研究論文

畳み込みニューラルネットワークに基づくうつ病の弁別†

万志江*、鍾寧**

A Convolutional Neural Network for Depression Discrimination †

Zhijiang Wan*, Ning Zhong**

Electroencephalogram (EEG) measurement, being an appropriate approach to understanding the underlying mechanisms of the major depressive disorder (MDD), is used to discriminate between depressive and normal control. With the advancement of deep learning methods, many studies have designed deep learning models to improve the classification accuracy of depression discrimination. However, few of them have focused on designing a convolutional filter to learn features according to EEG activity characteristics. In this study, a novel convolutional neural network named HybridEEGNet that is composed of two parallel lines is proposed to learn the synchronous and regional EEG features, and further differentiate normal controls from medicated and unmedicated MDD patients. A ten-fold cross validation method is used to train and test the model. The results show that HybridEEGNet achieves a sensitivity of 68.78%, a specificity of 84.45%, and an accuracy of 79.08% in three-category classification.

Key words: convolutional neural network, depression discrimination, EEG

1 Introduction

Major depressive disorder (MDD, also known as unipolar depression) is widely distributed in populations worldwide and is one of the leading causes of disability in both adolescents and adults. According to the World Health Organization's statistics, over 300 million individuals suffer from depression worldwide, and approximately 800,000 people die due to it every year [1-4]. An accurate diagnosis of depression in an early stage is critical and beneficial for depressed people who need to receive clinical treatment in time.

Based on the various physiological measurement tools, such as functional magnetic resonance imaging (fMRI), electroencephalogram (EEG) and positron emission tomography (PET), many studies have tried to measure the psychological data and develop an adjunctive diagnostic approach in clinical practice [5-9]. One of the measurement tools, namely, quantitative measurement of a brain's electrical signals taken from the EEG, is a neuroimaging technique with clear practical advantages because it does not involve invasive procedures, is easy to administer, is tolerated

† 原稿受理 令和２年２月２８日 Received February 28, 2020
* 前橋工科大学 客員研究員
** 生命情報学科 (Department of Life Science and Informatics)

well, and has a relatively low cost. Furthermore, the pervasive and persistent nature of depressive symptoms has made scalp-recorded EEG an appropriate approach to understanding the underlying mechanisms of the major depressive disorder. To this end, many studies proposed various EEG data-based methods for depression discrimination in recent years [7-12]. For instance, their findings showed that low-frequency bands, such as alpha and beta, are promising candidate biomarkers. This suggested that the resting-state EEG might conceal a biomarker for pathophysiology in neurodevelopmental disorders.

Recent advances in EEG acquisition and processing for discriminating depression have been paralleled by the increased availability of machine learning methods [13-16]. Despite their promise as a supplementary computer-aided diagnostic approach to studying depression, these analytic methods are semiautomatic because their methods require manual feature extraction and selection that are time-consuming and labor-intensive. Few studies used the raw EEG data as the model input directly for classifying depressive patients and healthy controls. In other words, a classification method that can directly learn from raw EEG data and automatically extract EEG features is more suitable for constructing an automated EEG analysis method for depression discrimination. Deep learning is such kind of machine learning method that is commonly used in many EEG data-based classification scenarios. As one of the deep learning methods, the convolutional neural network (CNN) is able to directly learn EEG features from raw data, and does not require a handcrafted set of features for classification [17-25]. For example, Acharya and colleagues presented the first application of CNN-based depression discrimination [24]. A novel CNN model named DeepConvNet was proposed to decode imagined or executed movements from raw EEG [18]. EEGNet introduced a compact convolutional neural network for EEG-based brain-computer interfaces (BCIs) used depthwise and separable convolutions to construct an EEG-specific model [25].

Although many studies have used CNNs to perform EEG-based classification tasks including depression discrimination, few of them designed a convolutional filter to learn EEG features according to the EEG activity pattern [23-25]. All of the existing methods mix EEG data of multiple channels into a row, which results in the following layers being unable to learn the spatial distribution characteristics of multichannel EEG data, and a partial loss of the synchronous characteristics of multiple brain regions.

As EEG activity characteristics always reflect the summation of the synchronous activity pattern over a network including several brain regions with similar spatial orientations, those methods might not fully use synchronous EEG characteristics to design the model and perform the classification task. Considering that EEG network irregularity is one of physiological symptoms that could be caused by depression, we have reasons to believe that the summation of the synchronous activity characteristic over a network might contain useful depression-specific information. In addition, since EEG activity possesses regional characteristics originating from different brain regions, regional EEG characteristics extracted from different EEG channels could also be used for depression discrimination. In other words, the synchronous and regional characteristics tend to reflect different aspects of depression-specific information. It is expected that richer and more accurate depression discrimination maybe achieved by fusing the two kinds of characteristics that synthesize the hybrid information. The above observations motivate us to design different

convolutional filters to learn the EEG's synchronous and regional characteristics and construct a CNN model to distinguish depression.

In this study, we propose a novel CNN model named HybridEEGNet to capture more depression-specific information. Specially, there are two kinds of convolutional filters in HybridEEGNet, which are used to simultaneously learn the synchronous EEG characteristic and the regional EEG characteristic. We also evaluate the proposed HybridEEGNet model with a three-category classification task of making a distinction among medicated depressive patients, unmedicated depressive patients and normal controls. The results demonstrate that our method constantly outperforms other state-of-the-art approaches for the task.

2 Methods

2・1 Input Data Representation

Inspired by successful architectures in computer vision, the input data of the first layer is represented as a 2D matrix. Every EEG record is a data matrix of size C * D, where C denotes the number of channels, and D represents the length of a record. Then, all records are split up into a sequence of fragments X1, X2, ..., XT. The size of each fragment Xt is C * d, where d is the ratio of D/(number of fragments). Each fragment Xt also has a corresponding category label Lt, and fragments belonging to the same record have the same category label. The one-hot coding strategy for three-category classification is used to represent Lt. Using the above notation, the inputs of the proposed CNN model are a set of time-ordered sequences X_1, X_2, ..., X_T with a set of corresponding labels L_1, L_2, ..., L_T.

2・2 HybridEEGNet Model Construction

Figure 1 shows the architecture of the HybridEEGNet model, in which two parallel lines are designed to run two independent TensorFlow graphs. Specifically, the HybridEEGNet model consists of two independent CNN submodels. Every independent CNN model comprises 8 convolutional layers and 8 max-pooling layers. Four fully connected layers and one softmax layer are shared by the two submodels. We refer to the submodel learning the EEG's synchronous characteristic as the SynEEGNet part, and the submodel learning the EEG's regional characteristic as the RegEEGNet part. Table 1 illustrates the parameters of the HybridEEGNet model. Columns under Layer Size provide information on input and output sizes of each layer, and the input or output of a layer contains the number of neurons equal to channels * data points * feature maps. For each layer of the SynEEGNet part, the input and output sizes are as same as those of the corresponding layer in the RegEEGNet part. To avoid a repetitive representation, we use one column Input Size or Output Size to describe the input size or the output size of the layer in different parts.

For each feature map, we define the direction along the data points as the y-axis, and the direction along the channels as the x-axis. *Filter Size* means the size of the convolutional filter or the max-pooling filter. *Syn* represents the convolutional filter used for learning the synchronous EEG characteristic, and *Reg* represents the convolutional filter used for learning the regional EEG characteristic. $Stride$ indicates how the filter shifts along the x-axis and the y-axis. The input data of the first layer is represented as a 2D matrix that comprises C channels; each channel contains D data points. The convolutional filters shift along the x-axis and the y-axis by one unit each time. The zero-padding method is utilized to pad the input of convolutional layer if the filter does not fit the input.

The max-pooling filters shift along the x-axis and the y-axis by one unit each time and by two units each time, respectively. The specific operations are illustrated as follows:

Convolutional layers. Two kinds of convolutional filters are designed to learn separately the EEG's synchronous and regional characteristics. The former could be regarded as the EEG characteristic of multiple channels, and the convolution operation is split into the first convolution across the data of multiple channels and the second convolution across time. The regional EEG characteristic could be regarded as the EEG characteristic of a single channel, and the convolution operation is split into the first convolution across the data of a single channel and the second convolution across time. To facilitate the following description, we use uppercase and lowercase letters to distinguish the layer operations in the two submodels. If $l = 0$, layer l is the input layer with the input being EEG fragment X_m. Let layer l ($l \geq 1$) be a convolutional layer. Then, the input of layer l comprises m^{l-1} feature maps from the previous layer. The output of layer l consists of m^l feature maps. The i^{th} feature map in layer l of two submodels, denoted Y_i^l and y_i^l, are computed as follows:

$$Y_i^l = f\left(B_i^{(l)} + \sum_{j=1}^{m_1^{(l-1)}} K_{i,j}^{(l)} * Y_i^{(l-1)}\right) \quad (l \geq 1)$$

$$y_i^l = f\left(b_i^{(l)} + \sum_{j=1}^{m_1^{(l-1)}} k_{i,j}^{(l)} * y_i^{(l-1)}\right) \quad (l \geq 1)$$

where $B_i^{(l)}$ and $b_i^{(l)}$ are bias matrices, and $K_{i,j}^{(l)}$ and $k_{i,j}^{(l)}$ are the convolutional filters connecting the jth feature map in layer (l-1) with the ith feature map in layer l. The leaky rectified linear unit (LeakyReLU) is used as the activation function f(•) after the convolution operation. In Figure 1, the convolutional filter is marked by a red rectangle, the size of the filter for learning the EEG's synchronous characteristic is C * 8, the size of the filter for learning the EEG's regional characteristic is 1 * 8. The corresponding convolution result is marked by a black rectangle, its size of 1 * 1.

Pooling layer. Let layer l be a pooling layer. Its output comprises $m_1^l = m_1^{l-1}$ feature maps of reduced size. Max-pooling is used as the downsampling operation. In Figure 1, the max-pooling filter is marked by a green rectangle, and the filter size is 1* 2. The corresponding downsampling operation's result is marked by a black rectangle, its size of 1 * 1.

Concatenation layer. The concatenation layer is also the first fully connected layer. The TensorFlow outputs of the last pooling layer for two submodels are concatenated into one vector, and feed that vector into the first fully connected layer.

Fully connected layer. Let layer l be a fully connected layer but not the first fully connected layer; then, the input of layer l is in the form of m_1^{l-1} feature maps. The identity activation function is utilized as activation function f(•), and the output of the ith unit in layer l is computed as follows:

$$Z_i^{(l)} = g\left(\sum_{j=1}^{m^{(l-1)}} w_{i,j}^{(l)} Z_j^{(l-1)}\right)$$

where $w_{i,j}^l$ and Z_j^{l-1} denote the corresponding weights of the i^{th} unit in layer l and the outputs of layer $(l-1)$, respectively.

Softmax layer. As shown in Figure 1, the last fully connected layer connects with the softmax layer that contains 3 neurons. It is noteworthy that the last fully connected layer also contains 3 neurons and connects with the softmax layer by the one-on-one method. Since the softmax layer corresponds to output classes (normal, medicated depressive patients, and unmedicated depressive patients), the feature matrix

learned by the last fully connected layer could be used to analyze the feature differences among the samples of three categories.

Loss function. The categorical cross-entropy is used as the loss function to compare the probability distribution with the true distribution $\{L_1, L_2, ..., L_T\}$ represented by the one-hot coding strategy. The loss function is computed as follows:

$$\text{Loss} = -\sum_{i=1}^{T}\sum_{j=1}^{M} L_{i,j} * \log(p_{i,j})$$

where T is the number of verification data samples, M is the number of classes, $p_{i,j}$ is the predicted value obtained from the fully connected layer, and $L_{i,j}$ is the true value.

3 Experiment and Results

3・1 Data Collection

All depressive patients were recruited from Beijing Anding Hospital, China. Every patient willing to participate in this project had to meet the inclusion and exclusion criteria specified by a clinician. The normal control group of the experiment was required to have no psychiatric disorders in the past and was also screened by a clinician. Ultimately, 35 subjects were recruited, who included 12 normal controls (6 females and 6 males) aged from 21 to 55 (with mean ± standard deviation (Std.) being 26.4 ± 9.8 years), 12 unmedicated patients (6 females and 6 males) aged from 25 to 54 (with mean ± Std. being 28.6±7.3 years), and 11 medicated patients (6 females and 5 males) aged from 20 to 56 (with mean ± Std. being 29.8±10.6 years).

In the experiment, subjects are asked to record their EEG data in the resting state. Specifically, subjects would sit on a sofa and keep eyes closed for 8 minutes while not intentionally thinking of anything in a dimly illuminated and soundproof room. They are also asked to maintain a minimum arousal level without falling asleep. In the data collection process, we select several representative brain regions from the prefrontal cortex (PFC), the frontal cortex, and the parietal cortex as EEG-collecting locations that were demonstrated to be closely related to depression.

To record multichannel EEG data, six surface electrodes (Fp1, Fp2, F3, F4, P3 and P4) are placed on the scalp according to the 10-20 international electrode system. EEG recordings are acquired using a platform (Brain Products Ltd., Germany) with BrainAmp 16-bit A/D convertor (ADC). The data collection software named BrainVision Analyzer provides a head model for volume conduction properties and supports localization of the signal sources of different EEG channels. This function is used to mitigate the effects of volume conduction on raw EEG data samples before exporting the latter. The down sampling rate of 500 Hz is used to downsample the collected EEG data. The software used for EEG data analysis is written in Python 3.0 configured with TensorFlow.

3・2 Data Preprocessing

The EEG data recorded in one trial is cut into three snippets, and the median snippet with the duration of 5 minutes is kept for analysis. The time snippets of the beginning 30 seconds and the last 2 minutes and 30 seconds are removed. Z-score normalization is used to overcome the amplitude scaling problem and remove the offset effect. For every median snippet, the data record of 5 minutes is fragmented into 50 data samples. Every sample contains 3072 sampling points (covering approximately 6.144 seconds). Based on the channel order of Fp1, Fp2, F3, F4, P3 and P4, the fragments of 6 channels are realigned into a data matrix. Every data matrix is fed into the CNN model as a new independent data sample.

The dataset ultimately used in this study includes

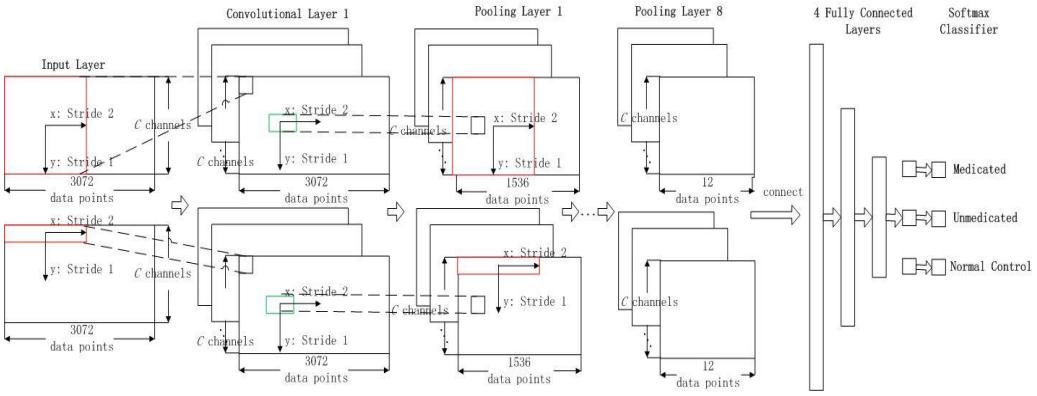

Fig.1 Architecture of the HybridEEGNet model for three-category classification

Table 1 Parameters fixed in each layer of the HybridEEGNet model

Layers	Type	Input size	Output size	Filter size		Stride	
				'syn'	'sin'	X	y
0	Input	6*3072	6*3072	-	-	-	-
1	Convolution	6*3072	6*3072*6	6*8	1*8	1	1
2	Max-pooling	6*3072*6	6*1536*6	1*2	1*2	1	2
3	Convolution	6*1536*6	6*1536*6	6*8	1*8	1	1
4	Max-pooling	6*1536*6	6*768*6	1*2	1*2	1	2
5	Convolution	6*768*6	6*768*6	6*8	1*8	1	1
6	Max-pooling	6*768*6	6*384*6	1*2	1*2	1	2
7	Convolution	6*384*6	6*384*6	6*8	1*8	1	1
8	Max-pooling	6*384*6	6*192*6	1*2	1*2	1	2
9	Convolution	6*192*6	6*192*12	6*8	1*8	1	1
10	Max-pooling	6*192*12	6*96*12	1*2	1*2	1	2
11	Convolution	6*96*12	6*96*12	6*8	1*8	1	1
12	Max-pooling	6*96*12	6*48*12	1*2	1*2	1	2
13	Convolution	6*48*12	6*48*12	6*8	1*8	1	1
14	Max-pooling	6*48*12	6*24*12	1*2	1*2	1	2
15	Convolution	6*24*12	6*24*12	6*8	1*8	1	1
16	Max-pooling	6*24*12	6*12*12	1*2	1*2	1	2
17	Concatenation	6*12*12*12	32	-	-	-	-
18	Full connection	32	16	-	-	-	-
19	Full connection	16	3	-	-	-	-
20	Full connection	3	3	-	-	-	-
21	Softmax	3	3	-	-	-	-

a total of 1750 data samples (consisting of 600 data samples of normal control, 600 data samples of unmedicated depressive patients, and 550 data samples of medicated depressive patients).

3・3 Comparison Baselines

Five kinds of CNN models are constructed as baseline approaches for result comparison. The simplified descriptions of model features of every baseline approach are as follows:

SynEEGNet. Compared with HybridEEGNet, we only keep the submodel part that learns the EEG's synchronous characteristic and remove the concatenation operation in the first fully connected layer. The last three fully connected layers, the softmax layer, and the model input are the same as those of HybridEEGNet.

RegEEGNet. Similarly to the construction of the SynEEGNet model, we only keep the submodel part that learns the EEG's regional characteristic and remove the concatenation operation in the first fully connected layer. The last three fully connected layers, the softmax layer, and the model input are also the same as in HybridEEGNet.

DeepConvNet [18]. The model uses a convolutional filter to mix the EEG data of multiple channels into a row, which might make it unable to fully use the spatial distribution characteristics of EEG activities originating from multiple brain regions.

AchCNN [24]. We refer to the CNN model constructed by Acharya et al. as AchCNN because we did not find any specific model name in the paper. Since the researchers demonstrated the advantage of CNN-based depression discrimination by comparing results of their model with those of several traditional methods (SVM, logistic regression, bagged tree, etc.), it is a valuable comparison baseline in this paper.

EEGNet [25]. An operation similar to that of [18] is used by EEGNet to merge the EEG data of multiple channels into one row as the output of the first convolutional layer.

For the baselines, namely, the SynEEGNet and RegEEGNet models that are bases of our model, the respective hyperparameters are tuned until we obtained the optimal classification accuracy. Afterwards, for a fair comparison, the HybridEEGNet used those parameters. For DeepConvNet and EEGNet, as their inputs and outputs are different from those of our model, we did not directly use their publicly released software implementations. Instead, we referred to the latter and subsequently built them and tuned the hyperparameters the same way as in [18]. For AchCNN, we reproduced the CNN network and tuned the hyperparameters the same way as in [24].

3・4 Evaluation Metrics

Evaluation metrics including sensitivity (*Sen*), specificity (*Spc*) and recognition accuracy (*Acc*) are used to evaluate the classification performance of models. Sensitivity refers to the ability of a classifier to correctly detect positive samples. Specificity refers to the ability of a classifier to correctly detect negative samples. Recognition accuracy refers to the ability of a classifier to correctly detect the samples with different labels. The above metrics are calculated using the following formulas:

$$\mathrm{SEN} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN}),$$

$$\mathrm{SPC} = \mathrm{TN}/(\mathrm{TN} + \mathrm{FP}),$$

$$\mathrm{ACC} = (\mathrm{TP} + \mathrm{TN})/(\mathrm{TP} + \mathrm{FP} + \mathrm{FN} + \mathrm{TN}).$$

where TP means true positives, TN denotes true negatives, FP represents false positives, and FN corresponds to false negatives. For three-category classification, the one-against-all approach is utilized to calculate the evaluation metrics. In other words, we take turns at using one of the three categories as the positive label and the remaining two categories as the

negative label. The results of 3 evaluations are averaged for estimating the model performance.

3・5 Parameter Settings

The tenfold cross-validation method is used to validate the classification performance of each model. In the training phase of each fold, all variables are initialized with random values following Gaussian distributions and trained for 1000 epochs; the batch size of every epoch is 300 data fragments. In each epoch, a random resampling strategy for selecting the training data is used to avoid the model performance decrease caused by the sample disequilibrium. A total of 100 medicated depressive patient data fragments, 100 unmedicated depressive patient data fragments, and 100 normal control data fragments are used to train the model in every iteration. The other parameters are momentum of 0.9, weight decay of 0.0005, the (base) learning rate of 0.001, and dropout rate of 0.9. A regularization term is added into the loss function of the proposed model so as to avoid overfitting. In the testing phase of each fold, 60 data samples of normal control, 60 data samples of unmedicated depressive patients, and 55 data samples of medicated depressive patients are utilized as the testing sample. The metrics' values resulting from 10 folds are averaged for estimating the model performance to prevent any model from obtaining good results by chance.

4 Results

Table 2 compares the classification results of HybridEEGNet and baseline approaches using the confusion matrix and evaluation metrics. In the table, *MD*, *UnMD* and *NC* represent the medicated depressive patient, unmedicated depressive patient and normal control categories. The *Positive* column shows that the samples of each category are used as positive samples in turn to calculate the evaluation metrics given in columns *Sen* (%), *Spe* (%) and *Acc* (%).

For each model, the results of average evaluation metrics are also given. From Table 2, we can see that HybridEEGNet achieves the best performance. The average sensitivity, specificity and accuracy are 68.78%, 84.45% and 79.08%, respectively. This result demonstrates that the HybridEEGNet model that merges the EEG's synchronous and regional characteristics is more suitable for distinguishing depressive patients than other kinds of CNN architectures. Examining the confusion matrix, we observe that in the condition of considering the unmedicated depressive patient samples as the positive samples, HybridEEGNet attains the sensitivity of 58.83%, i.e., it does not distinguish well the unmedicated depressive patient samples from the samples of the other two categories. The sensitivity results of other baseline models are also lower than in the conditions of considering medicated depressive patient or normal control samples as the positive samples. This might be caused by the EEG activity state of unmedicated depressive patients being a median state between a medicated state and the normal state.

Additionally, the table shows that most EEG samples of unmedicated depressive patient are classified as EEG samples of medicated depressive patient. This finding indicates that EEG samples of unmedicated depressive patients and medicated depressive patients have common EEG characteristics even though medicated depressive patients received medical treatment.

The classification performance of HybridEEGNet is compared with that of baseline models from the perspective of model structure. On the one hand, compared with SynEEGNet and RegEEGNet, HybridEEGNet integrates the feature extraction part

of the two models, and attains a higher classification performance, which demonstrates that the integration of the synchronous and regional EEG characteristics improves depression discrimination ability of the CNN model. On the other hand, compared with models that ignore synchronous EEG characteristics, HybridEEGNet also attains a better classification performance, which indicates the significance of learning spatial distribution characteristics of EEG activity generated by multiple brain regions for the task of depression discrimination.

Table 2 Comparison results of HybridEEGNet and baselines using the confusion matrix and evaluation metrics

Model		Actual			Evaluation metrics			
	Predicted	MD	UnMD	NC	Positive	Sen (%)	Spe (%)	ACC (%)
HybridNet	MD	40.8	18.6	12.1	MD	74.18	74.41	74.34
	UnMD	5.7	35.3	3.9	UnMD	58.83	91.65	80.4
	NC	8.5	6.1	44	NC	73.33	87.3	82.51
					average	68.78	84.45	79.08
SynEEGNet	MD	37.1	23.5	5.5	MD	67.45	75.83	73.2
	UnMD	5	27.1	4.2	UnMD	45.16	92	75.94
	NC	12.9	9.4	50.3	NC	83.83	80.6	81.71
					average	65.48	82.81	76.95
RegEEGNet	MD	38.1	29.1	13.8	MD	69.27	64.25	65.82
	UnMD	5.4	22.4	6.5	UnMD	37.33	89.65	71.71
	NC	11.5	8.5	39.7	NC	66.16	82.6	76.97
					average	57.5	78.83	71.5
DeepConvNet	MD	38.3	45	9.3	MD	69.63	54.75	59.42
	UnMD	2	4.1	1.6	UnMD	6.83	96.86	66
	NC	14.7	10.9	49.1	NC	81.83	77.73	79.14
					average	52.76	76.45	68.19
AchCNN	MD	30.3	37	12.1	MD	55.09	59.08	57.82
	UnMD	6	8.1	5.6	UnMD	13.5	89.91	63.71
	NC	18.7	14.9	42.3	NC	70.5	70.78	70.68
					average	46.36	73.25	64.07
EEGNet	MD	41.1	23.1	10.8	MD	74.72	71.75	72.68
	UnMD	9.4	25.4	7.5	UnMD	42.33	85.3	70.57
	NC	4.5	11.5	41.7	NC	69.5	86.08	80.4
					average	62.18	81.04	74.55

5 Discussion

This study is the first attempt to utilize a CNN model for differentiating normal controls from medicated and unmedicated depressive patients, and furthermore to analyze and compare the differences between the features learned by the convolutional layer or the fully connected layer using the deep-dreaming algorithm. In other words, the primary focus in this paper is mainly on two aspects: HybridEEGNet model construction and feature analysis.

In the constructed HybridEEGNet model, the synchronous EEG characteristic and the regional EEG characteristic are learned by different filters, and the fusion feature is used to distinguish depression. Although the results show that the spatially global voltage patterns contain effective depression-specific information, the filter for learning the synchronous EEG pattern did not strictly follow the approach of learning such global patterns. Specifically, The filter of size 6 * 8 is used to learn the synchronous characteristic in the convolutional layer. The first row of the feature matrix of a convolutional layer could be regarded as the joint result of processing the EEG data of 6 channels. However, with the movement of the filter in the channel direction, the zero-padding method is utilized to pad the input of a convolutional layer if the filter does not fit the input, i.e., the remaining rows of the feature matrix could only be regarded as the joint result of EEG data and zeros. One alternative way to improve this is by designing different filters with different sizes. Specifically, a filter of size 5 * 8 could be designed to process the input data of five EEG channels in a convolutional layer, and a filter of size 4 * 8 could be used to process the input data of four EEG channels in a convolutional layer. Furthermore, the characteristics of functional brain networks in depressive patients have been investigated by many studies that examined the resting-state scalp EEG data, i.e., the EEG activity in functional brain networks is affected by depression [26-29]. In this context, different convolutional filters could be used to process the EEG data of different functional brain networks. The inception network [30] that uses different filters to extract features and fuses them to obtain more abstract and effective features seems to be appropriate for this task. Although our results have similarities with the findings of previous studies, these results are still in the experimental stage. In future research, it would be more appropriate to use a dataset with a larger scale and more EEG channels for validating the reliability of our results. Additionally, those results could be used in reverse to discriminate normal controls from medicated and unmedicated depressive patients. Specifically, we can go back to a traditional approach and extract frequency power features from the raw EEG data based on the above findings.

Exploiting the spatial distribution and amplitude range difference of the EEG rhythm between depressive patients and normal controls, other classifiers could be used to make our results more interpretable and provide a neurobiological interpretation of the relationship between the findings and the neuropathology of the depressive disorder.

References

1) R. Belmaker, and G. Agam, *Major Depressive Disorder*. New England Journal of Medicine, vol. 358, no. 1, pp. 55-68, 2008.

2) J. Olesen, A. Gustavsson, M. Svensson et al., *The economic cost of brain disorders in Europe*. European Journal of Neurology, vol. 19, no. 1, p. 155, 2012.

3) H. Whiteford, L. Degenhardt, J. Rehm et al., *Global burden of disease attributable to mental and*

substance use disorders: findings from the Global Burden of Disease Study 2010. The Lancet, vol. 382, no. 9904, pp. 1575-1586, 2013.

4) C. Mathers, and D. Loncar, *Projections of global mortality and burden of disease from 2002 to 2030*. Plos Medicine, vol. 3, no. 11, p. e442, 2006.

5) R. Kerestes, C. Davey, K. Stephanou et al., *Functional brain imaging studies of youth depression: A systematic review*. NeuroImage : Clinical, vol. 4, pp.209-231, 2014.

6) O. Stelt, and A. Belger, *Application of Electroencephalography to the Study of Cognitive and Brain Functions in Schizophrenia*. Schizophrenia Bulletin, vol. 33, no. 4, pp. 955-970, 2007.

7) S. Olbrich, and M. Arns, *EEG biomarkers in major depressive disorder: Discriminative power and prediction of treatment response*. International Review of Psychiatry, vol. 25, no. 5, pp. 604-618 , 2013.

8) S. Salle, J. Choueiry, D. Shah et al., *Effects of Ketamine on RestingState EEG Activity and Their Relationship to Perceptual/Dissociative Symptoms in Healthy Humans*. Front Pharmacol, vol. 7, p. 348, 2016.

9) C. Michel, and M. Murray, *Towards the utilization of EEG as a brain imaging tool*. NeuroImage, vol. 61, no. 2, p. 371-385, 2012.

10) M.J. Schiller, *Quantitative Electroencephalography in Guiding Treatment of Major Depression*. Front Psychiatry, vol. 9, no. 779, 2019.

11) D. Iosifescu, *Electroencephalography derived biomarkers of antidepressant response*. Harv Rev Psychiatry. vol.19, no. 3, pp.144-54, 2011.

12) A. Baskaran, R. Milev and R.S. McIntyre, *The neurobiology of the EEG biomarker as a predictor of treatment response in depression*. Neuropharmacology. vol.63, no.4, pp.507-513, 2012.

13) D. Kim, A. Bolbecker, J. Howell et al., *Disturbed resting state EEG synchronization in bipolar disorder: A graph-theoretic analysis*. NeuroImage: Clinical, vol. 2, pp. 414-423, 2013.

14) H. Helgadóttir, Ó. Gudmundsson et al., *Electroencephalography as a clinical tool for diagnosing and monitoring attention deficit hyperactivity disorder: a cross-sectional study*. BMJ Open, vol. 5, no. 1, p. e55, 2015.

15) S. Liao, C. Wu, H. Huang et al., *Major Depression Detection from EEG Signals Using Kernel Eigen-Filter-Bank Common Spatial Patterns*. Sensors, vol. 17, no. 6, p. 1385, 2017.

16) S. Freitas, A. Marques, M. Bevilaqua et al., *Electroencephalographic findings in patients with major depressive disorder during cognitive or emotional tasks: a systematic review*. Braz J Psychiatry, vol. 4, no.38, pp. 338-346, 2016.

17) I. Arel, D. Rose, and T. Karnowski, *Deep Machine Learning - A New Frontier in Artificial Intelligence Research*. Computational Intelligence Magazine IEEE, vol. 5, no. 4, pp. 13-18, 2010.

18) R. Schirrmeister, J. Springenberg, L. Fiederer et al., *Deep learning with convolutional neural networks for EEG decoding and visualization*. Human Brain Mapping, vol. 38, no. 11, pp. 5391-5420, 2017.

19) M. Putten, S. Olbrich, and M. Arns, *Predicting sex from brain rhythms with deep learning*. Scientific Reports, vol. 8, no. 1, pp. 1-7, 2018.

20) X. Li, D. Zhang, P. Hou and B. Hu, *Deep fusion of multi-channel neurophysiological signal for emotion recognition and monitoring*. International Journal of Data Mining and Bioinformatics, vol. 18, no. 1, pp. 1-27, 2017.

21) S. Stober, A. Sternin, A. M. Owen et al., *Deep Feature Learning for EEG Recordings*. Computer

Science, vol. 165, pp. 23–31, 2016.

22) F. Morabito, M. Campolo, N. Mammone et al., *Deep Learning Representation from Electroencephalography of Early-Stage Creutzfeldt-Jakob Disease and Features for Differentiation from Rapidly Progressive Dementia*. International Journal of Neural Systems, vol. 27, no. 02, p. 1650039, 2017.

23) D. Wulsin, R. Gupta et al., *Modeling electroencephalography waveforms with semi-supervised deep belief nets: fast classification and anomaly measurement*. Journal of Neural Engineering. vol. 3, no. 08, p.036015, 2011.

24) U. Acharya, S. Oh, Y. Hagiwara et al., *Deep convolutional neural network for the automated detection and diagnosis of seizure using EEG signals*. Comput Biol Med, vol. 100, no. 534, pp. 270-278, 2018.

25) V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, B. J. Lance, EEGNet: *A Compact Convolutional Network for EEG-based Brain-Computer Interfaces*. Journal of Neural Engineering, vol. 15, no. 5, p. 056013, 2018.

26) M. Zhang, H. Zhou, L. Liu, L. Feng, J. Yang, G. Wang, N. Zhong *Randomized EEG functional brain networks in major depressive disorders with greater resilience and lower rich-club coefficient*. Clinical Neurophysiology, vol. 129, no. 4, pp. 743-758, 2018.

27) M. Shim, C. H. Im, Y. W. Kim, S. H. Lee, *Altered cortical functional network in major depressive disorder: A resting-state electroencephalogram study*. NeuroImage: Clinical, vol. 19, pp. 1000-1007, 2018.

28) A. Damborská, M. Tomescu, E. Honzírková, et al., *EEG Resting-State Large-Scale Brain Network Dynamics Are Related to Depressive Symptoms*. Frontiers in Psychiatry, vol. 10, no. 548, 2019.

29) F. Liu, J. Rosenberger, Y. Lou, R. Hosseini, J. Su, S. Wang, *Graph regularized EEG source imaging with in-class consistency and out-class discrimination*. IEEE Transactions on Big Data, vol. 3, no. 4, pp. 378-391, 2017.

30) C. Szegedy et al., *Going deeper with convolutions*. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1-9, 2015.