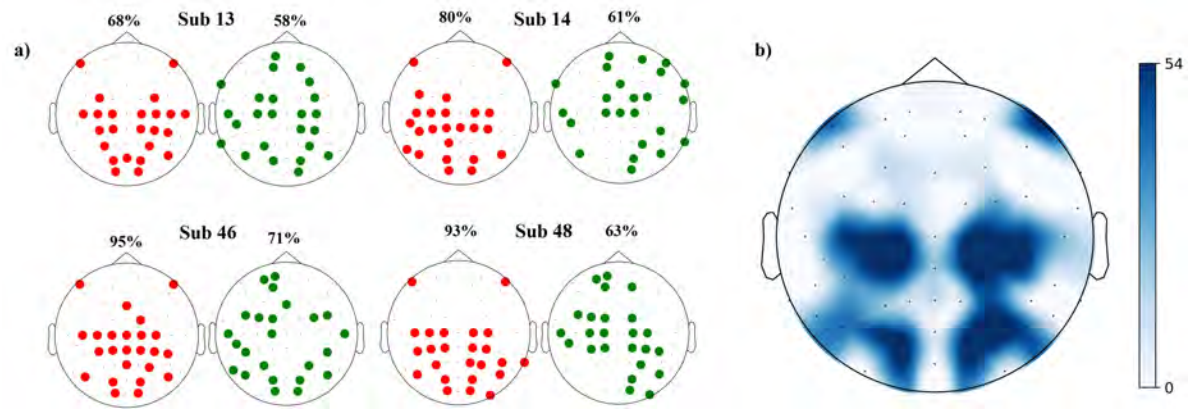

Figure 5: Top 24 channels selected based on LRP relevance score versus magnitude of weights, for exemplary subjects. Channels selected based on relevance are indicated in red, and those selected based on weight are indicated in green. The corresponding accuracies using selected channels are mentioned above each topoplot. In part a, the channel selection for example subjects where top24-LRP outperformed top24-weight is illustrated. In part b, the topomap visualizes the number of occurrences of every channel amongst the top24-LRP selected for 54 target subjects. Channels from motor and parieto-occipital areas are predominantly selected across all subjects.

the top24 channels that were selected in both cases for each of the 54 subject-independent models as individual topomaps. Figure 5a displays these plots for exemplary subjects, for whom top24-LRP outperformed top24-weight. The top24 channels are chosen here for exemplification since $N = 24$ is the minimum number of channels required for our proposed method to achieve subject-independent accuracy not significantly different from the baseline 62-channel LOSO-CV accuracy. The 24-channel accuracy for weight based selection is 80.67% and for relevance based selection is 83.59%. Please see Appendix A, for the complete set of top24-LRP and top24-weight topoplots for all 54 subjects.

From Figure 5a, we note that top24-LRP contains channels from the motor as well as parietal and occipital areas. In addition, the two frontal channels F9 and F10 have been selected. This pattern of selection in top24-LRP is consistent across most target subjects (please see Appendix A). The selection of electrodes from the occipital region, which is dedicated to vision, could be due to the fact that KU MI-EEG data partly consists of trials that used visual feedback. It is also possible that the subjects visually imagined the movement (visual imagery) rather than performing a mental rehearsal of the movement (kinesthetic imagery) [46, 47]. The top24 channels overlap with the electrodes marked as highly relevant in Figure 4. The common set of channels from the top24-LRP across 54 subject-independent models are C1, C3, C2, C4, CP1, CP3, CP4, CP2, and PO4. Except PO4, the remaining channels in this list are part of the motor channel set. However, the channels selected by weight show variability across different models, potentially caused by the subject-independent training of the network

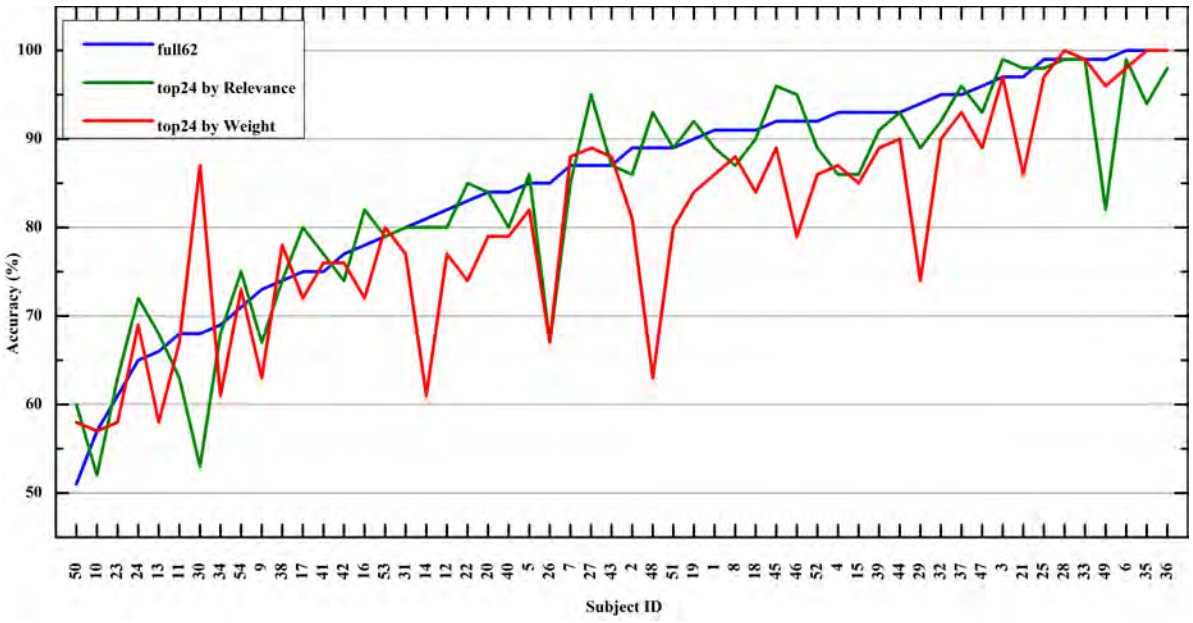

Figure 6: Subject-wise LOSO-CV classification accuracies of Deep ConvNet, using all 62 channels versus top24 channels selected by LRP relevance and magnitude of weight.

that impact the model weights. Selected channels are observed to be coming from different areas of the cortex, including the motor region. Only two channels, P4 and PO4, were common among the top24-weight across all 54 models. Interestingly, PO4 is also present in the common set identified for relevance based selection.

Figure 5b marks the number of occurrences of each channel, amongst the top24 selected for all target subjects based on LRP relevance scores. The topomap clearly demonstrates the predominance of motor, parietal and occipital channels, along with the two frontal channels F9 and F10.

5.6. Subject-wise Classification Accuracies using top24 Channels

In Figure 6, we plot the LOSO-CV classification accuracies individually for all 54 subjects using baseline models, as well as sparse models in which top24 channels were selected based on LRP relevances (top24-LRP) and magnitude of weights (top24-weight). The subjects are sorted by baseline accuracies to gain some insights regarding the performance differences between different subject groups.

The accuracies achieved using top24-LRP are indicated in green. For most subjects, the performance is similar or even better compared to that of the baseline model. Out of 54, 18 subjects illustrate an improvement in performance and 9 subjects show no change in performance with reference to the baseline. Specifically, for subjects 17, 24, 27, and 50, the accuracies are enhanced by 5% or higher. For subject 50, the performance actually improved from close to chance-level to 60%. Out of the remaining subjects, 17 suffered a performance decline of less than 5% and 10 showed a decline of greater than or equal to 5%. In particular, the accuracies of subjects 26, 30 and 49 deteriorated by 18%,

15% and 17%, respectively. We highlight here that the minimum number of channels required for subjects 26, 30, and 49 to achieve an optimal accuracy are 50, 14 and 44, respectively, for which the accuracies obtained are 82%, 95% and 97%, respectively. It seems that subjects 26 and 49 require data from more channels for better classification performance, and vice versa for subject 30.

On the contrary, the subject-independent classification accuracies acquired using top24-weight are lower than the baseline for most subjects. Out of 54, 11 subjects illustrate an improvement in performance and 5 subjects show no change in performance with respect to the baseline. Specifically, for subjects 30 and 50, the accuracies are enhanced by 5% or higher. It is interesting to note that the performance of subject 50 improved significantly using both top24-LRP and top24-weight, indicating the importance of channel selection. Out of the remaining subjects, 14 suffered a performance decline of less than 5% and 24 showed a decline which is 5% or greater. In particular, the accuracies of subjects 9, 14, 21, 26, 29, 46, and 46 declined by 10% or more.

5.7. Model Adaptation Results

The main objective of the adaptation experiments was to compare the performances of adapted 62-channel and 20 motor channel models versus adapted sparse models, that were sparsified by channel selection using LRP relevances (sparse-LRP) and weights (sparse-weight). For this comparison, we used sparse models created by channel selections ranging from $N = 6 - 30$. This channel selection range was considered in order to better understand the impact of channel selection (with fewer channels) on adaptation performance. The adaptation results for the aforementioned representative models obtained from channel selection, can be found in Tables 4 and 5. The average subject-independent accuracy using 20 motor channels is similar with and without adaptation, as reported in Table 3.

From Table 4, we notice that the adaptation results indicate significant performance improvements in sparse-LRP models for $N = 6$, 10, 16-24, 28 and 30. From $N = 10$ onwards, the sparse-LRP models achieved adaptation accuracies comparable with that of the baseline. These results are highlighted in bold in Table 4. Sparse-LRP with $N = 24$ has the highest average subject-independent accuracy of 83.59%, however, sparse-LRP with $N = 22$ achieved the highest adaptation accuracy of 86.33% amongst all sparse models. Sparse-LRP with $N = 18$ gained the highest accuracy improvement of 3.43% post adaptation, compared to other sparse-LRP models.

Table 5 contains the adaptation results of sparse-weight models. Interestingly, all sparse-weight models have gained significant performance improvements from adaptation except $N = 12$, whose adaptation accuracy is also significantly lower than that of the corresponding sparse-LRP model. The adaptation accuracies of sparse-weight models with $N = 6$, 10, 12, 16, and 22 are significantly lower than that of the corresponding sparse-LRP models.

Table 3: Adaptation Results for Baseline and 20 Motor Channel Models.

	Full – 62 channels	Motor – 20 channels
Pre-training Accuracy±SD	84.82±12.07	81.72±12.61
Adaptation Accuracy±SD	85.80±11.88	81.63±13.50

Table 4: Adaptation Results for Sparse-LRP Models.

	N=6	N=8	N=10	N=12	N=14	N=16	N=18	N=20	N=22	N=24	N=26	N=28	N=30
Pre-training Accuracy±SD	77.94±12.91	78.41±12.92	79.93±13.63	81.52±12.25	82.48±12.22	82.57±12.08	81.83±12.29	83.19±11.60	82.94±12.17	83.59±12.15	83.26±12.50	82.85±12.76	82.93±13.13
Adaptation Accuracy±SD	80.17±11.89*	79.91±13.16	**82.28**±12.29*	**83.33**±11.92	**84.28**±11.27	**85.35**±11.26*	**85.26**±11.47*	**85.83**±10.82*	**86.33**±10.64*	**85.74**±11.50*	**85.11**±12.50	**85.59**±11.83*	**85.96**±11.26*

Table 5: Adaptation Results for Sparse-weight Models.

	N=6	N=8	N=10	N=12	N=14	N=16	N=18	N=20	N=22	N=24	N=26	N=28	N=30
Pre-training Accuracy±SD	71.98±12.03	74.98±11.30	77.11±11.57	78.07±12.30	79.00±12.13	80.43±11.44	79.94±12.51	80.82±11.98	81.22±11.85	80.67±12.20	81.67±12.06	81.98±11.97	82.80
Adaptation Accuracy±SD	76.17±12.95**†	77.98±11.93*	79.98±11.95*†	80.32±12.13†	**82.33**±11.36**	**82.74**±12.21*†	**84.39**±11.75**	**84.39**±11.49**	**84.78**±11.15**†	**84.59**±11.58**	**84.28**±11.93*	**85.30**±11.91*	**85.43**±11.91*

The †($p<0.05$) indicates that the post adaptation accuracy is significantly lower than the post-adaptation accuracy of the corresponding sparse-LRP model. The * ($p<0.05$), and **($p<0.001$) indicate that the post adaptation accuracy is significantly higher than the pre-adaptation accuracy of the corresponding sparse model. Adaptation results that are not significantly different from the adaptation result of the baseline 62-channel model, have been highlighted in bold in Tables 4 5.

The sparse-weight models have obtained adaptation accuracies not significantly different from that of the baseline model starting from $N = 14$ onwards. Similar to sparse-LRP, maximum accuracy gain from adaptation was achieved by sparse-weight model using 18 channels.

6. Discussion

We further discuss the significant findings from this study and highlight the novel aspects and relative advantages of our proposed XAI with pruning based subject-independent channel selection method.

Our first set of experiments using the proposed method illustrated that channel selection performed using the LRP relevance scores can significantly reduce the number of channels required to achieve an optimal performance comparable with baseline, as reported in Figure 3 and Table 2. Our results obtained using KU dataset show that more than 60% of the total number of channels can be ignored ($N = 24$) with no significant impact on performance. In addition, the subject-independent classification accuracy suffers less than 1% drop by using only 54% of the channels and just about 3% drop by using only 19% of the total channels. It is also worth noting that the accuracy of these 19% channels selected by LRP relevance, which amounts to 12 out of the 62 channels in KU dataset, is similar to the accuracy obtained using the conventionally used 20 motor channels.

From our second set of experiments, we observed that the magnitude of weight criterion based channel selection could not achieve a subject-independent performance similar to the baseline. This method also needs at least 26% of the channels, in comparison with just 16% for our proposed method, to acquire an accuracy greater than 80%. Overall, the performance of weight magnitude based channel selection is not in par with LRP relevance based channel selection, as is evident from the significantly lower accuracies secured by this method when less than 40% of the total number of channels are selected. This is observable in Figure 3.

Next, by using the relevance maps in Figure 4, we illustrated the neurophysiological plausibility of the channel-wise relevance scores. This further ascertains the advantage of applying an XAI method for channel selection and the resulting interpretability.

In addition, by analyzing the top24 channels selected by LRP relevance and magnitude of weight, as illustrated in Figure 5, we identified and highlighted the consistency in selected channels in case of using LRP relevances and the variability of selection while using magnitude of weights, across all the subject-independent models. We also observe that most channels, selected using LRP relevances, are from the motor, parietal and occipital areas of the cortex, and the channels selected using weights are distributed all over without any clear indication of localization. It should be noted here that as per literature, only the sensorimotor channels are usually expected to contain information related to MI [10, 25]. The nine common channels amongst the top24-LRP selected across 54 subject-independent models come from the central, central parietal

and parieto-occipital regions, for LRP relevance based selection. On the other hand, the two common channels in weight based selection come from parietal and parieto-occipital region, but not the motor region which is involved during ERD/S related to MI [48]. It is to be noted that channel PO4 is present in both common sets.

Visualizing the subject-wise LOSO-CV classification accuracies using the top24 channels selected by the two methods (Figure 6), we observe that LRP relevance based selection illustrates a performance on par with the baseline for most subjects, while weight based selection leads to significant drop in accuracies for several subjects. This is especially evident in subjects with baseline accuracies ranging from 80% to 95%, for whom we notice large accuracy differences between top24-LRP versus top24-weight in Figure 5. This is an interesting finding that requires further investigations. It is possible that those subjects with higher baseline accuracies, above 95%, produce clearer MI related activations and a cleaner EEG with less noise, thus are able to perform well even with fewer channels selected based on weights. However, it is unclear as to how weights based channel selection works well for certain subjects whose baseline accuracies are less than 80%, in comparison with those with baseline performance in the range of 80-95%. We note that certain subjects did not benefit from the proposed relevance based channel selection method, possibly due to large distinction in their EEG features compared to other subjects. Nevertheless, for the remaining subjects we see clear indications of good performance using top24 channels selected by LRP relevance, that are either on par or sometimes even exceeding the baseline.

Next, with the help of Figure 3, we were able to highlight clearly the performance advantages of our proposed LRP relevance based channel selection in comparison with weight magnitude based selection, random selection, and fixed selection of 20 motor channels. By using our proposed method, we are able to achieve an average subject-independent accuracy that is as good as the baseline ($p=0.09$) by using only 38% of the total channel count. In addition, this performance was maintained across several other channel subsets while increasing the number of channels. Furthermore, our proposed method illustrates clear dominance over both weight based selection and random selection, for the same number of channels. This is apparent from the significantly higher accuracies obtained using LRP relevance based selection for smaller number of channels, in comparison with the other two methods. Interestingly, while weight based selection shows variability in performance as the number of channels are increased, random selection demonstrates a relatively smoother improvement in performance with increase in channel count. All three types of channel selection methods are found to saturate in performance beyond 64% of the total number of channels. Accuracy of fixed set of motor channels, although greater than that of random selection and weight based selection, does not indicate a clear superiority over relevance based selection of 20 channels.

Last but not the least, our model adaptation experiments showcased the robust performance of adapted sparse subject-independent models created by channel selection, in comparison with the baseline 62-channel and 20 motor channel subject-independent

models. These results are reported in Tables 3, 4 and 5. The pre- and post-adaptation accuracies of sparse-LRP models are higher than the corresponding accuracies of sparse-weight models. The performance of the adapted sparse-LRP model using only 16% of the total number of channels is similar to that of the adapted baseline model (p=0.13). The accuracy of the adapted sparse-LRP model using only 35% of the total number of channels is 0.53% higher (p=0.81) than that of the adapted baseline model using all channels. The adapted sparse-weight models exhibited good performance, as well. All adaptation accuracies obtained using sparse-weight models were significantly better than their corresponding pre-adaptation performances. However, the sparse-weight models required at least 22% of total channels to acquire a similar adaptation accuracy as that of the baseline. In addition, the adaptation accuracies of sparse-weight models, for certain values of N, were significantly lower compared to the corresponding performances of sparse-LRP models, as reported in Table 5. We identified that sparse models using 29% of the total channels, selected either by LRP or weight, were robust to adaptation and gained maximum improvement in accuracy. Although the adapted sparse-LRP models have secured higher performances in comparison with the adapted sparse-weight models, it should be noted that the former begin with a higher accuracy before adaptation in first place. In addition, the sparse-weight models achieved significantly higher post-adaptation accuracies with respect to their corresponding pre-adaptation accuracies, for all channel subsets from $N = 6 - 30$, while the sparse-LRP models did not consistently show a significant improvement. Further investigations are required to better understand the differences between sparse-LRP and sparse-weight models with respect to robustness towards adaptation. Overall, the model adaptation results highlight the usefulness of having sparse models that are more robust to adaptation and also illustrate the dominance of sparse models obtained using our proposed novel subject-independent channel selection method that combines XAI based relevance with neural network pruning.

Altogether, the results indicate that by using the channel-wise relevances computed by LRP, one can select a channel subset that would offer comparable performance for an unknown subject. Our proposed channel selection approach also allows us to control the selected number of channels for subject-independent classification. Using our proposed method with Deep ConvNet and KU dataset, we achieved a subject-independent accuracy as good as 62-channel baseline by using only 38% of the channels (p=0.09).

Limitations and Future Work: Our proposed channel selection method has demonstrated good performance while being highly interpretable. Nevertheless, our method has its own limitations. Although our proposed method selects channels in a subject-independent manner, the selected channels actually vary between different target subjects based on the training data used for estimating the relevances. For the channel selection to be truly subject-independent, we believe that identifying a common subset that can work optimally across all subjects is crucial.

We have applied the basic LRP-epsilon rule for estimating the channel-wise

relevance scores in this study. Future investigations may include comparisons with other LRP rules, different XAI metrics or even custom rules to define the calculation of the relevance scores for a more robust selection of channels. We have conducted this subject-independent channel selection study using a single EEG dataset that contains MI data from 54 subjects. Evaluations of the proposed method using multiple datasets is a potential future work, and will help to gather more generalizable insights regarding channel selection by relevance. In addition, our channel selection method needs to be validated using other DL based BCI models. Real-time evaluations of our proposed method will be performed in future.

7. Conclusion

In summary, the aim of this study was to propose and evaluate a novel methodology for subject-independent channel selection in DL based MI-BCI. We applied sophisticated techniques from deep learning for the purpose of channel selection, such as LRP and neural network pruning. In addition, we evaluated the performances of adapted models before and after channel selection.

Our main results from the channel selection experiments using Deep ConvNet and KU dataset are: (i) by selecting just about 38% of the total number of channels using our proposed method, an average subject-independent accuracy which is not significantly different from the baseline (p=0.09) can be achieved, with a minor drop of 1.23% (ii) by selecting as few as 12 out of the 62 channels using the channel-wise LRP relevances, a subject-independent classification accuracy that is similar to the 20 motor channel accuracy, and which is only 3.3% lower than the baseline can be obtained (iii) channels selected using LRP relevance indicate the influence of motor, parietal and occipital regions in MI classification (iv) the channel-wise average LRP relevances illustrate an ERD/S like pattern, signifying high neurophysiological plausibility in relevance estimation and thereby channel selection for MI-BCI (v) the subject-wise classification performances using the top 38% channels selected by our proposed method, are on par with the baseline accuracies for most subjects (vi) in comparison with weight based and random channel selections, LRP relevance based selections provide with significantly better accuracies especially for fewer channels (vii) the performance of the adapted sparse-LRP model using only 16% of the total number of channels is similar to that of the adapted baseline model (p=0.13), and the accuracy of the adapted sparse-LRP model using only 35% of the total number of channels exceeded that of the adapted baseline model by 0.53% (p=0.81). Through our results, we establish the significance of combining XAI based relevance with pruning to perform selection of highly relevant channels for subject-independent MI classification. Through this study, we signify the importance of channel selection for user comfort and convenience during data recording, for removal of noisy and redundant channels, and for realizing sparse subject-independent deep learning models that are robust, interpretable, and efficient.

Acknowledgment

This work was partially supported by the RIE2020 AME Programmatic Fund, Singapore (No. A20G8b0102). The computational work for this article was partially performed on the resources provided by the National Supercomputing Centre, Singapore.

References

[1] Lécuyer A, Lotte F, Reilly R B, Leeb R, Hirose M and Slater M 2008 Brain-computer interfaces, virtual reality, and videogames *Computer* ISSN 00189162

[2] Mak J N and Wolpaw J R 2009 Clinical Applications of Brain—Computer Interfaces: Current State and Future Prospects *IEEE Reviews in Biomedical Engineering* ISSN 19411189

[3] Mulder T 2007 Motor imagery and action observation: Cognitive tools for rehabilitation *Journal of Neural Transmission* ISSN 03009564

[4] Kübler A, Nijboer F, Mellinger J, Vaughan T M, Pawelzik H, Schalk G, McFarland D J, Birbaumer N and Wolpaw J R 2005 Patients with als can use sensorimotor rhythms to operate a brain-computer interface *Neurology* ISSN 00283878

[5] Ang K K, Guan C, Chua K S G, Ang B T, Kuah C, Wang C, Phua K S, Chin Y Z and Zhang H 2010 Clinical study of neurorehabilitation in stroke using eeg-based motor imagery brain-computer interface with robotic feedback ISBN 9781424441235

[6] Blankertz B, Tomioka R, Lemm S, Kawanabe M and Müller K R 2008 Optimizing spatial filters for robust EEG single-trial analysis *IEEE Signal Processing Magazine* ISSN 10535888

[7] Ang K K, Chin Z Y, Zhang H and Guan C 2008 Filter Bank Common Spatial Pattern (FBCSP) in brain-computer interface *Proceedings of the International Joint Conference on Neural Networks* ISBN 9781424418213

[8] Schirrmeister R T, Springenberg J T, Fiederer L D J, Glasstetter M, Eggensperger K, Tangermann M, Hutter F, Burgard W and Ball T 2017 Deep learning with convolutional neural networks for eeg decoding and visualization *Human Brain Mapping* ISSN 10970193

[9] Lawhern V J, Solon A J, Waytowich N R, Gordon S M, Hung C P and Lance B J 2018 EEGNet: A compact convolutional neural network for EEG-based brain-computer interfaces *Journal of Neural Engineering* ISSN 17412552 (*Preprint* arXiv:1611.08024)

[10] Mane R, Robinson N, Vinod A P, Lee S W and Guan C 2020 A Multi-view CNN with Novel Variance Layer for Motor Imagery Brain Computer Interface *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS* ISBN 9781728119908 ISSN 1557170X

[11] Nagarajan A, Robinson N and Guan C 2021 Investigation on robustness of eeg-based brain-computer interfaces *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference* **2021** ISSN 26940604

[12] Shan H, Xu H, Zhu S and He B 2015 A novel channel selection method for optimal classification in different motor imagery bci paradigms *BioMedical Engineering Online* **14**(1) ISSN 1475925X

[13] Jin J, Miao Y, Daly I, Zuo C, Hu D and Cichocki A 2019 Correlation-based channel selection and regularized feature optimization for mi-based bci *Neural Networks* **118** ISSN 18792782

[14] Gaur P, McCreadie K, Pachori R B, Wang H and Prasad G 2021 An automatic subject specific channel selection method for enhancing motor imagery classification in eeg-bci using correlation *Biomedical Signal Processing and Control* **68** ISSN 17468108

[15] Varsehi H and Firoozabadi S M P 2021 An eeg channel selection method for motor imagery based brain–computer interface and neurofeedback using granger causality *Neural Networks* **133** ISSN 18792782

[16] Schröder M, Lal T N, Hinterberger T, Bogdan M, Hill N J, Birbaumer N, Rosenstiel W and

Schölkopf B 2005 Robust eeg channel selection across subjects for brain-computer interfaces *Eurasip Journal on Applied Signal Processing* **2005**(19) ISSN 11108657

[17] Arpaia P, Donnarumma F, Esposito A and Parvis M 2021 Channel selection for optimal eeg measurement in motor imagery-based brain-computer interfaces *International Journal of Neural Systems* **31**(3) ISSN 17936462

[18] Feng J K, Jin J, Daly I, Zhou J, Niu Y, Wang X and Cichocki A 2019 An optimized channel selection method based on multifrequency csp-rank for motor imagery-based bci system *Computational Intelligence and Neuroscience* **2019** ISSN 16875273

[19] Arvaneh M, Guan C, Ang K K and Quek C 2011 Optimizing the channel selection and classification accuracy in eeg-based bci *IEEE Transactions on Biomedical Engineering* **58**(6) ISSN 00189294

[20] Parashiva P K and Vinod A P 2019 A new channel selection method using autoencoder for motor imagery based brain computer interface vol 2019-October ISSN 1062922X

[21] Arvaneh M, Guan C, Ang K K and Quek H C 2012 Robust eeg channel selection across sessions in brain-computer interface involving stroke patients *The 2012 International Joint Conference on Neural Networks (IJCNN)* 1–6

[22] Gunning D and Aha D W 2019 Darpa's explainable artificial intelligence program *AI Magazine* **40**(2) ISSN 07384602

[23] Arrieta A B, Díaz-Rodríguez N, Ser J D, Bennetot A, Tabik S, Barbado A, Garcia S, Gil-Lopez S, Molina D, Benjamins R, Chatila R and Herrera F 2020 Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai *Information Fusion* **58** ISSN 15662535

[24] Bach S, Binder A, Montavon G, Klauschen F, Müller K R and Samek W 2015 On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation *PLoS ONE* **10**(7) ISSN 19326203

[25] Lee M H, Kwon O Y, Kim Y J, Kim H K, Lee Y E, Williamson J, Fazli S and Lee S W 2019 EEG dataset and OpenBMI toolbox for three BCI paradigms: An investigation into BCI illiteracy *GigaScience* ISSN 2047217X

[26] Tjoa E and Guan C 2021 A survey on explainable artificial intelligence (xai): Toward medical xai *IEEE transactions on neural networks and learning systems* **32** 4793—4813 ISSN 2162-237X URL `https://doi.org/10.1109/TNNLS.2020.3027314`

[27] Yeom S K, Seegerer P, Lapuschkin S, Binder A, Wiedemann S, Müller K R and Samek W 2021 Pruning by explaining: A novel criterion for deep neural network pruning *Pattern Recognition* **115** ISSN 00313203

[28] Molchanov P, Tyree S, Karras T, Aila T and Kautz J 2017 Pruning convolutional neural networks for resource efficient inference

[29] Li H, Samet H, Kadav A, Durdanovic I and Graf H P 2017 Pruning filters for efficient convnets

[30] Liu C and Wu H 2019 Channel pruning based on mean gradient for accelerating convolutional neural networks *Signal Processing* **156** ISSN 01651684

[31] Blalock D, Ortiz J J G, Frankle J and Guttag J 2020 What is the state of neural network pruning? URL `https://arxiv.org/abs/2003.03033`

[32] Han S, Pool J, Tran J and Dally W J 2015 Learning both weights and connections for efficient neural networks vol 2015-January ISSN 10495258

[33] Vishnupriya R, Robinson N, M R R and Guan C 2021 Performance evaluation of compressed deep cnn for motor imagery classification using eeg ISSN 1557170X

[34] LeCun Y, Bottou L, Bengio Y and Haffner P 1998 Gradient-based learning applied to document recognition *Proceedings of the IEEE* **86**(11) ISSN 00189219

[35] Clevert D A, Unterthiner T and Hochreiter S 2016 Fast and accurate deep network learning by exponential linear units (ELUs) *4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings* (*Preprint* arXiv:1511.07289)

[36] Montavon G, Lapuschkin S, Binder A, Samek W and Müller K R 2017 Explaining nonlinear classification decisions with deep taylor decomposition *Pattern Recognition* **65** ISSN 00313203

[37] Sturm I, Lapuschkin S, Samek W and Müller K R 2016 Interpretable deep neural networks for single-trial eeg classification *Journal of Neuroscience Methods* **274** ISSN 1872678X
[38] Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, Desmaison A, Kopf A, Yang E, DeVito Z, Raison M, Tejani A, Chilamkurthy S, Steiner B, Fang L, Bai J and Chintala S Pytorch: An imperative style, high-performance deep learning library
[39] Kokhlikyan N, Miglani V, Martin M, Wang E, Alsallakh B, Reynolds J, Melnikov A, Kliushkina N, Araya C, Yan S and Reblitz-Richardson O 2020 Captum: A unified and generic model interpretability library for pytorch (*Preprint* arXiv:2009.07896)
[40] Janowsky S A 1989 Pruning versus clipping in neural networks *Physical Review A* **39**(12) ISSN 10502947
[41] Mozer M C and Smolensky P 1989 Skeletonization: A technique for trimming the fat from a network via relevance assessment vol 1
[42] Kingma D P and Ba J L 2015 Adam: A method for stochastic optimization
[43] Ioffe S and Szegedy C 2015 Batch normalization: Accelerating deep network training by reducing internal covariate shift vol 1
[44] Srivastava N, Hinton G, Krizhevsky A, Sutskever I and Salakhutdinov R 2014 Dropout: A simple way to prevent neural networks from overfitting *Journal of Machine Learning Research* **15** ISSN 15337928
[45] Zhang K, Robinson N, Lee S W and Guan C 2021 Adaptive transfer learning for eeg motor imagery classification with deep convolutional neural network *Neural Networks* ISSN 18792782
[46] Guillot A, Collet C, Nguyen V A, Malouin F, Richards C and Doyon J 2009 Brain activity during visual versus kinesthetic imagery: An fmri study *Human Brain Mapping* **30**(7) ISSN 10659471
[47] Solodkin A, Hlustik P, Chen E E and Small S L 2004 Fine modulation in network activation during motor execution and motor imagery *Cerebral Cortex* **14**(11) ISSN 10473211
[48] Pfurtscheller G, Brunner C, Schlögl A and da Silva F H L 2006 Mu rhythm (de)synchronization and eeg single-trial classification of different motor imagery tasks *NeuroImage* **31**(1) ISSN 10538119

Appendix A. Top 24 channels selected by relevance versus weight

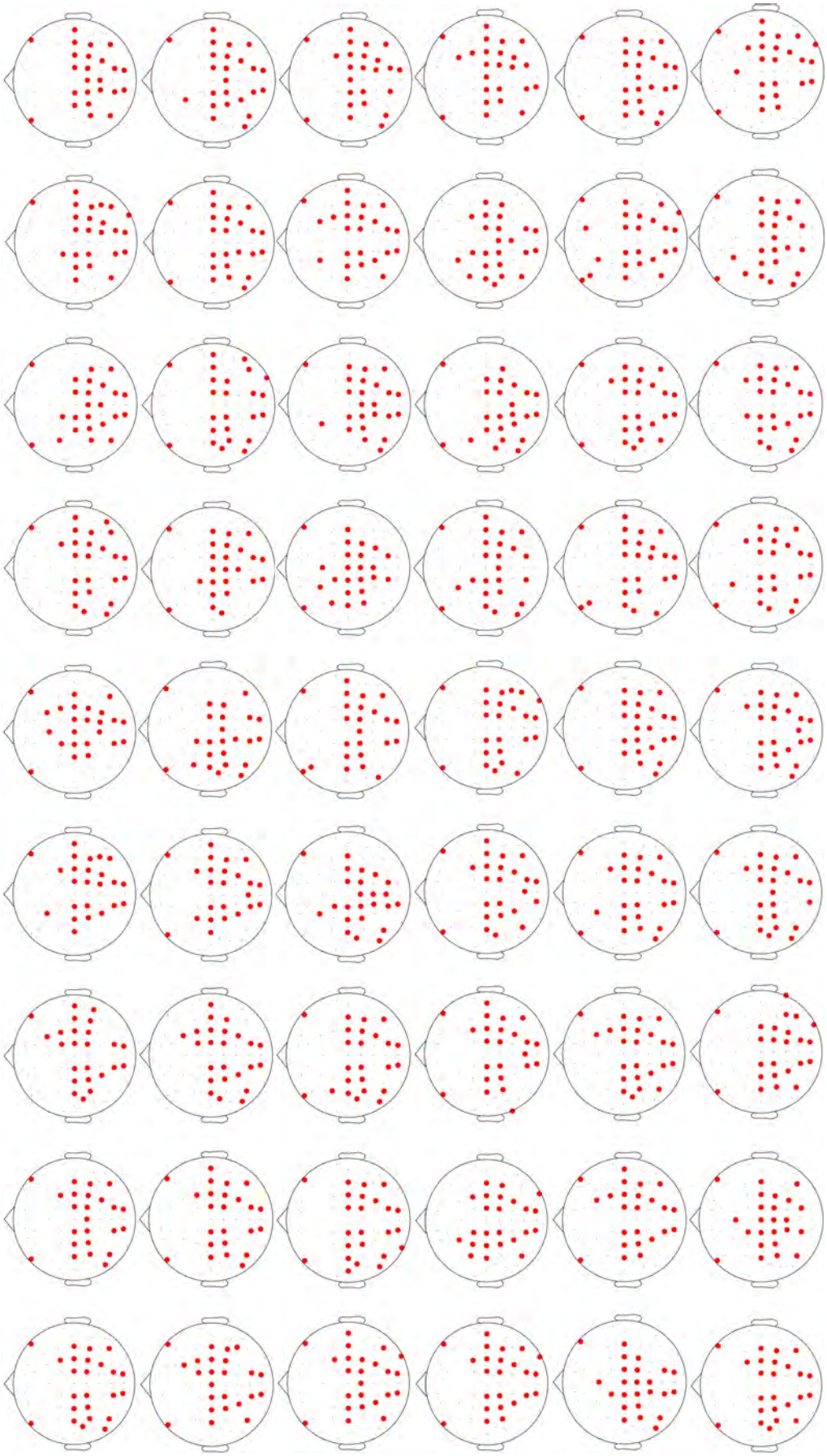

Figure A1: Top 24 channels selected based on LRP relevances, for 54 subject-independent models. Selected channels are highlighted in red.

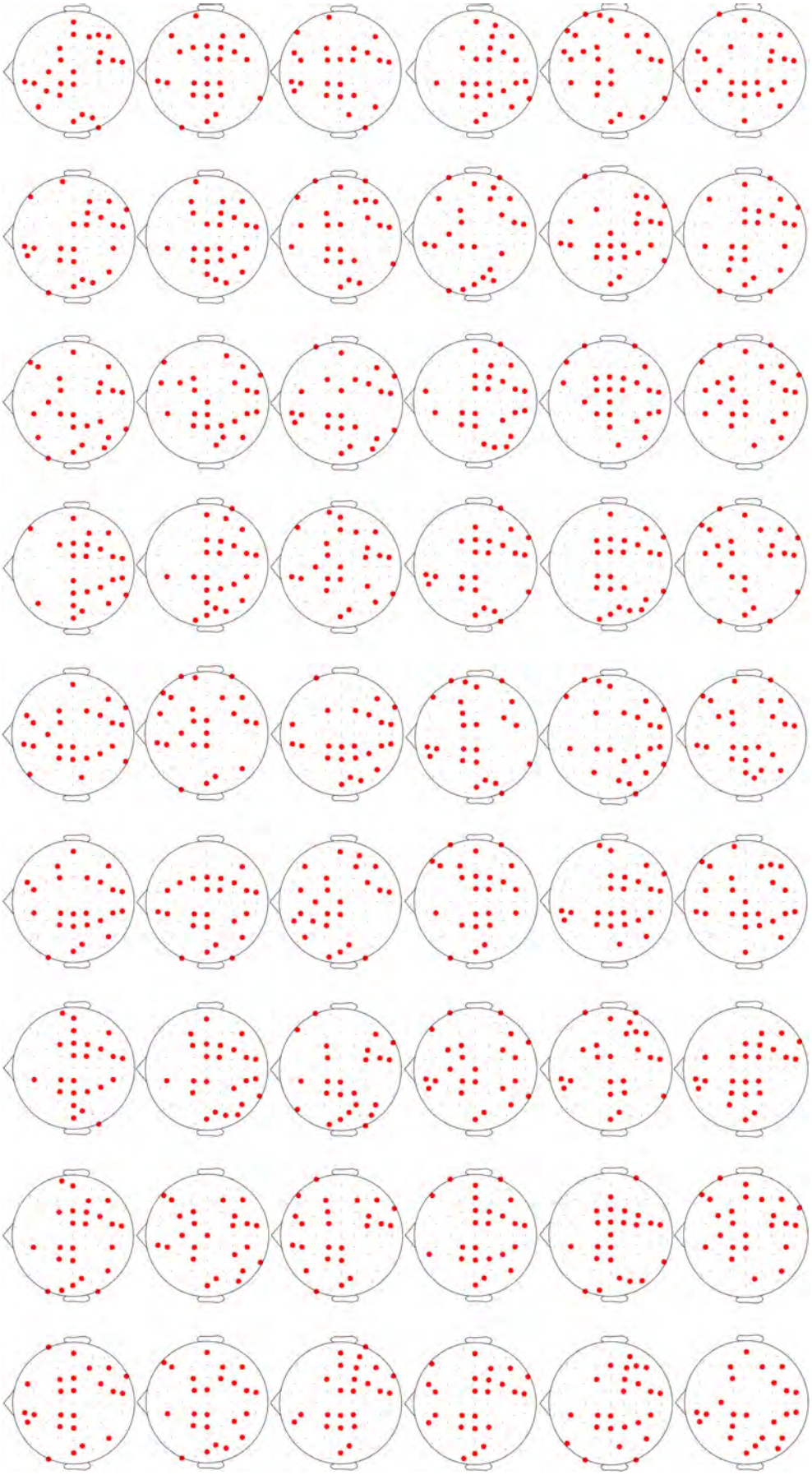

Figure A2: Top 24 channels selected based on magnitude of weight, for 54 subject-independent models. Selected channels are highlighted in red.