Deep Learning with Convolutional Neural Networks for Motor Brain-Computer Interfaces based on Stereo-electroencephalography (SEEG)

Xiaolong Wu, Shize Jiang, Guangye Li, Shengjie Liu, Benjamin Metcalfe, Liang Chen*, Dingguo Zhang*

Abstract*—*Objective: **Deep learning based on convolutional neural networks (CNN) has achieved success in brain-computer interfaces (BCIs) using scalp electroencephalography (EEG). However, the interpretation of the so-called 'black box' method and its application in stereo-electroencephalography (SEEG)-based BCIs remain largely unknown. Therefore, in this paper, an evaluation is performed on the decoding performance of deep learning methods on SEEG signals.** ***Methods:*** **Thirty epilepsy patients were recruited, and a paradigm including five hand and forearm motion types was designed. Six methods, including filter bank common spatial pattern (FBCSP) and five deep learning methods (EEGNet, shallow and deep CNN, ResNet, and a deep CNN variant named STSCNN), were used to classify the SEEG data. Various experiments were conducted to investigate the effect of windowing, model structure, and the decoding process of ResNet and STSCNN.** ***Results:*** **The average classification accuracy for EEGNet, FBCSP, shallow CNN, deep CNN, STSCNN, and ResNet were 35 $\pm$ 6.1%, 38 $\pm$ 4.9%, 60 $\pm$ 3.9%, 60 $\pm$ 3.3%, 61 $\pm$ 3.2%, and 63 $\pm$ 3.1% respectively. Further analysis of the proposed method demonstrated clear separability between different classes in the spectral domain.** ***Conclusion:*** **ResNet and STSCNN achieved the first- and second-highest decoding accuracy, respectively. The STSCNN demonstrated that an extra spatial convolution layer was beneficial, and the decoding process can be partially interpreted from spatial and spectral perspectives.** ***Significance:*** **This study is the first to investigate the performance of deep learning on SEEG signals. In addition, this paper demonstrated that the so-called 'black-box' method can be partially interpreted.**

***Index Terms*—stereo-electroencephalography (SEEG), brain-computer interface (BCI), forearm and hand motion, deep learning, convolutional neural networks (CNN)**

I. Introduction

Brain-computer interfaces (BCIs) can be used to translate brain signals into commands to control external devices. BCIs can be broadly categorized by recording methodology into two classes: non-invasive and invasive. For noninvasive methods, Electroencephalography (EEG) [1], which records the signal from the surface of the scalp, is the most popular method due to its low cost and ease of acquisition. Invasive BCIs record signals using different methods, such as electrocorticography (ECoG) [2], spiking activities [3] and stereo-electroencephalography (SEEG) [4], [5]. Invasive methods can provide a higher signal-to-noise ratio than on-invasive methods and so are of significant interest in the development of high-performance BCIs BCIs using an invasive SEEG paradigm have recently been demonstrated. For example, Murphy et al. showed that using a support vector machine (SVM), signals from deep brain regions, including the central sulcus and insular cortex, can be used to differentiate grasp levels [6]. Tan et al. recorded SEEG from 9 individuals with Parkinsonism performing a grasping task at different force amplitudes for two seconds and demonstrated that the subthalamic nucleus showed different spectral responses at different grasp levels [7]. Two-dimensional cursor control and hand gesture classification have also been demonstrated using SEEG [8]. The control of a prosthetic hand using SEEG electrodes on epileptic subjects has also been investigated, by decoding three different hand movements and a resting state [9]. A recent study demonstrated that continuous changing grasp force decoding can be achieved [10] and it is also possible to decode perceived speech from SEEG electrodes located within the auditory cortex [11].

The general processing pipeline of BCIs includes three steps: 1) signal pre-processing and filtering, 2) feature extraction, and 3) classification. Pre-processing is largely dependent on the mode of data acquisition but will include broad-band filtering, trend removal, and potentially re-referencing. Feature engineering is performed to extract useful information based on some template or statistical feature. For example, a running average window was employed to extract the local motor potential (LMP) [12]. Frequency representation is another commonly used feature [13]. For example, the power of the bands 1-4 Hz, 4-8 Hz, 8-13 Hz, 13-30 Hz, and 60-195 Hz was extracted and concatenated as input to an SVM classifier [14]. Other commonly used features are statistical features, such as mean, median, standard deviation, etc. [15]. Dimensionality reduction is also used commonly to reduce the computation time and avoid over-fitting using principal component analysis (PCA) and independent component analysis (ICA) [16]. Classification is achieved using methods including support vector machines (SVM) (e.g. to classify healthy and seizure EEG signals [17], linear discriminant analysis (LDA) (to classify patients with dementia from a healthy control group [18]), and filter bank common spatial filter (FBCSP) (which in this application is often superior to the traditional CSP algorithm [13]).

For all classifiers, feature engineering is a critical step to maximize the overall decoding accuracy. However, the feature engineering and subsequent classification steps all depend heavily on expert experience. Therefore, the optimal feature extraction and classification model might not be guaranteed. In conclusion, this 'traditional' decoding method is sub-optimal in the general case.

Compared to these 'traditional' decoding methods, the equal and often superior performance of deep learning methods, especially convolutional neural networks (CNNs) has been demonstrated in various EEG studies. For example, a CNN model called the TSception network, which was composed of a temporal block and a spatial block showed a superior result in an emotion decoding task [19]. In their design, multiple 1D convolution layers of different kernel

This work is supported by the EPSRC New Horizons Grant of UK (EP/X018342/1), the National Natural Science Foundation of China (No. 91848112 and No. 52105030), the China Postdoctoral Science Foundation (No. 20Z102060158), and the Medical & Engineering Cross Foundation of SJTU (No. AH0200003). (Co-corresponding authors(*): Liang Chen and Dingguo Zhang (email: d.zhang@bath.ac.uk).)

Xiaolong Wu, Benjamin Metcalfe, and Dingguo Zhang are with the Centre for Autonomous Robotics (CENTAUR), Department of Electronic & Electrical Engineering, University of Bath, UK

Guangye Li and Shengjie Liu are with the School of Mechanical Engineering, Shanghai Jiao Tong University, China

Shize Jiang and Liang Chen are with Huashan Hospital, Fudan University, China.

shapes are used along the temporal dimension to extract information in different frequencies. Then in the spatial layer, another set of 1D convolutional kernels of different lengths are used along the channel/spatial dimension to simulate a spatial filter. The same principle was used by the EEGNet to extract temporal and spatial information [20]. ResNet is another widely used deep learning model which has been proven to be superior in image processing, such as image classification, annotation, and captioning. Recently, it has also been tested on the scalp EEG single. [21]–[23].

Despite the recent applications of deep learning to EEG analysis, the application of deep learning (CNNs in particular) in SEEG-based BCI remains largely unknown. To evaluate CNNs on invasive SEEG signals, this work will attempt to decode hand and forearm motion using different methods, including a filter-bank common spatial pattern (FBCSP) [13], EEGNet [20], shallow convolution network (shallow CNN) [24], deep convolution network (deep CNN) [24], ResNet [25], and a proposed novel Spatial-Temporal-Spatial convolutional network (STSCNN). Then, the impact of parameters such as the running window length, stride (window overlap), and model depth, will be investigated. In addition, two more analyses will be conducted to gain an understanding of the so-called 'black-box' methods. For example, Gradient-weighted Class Activation Mapping (Grad-CAM) will be used on ResNet to interpret the contribution of different SEEG electrodes to the decoding task. Then, the spectral analysis will be applied to STSCNN to partially interpret the decoding process.

The novelties of this work are four-fold. Firstly, this work will demonstrate the feasibility of using deep learning on SEEG data. Although the SEEG data was sparse, this work will demonstrate that deep learning can be applied to the SEEG data using one data augmentation method (sliding windows) and different network architectures. Secondly, the spatial-temporal-spatial convolutional layers configuration will be shown to be superior to the temporal-spatial configuration. Thirdly, this work will interpret the so-called 'black-box' deep learning model from the spectral domain, which is useful to reveal possible neuroscientific meanings, such as a bio-marker of a specific frequency band. Finally, it will be shown that the decoding accuracy variation was positively related to the spectral response in both low and high-frequency ranges.

II. Experiment Setup

A. Participants and Data Recording

A total of 30 human participants (participants *1*, *2*...*30*) were recruited in this study. The participants were patients with intractable epilepsy implanted with SEEG electrodes for pre-surgical assessment of seizure focus, and all were enrolled with written consent. To be included in this study, participants must have normal cognitive ability and normal arm movement. The clinical profile of the participants is shown in **Table** I. All implantation parameters were solely determined by clinical need as part of the pre-surgical assessment. SEEG signals were acquired using a clinical recording system (EEG-1200C, Nihon Kohden, Irvine, CA).

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. 2, and is the same as used in [26]. During the experiment, the participants were reclined on the bed and visually cued to perform 5 types of hand or forearm movements in random order. The movement was performed using the hand contralateral to the hemisphere with the majority of the implanted SEEG electrodes. There were three stages in each trial: during the 4s rest stage the participant kept still while resting their arm on the bed; in the cue stage, a cue (a cross) was shown on an LCD screen for 1 second; in the task stage, a picture of a particular motion appeared on the screen and the participant performed the indicated movement (grasp, scissors gesture, elbow flexion, wrist supination, thumb flexion) repetitively for 5s. The subject executed each of the 5 tasks 20 times, resulting in a total of 100 trials per participant (16.67 min total).

C. Electrode Localization

The participants had a total of 4057 electrodes (rounded mean $\pm$ std: 135 $\pm$ 42 per subject) implanted. Each electrode shaft was 0.8 mm in diameter and contained 8–16 contacts (contact length 2 mm) with centre-to-centre spacing of 3.5 mm. To locate electrodes in brain space, we first segmented the individual brain of each participant using the pre-surgical MRI using Freesurfer software [27]. Then, the anatomical location of each electrode contact was obtained using an open-source toolbox, iEEGview [28]. Finally, the extracted contacts from each subject were projected onto a standard brain model (Montreal Neurological Institute, MNI). The location of the contacts is illustrated in Fig. 1. Subplot figures *1*, *2*, *3*, *4*, *5* are electrodes location for five example participants. Subplot *6* shows the locations of electrodes from all participants projected into the MNI standard brain model.

III. Methods

In this paper, a participant-specific model will be trained for every decoding method (within-participant decoding). To train the models, the signals were first pre-processed, and then six decoding methods, including the FBCSP algorithm and five deep learning models, were trained to classify data into five movements.

After the initial decoding, investigations of the CNN-based models were conducted from two perspectives. In the first investigation, the effect of data cropping strategies and model depth on decoding performance was studied. The second investigation, conducted using ResNet and the proposed STSCNN, tried to interpret and visualize the so-called 'black-box' deep learning algorithm. Finally, the relationship between decoding accuracy and frequency modulation was examined.

A. Signal Pre-Processing

First, SEEG data were down-sampled to 500Hz using `resample` in [29]. The signals were then band-pass filtered from 0.5 Hz to 200 Hz using a 4^{th} order Butterworth filter, similar to previous SEEG studies [9]. Then, a notch filter was used to eliminate 50 Hz power-line interference. Next, channels with extensive 50 Hz interference were identified and excluded in the following calculation using the same procedure as in [10], and a total of 32 out of 4606 electrodes were removed.

For FBCSP methods, an extra re-reference was applied using the Laplacian re-reference method, which proved to be beneficial and superior compared to other re-reference methods [26]. No re-reference was applied for deep learning methods because deep learning methods were proved to be able to learn a spatial filter after the training [20].

B. FBCSP Algorithm

For the FBCSP method, the algorithm proposed in [13] was used. In brief, multiple band-pass filters were used to extract signals in

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	500
10	right superior parietal lobe	R	L	M	31	RH	6	94	500
11	right superior parietal lobe	R	R	M	31	RH	6	102	2000
12	right ACC	R	L	M	19	BI	9	130	2000
13	left temporal & insular lobe	R	R	F	30	BI	13	170	2000
14	left temporal lobe	R	R	F	31	LH	10	144	2000
15	left occipital & parietal lobe	R	R	M	27	BI	10	144	2000
16	None	R	R	M	16	BI	13	137	2000
17	right temporal lobe	R	L	M	24	RH	8	108	2000
18	left temporal lobe	R	R	F	30	LH	9	118	2000
19	left temporal lobe	R	R	F	33	LH	12	150	2000
20	None	R	R	F	23	BI	15	198	2000
21	right temporal lobe	R	L	F	23	RH	10	130	2000
22	left temporal lobe	R	R	F	42	LH	10	137	2000
23	left temporal lobe	R	R	M	33	BI	11	154	2000
24	left SMA	R	R	M	15	LH	8	110	2000
25	right occipital	R	L	M	25	BI	8	108	2000
26	None	R	R	M	29	BI	5	72	2000
27	right temporal lobe	R	L	M	22	BI	6	56	2000
28	right parietal lobe	R	L	M	15	RH	7	102	2000
29	None	R	R	M	26	LH	10	136	2000
30	left temporal lobe	R	R	F	27	BI	10	117	2000

Abbreviations for this Table: SID: Subject ID; EZ, Epileptogenic Zone; RH, Recording Hemisphere; BI, bilateral; SR, Sampling Rate; SMA, Supplementary Motor Area; ACC, anterior cingulate cortex; EL, Number of Electrode Shafts; NC: Number of Contacts; DH, dominant hand; EH, experiment hand; None: epilepsy zone unknown.

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. Then spatial filtering was performed using the CSP algorithm. Next, the mutual information-based feature selection algorithm was used to select the discriminative CSP features. Finally, an `SVM` classifier was used to classify the signal into 5 tasks [30].

C. Deep Learning Models

Five deep learning networks were implemented and compared in this study, including EEGNet [20], deep convolutional network (deep CNN) [31], shallow convolutional network (shallow CNN) [31], ResNet [25], and STSCNN. Since detailed information on other deep learning models can be found in their original papers, only details of ResNet and STSCNN will be presented in this section.

1) ResNet Model: The ResNet architecture used in this work is similar to other ResNet variants, the only difference was the number of the convolution layers [25]. The schematic plot of the ResNet model implemented in this paper is presented in Fig. 3. In the first block, the 2-dimensional raw SEEG input, in the shape of (channel, time points), was expanded to a 3D cube using a convolution operation with multiple 1D kernels. The second block is a residual block with an identity connection. The residual block is composed of two 3×3 convolutional layers, Batch normalization, and ReLu non-linearity. The third block is similar to the second one, except a 1×1 convolutional connection was used instead of the identity connection. In blocks 2 and 3, the depth of the feature map increases, while the width and height decrease. In the last block, the intermediate 3D cube will be collapsed into a 1D vector using a pooling operation and then transferred to a 1×5 vector using a linear layer. To complete the classification, a softmax operation was applied to yield a probability prediction.

2) STSCNN Model: The STSCNN was inspired by a deep CNN model. In the original deep CNN, the raw data is first fed into a temporal and then spatial convolutional layer. The temporal and spatial layers serve as temporal and spatial filters, respectively. However, SEEG electrodes penetrate through a wide range of brain areas and only part of them contain information useful for decoding within the proposed paradigm. Therefore, it is reasonable to hypothesize that it would be beneficial to have an extra spatial layer before the temporal-spatial layers. In the proposed model, a spatial layer was introduced on top of the original deep CNN (taken from [31]) to attenuate the noisy channels. The attenuation can be achieved by learning and assigning low weights to channels. The major components of the proposed STSCNN are presented in Fig. 3.

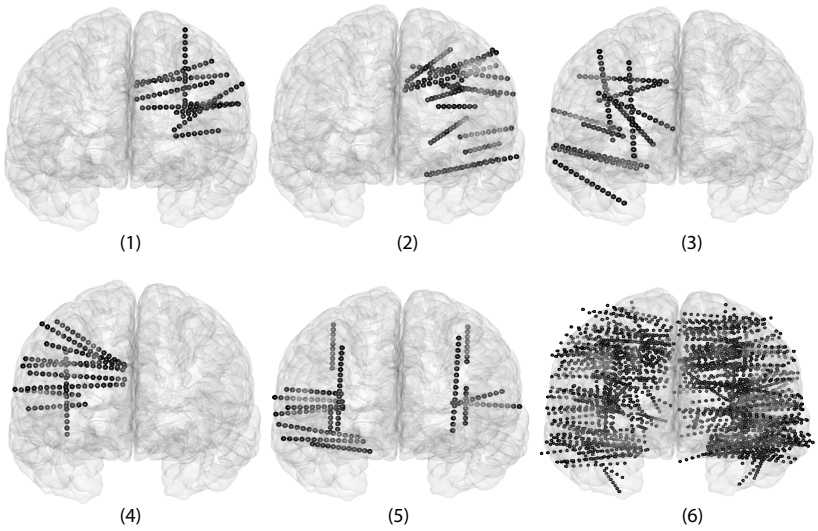

Fig. 1. The 3D locations of the SEEG electrodes. Subplots 1,2,3,4 and 5 illustrate the electrodes' location in the individual brain space for 5 example participants. The lower right subplot represents the electrodes aggregated from all subjects, projected into the MNI standard brain model.

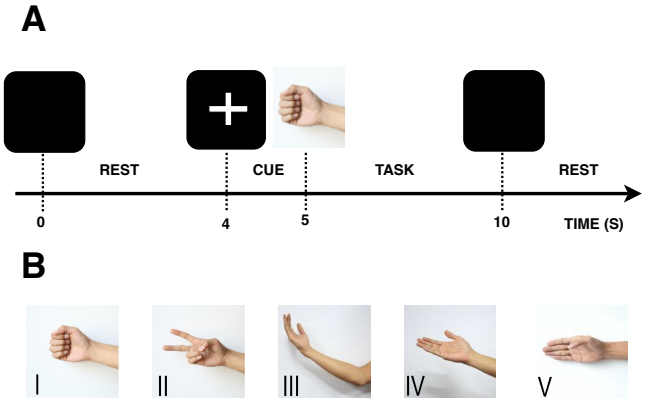

Fig. 2. The experimental paradigm. (A) There are three stages in one trial: rest, cue, and task. In the resting stage, the participant remains still for 4 seconds. Then a cross cue will appear, lasting for 1 second, to prepare the subject for the upcoming task. In the task stage, a picture of one of five motions will appear, and the participant will perform the corresponding task repeatedly for 5 seconds before the screen turns dark again. (B) Each participant performed five different types of forearm and hand motions (grasp, scissor gesture, elbow flexion, wrist supination, thumb flexion). The pictures were randomly presented while ensuring an equal number of appearances of each gesture.

D. Training Procedure

A stratified five-fold cross-validation procedure was employed for both the FBCSP and deep learning model training. However, the data splitting procedure was different between these two methods. For FBCSP, the entire data were split into training and testing datasets, while for the deep learning method, an extra validation dataset was needed for hyperparameter tuning and early stopping. In detail, for FBCSP, data were split in an 80/20 manner in each fold, in which the model was trained on an 80% training set and then tested on the remaining 20% testing set. For the deep learning model, data were first split in an 80/20 manner in each fold, and then the original 80% dataset was further split into an 80% training dataset for model training and a 20% validation dataset for hyper-parameter tuning and early stopping, while the original 20% data were used for model testing. Therefore, the final training, validation, and testing dataset contained 64%, 16%, and 20% of the entire dataset for each fold. The training was carried out on a desktop computer with an Intel(R) Xeon(R) Gold 5118 CPU, 64.0 GB RAM, and one NVIDIA Quadro P5000 GPU card. During the validation, the mean decoding accuracy averaged across all participants and all folds were reported.

In addition, since there are only 20 trials for each task, a data augmentation procedure (cropped training) was implemented after data splitting. Cropped training is a common strategy for data augmentation for deep learning, and it has been proven to be effective to enhance the decoding performance of object recognition in image recognition [25] and EEG decoding tasks [31]. In detail, for each trial in the original data, a sliding rectangle window of 500 ms was used to slide along the time dimension with a 200 ms step size (stride). This windowing process will partition the original trials (10 s long) into multiple shorter trials (500 ms long). In the end, there were 1900 trials for each subject, organized in the shape of (1900, N, 500) where N represented the channel number. Using the partitioning schema presented in the previous content, there were 1216 (1900 $\times$ 64%), 304 (1900 $\times$ 16%) and 380 (1900 $\times$ 20%) trials in the training, validation and testing set, respectively. To train all five deep learning models, the loss was calculated as the cross entropy loss and the Adam optimizer [32] was used to update model parameters at a learning rate of 0.001. The maximum training epochs were set to 500 and the training process will be stopped if the validation loss stopped decreasing for 20 epochs (early stopping).

After the initial decoding, investigations of the CNN-based models were conducted from two perspectives. In the first investigation, the effect of data cropping strategies and model depth on decoding performance was studied. The second investigation, conducted on the ResNet and the proposed STSCNN, tried to interpret and visualize the so-called 'black-box' deep learning algorithm.

E. Analysis 1: Cropped Training and Model Depth

After the initial classification, two extra experiments were conducted on the best model to explore the effect of cropping strategy and model depth on decoding accuracy.

1) Cropped Training Process: In the initial cropped training, each data set was augmented using a window of 500 ms and a step size of 200 ms. The best combination of these two parameters will be

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2023.3242262

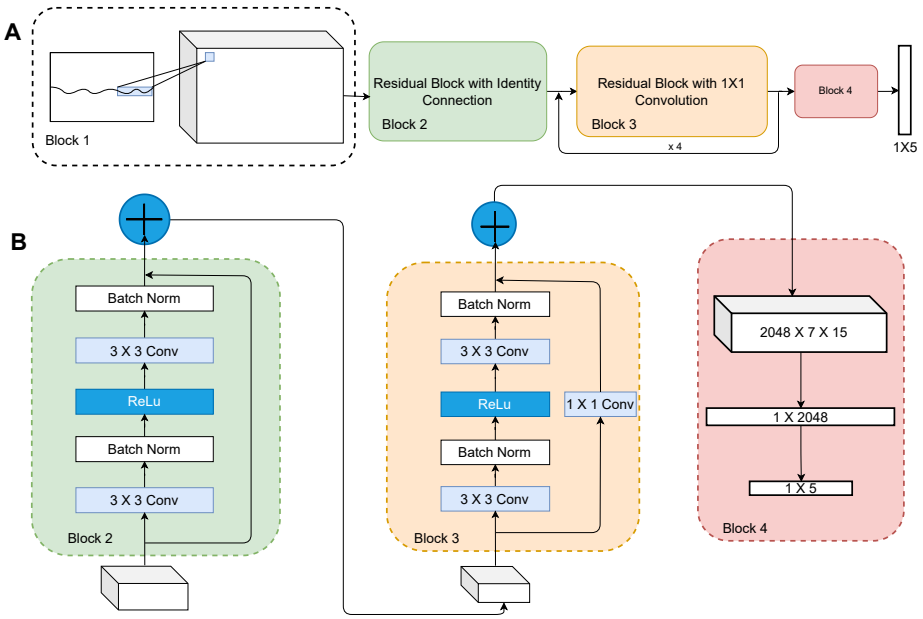

Fig. 3. ResNet model architecture. It consists of 4 blocks, as shown in upper part A. The first block uses multiple 1D kernels to expand the 2D data to a 3D cube. The second block is a residual block with an identity connection. The third block is a concatenation of three residual blocks with a 1×1 convolution connection. In the last block, the 3D cube was converted into a 1D vector using a pooling operation. In lower part B, the detailed structures of blocks 2, 3, and 4 were presented.

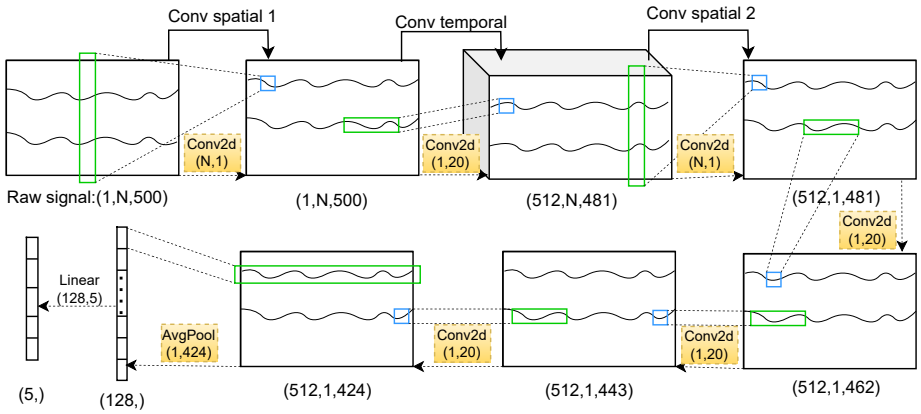

Fig. 4. The proposed STSCNN model architecture. Input and features were represented as black rectangles with shapes indicated below. Kernels and their products are denoted as green and blue rectangles while the kernel sizes were indicated in the middle yellow box. ReLU was used as the non-linear function. Two spatial filters and one temporal filter were denoted as Conv spatial 1, Conv spatial 2, and Conv temporal in the plot. AvgPool: average pooling layer; Conv2d: 2-D convolutional layer;

investigated further in this section. Since the number of possible combinations is large, an alternating searching strategy was used. The decoding accuracy was first evaluated using window sizes from 200 ms to 1200 ms in 200 ms increments, while the step size was fixed at 200 ms. Then the optimal window size was identified which corresponded to the highest decoding accuracy. Next, step sizes of 20 ms, 50 ms, 100 ms, 200 ms, 300 ms, 400 ms, and 500 ms were tested using the best window size identified in the previous step. The best step size is the one that achieved the highest decoding accuracy. The best combination of window size and step size will be used for subsequent analysis.

2) ResNet With Various Depths: After the optimal cropping window size and stride size were determined, the effect of model depth on the decoding performance was evaluated. While more layers provide better representation ability, over-fitting is a major problem. Therefore, a balance must be achieved between model complexity and generalization. In this section, various depths were evaluated and compared using the ResNet model. As shown in Fig. 3, the original ResNet has 4 stacked residual blocks. The performance of ResNet using 3,4,5,6,7,8 stacked residual blocks will be further investigated.

F. Analysis 2: Visualization of The Decoding Process of ResNet and STSCNN

Both ResNet and STSCNN were re-trained using the previously obtained optimal cropping and depth hyper-parameters. Then, different methods were explored to visualize the so-called black box deep learning model.

For STSCNN, the first three layers were designed to mimic the spatial and temporal filter using 1-dimension convolution kernels along channel and time axes. Therefore, STSCNN will be interpreted from

both spatial and spectral perspectives. However, for ResNet, the 2D convolution kernels cannot be viewed as spatial or spectral filtering. Therefore, a common technology to visualize the decoding process of deep learning, Gradient-weighted Class Activation Mapping (Grad-CAM), was used to interpret the ResNet model [33].

1) Visualization of Decoding Process of STSCNN: For STSCNN, Conv_spatial_1/Conv_spatial_2 and Conv_temporal are designed to mimic spatial and temporal filtering operations, respectively. Three approaches were taken to examine the decoding process of STSCNN. First, the correlation between SEEG data of different tasks was examined before and after the Conv_spatial_1 layer to check if this layer mimics a spatial filter to maximize the distinguishability. Second, the correlation between different sub-band power traces of raw data and the output (feature map) of the Conv_spatial_2 layer was calculated to test if the network extracted band-specific features like a filter, similar to the EEG study [31]. Third, the spectral content was analyzed on the feature map of Conv_temporal to evaluate the separability in the spectral domain among different classes.

In the first investigation, we hypothesize that an extra spatial layer may enhance the distinguishability (quantified by correlations coefficient) among different task data in the temporal domain. The 2D (time * channel) raw data was denoted as bold $\boldsymbol{x_i}$, while the 2D feature map of Conv_spatial_1 was denoted as $\hat{\boldsymbol{x_i}}$, where *i* represent the class. The correlation coefficient between all possible pairs of classes before and after Conv_spatial_1 were calculated using the following equations, separately.

$$corr_{ij} = corr(\boldsymbol{x_i}, \boldsymbol{x_j}), i, j = 1, 2, 3, 4, 5 \tag{1}$$

$$\hat{corr}_{ij} = corr(\hat{\boldsymbol{x_i}}, \hat{\boldsymbol{x_j}}), i, j = 1, 2, 3, 4, 5 \tag{2}$$

If the model has learned a spatial filter, it is expected that there will be a decreased R between classes.

In the second analysis, to investigate the filtering behaviour of the model, the correlation between the sub-band power trace of input data and the feature map of Conv_spatial_2 was calculated for each channel separately. Firstly, the temporal-spectral representation $\hat{ts}_c^{t,f}$, in which *t,f,c* represents time stamp, frequency range, and channel, of each channel was calculated in the (0,200) Hz range. The frequency resolution was set to 4 Hz, which results in 200/4=50 frequency intervals. This 2D time-frequency representation was calculated using the FFT-based *tfr_morlet* in the *MNE* package [34]. Then, sub-band power was obtained by squaring the above representation. Next, to obtain the power trace $ts_c^{t,f}$, a moving-average window was used to slide along the temporal axis. The window length is set to be the same as the receptive field of a single value (unit) in the output feature map $fmap_{tk}$ (*t* and *k* correspond to feature map width and height, respectively) of Conv_spatial_2. The procedure was in line with the previous EEG study [31]. For example, in STSCNN, the receptive field of one unit in the Conv_spatial_2 feature map is 20 points in the raw SEEG data (Spatial convolution will not change the receptive field along the temporal dimension.). Therefore, the window length was set as 20 to slide along the power trace and calculate the mean value in each window. Next, a scalar correlation coefficient $R_c^{f,k}$ was obtained between the power trace $ts_c^{t,f}$, and a particular row *k* in the feature map $fmap_{t,k}$. This calculation was repeated for 50 frequency ranges and 512 rows to produce a 2D map $\boldsymbol{R}_c$ for a particular channel.

In the last analysis, the hypothesis is that the STSCNN learned to extract signals in specific frequency ranges to facilitate classification. To verify this, the frequency content was analyzed for each row of the 2D feature map (output) $fmap_{t,k}$ of Conv_spatial_2 layer using the FFT method.

2) Visualization of Decoding Process of ResNet: The Grad-CAM [33] method was used to visualize the decoding process of the ResNet model. Grad-CAM is commonly used in image recognition, and it produces a coarse localized heat map highlighting the key regions for decoding. However, the interpretability of the deep learning model using Grad-CAM depends on the model structure. For example, in an image classification task, the gradient heat map should be of a similar size as the raw input image. Then, the final Grad-CAM activation map can be produced by first rescaling the gradient map and overlapping it on top of the raw image. For the ResNet implemented in this work, the width and height of the feature map were halved at every residual block, which meant the shape of the gradient map was more and more different from that of the raw SEEG data. For example, for one SEEG data in the shape of (208, 500), the shapes of the feature map (omit the batch dimension) of block 1, block 2 and block 3 are (64, 208, 449), (128,108,225), (2048,7,15), respectively. When the Grad-CAM was applied on block 3, the weighted gradient of shape (7, 15) will be re-scaled to (208, 500) and overlapped on top. The meaningful interpretation of the Grad-CAM was lost because of this large re-scaling factor. Therefore, in this work, the visualization was conducted on the second block, in which the (108,225) gradient was re-scaled and then overlapped on top of the (208, 500) input.

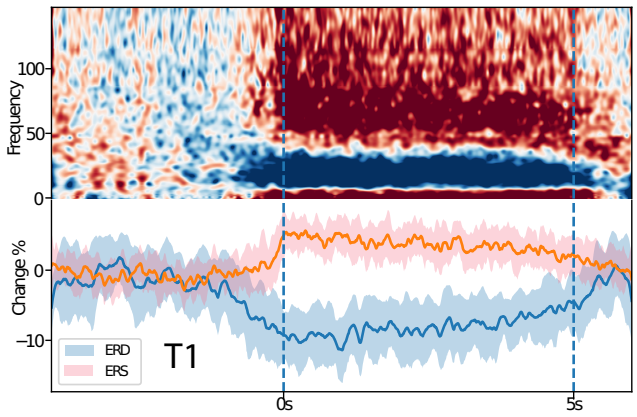

Fig. 5. Time-frequency representation (upper subplot) and its corresponding ERS/ERD plot (lower subplot) for task 1, generated from an example electrode of subject 30. Other task responses exhibit a similar response. The upper half is normalized amplitude values obtained against the baseline period (before the first dashed vertical line). The lower half contains amplitude of high frequency (60 to 150 Hz) and low frequency (10 to 30 Hz), and the thick lines are average amplitude. Two dashed vertical lines represent the beginning and the ending of the task, respectively.

G. Relationship Between Decoding Performance and Channel Reactivity

In the previous analysis, it is evident that STSCNN performs classification by extracting information in frequency bands. Therefore, the hypothesis is that subjects with high decoding accuracy have more electrodes that are reactive (responsive) to the movement task in the spectral domain. One of the methods to quantify task reactivity is the correlation coefficient between frequency power and task state (idle or moving). Event-related synchronization (ERS) and event-related desynchronization (ERD) are well-known power modulations found in brain signals. Therefore, the power of frequencies showing ERS/ERD was used to compute the correlation. To find the frequency range exhibiting ERS/ERD, the temporal-spectral representation of the SEEG signals was obtained by performing time-frequency decomposition using the public MNE toolbox [34] and normalized to the baseline (the resting stage), same as in another SEEG-based

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2023.3242262

BCI study [10]. From the temporal-spectral representation, a high-frequency range between (60 to 150 Hz) and a low-frequency range between (10 to 30 Hz) were identified which showed obvious ERS and ERD, respectively, as shown in Fig. 5. For a certain channel c, the normalized power traces of high and low-frequency ranges were obtained by averaging along the frequency axis and denoted as HFP_c and LFP_c, respectively. In addition, two number series, $label_{HFP} = [-1, ... -1, 1, ...1]$ (-1 and 1 represent idle and task state, respectively) and $label_{LFP} = [1, ..1, -1, .. -1]$ (1 and -1 represent idle and task state, respectively) were defined. Finally, correlation coefficients, $corr^c_{HFP}$ and $corr^c_{LFP}$, for a certain channel c, can be obtained using Eq. 3 and Eq. 4, respectively. In additional, the mean correlation coefficient $\overline{corr}^c$, between $corr^c_{HFP}$ and $corr^c_{LFP}$, was also obtained using Eq. 5.

$$corr^c_{HFP} = corr(label_{HFP}, HFP_c) \tag{3}$$

$$corr^c_{LFP} = corr(label_{erd}, LFP_c) \tag{4}$$

$$\overline{corr}^c = mean(corr^c_{HFP}, corr^c_{LFP}) \tag{5}$$

Next, the participant-specific reactivity indicator calculated from the high-frequency band, denoted as $reac^{HFP}_{sid}$ (sid is the subject ID), was obtained by averaging $corr^c_{HFP}$ across all channels for that participant. The same procedure was used to obtain $reac^{LFP}_{sid}$ and $\overline{reac}_{sid}$, which represents the low-frequency band indicator and the mean indicator, respectively.

IV. Results

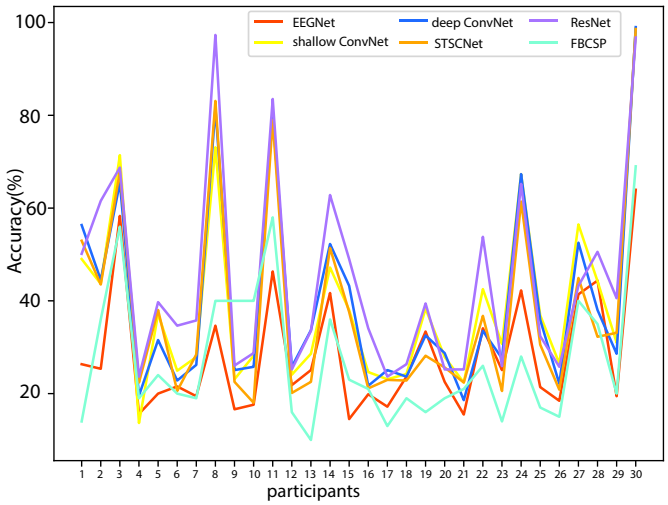

Fig. 6. Decoding accuracy of all participants using all models. The X-axis is the subject ID, and the Y-axis is decoding accuracy. Different decoding methods were represented in different colours.

A. Decoding Result

The decoding result of 30 participants is presented in Fig. 6 using a 500 ms window and step size of 200 ms. The mean decoding accuracy of EEGNet, FBCSP, shallow CNN, deep CNN, STSCNN, and ResNet are 27 ± 6.5%, 37 ± 5.2%, 39 ± 4.1%, 42 ± 4.0%, 44 ± 3.3%, and 51 ± 3.2%, respectively, averaged across all 30 subjects. This overall decoding accuracy is low compared to other invasive studies. This is because the SEEG electrodes were placed strictly according to the clinical needs of epilepsy treatment. In many cases, the electrodes were placed in regions that are not related to movement activity, this likely leads to poor decoding performance. Since there is evidence that signals from the sensorimotor area of the cortex contain motor control information, participants that do not have electrodes in these areas were excluded. Thus, five participants (9, 21, 22, 25, 29) were identified and excluded from the subsequent content. For the remaining subjects, the average decoding accuracy was 35 ± 6.1%, 38 ± 4.9%, 60 ± 3.9%, 60 ± 3.3%, 61 ± 3.2%, and 63 ± 3.1% for EEGNet, FBCSP, shallow CNN, deep CNN, STSCNN, ResNet, as shown in Fig. 7. ResNet obtained the highest mean decoding accuracy, while EEGNet was the worst. The decoding accuracy of shallow CNN, deep CNN, STSCNN, and ResNet were significantly higher than that of EEGNet and FBCSP (Wilcoxon signed-rank test, p=0.000015). However, no significant difference was found between EEGNet and FBCSP, and among shallow CNN, deep CNN, STSCNN, and ResNet.

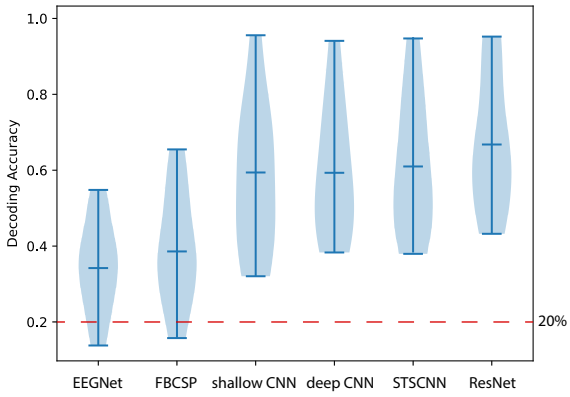

Fig. 7. Violin plot of the decoding accuracy obtained from the selected subjects using all models. A Violin plot can show the decoding accuracy distribution. The top, bottom, and middle bars represent the min, max, and mean accuracy, respectively. The dashed line indicates the chance level accuracy.

B. Analysis 1: Cropped Training and Model Depth

In the above decoding task, the ResNet model was initially constructed with 3 stacked residual blocks and trained using an arbitrary 500 ms window and 200 ms step length. In the subsequent section, the effect of residual block number, window size, and step size on decoding accuracy will be analyzed.

1) Cropped Training: In the previous result, two cropping parameters, window size, and step size were set to 500 ms and 200 ms arbitrarily. To find the optimal cropping strategy, these two parameters were searched in turn, as presented in the Method section: III-E1.

The decoding accuracy averaged across the selected participants, using different window sizes and step sizes are presented in Fig. 8. The blue dots are decoding accuracy obtained with different window sizes, as indexed by the lower axis. The highest decoding accuracy was obtained with a window size of 400 ms. The red dots, indexed by the upper axis, represent the decoding accuracy obtained with different step sizes while the window size was fixed at 400 ms. The highest decoding accuracy was obtained with a step size equal to 100 ms using a window size of 400 ms. Therefore, the optimal combination is a 400 ms window size and a 100 ms step size. The same combination will be used in the subsequent analysis.

2) Accuracy Varied With Depth: After the optimal window size and step size hyper-parameters were determined, the effect of depth (number of residual blocks) on decoding accuracy was explored. Using the remaining participants, the average performance of ResNet

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2023.3242262

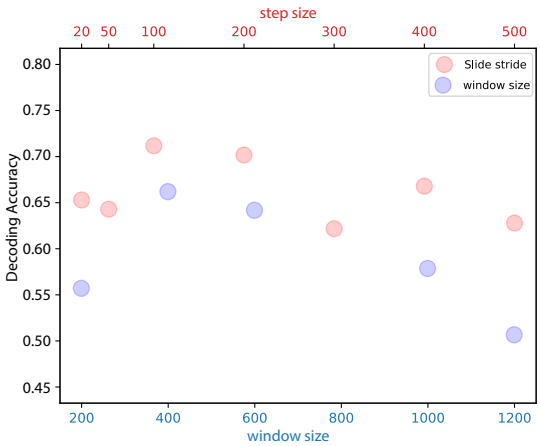

Fig. 8. Decoding accuracy using different window sizes and step sizes on the ResNet model. The blue dots were decoding accuracy obtained with different window sizes, as indicated in the lower axis. The highest decoding accuracy was obtained using a window size of 400 ms. The red dots represented decoding accuracy obtained with different step sizes, indexed by the upper axis, while the window size was fixed at 400 ms. The highest decoding accuracy was obtained with a step size equal to 100 ms.

was evaluated using 3,4,5,6,7,8 residual blocks. The result is presented in Fig. 9. It shows that ResNet performance increased with depth and plateaued at 5 layers. Another point worth noting was that the decoding performance did not decline with more layers. This point is in line with other image recognition studies, which demonstrated that over-fitting was mitigated in a deeper residual network [25].

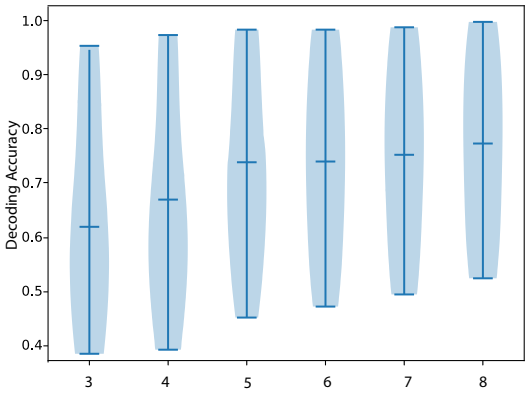

Fig. 9. Violin plot of the decoding accuracy of the selected subjects using ResNet with various depths. Decoding accuracy plateaued at five residual blocks.

C. Analysis 2: Visualization of The Decoding Process of ResNet and STSCNN

The subsequent two subsections will present the visualization of the decoding process of ResNet and STSCNN, respectively. Both models were re-trained using the previously obtained cropping and depth hyper-parameters.

1) Visualization of Decoding Process of ResNet: An Grad-CAM heat map, generated from an example trial of participant *11*, is presented in Fig. 10. The heat map shows a clear stratified distribution of high gradient amplitude along the channel axis. To better visualize the different weights assigned by ResNet, the mean gradient was calculated along the channel axis and plotted to the right side. It is clear that ResNet has assigned different importance to channels: high values denote the informative channels for the decoding task, as illustrated in the right subplot.

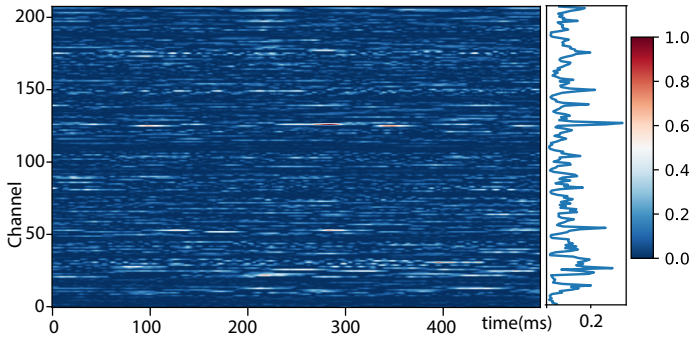

Fig. 10. Grad-CAM heat map generated from the second residual block of ResNet using an example trail from movement task 1 of participant *11*, which has 208 electrodes. X and Y axes are the time points (ms) and channel index, respectively. The right subplot is the mean activation strength averaged along the channel axis.

2) Visualization of Decoding Process of STSCNN: Three approaches were used in this section to interpret the decoding process of STSCNN. The analysis of these three approaches was conducted on the fully trained network with window and step sizes of 400 ms and 100 ms.

In the first approach, the hypothesis that an extra spatial layer may enhance the distinctiveness of data from different classes was examined. The correlation coefficient R was used to qualitatively measure the distinctiveness between data from the two classes. The resulting R matrix, calculated using the method described in section III-F1 for subject 11, was presented in Fig. 11. It is obvious from these two subplots that the correlation between two different classes decreased after the first spatial layer. This enhancement brought by the spatial layer was similar to that obtained by a spatial filter (Laplacian reference) methods [14], [26].

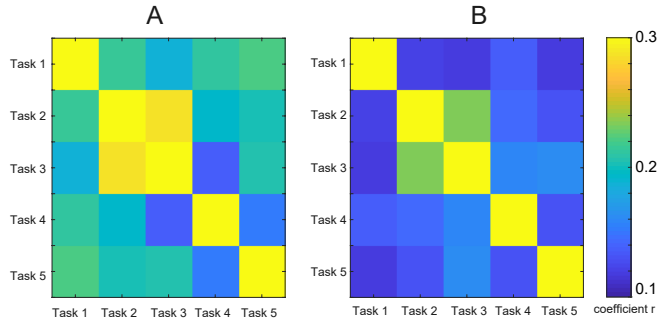

Fig. 11. Matrix of correlation coefficient (r) of all possible pairs of the task data. Subplots A and B represent r before and after the spatial layer, respectively. This plot was obtained from an example participant 11. Other subjects exhibited a similar result.

Next, to find out if the network extracted band-specific features like a filter, the testing data set was fed into the trained network to obtain the intermediate feature map of Conv_spatial_2. Then, a scalar value was obtained by correlating the sub-band power trace and rows of the feature map. For the 512 rows and 50 frequencies, this correlation was repeated 512×50 times to obtain a 2D map, as shown in Fig. 12 for subject 11.

For improved visualization, the horizontal axis has been sorted by the average correlation coefficient within 60-200 Hz. After calculating

the same 2D correlation map for all channels, it is clear that there were around 47% of total channels that showed a strong correlation similar to Fig. 12. The right subplot shows the mean trace (blue dashed line) of the absolute value averaged along the x-axis. To avoid the possibility that the strong correlation resulted from network random initialization, the same correlation map was calculated using a randomly initialized model, then the mean trace was plotted in the right subplot as the green dashed line. The red dashed line represents the difference between the blue and green lines. As shown in the plot, there is an obvious difference between correlations calculated using feature maps on the trained and untrained models, demonstrating that STSCNN has learned frequency-specific features after training.

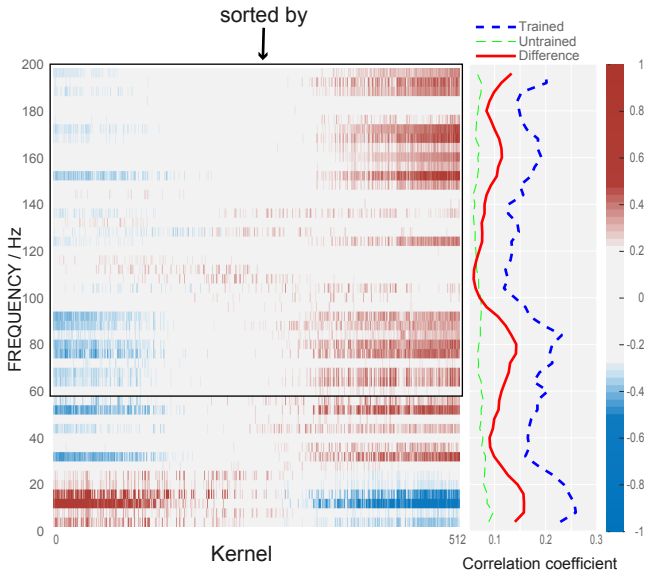

Fig. 12. Correlation map obtained from an example subject 30. The left subplot is the correlation coefficient map between the input sub-band power trace and the Conv_spatial_2 feature map. Each row of the feature map corresponds to a convolutional kernel. The kernels were sorted by the mean correlation in the range of 60 and 200 Hz. The right subplot shows the mean correlation obtained from the trained (blue dashed line) and untrained model (green dashed line). The red dashed line indicates the difference between the blue and green lines.

In the last analysis, the hypothesis that the model was trained to perform classification in the spectral domain was examined by performing a frequency analysis of the feature map of Conv_spatial_2. Similar to the previous procedure, data from different tasks were fed into the trained model separately, and the frequency analysis was performed using `scipy.fft` on each row of the feature map of Conv_spatial_2. The result from the FFT analysis was plotted with different colours for different tasks. The mean frequency amplitude was denoted as solid lines, and the region within one standard deviation was denoted as the shaded area. Results from four example rows were presented in Fig. 13 for subject 11. As can be seen from the plot, there is a clear separation among tasks, as indicated by different colours. To quantify the separation, the number of distinguishable class pairs was calculated using the Wilcoxon rank-sum test (critical p-value was 0.05) at each frequency value.

To highlight the frequencies that exhibit the most significant difference between classes, the number of distinguishable class pairs was raised to Euler's number e (exp() transformed). For a better visualization of kernel 35 which exhibits separability in a very narrow sub-band, the frequency axis was log-transformed and plotted in the upper-right corner.

In addition, the subplots demonstrate that frequency ranges which exhibit obvious separation are different for different rows in the feature map. This heterogeneous filtering can further enhance the decoding performance, similar to the ensemble methods, which use multiple learning algorithms to obtain better predictive performance than could be obtained from any of the constituent learning algorithms alone [35].

D. Relationship Between Decoding Performance and Frequency Response

Since the STSCNN was shown to perform classification in the spectral domain, this section will further investigate the relationship between decoding accuracy and the spectral response.

In a preliminary analysis, the electrode distribution according to the reactivity measured by $\overline{corr}^c$, from four example subjects was plotted in Fig. 14. By comparing the distribution between two columns, it is clear that subjects *11* and *30*, who have higher decoding accuracy, tend to have more electrodes with high reactivity. While subjects *4* and *5*, who achieved chance level decoding accuracy, have more electrodes distributed in the low reactivity range.

To have a clearer view of the relationship between decoding accuracy and electrode reactivity, a scatter plot was used to show $reac_{sid}^{HFP}$, $reac_{sid}^{LFP}$, and $\overline{reac}_{sid}$ against decoding accuracy for all participants, as shown in Fig. 15. The scatter plots were fitted using linear and 2nd order polynomial curves as shown in the blue and the red lines, respectively. As can be observed in the plot, there is a positive relationship between the spectral response in both low and high-frequency bands and the decoding accuracy. Visual inspection indicates that polynomial curves have a better fitting.

V. Discussion

A. Extra Spatial Filter is Beneficial

This work demonstrated that the STSCNN performs better than the original deep CNN model by adding an extra spatial filter layer. This enhancement may come from eliminating irrelevant channels by projecting the raw SEEG data into another space (much like LDA). A further question is what is the best dimension in the projection space. In this work, the raw data were projected into a space with the same dimensions as the channel number (the dimension number remained the same). This strategy may be sub-optimal when considering the fact that SEEG electrodes were arranged along a shaft and a strong correlation might exist between adjacent electrodes, which means only a subset of channels is required for the decoding. Therefore, further study is required to understand the optimal projection space dimension.

B. Application of ResNet on Time Series Data

Neural networks are a universal approximation method, and when given sufficient capacity, a feed-forward network with a single hidden layer is able to approximate any function. However, such a complex layer is prone to over-fitting, and a common workaround is to add more layers (make a deeper network). While AlexNet has only 5 convolutional layers, VGG net [36] and GoogleNet [37] have 19 and 22 layers, respectively. However, increasing depth simply by stacking more layers may cause the notorious vanishing gradient problem. As a result, as the network gets deeper, the overall performance saturates or even begins to degrade. To solve the gradient vanishing problem, gated shortcut connections were introduced as a 'highway', hence the Highway Network in which parameterized gates were introduced to control information flow through the shortcut [38]. This gated flow of information can also be found in the Long Term Short Memory

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2023.3242262

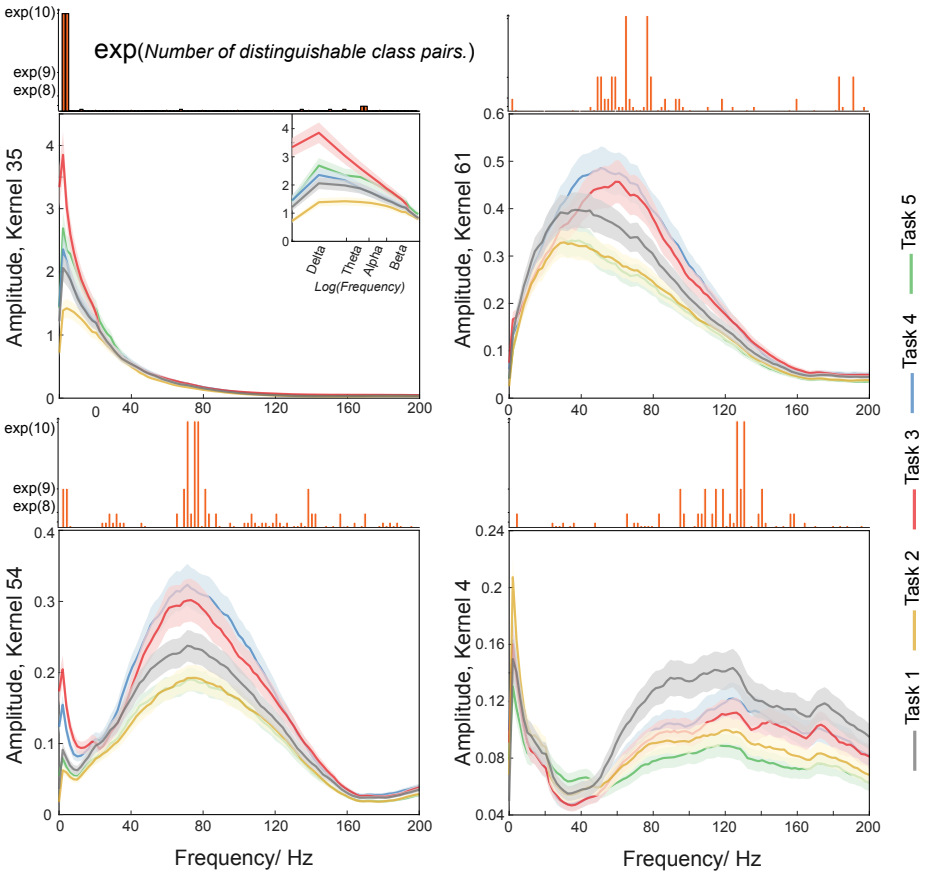

Fig. 13. Frequency representation of four example rows taken from the feature map of Conv_spatial_2 using testing data obtained from an example subject 30. Each subplot corresponds to one row from the feature map, and thus corresponds to one convolutional kernel. Different colours represent different tasks. The shaded area denotes the area within one standard deviation from the mean value (solid line). To have a better visualization for the first subplot (upper left), the frequency axis was log-transformed as an additional inset plot. The bars on top of each subplot indicate the exponent of the number of distinguishable class pairs.

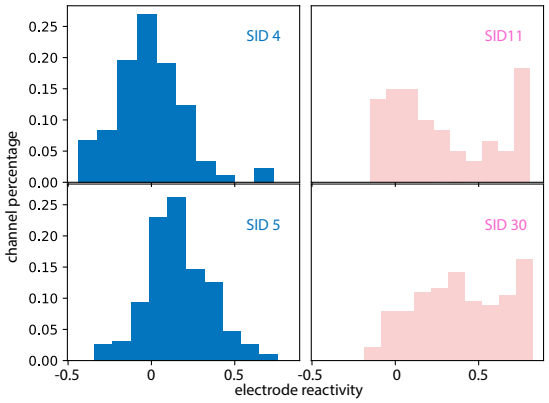

Fig. 14. Electrodes distribution according to the reactivity measured by $\overline{corr}^c$. Subjects *4* and *5* on the left column achieved lower decoding accuracy compared to subject *11* and *30* on the right column. It shows that subjects tend to have more reactive electrodes if the subject achieves higher decoding accuracy.

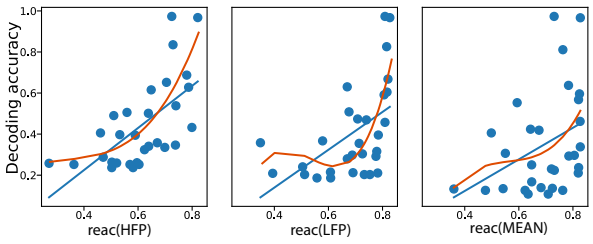

Fig. 15. Relationship between subject electrode reactivity and decoding accuracy. Three subplots denote the scatter plot using $reac_{sid}^{HFP}$, $reac_{sid}^{LFP}$, and $\overline{reac}_{sid}$, respectively. The blue and red lines represent the linear fitting and polynomial fitting, respectively.

(LSTM) cell [39]. In this sense, ResNet can be seen as a special case of a Highway Network that uses an identity shortcut connection to bypass layers. Because of the residual connection, many independent effective paths are available. One counter-intuitive result from these multiple paths is that ResNet can achieve comparable performance even if some layers of a trained network were dropped [40]. However, its usage in BCI is rare, a possible reason is that ResNet requires a 3D cube as input while EEG is 2D time series. A common practice in these few EEG studies is to apply an extra feature engineering step before the ResNet model. For example, the direct Directed Transfer Function (dDTF) method was used to represent EEG as images before the ResNet model [23]. In another EEG study, short-time Fourier transform (STFT) was applied to EEG data first to yield the 3D input for the ResNet model. In this paper, a full decoding pipeline was proposed by first expanding the 2D SEEG data into 3D data using multiple 1D convolutional kernels. The resulting 3D data will be used

in the following ResNet model, similar to the image classification task.

In addition, this work introduces the first application of grad-CAM in BCI decoding tasks. The grad-CAM technique is commonly used in image recognition tasks, in which regions with high contribution can be identified with high gradient values. For time-series data (SEEG data in this work), the heat map exhibited a clear stratified distribution of high value along the channel dimension (Fig. 10). This implies the different contributions of channels to the decoding task, and therefore it could be a potential method for channel selection.

C. Filtering Effect of CNN

In the analysis of STSCNN, the feature map of Conv_spatial_2 demonstrated high separability among different motion types in certain frequency ranges. More interesting is that the common 1/f power law does not apply to the feature map anymore, as demonstrated in Fig. 13. This means that the STSCNN was trained to amplify certain frequencies while suppressing others. In addition, the frequency ranges that have been amplified or suppressed are different for different kernels, which imply that STSCNN might has learned a set of heterogeneous filters. These heterogeneous filters were beneficial to the classification task. For example, in the four subplots of Fig. 13, the discriminative frequency was around 5 Hz, 70 Hz, 80 Hz, and 120 Hz respectively. Therefore, higher decoding accuracy can be obtained by aggregating decoding results from the above different frequency ranges, similar to the idea in the ensemble learning. However, not every channel exhibits such spectral separation as in Fig. 13, and visual inspection found around 10% of total channels behaved in this way. This means the spectral information was only part of the whole information used for the classification. Therefore, this approach only provides a partial explanation of the decoding process of the 'black-box' model.

D. Limitation and Future Work

The vast array of deep learning models and their variants prohibits a thorough evaluation of deep learning methods in BCI applications. For example, many variants of ResNet exist, such as ResNet-50, ResNet-110, and ResNet-152. In addition, even though many of these models have been proven to be effective in some applications, a satisfactory result can not be guaranteed in others. For example, while a superior result can be obtained using ResNet, a contrary conclusion from another BCI study showed worse performance compared to deep CNN [31]. As a result, the applicability of a model in a particular task needs to be evaluated on a case-by-case basis. To further complicate the issue, there are many possible options in every aspect of the deep learning model, including, for example, the kernel shape of the convolutional layer, the non-linearity function, the pooling operation, etc. The enormous range of models and possible design options mentioned above prohibit a thorough analysis of the deep learning method for SEEG signal decoding. Therefore, this paper only presents a subset of possible deep learning architectures, and as a consequence, the optimal model can not be guaranteed. In this work, the goal was the gain an insight into the impact of different network structures on decoding performance, rather than to find the optimal solution. The STSCNN was studied for two reasons. Firstly, it showed that an additional spatial filter is beneficial as it helps to separate different classes (Fig.11). Secondly, it demonstrated that the CNN-based deep learning models can be partially understood from the spectral perspective (Fig.13).

Another limitation is the interpretation of STSCNN conducted in this paper. In the spectral analysis of the feature map of Conv_spatial_2, it was demonstrated that the classification was possibly conducted in the spectral domain. However, it only can be confirmed if the next convolution layer, which consumes the feature map, indeed mimics a filter that band passes the differentiating frequency ranges identified in the previous step. Without knowing the filter property of the learned kernel coefficient, it is still unknown if ResNet performs classification in the spectral domain exclusively, or looks at other aspects. While this analysis only provides a possible explanation of the decoding process, further investigation is needed to gain a full understanding.

VI. Conclusion

In this paper, a comparative study of CNN-based deep learning methods was conducted on SEEG signals for the first time. Five types of movements were classified, using both machine learning and deep learning methods, based on SEEG recordings from 30 participants with intractable epilepsy. This work demonstrated the feasibility of using deep learning on SEEG data. Compared with other methods, ResNet achieved the best decoding accuracy, while the STSCNN demonstrated that the spatial-temporal-spatial convolutional layers configuration is better than the temporal-spatial configuration. Further, various experiments were conducted to better understand this so-called 'black box' method from the spectral domain, which might be useful to reveal possible neuroscientific meanings, such as a biomarker of a specific frequency band. Finally, it is demonstrated that the decoding accuracy variation was positively related to the spectral response in both low and high-frequency ranges. In conclusion, this paper showed that deep learning methods have a high potential for SEEG signal decoding and the decoding process can be partially understood from spatial and spectral perspectives.

References

[1] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan, "Brain–computer interfaces for communication and control," *Clin Neurophysiol*, vol. 113, no. 6, pp. 767–791.

[2] G. Schalk, J. Kubanek, K. J. Miller, N. R. Anderson, E. C. Leuthardt, J. G. Ojemann, D. Limbrick, D. Moran, L. A. Gerhardt, and J. R. Wolpaw, "Decoding two-dimensional movement trajectories using electrocorticographic signals in humans," *J Neural Eng*, vol. 4, no. 3, pp. 264–75.

[3] A. L. Orsborn, C. Wang, K. Chiang, M. M. Maharbiz, J. Viventi, and B. Pesaran, "Semi-chronic chamber system for simultaneous subdural electrocorticography, local field potentials, and spike recordings," in *2015 7th International IEEE/EMBS Conference on Neural Engineering (NER)*, pp. 398–401.

[4] G. Li, S. Jiang, S. Paraskevopoulou, G. Chai, Z. Wei, S. Liu, M. Wang, Y. Xu, Z. Fan, Z. Wu, L. Chen, D. Zhang, and X. Zhu, "Detection of human white matter activation and evaluation of its function in movement decoding using stereo-electroencephalography (SEEG)," *J of Neural Eng*.

[5] M. Wang, G. Li, S. Jiang, Z. Wei, J. Hu, L. Chen, and D. Zhang, "Enhancing gesture decoding performance using signals from posterior parietal cortex: a stereo-electroencephalograhy (SEEG) study," *J Neural Eng*, vol. 17, no. 4, p. 046043.

[6] B. A. Murphy, J. P. Miller, K. Gunalan, and A. B. Ajiboye, "Contributions of subsurface cortical modulations to discrimination of executed and imagined grasp forces through stereoelectroencephalography," *PLoS One*, vol. 11, no. 3.

[7] H. Tan, A. Pogosyan, A. Anzak, K. Ashkan, M. Bogdanovic, A. L. Green, T. Aziz, T. Foltynie, P. Limousin, L. Zrinzo, and P. Brown, "Complementary roles of different oscillatory activities in the subthalamic nucleus in coding motor effort in Parkinsonism," *Exp Neurol*, vol. 248, pp. 187–95.

[8] S. Vadera, A. R. Marathe, J. Gonzalez-Martinez, and D. M. Taylor, "Stereoelectroencephalography for continuous two-dimensional cursor control in a brain-machine interface," *Neurosurg Focus*, vol. 34, no. 6, p. E3.

[9] 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," pp. 375–378, 2017 8th International IEEE/EMBS Conference on Neural Engineering (NER).

[10] X. Wu, G. Li, S. Jiang, S. Wellington, S. Liu, Z. Wu, B. Metcalfe, L. Chen, and D. Zhang, "Decoding continuous kinetic information of grasp from stereo-electroencephalographic (SEEG) recordings," *J of Neural Eng*, vol. 19, no. 2, p. 026047.

[11] H. Akbari, B. Khalighinejad, J. L. Herrero, A. D. Mehta, and N. Mesgarani, "Towards reconstructing intelligible speech from the human auditory cortex," *Scientific Reports*, vol. 9, no. 1, p. 874.

[12] Z. Li, J. E. O'Doherty, T. L. Hanson, M. A. Lebedev, C. S. Henriquez, and M. A. Nicolelis, "Unscented kalman filter for brain-machine interfaces," *PLoS One*, vol. 4, no. 7, p. e6243.

[13] K. K. Ang, "Filter bank common spatial pattern (FBCSP) in brain-computer interface," pp. 2390–2397, 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence).

[14] S. Liu, G. Li, S. Jiang, X. Wu, J. Hu, D. Zhang, and L. Chen, "Investigating data cleaning methods to improve performance of brain–computer interfaces based on stereo-electroencephalography," *Front Neurosci*, vol. 15, 2021.

[15] A. Subasi and M. Ismail Gursoy, "EEG signal classification using PCA, ICA, LDA and support vector machines," *Expert Syst Appl*, vol. 37, no. 12, pp. 8659–8666.

[16] L. J. Cao, K. S. Chua, W. K. Chong, H. P. Lee, and Q. M. Gu, "A comparison of PCA, KPCA and ICA for dimensionality reduction in support vector machine," *Neurocomputing*, vol. 55, no. 1, pp. 321–336.

[17] B. Richhariya and M. Tanveer, "EEG signal classification using universum support vector machine," *Expert Syst Appl*, vol. 106, pp. 169–182, 2018.

[18] E. Neto, F. Biessmann, H. Aurlien, H. Nordby, and T. Eichele, "Regularized linear discriminant analysis of eeg features in dementia patients," *Front in Aging Neurosci*, vol. 8, 2016.

[19] Y. Ding, N. Robinson, Q. Zeng, D. Chen, A. a. Phyo wai, T.-S. Lee, and C. Guan, "TSception:a deep learning framework for emotion detection using EEG," *2020 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–7, 07 2020.

[20] 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, p. 056013.

[21] K. H. Cheah, H. Nisar, V. V. Yap, C.-Y. Lee, and G. R. Sinha, "Optimizing residual networks and VGG for classification of EEG signals: Identifying ideal channels for emotion recognition," *J of Healthcare Eng*, vol. 2021, p. e5599615.

[22] M. J. Hasan, D. Shon, K. Im, H.-K. Choi, D.-S. Yoo, and J.-M. Kim, "Sleep state classification using power spectral density and residual neural network with multichannel EEG signals," *Appl Sci*, vol. 10, no. 21, p. 7639.

[23] S. Bagherzadeh, K. Maghooli, A. Shalbaf, and A. Maghsoudi, "Emotion recognition using effective connectivity and pre-trained convolutional neural networks in EEG signals," *Cogn Neurodynamics*, vol. 106, pp. 169–182.

[24] R. T. Schirrmeister, J. T. Springenberg, L. D. J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, and T. Ball, "Deep learning with convolutional neural networks for EEG decoding and visualization," *Human Brain Mapping*, vol. 38, no. 11, pp. 5391–5420.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778.

[26] G. Li, S. Jiang, S. E. Paraskevopoulou, M. Wang, Y. Xu, Z. Wu, L. Chen, D. Zhang, and G. Schalk, "Optimal referencing for stereo-electroencephalographic (SEEG) recordings," *Neuroimage*, vol. 183, pp. 327–335.

[27] B. Fischl, "FreeSurfer," *Neuroimage*, vol. 62, no. 2, pp. 774–781, 2012.

[28] G. Li, S. Jiang, C. Chen, P. Brunner, Z. Wu, G. Schalk, L. Chen, and D. Zhang, "iEEGview: an open-source multifunction GUI-based Matlab toolbox for localization and visualization of human intracranial electrodes," *J Neural Eng*, vol. 17, no. 1, p. 016016, 2019.

[29] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors, "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python," *Nature Methods*, vol. 17, pp. 261–272, 2020.

[30] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *J Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[31] R. T. Schirrmeister, J. T. Springenberg, L. D. J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, and T. Ball, "Deep learning with convolutional neural networks for EEG decoding and visualization," *Human Brain Mapping*, vol. 38, no. 11, pp. 5391–5420.

[32] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *arXiv e-prints*, p. arXiv:1412.6980, Dec. 2014.

[33] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," *International Journal of Computer Vision*, vol. 128, no. 2, pp. 336–359.

[34] A. Gramfort, M. Luessi, E. Larson, D. Engemann, D. Strohmeier, C. Brodbeck, R. Goj, M. Jas, T. Brooks, L. Parkkonen, and M. Hämäläinen, "MEG and EEG data analysis with MNE-python," *Front in Neurosci*, vol. 7, p. 267.

[35] L. Rokach, "Ensemble-based classifiers," *Artif Intell Review*, vol. 33, no. 1, pp. 1–39.

[36] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *3rd International Conference on Learning Representations, ICLR 2015*, 2015.

[37] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–9, 2015.

[38] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Training Very Deep Networks," *arXiv e-prints*, p. arXiv:1507.06228, Jul. 2015.

[39] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780.

[40] A. Veit, M. Wilber, and S. Belongie, "Residual networks behave like ensembles of relatively shallow networks," *Proceedings of the 30th International Conference on Neural Information Processing Systems*, p. 550–558, 2016.