Channel Selection for Stereo-Electroencephalography (SEEG)-Based Invasive Brain-Computer Interfaces Using Deep Learning Methods

Xiaolong Wu, Guangye Li, Xin Gao, Benjamin Metcalfe, and Dingguo Zhang, *Senior Member, IEEE*

***Abstract*— Brain-computer interfaces (BCIs) can enable direct communication with assistive devices by recording and decoding signals from the brain. To achieve high performance, many electrodes will be used, such as the recently developed invasive BCIs with channel numbers up to hundreds or even thousands. For those high-throughput BCIs, channel selection is important to reduce signal redundancy and invasiveness while maintaining decoding performance. However, such endeavour is rarely reported for invasive BCIs, especially those using deep learning methods. Two deep learning-based methods, referred to as Gumbel and STG, were proposed in this paper. They were evaluated using the Stereo-electroencephalography (SEEG) signals, and compared with three other methods, including manual selection, mutual information-based method (MI), and all channels (all channels without selection). The task is to classify the SEEG signals into five movements using channels selected by each method. When 10 channels were selected, the mean classification accuracies using Gumbel, STG (referred to as STG-10), manual selection, and MI selection were 65%, 60%, 60%, and 47%, respectively, whilst the accuracy was 59% using all channels (no selection). In addition, an investigation of the selected channels showed that Gumbel and STG have successfully identified the pre-central and post-central areas, which are closely related to motor control. Both Gumbel and STG successfully selected the informative channels in SEEG recordings while maintaining decoding accuracy. This study enables future high-throughput BCIs using deep learning methods, to identify useful channels and reduce computing and wireless transmission pressure.**

Manuscript received 20 July 2023; revised 29 September 2023 and 1 December 2023; accepted 7 February 2024. Date of publication 13 February 2024; date of current version 20 February 2024. This work was supported in part by the EPSRC New Horizons Grant of U.K. under Grant EP/X018342/1 and in part by the China Postdoctoral Science Foundation under Grant 20Z102060158. *(Corresponding author: Dingguo Zhang.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Ethical Committee of the University of Bath under Reference No. EP 20/21 050 and the Ethics Committee of Huashan Hospital, Shanghai, China, under Reference No. KY2019518.

Xiaolong Wu, Xin Gao, Benjamin Metcalfe, and Dingguo Zhang are with the Department of Electronic and Electrical Engineering, University of Bath, BA2 7AY Bath, U.K. (e-mail: d.zhang@bath.ac.uk).

Guangye Li is with the School of Mechanical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China.

Digital Object Identifier 10.1109/TNSRE.2024.3364752

***Index Terms*— Stereo-electroencephalography (SEEG), brain-computer interface (BCI), channel selection, deep learning, movement classification.**

I. Introduction

BRAIN-COMPUTER interfaces (BCIs) are technologies that decode brain signals to intentions and hold the promising potential to restore lost functionality. Recording interfaces usually contain many electrodes that in turn give rise to multiple channels. For example, there could be between 19 and 25 electrodes in a standard scalp EEG, and in high-density EEG this number can rise to 256 [1], [2], [3]. The channel number can be even higher for invasive BCIs. For example, recent developments in invasive BCI have witnessed a trend of high-density and high-throughput recording devices that can simultaneously record thousands of channels [4], [5], [6]. This high channel count design may enhance BCI performance, but it also brings challenges associated with high data dimensionality and the sheer quantity of data to be processed and communicated [7]. For example, high-density electrodes may introduce redundancy between neighbouring electrodes, and the processing of big data requires sophisticated signal processing steps and leads to more time delay. For invasive BCIs, the implantation of BCI with high throughput would require a more invasive and complex surgical operation, which brings higher safety issues. In addition, for wireless invasive BCIs which perform signal pre-processing using chips inside the brain, the heat generated by heavy computing might bring damage to the brain tissue [8]. Therefore, channel selection can be used to select the most informative channels to achieve a balance between decoding accuracy and the number of implanted channels.

On the other hand, signals not related to the decoding target (noise) might be picked up if the location of the electrode is suboptimal. For example, the emerging Stereo-electroencephalography (SEEG) based BCIs use approximately hundreds of depth electrodes to capture large-scale intracranial neural activity from various cortical and subcortical areas [9], [10]. It has been demonstrated that these areas contain very different signals when performing a motor task [11]. Therefore, channel selection is helpful to extract useful information.

A. Channel Selection Methods

Methods to select channels can be categorized into four strategies: filtering, wrapping, embedded, and human-based techniques. These methods have been extensively studied in non-invasive BCI (see [12]). In filtering methods, channels are selected based on an independent evaluation criterion, such as a distance measure, an information measure, a dependency measure, or a consistency measure. For example, He et al. presented a statistical channel selection method for classifying motor imagery using a sequential fast-forward search strategy to find the optimal combination of channels [13]. In the case of wrapping techniques, a classification algorithm is used to evaluate the candidate channel subsets, which are generated by a search algorithm. For example, a wrapper approach with a random search strategy for subset channel selection has been adopted in a study of motor imagery classification tasks [14]. However, this method needs to retrain the decoder every time a subset of channels is chosen, and consequently, they are generally more computationally expensive than filtering techniques. In the embedded techniques, the channels are selected simultaneously with the decoding process. For example, channel selection was achieved using a recursive feature elimination (RFE) method during the training of an SVM model in a motor imagery classification task [15]. Embedded methods based on deep learning have also been studied. For example, a squeeze-and-excitation block was incorporated into a convolution neural network (CNN) model to perform automatic channel selection in a motor imagery recognition task [16]. In another EEG-based BCI study, channel selection was achieved using a deep learning method by re-parametrizing the discrete channel sampling problem using the Gumbel-softmax trick [17].

Compared to the aforementioned studies about channel selection in non-invasive signals, channel selection in invasive BCI has attracted much less attention, especially using the deep learning method. Although such studies are rare, it is more important for the invasive BCI to achieve a balance between decoding accuracy and invasiveness by choosing subset effective channels. In an SEEG BCI research, Li et al. used the correlation between the power of each frequency band and the task state to select channels [18]. As a wrapper method, the forward optimal feature selection (fOFS) method was used in a hand gesture classification task using SEEG signals [19]. Genetic-based methods have also been explored [20].

The aforementioned conventional methods are often suboptimal in that only a subset of channel combinations can be explored by the heuristic searching strategies. It also faces difficulties in the high throughput context because of the over-fitting problem when the channel count is high [21]. To solve these problems, in this paper, two end-to-end deep learning-based channel selection methods will be evaluated using the SEEG data recorded during participants performing arm or hand movements. The task is to identify a channel subset that can be used to classify the signals into five classes with decoding accuracy comparable to that using all channels. This paper focused on the deep learning-based method because deep learning has attracted increasing attention in BCIs, and it has been proven to be comparable, most time superior to the traditional methods [10], [22], [23], [24], [25]. Compared to the conventional methods, integrating the channel selection function into the deep learning framework is advantageous in several aspects. First, the global optimum solution (the best channel combination) can be obtained using the gradient descent strategy to search over the entire space. Second, channel selection can be achieved simultaneously with the decoding task at hand during the end-to-end training process, and therefore, no need to retrain the decoder multiple times. Third, the proposed methods are modules that can be used in a plug-and-play fashion. This is highly desirable because they can be used in any other deep-learning architectures, including some of the most advanced networks, such as the *Transformer* model [26].

The novelty of this manuscript is three-folded:

- To our best knowledge, it is the first time that the deep learning-based channel selection method has been studied on the SEEG data.
- To our best knowledge, it is the first time the STG method has been studied on BCIs.
- Compared to the previous non-invasive BCI channel selection studies, this manuscript studied the selected electrodes in different brain regions. This analysis helps to gain new insight into the movement representation inside the brain.

B. Deep Learning Method

Under the deep learning framework, one strategy for channel selection is to learn one-hot weights, in which the ones represented the chosen channels and the zeros are the unselected ones. However, this causes an inability to perform backpropagation due to the categorical latent variables (corresponding to the channel weights). A common strategy to tackle this problem is to reparametrize the discrete distribution with continuous relaxation. This paper will evaluate two example relaxation methods. The first one referred to as the Gumbel selection, is to use a gradient estimator to replace the non-differentiable sample from a categorical distribution with a differentiable sample from a novel Gumbel-Softmax distribution [27]. This method has been shown to be efficient to build a concrete autoencoder, an end-to-end differentiable method for global feature selection [28], and it has been reported in a recent scalp EEG study [17]. The second method, referred to as the STG selection, uses a Gaussian-based continuous relaxation of the Bernoulli variables, which represent the stochastic gating (hence STG) of each discrete variable [29]. This paper will implement these two methods as deep learning networks (referred to as the selection subnets). The selection subnet (Gumbel or STG) will be stacked on top of a decoding subnet, which can be either a classification or regression subnet. Then, the stacked networks will be jointly trained to extract the informative channels and classify the SEEG signals simultaneously.

II. Experiment Setup

To evaluate the proposed methods, the SEEG signals which were acquired during participants performing five different hand or forearm movements will be used. The task in this experiment was to classify SEEG signals into five different movement classes using subset channels. The detailed experiment paradigm and decoding algorithm will be presented in the following sections.

TABLE I
CLINICAL PROFILES OF PARTICIPANTS IN THE STUDY

SID	EZ	DH	EH	Gender	Age	RH	EL	NC	SR (Hz)
1	inferior frontal gyrus	R	R	F	23	LH	10	121	1000
2	left occipital lobe	R	R	M	33	LH	15	180	1000
3	right central region	R	L	F	30	RH	7	60	1000
4	right temporal lobe	R	L	M	26	RH	13	178	1000
5	right inferior frontal gyrus	R	L	M	25	RH	10	143	1000
6	right temporal & insular lobe	R	L	F	17	BI	10	169	1000
7	right frontal	R	L	F	28	RH	9	114	1000
8	left temporal parietal lobe	R	R	M	27	LH	16	208	2000
9	right temporal lobe	R	L	M	15	BI	13	194	1000
10	right superior parietal lobe	R	L	M	31	RH	6	94	1000
11	right superior parietal lobe	R	L	M	31	RH	6	102	2000
12	right ACC	R	L	M	19	BI	9	130	2000

Abbreviations for this Table: SID: Participant ID; EZ, Epileptogenic zone; DH, Dominant hand; EH, Experiment hand; RH, Recording hemisphere; EL, Number of electrode shafts; NC: Number of contacts; SR, Sampling rate; ACC, Anterior cingulate cortex; RH, Right hemisphere; LH, Left hemisphere; BI, Bilateral;

A. Participants and Data Recording

There were 12 human participants (referred to as *1*, *2*...,*12*) recruited in this study. The participants were patients with intractable epilepsy and were implanted with SEEG electrodes for pre-surgical assessment of seizure focus. All participants were enrolled with written consent. The clinical profile of all participants is shown in **Table I**. All implantation parameters were determined solely by clinical needs. SEEG signals were acquired using a clinical recording system (EEG-1200C, Nihon Kohden, Irvine, CA) and sampled at 1000 or 2000 Hz. Each electrode shaft was 0.8 mm in diameter with 8–16 contacts (Huake Hengsheng Medical Corp., Beijing, CN).

This study was reviewed and approved by the Ethical Committee of the University of Bath (Ethical approval reference №: EP 20/21 050) and the Ethics Committee of Huashan Hospital (Shanghai, China) (Ethical approval reference №: KY2019518).

B. Experimental Protocol

The experimental paradigm is shown in Fig. 1. The participants were reclined on a hospital bed during the whole experiment. One trial lasts for 10 s (4 s rest, 1 s cue, and 5 s task). To begin the trial, the participants kept still for 4 seconds (resting stage). Then, a visual cue (a cross) was shown on an LCD screen for 1 s (cue stage). When the cue stage ended, the cross disappeared and a picture of one of five tasks was presented (grasp, scissor gesture, elbow flexion, wrist supination, thumb flexion). The participant performed the specified task repeatedly for 5 s, using the hand contralateral to the hemisphere with the majority of the implanted SEEG electrodes. The five tasks were randomly presented for a total of 20 times per task. In the end, there are a total of 100 trials per participant (16.67 mins total).

C. Electrode Localization

The 12 participants had a total of 1554 contacts (rounded mean ± std: 129 ± 32 per participant) implanted. The electrode locations, in a standard brain model (Montreal Neurological Institute (MNI)), were obtained using an open-source toolbox, iEEGview [30]. The anatomical label of each contact was identified using Freesurfer's cortical parcellation and subcortical segmentation [31].

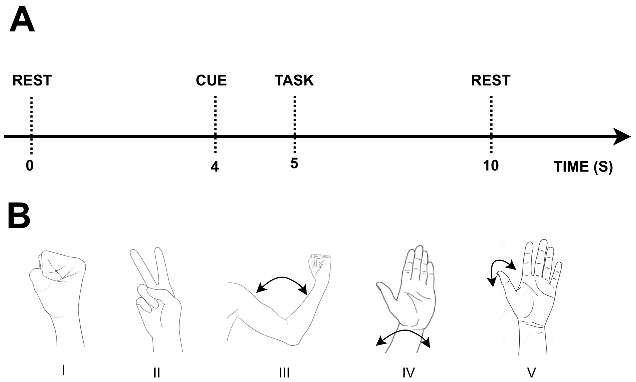

Fig. 1. The experiment paradigm. (A) A trial consisted of three stages: rest, cue, and task. In each trial, participants will keep still during the 4-second resting stage. Then, a cue (a white cross on the screen) will appear for 1 second. After the cue disappears, a picture of one of five movements will appear, and the participants will perform the corresponding task repeatedly for 5 seconds. (B) Five movement tasks were used in this experiment: grasp, scissor gesture, elbow flexion, wrist supination, and thumb flexion.

III. METHODS

In this section, the data preprocessing step and the calculation of temporal-spectral representation will be briefly introduced. Then, detailed information about two proposed methods (Gumbel and STG selection) will be presented. Both implementations of these two methods used a stacked architecture, consisting of a selection subnet and a classification subnet. Whilst the classification subnet was the same, the implementation of the selection subnet was different. Therefore, the stacking architecture will be presented first, and then the selection subnet, specific to each method, will be further introduced respectively.

A. Signal Pre-Processing & Temporal-Spectral Representation

First, the SEEG data was down-sampled to 1000 Hz. The SEEG signals were then band-pass filtered from 0.5 Hz

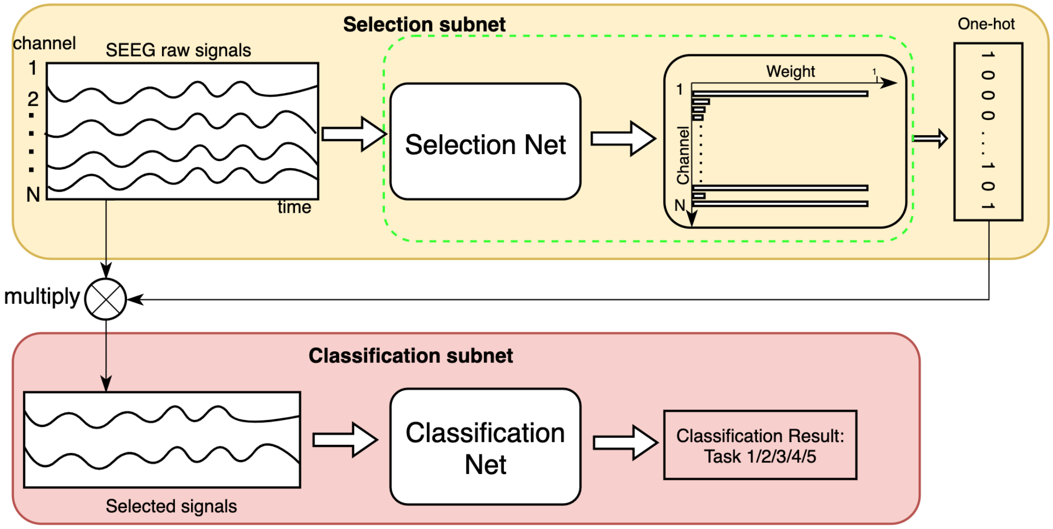

Fig. 2. Plot of the stacked channel selection and classification subnets. For input of $\boldsymbol{X}$, in the shape of $N \times T$ (N, T denoted channel number and temporal sampling points, respectively.), the selection subnet will learn one-hot-like weights. The channel will be selected by multiplying the input data with the one-hot-like vector. The resulting selected channel will be fed into a classification net to classify the SEEG signals. The part framed in the green area was implemented differently using the proposed Gumbel and STG methods, while other components were kept the same.

to 400 Hz using a 4^{th} order Butterworth filter, and a notch filter was used to eliminate 50 Hz line noise. Next, channels with extensive line noise were identified and excluded in the following calculation using the same method from our previous work [10].

The temporal spectral representation of the SEEG signals was obtained using the MNE toolbox [32]. Then, event-related desynchronization (ERD) and event-related synchronization (ERS) were calculated in frequency bands 0.5-30 Hz and 60-150 Hz, respectively, as described in our previous work [10].

B. Channel Selection Methods

The two proposed methods follow a similar architecture, in which a selection subnet is stacked on top of a classification subnet. In this stacked architecture, the first channel selection subnet performs channel selection, and the classification subnet will classify the selected data into five movements. An illustration of this stacked architecture is presented in Fig. 2.

In this stacking arrangement, the complete network can be represented by:

$$y = \mathcal{C}(\mathcal{S}(\boldsymbol{X})) \tag{1}$$

in which $\mathcal{S}$ and $\mathcal{C}$ denote the selection subnet and the classification subnet, respectively. Raw input signals were represented by $\boldsymbol{X}$ which were in the shape of $N \times T$ (N, T denoted channel number and temporal sampling points, respectively). For the two selection methods, both try to learn one-hot-like weights representing the importance of each channel to the decoding task. However, these two methods differ in how the one-hot weights were learned. For the Gumbel method, the channel number to be selected needs to be supplied to the model as *a priori*, while, for the STG selection method, the channel number will be determined by a weight sparsity hyperparameter. Detailed information on these two methods will be presented below.

1) Gumbel Selection: The Gumbel method was first introduced in the EEG decoding task in a previous work [17], with detailed information on the algorithm and implementation, and the general idea will be given in this manuscript. In the Gumbel selection, multiple neurons were employed, in which each neuron select one channel. Therefore, the channel number to be selected must be provided as *a priori* (hyperparameter). For example, when M channels need to be selected, M neurons will be used. For each neuron, the channel selection task can be viewed as a categorical sampling problem and this sampling process can be approximated with Gumbel-argmax trick. Next, to make the process differentiable, the argmax operation is replaced with softmax, hence the Gumbel-softmax trick. During the approximation, an extra temperature parameter β controls the extent to which Gumbel-Softmax approximates categorical distribution. As β approaches 0, samples from the Gumbel-Softmax distribution become one-hot and the Gumbel-Softmax distribution becomes identical to the categorical distribution. Therefore, channels corresponding to the '1's are the selected channels.

To select M channels from a total of N inputs, a probability weight can be written as 2D matrix: $\boldsymbol{W} = [\boldsymbol{w_1}, \boldsymbol{w_2} \ldots, \boldsymbol{w_M}]$, in which for each neuron m the vector $\boldsymbol{w_m} = w_1^m, w_2^m, .., w_N^m$ represents the probabilities to select N channels $i = 1, 2, \ldots N$.

The channel selection task can be viewed as a categorical sampling problem, in which channels are the categories (being selected or not). The Gumbel-Max trick provides a simple way to draw a sample $\widehat{w_n^m}$ (selection neuron m select the n-th channel) from a categorical distribution via:

$$\widehat{w_n^m} = one\_hot(\underset{n}{argmax}\left(g_n^m + log\left(\alpha_n^m\right)\right)) \tag{2}$$

where g_1^m, g_2^m, $\ldots g_N^m$ are independent and identically distributed (i.i.d) samples drawn from Gumbel(0, 1) [33], and $\boldsymbol{\alpha}$ is the learnable parameter. Next, to make the process differentiable, the argmax operation is replaced with softmax, and $\widehat{\boldsymbol{w}}$

can be rewritten as:

$$w_n^m = \frac{exp((log(\alpha_n^m) + g_n^m)/\beta)}{\sum_{n=1}^{N} exp((log(\alpha_n^m) + g_n^m)/\beta)} \tag{3}$$

The temperature parameter β controls the extent to which Gumbel-Softmax approximates categorical distribution. As β approaches 0, samples from the Gumbel-Softmax distribution become one-hot and the Gumbel-Softmax distribution becomes identical to the categorical distribution.

Then, given the original input $\boldsymbol{X}$ in the shape of (N × T), each neuron computes its output channel as $\boldsymbol{z_m} = \boldsymbol{w^m X}$, and the final out of the selection layer would be $\boldsymbol{Z} = [\boldsymbol{z_1}, \boldsymbol{z_2}, \ldots \boldsymbol{z_M}]$ in the shape of (M × T).

As β approaches 0, the Gumbel-Softmax distribution becomes identical to the categorical distribution, and the probability of each channel, denoted as matrix $\boldsymbol{P}$, can be obtained by normalizing the weights in the below equation:

$$p_n^m = \frac{\alpha_n^m}{\sum_{n=1}^{N} \alpha_n^m} \tag{4}$$

in which p_n^m represents the probability of channel n selected by neuron m.

In addition, the channel entropy is calculated to monitor the convergence of the selection process, which can be computed as below equation:

$$H_n = -\frac{1}{logN} \sum_{j=1}^{N} \alpha_n^m log(\alpha_n^m) \tag{5}$$

2) STG Selection Method: The STG method was first proposed by Yamada et al. [29], which is based on the probabilistic relaxation of discrete Bernoulli variables. While detailed information can be obtained in the original paper, the adaptation in channel selection using SEEG signals will be given below.

For raw SEEG data with N channels, the Bernoulli gates are applied to each of the N channels to activate or inactivate the input feature. These Bernoulli gates are represented by a random vector $\boldsymbol{s} = [s_1, s_2 \ldots s_d \ldots s_N]$ whose entries are independent and satisfy $\mathbb{P}(\boldsymbol{s}_i = 1) = \pi_i$ for $i \in [1, N]$. For a dataset compromising data $\boldsymbol{X}$ and corresponding label $\boldsymbol{y}$, the channel selection can be implemented as $\boldsymbol{X} \odot \boldsymbol{s}$, where the $\odot$ is the point-wise product. Then, the corresponding risk (loss) can be written as:

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{s}) = \mathbb{E}_{X,y}\mathbb{E}_{\boldsymbol{s}}[L(f_\theta(\boldsymbol{X} \odot \boldsymbol{s}), y) + \lambda\|\boldsymbol{s}\|_0] \tag{6}$$

in which the $\boldsymbol{\theta}$ and $\boldsymbol{s}$ represent the model parameter and the gating variables, and f_θ denotes the whole neural network that predicts label $\boldsymbol{y}$. λ is the hyperparameter that controls the portion of the channel to be selected. With Eq.6, the task is to search for parameter $\boldsymbol{\theta}$ and $\boldsymbol{s}$ to minimize loss $\mathcal{L}$ such that $\|\boldsymbol{s}\|_0$ is small compare to N. However, the optimization of the loss function containing discrete Bernoulli variables is unstable. To address this problem, the Bernoulli variables $\boldsymbol{s}$ were replaced by a Gaussian-based relaxation. The relaxation was referred to as the stochastic gate (STG), and defined as $s_i = max(0, min(1, \mu_i + \epsilon_i))$ for $i \in [1, N]$, where ϵ_d was drawn from $\mathcal{N}(0, \sigma^2)$ and σ was fixed during the training. This approximation can be viewed as a clipped, mean-shifted, Gaussian random variable.

At this point, the whole net can be optimized through gradient backpropagation. After the model is fully trained, the stochastic gating variable will be set as $s_i = max(0, min(1, \mu_i))$ to remove the stochasticity.

This procedure introduces a hyperparameter, λ, that controls approximately how many channels will be selected: more channels will be selected with lower λ, and vice versa. Therefore, the best λ will be searched in this paper in section III-D.1.

In summary, for both the selection subnet and classification subnet, the subject-dependent training was performed, i.e. each subject will train their own selection subnet and classification subnet. During the training, the proposed two channel-selection methods learned a weight vector compromising of 0s and 1s, in which the 1s correspond to the selected channels while the 0s correspond to the useless channels. The number of 1s in the weight vector could be different, and therefore, a different number of channels could be identified for different subjects. Next, the identified channel corresponding to the 1s will be extracted through the multiply operation and sent into the subsequent classification network to predict the class label, while the channels corresponding to the 0s will be discarded.

3) Mutual Information Channel Selection Method: The MI channel selection method was first proposed in [34]. While the full explanation of the method is provided in the original work, we provide a detailed summary in this manuscript. In this method, starting from an empty selected channel set, the MI between each channel and the class labels is calculated and the channel with the highest MI is added to the selected channels. Then, MI will be computed again between class labels and the combination of the selected channel set and each remaining channel, while the channel with the highest MI will be added to the selected channels. This process is repeated until the desired number of channels is selected. To extract features during the computation of the MI, we extract spectral features in 0.5-4 Hz, 4-8 Hz, 8-13 Hz, 13-30 Hz, 60-75 Hz, 75-95 Hz, 105-125 Hz, and 125-150 Hz, as in our previous study [35]. Note that these extracted features were only used to select channels, while movement classification was performed using raw SEEG signals of the selection channels.

4) Manual Channel Selection Method: Aside from automatic channel selection, manual selection was also performed to make a comparison. In detail, for manual selection, the temporal-spectral representation was calculated first. Then, channels showing high ERS or ERD by visual inspection during the task stage were chosen. The calculation of ERS/ERD and the procedure to sort the channels according to the electrode reactivity was described in our previous work [35].

C. The Classification Subnet

The classification subnet in the stacked architecture will classify the selected SEEG signals into five tasks. In this paper, the deep ConvNet model proposed in an EEG study will be used as the classification subnet [36]. However, it should be noted that this subnet does not necessarily have to be a classification network, and it could be any deep learning network, including classification and regression.

D. Training Process

The training process was performed separately for different subjects. This is because the SEEG data varies among subjects. These variations come from two aspects. First is the implantation location. The subjects are epileptic patients who have SEEG implanted in the possible seizure onset zone (SOZ) during the seizure monitoring. Since the SOZs are not the same, the recorded signals reflect very different biological and physical processes. Second, the number of implanted electrodes was also different.

To train the network for each subject, we used a strategy similar to the five-fold cross-validation process. In detail, for each fold, 20% of the entire data was used as a testing dataset, while the remaining 80% of the entire data was used as training and validation. For this 80% remaining data, 80% was the training set while 20% was the validation set. The validation dataset was used for hyperparameter tuning and early stopping. Therefore, the final training, evaluation, and testing dataset contained 64% (80% x 80%), 16% (80% x 20%), and 20% of the entire dataset for each fold. The split was performed within each task. During the evaluation, the mean decoding accuracies averaged across all participants and all folds were reported. For both proposed methods, the model parameters were initialized randomly and trained through gradient descent using the Adam optimizer [37]. Since there are only 100 trials for each participant, to obtain sufficient data to train the deep learning network, a sliding-window strategy was used to augment the SEEG data [36]. The window and sliding step were set to 500 ms and 100 ms respectively, which proved to be able to achieve satisfactory results in our previous classification experiment (without channel selection). The 3D windowed data, in the shape of $B\ \times N \times\ T$ (B, N, T denoted batch size, channel number and temporal sampling points, and $T = 500$), will be used as input by all methods during the training. In the end, the mean accuracy obtained by averaging these five folds was reported in the manuscript.

For both methods, the selection and classification subnets will be trained simultaneously. The training can be stopped when two conditions are met: the channel weights (probabilities) approximate one-hot (low entropy) and classification accuracy does not increase for certain epochs (early stopping). The first and second conditions indicated the completion of training for the selection subnet (phase one) and classification subnet (phase two), respectively. Particularly, further training experiments demonstrated that the channel weights always achieve one-hot in the first few training epochs and stay stable during the subsequent training of the classification subnet for both methods. Therefore, the whole network was trained simultaneously and stopped when the classification accuracy plateaued for certain epochs (early stopping).

In addition, for the Gumbel method, the threshold τ and the temperature β, should be adjusted to facilitate the training process. In general, both β and τ should be set with large values at the beginning of the training, so that the selection subnet can exploit all possible channel combinations with no penalty imposed. As the training continues, β should be decreased to approximate one-hot discrete sampling. As each column of $\boldsymbol{P}$ approaches one-hot, the corresponding entropy will decrease. At the same time, τ should be decreased during training to punish the duplicated selection made by different selection neurons. In this paper, the temperature β and threshold τ were scheduled to exponentially decay from 10 and 2 to 0.1 and 0.1, respectively, similar to the schedule used in [17].

1) Hyperparameter Search: There is a hyperparameter in each method to control how many channels will be selected. For Gumbel selection, the channel number to be selected (M) was set as *a prior*. The effect of selected channel numbers on the decoding accuracy was studied using M from [2, 4, 6, 8, 10, 12]. The value corresponding to the highest decoding accuracy was used for subsequent analysis.

For STG selection, the hyperparameter λ determines approximately how many channels will be selected. A lower value will loosen the restriction of selected channels and result in more channels being selected, which means even channels that contribute little to the decoding will be selected. To find the best value, λ was searched from [0.01, 0.1, 0.2, 0.4, 0.6, 0.8] by evaluating the average decoding accuracy of all participants. The value corresponding to the highest decoding accuracy will be used.

E. Comparison Between Selection Methods

After the best hyperparameters were obtained, the decoding accuracy of selection methods will be compared. However, the STG method will select any channel that is useful for the classification task and the selected number was not the same for different participants. This is different from the Gumbel selection, in which the selected channel count was the same for all participants. To make a fair comparison between these two methods, the same channel count as used in the Gumbel method was picked from the STG selection result, which corresponds to the highest weight in $\boldsymbol{\mu}$. For example, if 10 channels were used for Gumbel selection, 10 channels will be picked corresponding to the 10 highest values in $\boldsymbol{\mu}$.

In total, six channel selection strategies will be used and compared in this section:

1) The STG selection without channel number restriction (referred to as STG)
2) 10 channels selected by the Gumbel selection
3) 10 top channels (highest weight) selected by STG selection (referred to as STG-10)
4) 10 Mutual Information (MI) selected channels
5) 10 channels selected manually
6) all available channels (no selection was performed)

All statistical analyses were performed with the Wilcoxon rank-sum test from *SciPy* [38].

IV. Results

A. Decoding Result

1) Hyperparameter Selection: The highest decoding accuracy obtained with STG selection corresponds to λ of 0.2. The average decoding accuracy declined when using more or fewer channels by decreasing or increasing λ, respectively. Therefore, the λ was set as 0.2 in the subsequent content.

The decoding accuracy using Gumbel selection with various channel counts is presented in Fig. 4. From the plot, the decoding accuracy plateaued around 10 channels. Therefore, 10 channels will be used in the subsequent content.

2) Decoding Result: The decoding result of 12 participants, with and without the selection network, is presented in Fig. 5. The decoding accuracies (mean ± standard deviation) were 67 ± 4.2%, 65 ± 3.5%, 60 ± 3.1%, 60 ± 4.5%, 59 ± 3.6%, and 47 ± 1.9% when using STG, Gumbel, STG-10, manual selection, all channels, and MI method, respectively.

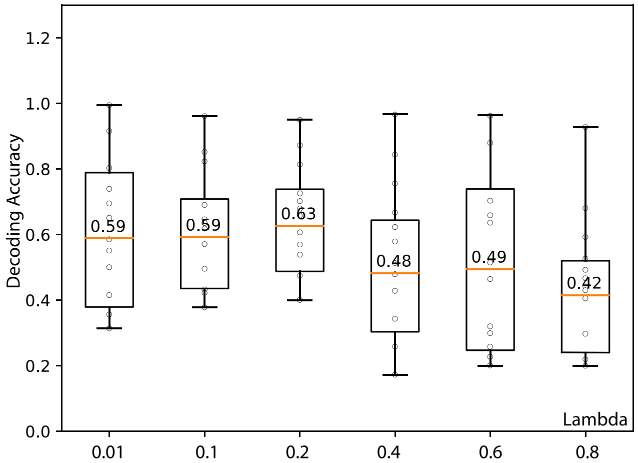

Fig. 3. Decoding results of 12 participants using STG method with various λ values. The box plot showed the highest decoding accuracy was obtained with λ of 0.2. The upper and lower end of the box represent the 25th and 75th percentile, while the middle orange line denotes the median value.

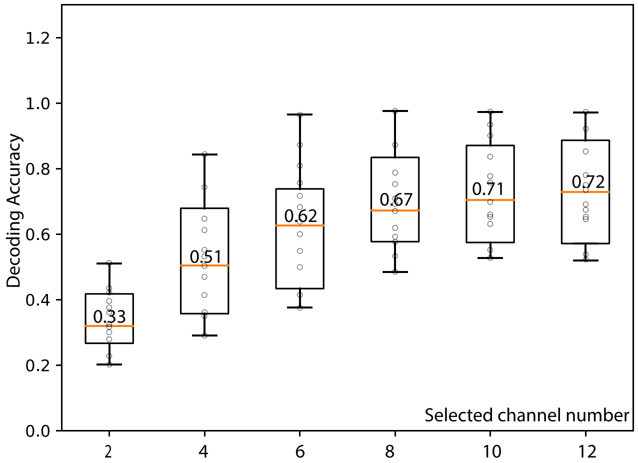

Fig. 4. Decoding results with various channel counts using the Gumbel selection method. The middle orange line denotes the median value.

There is no significant difference between decoding accuracies using Gumbel and STG, though they both significantly outperformed STG with 10 channels, manual selection, all channels, and MI selection ($p < .001$, Wilcoxon rank-sum test). The significantly better performances of the two deep learning-based methods, compared to the MI method, indicate they have learned to identify the informative channels. In addition, it should be noted that there are two outliers, in the dashed rectangle, using the manual and MI methods. This demonstrated that the proposed automated selection methods can alleviate possible human error and are more robust than the manual method.

The highest mean decoding accuracy came from STG selection (67 ± 4.2%). Using STG selection, a total of 282 channels were selected from 12 participants, while there are 120 channels selected by Gumbel selection. However, when only using the top 10 channels selected by STG selection (STG-10), Gumbel outperformed STG-10 (65 ± 3.5% and 60 ± 3.1%, respectively, $p < .005$). Therefore, Gumbel selection outperformed STG in obtaining comparable accuracy using a few channels. In the next section, the relationship between the channels selected by the two methods will be explored.

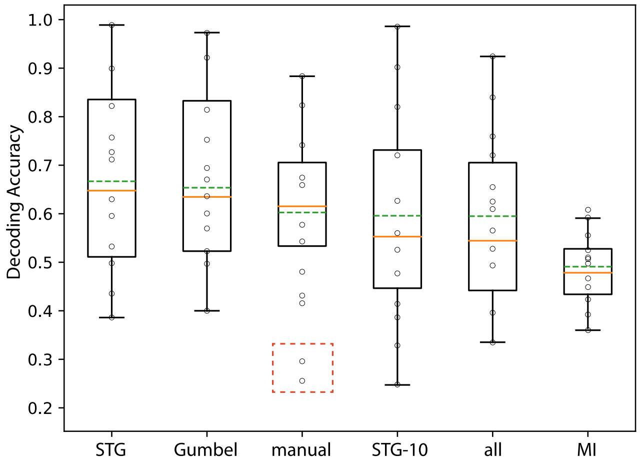

Fig. 5. Decoding accuracies of 12 participants using various selection methods. Six experiments were reported, including STG (decoding using all channels selected by STG selection), Gumbel (decoding using 10 channels selected by Gumbel-Softmax selection), STG-10 (decoding using top 10 channels selected by the STG methods), manual (decoding using 10 manually selected channels), all (decoding using all channels) and MI (decoding using 10 channels selected using MI method). Orange and green dashed lines represent the median and mean, respectively. The dashed rectangle indicated the two outliers using the manual selection method.

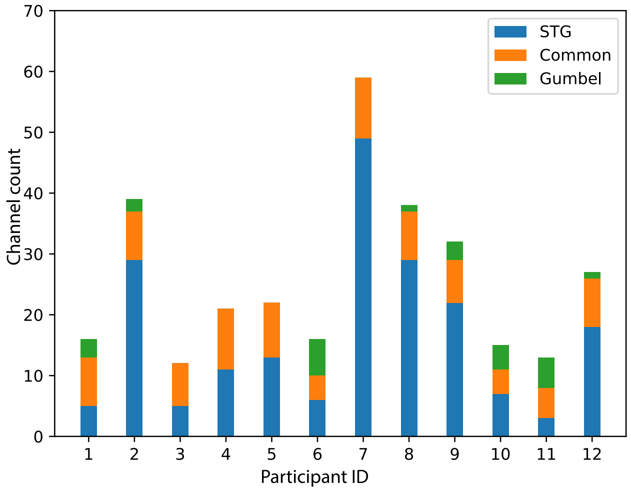

Fig. 6. Channels count selected by Gumbel and STG selection. Channels selected by different methods were indicated by different colours, while the middle orange bar represented channels selected by both methods.

B. Selected Channels

The relationship between the channels selected by Gumbel and STG is compared and presented in Fig. 6.

The plot indicated that most channels selected by Gumbel selection have also been selected by STG. The considerable overlap between channels selected by the two methods indicates both can be used to select the informative channels. Because both selected very similar channels, the following analysis of the selected channels will focus on the Gumbel method. It is worth noting that more channels were selected

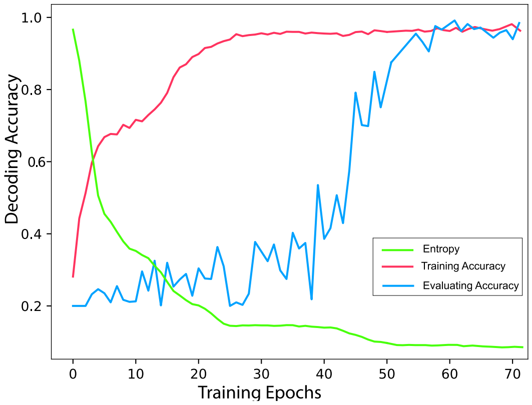

Fig. 7. The training process from participant 10. The entropy declined quickly in the first few epochs and stayed stable during the rest training.

by the STG method, which implies that the higher decoding accuracy obtained by STG selection might result from using more channels. On the other hand, the improvement obtained by STG over Gumbel is not significant (Wilcoxon rank-sum test, p = 0.8223). Next, the training process of the Gumbel method will be evaluated.

C. Training Process of Gumbel Selection

The proposed methods facilitate the selection process by approximating the categorical variables using certain relaxation. This section will examine the training process to verify the channel weights' stable convergence. An example training process of the stacked network from participant 10 is presented in Fig. 7. The entropy declined quickly within the first few epochs. Importantly, the monotonously declined entropy stayed stable during the rest training epochs. This plot indicated that the integration of the selection subnet didn't interrupt the training of the classification subnet and these two subnets can be trained simultaneously.

In addition, the training process of the selection net for participant 10 is presented in Fig. 8. It showed that the probability began to exhibit one-hot distribution from around the 10^{th} epoch. This is in line with Fig. 7, which demonstrated that the entropy decreased rapidly at the beginning of the training. The weight was then adjusted, which means the net was exploring other channel combinations. The monotonously decreased entropy from Fig. 7 demonstrated that the selection subnet was stable during the training. Other participants exhibited similar training processes.

D. Channel Selected By Gumbel Selection

In this section, the selected channels will be analyzed. Since the channels selected by the two proposed methods were considerable overlap and the Gumbel method can achieve comparable accuracy using few channels, channels selected by the Gumbel method will be used.

1) Anatomical Location of The Selected Channels: To evaluate the channel selection network, the selected channel and unselected channels, aggregated from all participants, were projected into an MNI standard brain model, denoted as green and black dots, respectively. The pre and post-central cortex were also indicated as shaded areas using the Desikan-Killiany Atlas, as presented in Fig. 9.

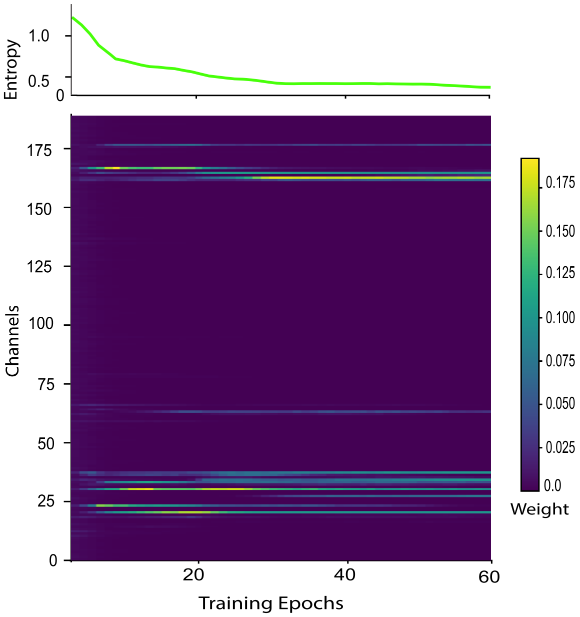

Fig. 8. Changes in channel weights as the training progressed. At the beginning of the training, all channels have about the same weights and weight entropy is high (around 1.0). As the training progressed, the selection network adjusted channel weights and the entropy decreased. When the training is finished, the selection network has learned an approximate one-hot weight and the entropy approached 0.1.

It demonstrated that the selected channels were mostly localized in the pre and post-central areas. This spatial distribution of the selected channels is in line with the conventional understanding of the neural representation of movement, which involves the primary motor and sensory cortex primarily [39], [40]. In addition, it is worth noting that channels in other areas, such as the posterior cingulate, and superior parietal, are also selected.

To better understand the decoding contribution of certain regions, channels were grouped according to the anatomic region they are located. Then, the number of the selected channels was plotted on top of the unselected channel count for each region, as presented in Fig. 10. From the plot, it is obvious that pre- and post-central cortex are most likely to be selected, and the ratios $((selected\_count) \div (unselected\_count))$ from these two regions are significantly higher than other regions. (To determine the significance of higher rations selected from these two regions, the interquartile range method was used. In detail, ratios that are outside of range $[Q1 - 1.5 \times IQR, Q3 + 1.5 \times IQR]$ are considered outliers (significantly higher), where *Q1*, *Q3* and *IQR* represented the first quartile, third quartile, and interquartile range.)

However, not every channel from the pre and post-central cortex was selected. This might be because the number of channels to be selected was limited to 10 for each participant. On the other hand, other regions that are not considered to be directly involved in motor control were also selected with a lower ratio. This might be caused by two possible reasons: 1) Other regions might also contain useful information, but contribute less to the decoding; 2) the random initialization and the stochastic nature of the deep learning training process imply that the channels, selected in every selection process, could be different [41].

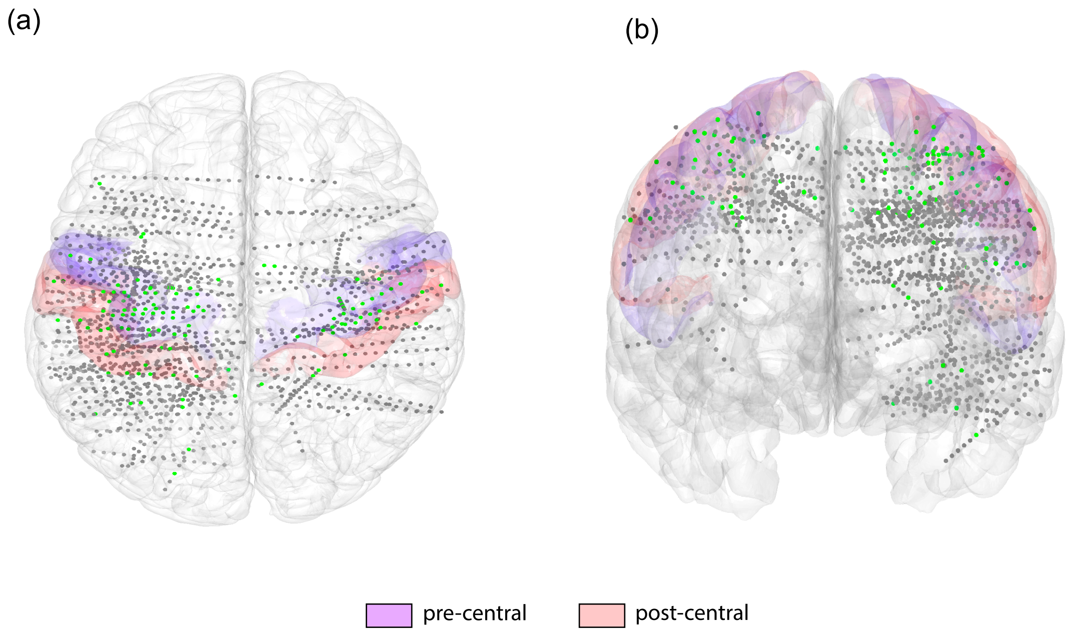

Fig. 9. The spatial location of selected (green dots) and unselected channels (grey dots) in the MNI standard brain model. This plot was generated using electrodes pooled from all participants. The pre and post-central cortex were also highlighted as colourful areas. Subplot *a* and *b* were plotted from the top and front view, respectively.

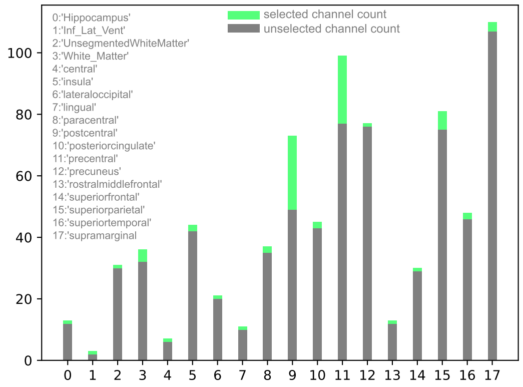

Fig. 10. The number of selected and unselected channels is grouped by regions. A total of 18 regions were included, and the anatomic names were presented beside the bar plot. The selection ratios of regions 9 and 11 are significantly higher than other regions.

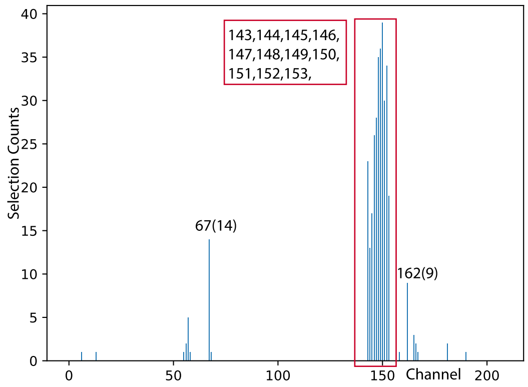

Fig. 11. The number of times each channel was selected in 40 experiments for participant 10. The most likely selected 13 channels were framed in a red rectangle.

On the other hand, channels selected by the manual method were more concentrated in the motor areas. This was actually in accordance with previous studies demonstrating that signals in motor areas exhibited strong ERS/ERD [42].

To evaluate the training stability and eliminate the possibility that high decoding accuracy was obtained by random network initialization, the training process (selection and classification) was repeated 40 times on participant 10. Then, the number of times each channel was selected in 40 experiments was plotted, as in Fig. 11. Other participants exhibited similar results.

It is obvious that these 11 channels framed in red rectangles were most frequently selected. However, other channels were also selected with lower occurrences. This implied that different channels might be selected in different training experiments. To better understand the selected channels, the anatomical labels of these 13 channels were obtained using the Desikan-Killiany Atlas and presented in Table II. It showed that selected channels were located in the cortex and white matter of precentral and postcentral regions. This result was in line with previous studies demonstrating that both precentral and postcentral were involved in the movement control [43], [44]. The identification of the white matter of

TABLE II
The Anatomic Labels of the Most Frequently Selected Channels

67	ctx-lh-postcentral	149	ctx-lh-precentral
143	wm-lh-posteriorcingulate	150	ctx-lh-postcentral
144	wm-lh-posteriorcingulate	151	ctx-lh-postcentral
145	Left-UnsegmentedWhiteMatter	152	wm-lh-postcentral
146	wm-lh-precentral	153	ctx-lh-postcentral
147	wm-lh-precentral	162	wm-lh-postcentral
148	wm-lh-precentral		

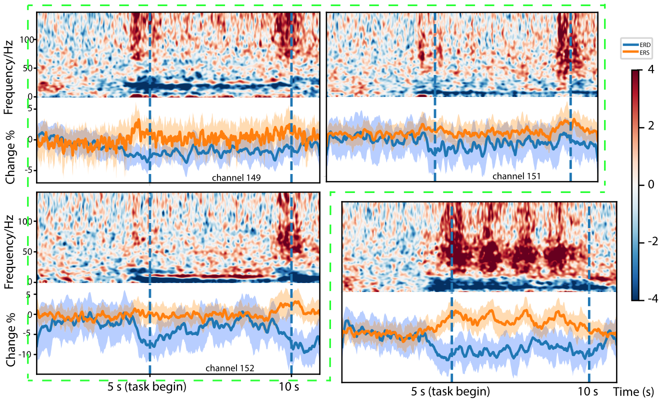

Fig. 12. The temporal-spectral representation of the deep-learning-selected (framed by the green dashed line) and the manual-selected channel (channel 167) from an example participant 10. Other participants have similar results.

precentral and postcentral confirmed a recent SEEG study showing that white matter was activated and contained useful information for movement control [45]. A channel located in unsegmented white matter (145) was also frequently selected, which might result from the volume conduction [46], [47]. In addition, two channels from the posterior cingulate (143, 144) were also frequently selected, as shown in Fig.11 during the repetition experiment. It is in line with the understanding that PCC was proved to show a complex and dynamic pattern that partially reflects activity in other brain networks during the reaction to a changing environment [48].

In summary, three conclusions can be drawn from the selection result: 1) randomness existed in the selection result; 2) the informative channels were selected at a much higher frequency; 3) the locations of the frequently selected channels confirmed previous understanding of motor control.

2) Gumbel Selection V.S. Manual Selection: In this manuscript, we showed that there is no significant difference between decoding accuracies obtained by the Gumbel and the manual selection methods. However, we also demonstrated that possible human error could dramatically impact the manual selection performance, as indicated by the two outliers in Fig. 5, while the Gumbel selection method exhibited very stable decoding accuracy in the 40-time-repeating experiment. In this sense, the deep learning-based method is superior to the manual selection method.

Next, to gain more understanding of the channels selected by the deep learning method, the channels selected by these two methods were compared. For the manual selection method, similar to a previous SEEG decoding study [18], channels were selected according to their response to the task in the spectral domain. More specifically, channels that exhibited magnitude or power responsive in certain frequency bands were selected.

Therefore, the temporal-spectral representation of the representative three channels of participant 10 (149, 151, 152), which are most often selected by both Gumbel and STG selection, were calculated, indicated as the subplots framed within the green dashed line in Fig. 12. The same representation of a typical manually selected channel (not selected by either method) from the same participant is presented in the lower right subplot.

In this plot, channel 167 is a typical channel showing strong spectral response (ERS/ERD), and is very likely to be chosen during the manual selection. The other three channels (149, 151, 152) which were consistently selected in many repetitions, as indicated in Fig. 11, exhibited, unexpectedly, much lower ERS/ERD response compared to the manually-selected channel. In summary, while manual selection would choose channels with a strong spectral response, the channels frequently selected by the proposed deep learning-based methods exhibited much weaker ERS/ERD. Similar selection results and spectral responses were found from all other participants.

V. Discussion

In this paper, two deep learning-based channel selection methods (Gumbel and STG) were presented and compared with other methods, including manual selection and MI-based selection. The experiment result indicated that some channels were consistently selected by both methods, and the decoding accuracy using these channels was comparable to that of using manually selected channels or using all channels. In this section, the implications of the selected channels will be discussed. Then, the differences between the proposed methods will be compared.

A. Anatomical Location of Selected Channels

In this work, the most likely selected channels were found to be located in the pre-central and post-central regions. The identified locations confirmed previous electrocorticography (ECoG) studies, which showed that both the primary motor and sensory cortex contain rich information related to movement [39], [40]. In addition, the proposed methods also frequently identified regions, such as PCC and white matter, as shown in Fig.11 and Tab.II. The identification of white matter and its lower selection frequency than that of the pre and post-central areas is in line with previous SEEG studies showing that white matter is also activated and contains useful information but contributes less than grey matter to the decoding task [45]. The identification of the PCC, which was proved to be activated during memory retrieval [49], might imply rehearsal was performed by the participant before actual movement when the cue was presented. Together with the comparable decoding accuracy using the selected channels, this paper proved that these two deep learning methods were capable of identifying task-related regions and channels. This ability might help to identify a distributed, yet closely connected neural network. For example, to facilitate motor control, two brain networks exist and both converge on the primary motor area [50], [51]. While the distributed neural response has yet to be fully understood, channel selection has the potential to identify distributed regions engaged in a network.

B. Gumbel Selection Versus STG Selection

From the previous analysis, it is clear that both methods have advantages and disadvantages. For Gumbel selection, it achieved higher decoding accuracy than STG when using the same channel count. In Gumbel selection, the selected channel count, which needs to be set *a priori*, is not obvious without iterative testing, which is time-consuming.

On the other hand, there is no fixed channel count that the STG method needs to select. By selecting any possible informative channels, it can achieve the highest decoding accuracy, but when using only a subset of the selected channels, its accuracy is inferior to the Gumbel selection. Therefore, STG selection would be more appropriate when there is no requirement for the channel number.

C. Characteristics of the Informative Channels

At this point, the characteristics of the informative channels remain elusive. The average decoding accuracies were 65% and 60% when using Gumbel and manual selection methods, respectively. The relatively higher decoding accuracy implies the Gumbel method has identified a better channel subset. The next question is, can the channels selected by the Gumbel method help to facilitate the manual selection? To answer that question, there is a need to understand what the Gumbel selection method looks at during the training. While the criteria of channels selected manually are the spectral response (ERS/ERD), the Gumbel selection looks at something more than that. As demonstrated in Fig. 12, while there are many channels that exhibited strong ERS/ERD, the channels selected by Gumbel selection showed a much weaker spectral response. To answer that question requires further study on the learning process of the so-called ‘black-box’ deep learning model.

D. Limitations and Prospects of The Study

This paper presented two channel selection networks based on deep learning methods. However, there are limitations to both methods. For Gumbel selection, the selected channel count needs to be set before the selection. For STG selection, there is also a hyperparameter to control the selected channel count. Both hyperparameters can only be obtained by multiple training experiments. These multiple experiments would violate the essential idea behind the embedded selection method.

However, repetitive training is, sometimes, inevitable in BCIs applications. In BCIs, the task-related signals are usually distributed into a wide area, which means not only directly related but also indirectly involved channels (electrodes) are all useful. Also, volume conduction means the useful information is not spatially confined but rather distributed. Therefore, the informative channels are not definitive, only the degree of usefulness is different. From this point of view, it is natural and inevitable to conduct multiple experiments to achieve a balance between decoding accuracy and channel count.

Another limitation is the inherent randomness of the deep learning model. This non-determinism could come from many aspects, such as random initialization, dropout, and noisy injection [41]. This randomness leads to inconsistent results in multiple training processes, as demonstrated in Fig.11 that showed that whilst most informative channels can be repetitively selected, there is still a certain chance to select other channels.

Except for the limitation, the algorithms proposed in this paper can be used in broader applications, not confined to SEEG-based BCIs or classification tasks. They can be integrated into any deep learning method for both invasive and non-invasive BCIs. For example, in ECoG-based BCIs, the standard recording devices have 64 (8×8 grid), 48 (6×8 grid), or 16 (2×8 strip) contacts [52], [53], whilst high-density ECoG with more than a thousand channels has also been proposed [6], [54]. The high channel count complicates the data transmission and hardware design, where the power consumption and heat need to be carefully handled [7]. Further, there might be redundancy between the neighboring electrodes of high-density ECoG (μ ECoG) [55]. In this context of high channel count and high-density devices, channel selection has paramount significance in achieving the balance between decoding accuracy and data volume.

VI. Conclusion

In this paper, two deep learning-based channel selection methods (Gumbel and STG) were introduced and compared with manual and Mutual Information-based channel selection. By using a plug-and-play approach, these selection networks can be stacked on top of any deep learning network (classification or regression) and trained simultaneously. The highest decoding accuracy arose from STG selection using all identified channels, while Gumbel selection was superior when only 10 channels were used. On the other hand, manual selection could lead to suboptimal classification accuracy because of human error. By comparing the channels selected by the two methods, it is proved that most channels selected by these two methods were overlapped, and the identified anatomic regions confirmed the previous understanding of movement control. The proposed methods can be used in future high-throughput BCIs to achieve a balance between invasiveness and decoding performance.

Acknowledgment

The authors would like to thank Prof. Liang Chen and his team at the Shanghai Huashan Hospital for their assistance in collecting the SEEG data.

References

[1] M. Seeck et al., “The standardized EEG electrode array of the IFCN,” *Clin. Neurophysiol.*, vol. 128, pp. 2070–2077, Oct. 2017.

[2] G. Toscano et al., “Visual analysis of high density EEG: As good as electrical source imaging?” *Clin. Neurophysiol. Pract.*, vol. 5, pp. 16–22, 2020.

[3] P. Fiedler, C. Fonseca, E. Supriyanto, F. Zanow, and J. Haueisen, “A high-density 256-channel cap for dry electroencephalography,” *Hum. Brain Mapping*, vol. 43, no. 4, pp. 1295–1308, Mar. 2022.

[4] T. Jiang et al., “Local spatial correlation analysis of hand flexion/extension using intraoperative high-density ECoG,” in *Proc. 37th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Aug. 2015, pp. 6190–6193.

[5] P. T. Wang et al., “Comparison of decoding resolution of standard and high-density electrocorticogram electrodes,” *J. Neural Eng.*, vol. 13, no. 2, Apr. 2016, Art. no. 026016.

[6] E. Musk, “An integrated brain–machine interface platform with thousands of channels,” *J. Med. Internet Res.*, vol. 21, no. 10, Oct. 2019, Art. no. e16194.

[7] G. Charvet et al., “A wireless 64-channel ECoG recording electronic for implantable monitoring and BCI applications: WIMAGINE,” in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Aug. 2012, pp. 783–786.

[8] C. S. Mestais, G. Charvet, F. Sauter-Starace, M. Foerster, D. Ratel, and A. L. Benabid, “WIMAGINE: Wireless 64-channel ECoG recording implant for long term clinical applications,” *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 23, no. 1, pp. 10–21, Jan. 2015.

[9] A. M. Lozano et al., “Deep brain stimulation: Current challenges and future directions,” *Nat. Rev. Neurol.*, vol. 15, no. 3, pp. 148–160, Mar. 2019.

[10] X. Wu et al., “Decoding continuous kinetic information of grasp from stereo-electroencephalographic (SEEG) recordings,” *J. Neural Eng.*, vol. 19, no. 2, Apr. 2022, Art. no. 026047.

[11] G. Li et al., "Assessing differential representation of hand movements in multiple domains using stereo-electroencephalographic recordings," *NeuroImage*, vol. 250, Apr. 2022, Art. no. 118969.

[12] T. Alotaiby, F. E. A. El-Samie, S. A. Alshebeili, and I. Ahmad, "A review of channel selection algorithms for EEG signal processing," *EURASIP J. Adv. Signal Process.*, vol. 2015, no. 1, p. 66, Dec. 2015.

[13] L. He, Z. Yu, Z. Gu, and Y. Li, "Bhattacharyya bound based channel selection for classification of motor imageries in EEG signals," in *Proc. Chin. Control Decis. Conf.*, Jun. 2009, pp. 2353–2356.

[14] Q. Wei and Y. Wang, "Binary multi-objective particle swarm optimization for channel selection in motor imagery based brain–computer interfaces," in *Proc. 4th Int. Conf. Biomed. Eng. Informat. (BMEI)*, vol. 2, 2011, pp. 667–670.

[15] T. N. Lal et al., "Support vector channel selection in BCI," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 6, pp. 1003–1010, Jun. 2004.

[16] H. Zhang, X. Zhao, Z. Wu, B. Sun, and T. Li, "Motor imagery recognition with automatic EEG channel selection and deep learning," *J. Neural Eng.*, vol. 18, Nov. 2020.

[17] T. Strypsteen and A. Bertrand, "End-to-end learnable EEG channel selection for deep neural networks with gumbel-softmax," *J. Neural Eng.*, vol. 18, no. 4, Aug. 2021, Art. no. 0460a9.

[18] G. Li, S. Jiang, Y. Xu, Z. Wu, L. Chen, and D. Zhang, "A preliminary study towards prosthetic hand control using human stereo-electroencephalography (SEEG) signals," in *Proc. 8th Int. IEEE/EMBS Conf. Neural Eng. (NER)*, May 2017, pp. 375–378.

[19] M. Wang et al., "Enhancing gesture decoding performance using signals from posterior parietal cortex: A stereo-electroencephalograhy (SEEG) study," *J. Neural Eng.*, vol. 17, no. 4, Aug. 2020, Art. no. 046043.

[20] Q. Wei and W. Tu, "Channel selection by genetic algorithms for classifying single-trial ECoG during motor imagery," in *Proc. 30th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Aug. 2008, pp. 624–627.

[21] F. Lotte and C. Guan, "Regularizing common spatial patterns to improve BCI designs: Unified theory and new algorithms," *IEEE Trans. Biomed. Eng.*, vol. 58, no. 2, pp. 355–362, Feb. 2011.

[22] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, "EEGNet: A compact convolutional neural network for EEG-based brain–computer interfaces," *J. Neural Eng.*, vol. 15, no. 5, Oct. 2018, Art. no. 056013.

[23] Y. Ding et al., "TSception: A deep learning framework for emotion detection using EEG," 2020, *arXiv:2004.02965*.

[24] A. Du, S. Yang, W. Liu, and H. Huang, "Decoding ECoG signal with deep learning model based on LSTM," in *Proc. IEEE Region 10 Conf.*, Oct. 2018, pp. 430–435.

[25] S. Mousavi, F. Afghah, and U. R. Acharya, "SleepEEGNet: Automated sleep stage scoring with sequence to sequence deep learning approach," *PLoS ONE*, vol. 14, no. 5, May 2019, Art. no. e0216456.

[26] A. Vaswani et al., "Attention is all you need," 2017, *arXiv:1706.03762*.

[27] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with gumbel-softmax," in *Proc. 5th Int. Conf. Learn. Represent.*, 2017.

[28] A. Abid, M. F. Balin, and J. Zou, "Concrete autoencoders for differentiable feature selection and reconstruction," in *Proc. 36th Int. Conf. Mach. Learn.*, 2019.

[29] Y. Yamada, O. Lindenbaum, S. Negahban, and Y. Kluger, "Feature selection using stochastic gates," in *Proc. 37th Int. Conf. Mach. Learn.*, vol. 119, 2018, pp. 10648–10659.

[30] G. Li et al., "IEEGview: An open-source multifunction GUI-based MATLAB toolbox for localization and visualization of human intracranial electrodes," *J. Neural Eng.*, vol. 17, no. 1, Dec. 2019, Art. no. 016016.

[31] R. S. Desikan et al., "An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest," *Neuroimage*, vol. 31, no. 3, pp. 968–980, 2006.

[32] A. Gramfort et al., "MEG and EEG data analysis with MNE-Python," *Frontiers Neurosci.*, vol. 7, p. 267, Mar. 2013.

[33] E. J. Gumbel, "Statistical theory of extreme values and some practical applications," *J. Roy. Aeronaut. Soc.*, vol. 58, no. 527, pp. 792–793, 1954.

[34] T. Lan, D. Erdogmus, A. Adami, M. Pavel, and S. Mathan, "Salient EEG channel selection in brain computer interfaces by mutual information maximization," in *Proc. IEEE Eng. Med. Biol. 27th Annu. Conf.*, 2005, pp. 7064–7067.

[35] X. Wu et al., "Deep learning with convolutional neural networks for motor brain–computer interfaces based on stereo-electroencephalography (SEEG)," *IEEE J. Biomed. Health Informat.*, vol. 27, no. 5, pp. 2387–2398, May 2023.

[36] R. T. Schirrmeister et al., "Deep learning with convolutional neural networks for EEG decoding and visualization," *Hum. Brain Mapping*, vol. 38, no. 11, pp. 5391–5420, Nov. 2017.

[37] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[38] P. Virtanen, "SciPy 1.0: Fundamental algorithms for scientific computing in Python," *Nature Methods*, vol. 17, pp. 261–272, Feb. 2020.

[39] J. N. Sanes and J. P. Donoghue, "Plasticity and primary motor cortex," *Annu. Rev. Neurosci.*, vol. 37, pp. 393–415, Jan. 2000.

[40] L. Miller and N. Hatsopoulos, "Neuronal activity in motor cortex and related areas," in *Brain–Computer Interfaces: Principles and Practice*, 2012, p. 15.

[41] D. Zhuang, X. Zhang, S. Leon Song, and S. Hooker, "Randomness in neural network training: Characterizing the impact of tooling," 2021, *arXiv:2106.11872*.

[42] B. Graimann, J. E. Huggins, S. P. Levine, and G. Pfurtscheller, "Visualization of significant ERD/ERS patterns in multichannel EEG and ECoG data," *Clin. Neurophysiol.*, vol. 113, no. 1, pp. 43–47, Jan. 2002.

[43] N. R. Anderson, T. Blakely, G. Schalk, E. C. Leuthardt, and D. W. Moran, "Electrocorticographic (ECoG) correlates of human arm movements," *Exp. Brain Res.*, vol. 223, no. 1, pp. 1–10, Nov. 2012.

[44] G. Schalk et al., "Decoding two-dimensional movement trajectories using electrocorticographic signals in humans," *J. Neural Eng.*, vol. 4, no. 3, pp. 264–275, Sep. 2007.

[45] G. Li et al., "Detection of human white matter activation and evaluation of its function in movement decoding using stereo-electroencephalography (SEEG)," *J. Neural Eng.*, vol. 18, no. 4, Aug. 2021, Art. no. 0460c6.

[46] G. Arnulfo, J. Hirvonen, L. Nobili, S. Palva, and J. M. Palva, "Phase and amplitude correlations in resting-state activity in human stereotactical EEG recordings," *NeuroImage*, vol. 112, pp. 114–127, May 2015.

[47] E. Landré, M. Chipaux, L. Maillard, W. Szurhaj, and A. Trébuchon, "Electrophysiological technical procedures," *Neurophysiologie Clinique*, vol. 48, no. 1, pp. 47–52, Feb. 2018.

[48] R. Leech, R. Braga, and D. J. Sharp, "Echoes of the brain within the posterior cingulate cortex," *J. Neurosci.*, vol. 32, no. 1, pp. 215–222, Jan. 2012.

[49] J. Parvizi and S. Kastner, "Promises and limitations of human intracranial electroencephalography," *Nature Neurosci.*, vol. 21, no. 4, pp. 474–483, Apr. 2018.

[50] P. Haggard, "Human volition: Towards a neuroscience of will," *Nature Rev. Neurosci.*, vol. 9, no. 12, pp. 934–946, Dec. 2008.

[51] N. Picard and P. L. Strick, "Motor areas of the medial wall: A review of their location and functional activation," *Cerebral Cortex*, vol. 6, no. 3, pp. 342–353, 1996.

[52] E. C. Leuthardt, G. Schalk, J. R. Wolpaw, J. G. Ojemann, and D. W. Moran, "A brain–computer interface using electrocorticographic signals in humans," *J. Neural Eng.*, vol. 1, no. 2, pp. 63–71, 2004.

[53] D. Moran, "Evolution of brain–computer interface: Action potentials, local field potentials and electrocorticograms," *Current Opinion Neurobiol.*, vol. 20, no. 6, pp. 741–745, Dec. 2010.

[54] T. Kaiju, M. Inoue, M. Hirata, and T. Suzuki, "High-density mapping of primate digit representations with a 1152-channel μECoG array," *J. Neural Eng.*, vol. 18, no. 3, Jun. 2021, Art. no. 036025.

[55] S. Kellis et al., "Multi-scale analysis of neural activity in humans: Implications for micro-scale electrocorticography," *Clin. Neurophysiol.*, vol. 127, no. 1, pp. 591–601, Jan. 2016.