EEG Deep Feature Learning and Significance Analysis for Alcoholism Classification

Yuhao Zhang

Research School of Computer Science, Australian National University, Canberra, ACT, 2601 Australia
u6829434@anu.edu.au

Abstract. Electroencephalogram (EEG) contains a wealth of human mental information and plays an important role in many applications such as human-computer interaction. However, the raw EEG signals are not quite proper for machine learning tasks. Many feature extraction methods have been proposed for EEG data, but failed to analyse the significance of the features. In this study, we use EEG-based alcoholism classification as an example task to propose a hand-crafted feature extraction method and also three deep learning-based feature extractors to try to get the salient EEG features. The hand-crafted features are based on EEG spectral power. A 2D convolutional neural network (Conv2D) is to extract deep spatial features. Besides, a Long Short-term Memory (LSTM) and a 1D convolutional neural network (Conv1D) are used to extract deep temporal EEG features. We evaluate these features on UCI EEG dataset, and the highest result is 67.31% achieved by Conv2D deep spatial features. The results also show that deep learning features outperform the hand-crafted features. Furthermore,we also perform a feature distinctiveness analysis using Gedeon Method on EEG spectral power features. We find that beta frequency band of EEG signal is the most significant in alcoholism classification task.

Keywords: Deep learning · EEG · Pattern recognition · Feature distinctness analysis.

1 Introduction

Brain-computer interfaces (BCIs) have a wide range of applications in the medical, education and gaming fields. Electroencephalogram (EEG), as an important physiological signal in BCIs, has attracted more attentions from researchers in recent years. EEG is able to reflect brain activities of human beings directly and convey much mental information such as emotion and intention [14, 9]. So, it is widely used in pattern recognition/classification tasks along with various machine learning techniques [11].

However, there are also some limitations of EEG signals when they are utilised to do machine learning tasks: ***(a)*** EEG signal formats vary a lot; they can be quite different according to different recording devices; the channel number and the signal length are also variable. ***(b)*** EEG data contain a huge amount of elements and it is hard to say which of them is more significant; and it is very computational expensive if we use all of EEG signals as input to some machine learning models such as decision tree or SVM. ***(c)*** EEG signals are sequential data along the time, so they have temporal information; and they also have spatial information based on the locations of recording electrodes; so using the raw signals may lose the spatio-temporal information of the EEG data. Due to the limitations above, feature extraction is very necessary for EEG signals. There are many EEG feature extraction methods and they can be categorised into two primary classes: hand-crafted and deep learning-based. A common hand-crafted feature extraction method is to apply Fast Fourier Transform (FFT) to extract frequency domain features such as spectral power. Other statistics such as maximum and mean of signal magnitudes are also used as hand-crafted features. Deep learning, a advanced machine learning technique, has become popular due to its superior performance in various classification problems. It can extract high-level features automatically without any manual processes. So, kinds of deep learning methods [15, 3], such as convolutional neural network (CNN), recurrent neural network (RNN) and transformer, are proposed to extract features from EEG signals. However, some deep learning-based methods fail to remain the spatial and temporal characteristics of EEG signals and do not explore which kind of features are more salient and discriminative.

It is also very interesting and meaningful to find that which feature values are more distinctive in a given feature vector. Garson [4] proposed a input contribution measure based on weight matrix, however it does not consider the neutralisation of positive and negative weights; Gedeon [5] used absolute value of neuron weights to calculate the contribution of an input neuron to an output neuron. It overcame the effect of the weight sign on the contribution and demonstrated its stability during training. Gedeon and Harris [6] proposed a distinctiveness analysis technique for hidden neurons based on the angle of two vectors (one for each hidden neuron) consisting of the activation

for various patterns. This is extended to input analysis by replacing the activation values with normalized neuron weights.

So, this study uses a specific task, alcoholism classification, as an example, to propose kinds of hand-crafted and deep learning based EEG feature learning methods. Firstly, we calculate the mean values of three frequency bands (theta, alpha and beta) as hand-crafted EEG features. Then, we propose three different neural networks (namely 2D CNN, LSTM and 1D CNN) aiming to extract spatial and temporal EEG deep features. We evaluate different EEG features using a 2-linear-layer network classifier on UCI EEG dataset [1]. We explore which EEG features are more significant by comparing the classification performance. Besides, I will apply the Gedeon method [5] to analyse the significance of EEG features and recognise the most important ones in the alcoholism classification.

In summary, the main contributions of our work are:

- We apply kinds of hand-crafted and deep learning-based methods to learn EEG features for alcoholism classification on UCI EEG dataset, and find that image-based spatial features are more salient and discriminative.
- We utilise magnitude and functional measures in Gedeon Method to analyse the significance of the hand-crafted EEG spectral power features, and recognise the most and least distinctive features in alcoholism classification problem.

The rest paper is organised as follows: Section 2 describes the data used in this study and details different feature extraction methods and experiment settings; Section 3 reports the experiment results using different features, and also discusses the significance of the input EEG features; Section 4 concludes the findings of this study and also discusses the future research plan.

2 Method

This section describes the data used in this study and the data pre-processing methods in detail. Besides, three deep learning-based EEG feature extractors are proposed to extract different kinds of EEG features for alcoholism identification. Furthermore, an effective input distinctiveness technique is also applied on EEG spectral power features.

2.1 Dataset & Pre-processing

This study uses a public EEG dataset, namely UCI EEG dataset, to evaluate the proposed feature learning methods. The dataset includes EEG signals of 122 subjects, 77 of which are diagnosed with alcoholism and the other 45 are control subjects. The 64-channel EEG signals of the subjects are recorded at 256 Hz while seeing the pictures from the Snodgrass and Vanderwart picture set. Eahc subject has 120 trials and each trial is one-second length. So, for each trial, the EEG data has the shape of 64×256. All the trials of a subject share the same label which can be 1 for alcoholism and 0 for control.

Since theta (4-7 Hz), alpha (8-13 Hz) and beta (13-30 Hz) frequency bands contain the most salient information of EEG when people are awake [1], the mean values for these three frequency bands of each electrode are calculated as the hand-crafted features. Thus we have a feature vector with the size of $64 \times 3 = 192$ for each trial. We will utilise these hand-crafted EEG spectral power features to classify whether a subject is alcoholism or control, and take the result as a baseline to compare with other deep learning features.

EEG signal has a significant inter-subject variability, which means they vary a lot across subjects [7]. As Figure1 shows, the maximum, minimum and mean of feature value of different subjects are quite different and distributed a lot demonstrating the inter-subject variability of EEG signal. The inter-subject variability of EEG data makes it hard to classify alcoholism across subjects. However, cross-subject prediction is much more meaningful in real life because we usually collect EEG signals of some subjects for training the model to predict other fresh subjects. So, in order to evaluate the generalisation ability of our networks, I run the cross-subject experiments using 11 subject rotating test set proposed in [10].

2.2 EEG Deep Feature Learning

EEG Image-based Feature Leaning using 2D CNN Multi-channel EEG signals are obtained from different electrodes which are located at different positions. So, the 64-channel EEG signals used in this study have spatial information. When I treat the EEG feature as a 1D vector, the spatial information of EEG signals will be lost. So,

[1] https://kdd.ics.uci.edu/databases/eeg/eeg.data.html

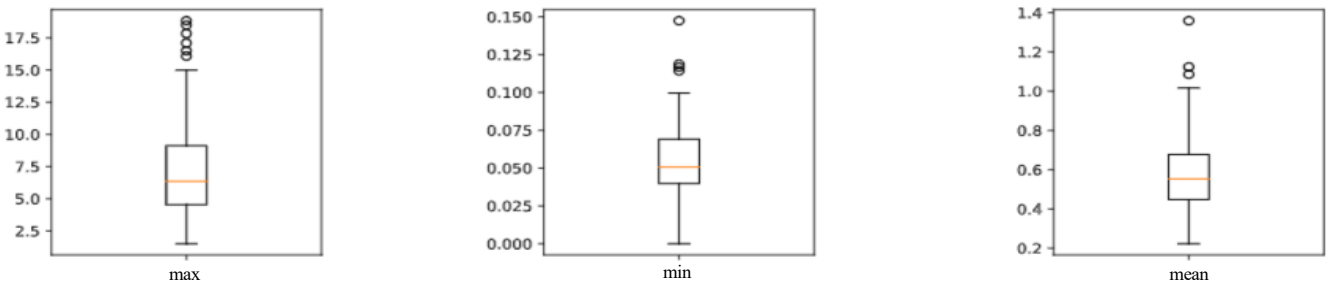
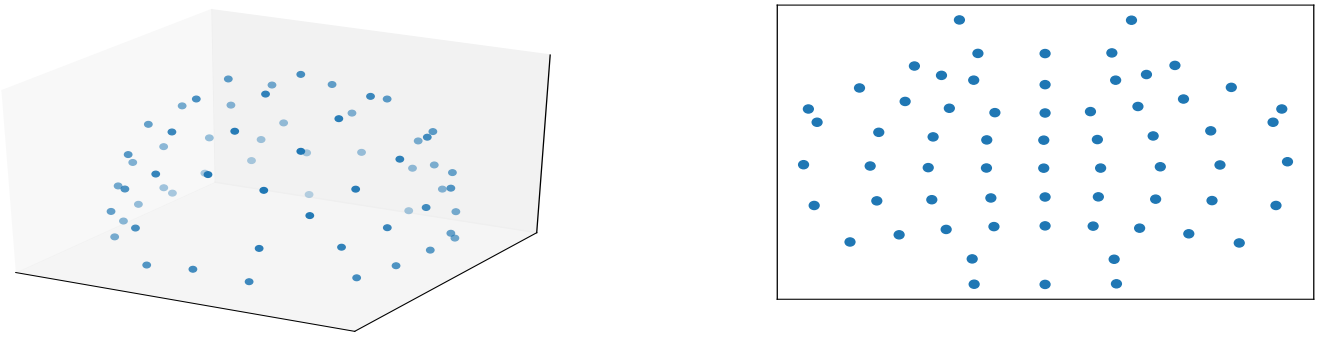


Fig. 1. The distribution of maximums, minimums and means of feature value from different subjects in UCI dataset

in this study, I transform 64-channel EEG signals to 2D images based on the electrode locations, and apply CNNs to extract high-level spatial EEG features.

The 64 EEG electrodes are located over the scalp as a sphere in a 3D space, I use Azimuthal Equidistant Projection (AEP) to project them to a 2D space (shown as Figure 2). Then, the electrode mean values of theta, alpha and beta frequency bands form RGB channels of the image, respectively. The generated images are 32×32, and Clough-Toucher scheme [2] is applied to estimate and interpolate the values between electrodes. Finally, three different frequency bands EEG features construct RGB color images.



(a) EEG electrode locations in the 3D space

(b) EEG electrode locations in the 2D space

Fig. 2. EEG electrode locations from 3D space to 2D space

I adopt a 3-layer CNN to extract spatial features from EEG images (shown as Figure 3). The inputs are EEG images with the shape of $b \times 3 \times 32 \times 32$. The kernel number of three convolutional layers are 4, 8, 16 respectively with the same kernel size of 3×3. The first two convolutional layers are followed by maxpooling layers with the kernel size of 2×2. Hence, the outputs of the convolutional layers have the shape of $b \times 16 \times 8 \times 8$. Then they are flattened to $b \times 1024$ and flow into the classifier later.

EEG Temporal Feature Learning using LSTM EEG signal is a kind of sequential data varying along time. Every signal magnitude in each time step is related to the previous and the afterward time steps, that means the trend how EEG signal changes contains a wealth of information. However, image-based CNN method focuses on spatial feature extraction but ignores the temporal feature of EEG data. Recurrent Neural Networks (RNNs) performs well in dealing with sequential data, so we apply Long Short-term Memory (LSTM), a variant of RNNs, to extract temporal features from EEG signals. The LSTM uses four gates to control different information: Forget gate (f) controls the output of the last time step, and aims to remain the important information and 'forget' the unnecessary information; input gate (i) controls the input of the current time step; cell gate (g) controls the information that flows into the memory cell; output gate (o) controls the output of current memory cell. The detailed equations are shown below, where $\odot$ denotes the Hadamard product, x_t denotes the input at the t time step. The gate design allows LSTM to have the memory of previous input and capture the temporal information from sequential data.

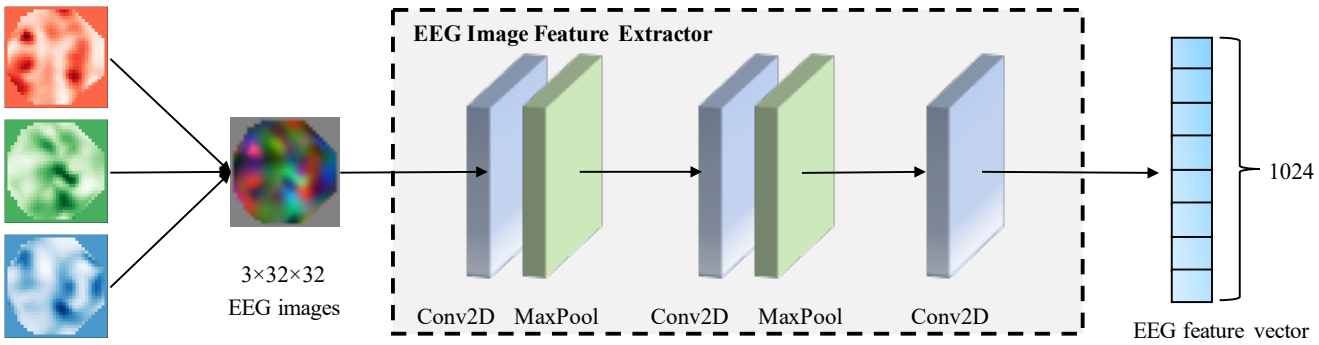


Fig. 3. The architecture of 2D convolutional neural network EEG image-based feature extractor for alcoholism classification

$$
\begin{aligned}
i_t &= \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i) \\
f_t &= \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f) \\
g_t &= \tanh\left(W_{xg}x_t + W_{hg}h_{t-1} + b_g\right) \\
o_t &= \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o) \\
c_t &= f_t \odot c_{t-1} + i_t \odot g_t \\
h_t &= o_t \odot \tanh(c_t)
\end{aligned}
\tag{1}
$$

In this paper, we use one-layer LSTM as EEG temporal feature extractor $\mathcal{E}_{\mathcal{L}}$. The input is the original 256Hz 64-channel EEG data with the size of $b \times 256 \times 64$, where b denotes batch size. The hidden size of LSTM is set as 64, and we use the output of the last hidden layer as the EEG temporal feature with the size of $b \times 64$.

EEG Temporal Feature Learning using 1D CNN In recent years, researches have adopted 1D convolutional neural networks (Conv1D) to deal with sequential data (such as text, audio and physiological signals including EEG) and achieved comparable performance[13]. Conv1D uses the 1D convolutional kernel with different sizes to aggregate the information of consecutive time steps so that it can also extract temporal features from EEG signals. The transition of a Conv1D layer is shown as Equation 2, where out_j and in_k denote the jth channel of output and the kth channel of input; w and b denote weights and bias while the subscripts represent the channel index; $\star$ denotes 1D convolutional operator.

$$
\text{out}_j = \left(\sum_{k=0}^{C_{\text{in}-1}} w_{j,k} \star \text{in}_k\right) + b_j \tag{2}
$$

In this study, we use a 3-layer Conv1D network as EEG temporal feature extractor $\mathcal{E}_{\mathcal{C}}$ (shown as Figure 4). The input is also the original 256Hz 64-channel EEG signals with the size of $b \times 64 \times 256$. The kernel sizes of three Conv1D layers are 5, 3, 3 respectively. We choose to decrease output channel numbers for CNN layers 24 16 and 8, to reduce the feature size and computation complexity. So, the output of convolutional layers is in the shape of $b \times 8 \times 248$. Then it is flattened into a $b \times 1984$ feature vector.

2.3 Neural Network Classifier

After feature extraction, we got EEG deep features from different DNN feature extractors. In this study, we use a 2-linear-layer network as alcoholism classifier to accept these features and predict alcoholism label separately. A dropout with 0.5 is applied after the first linear layer to avoid overfitting. The output size of the first layer, namely the number of hidden neurons, is λ regarded as a hyper-parameter of the model. The first layer is followed by a ReLu activator and the second layer is followed by a Sigmoid activator.

The classifier takes kinds of EEG features (including hand-crafted and deep learning features) and outputs a predicted score s. In the training phase, we use binary cross entropy as loss function and try to minimise it. In the testing phase, we can infer whether the subject is alcoholism or in control by the following formulation.

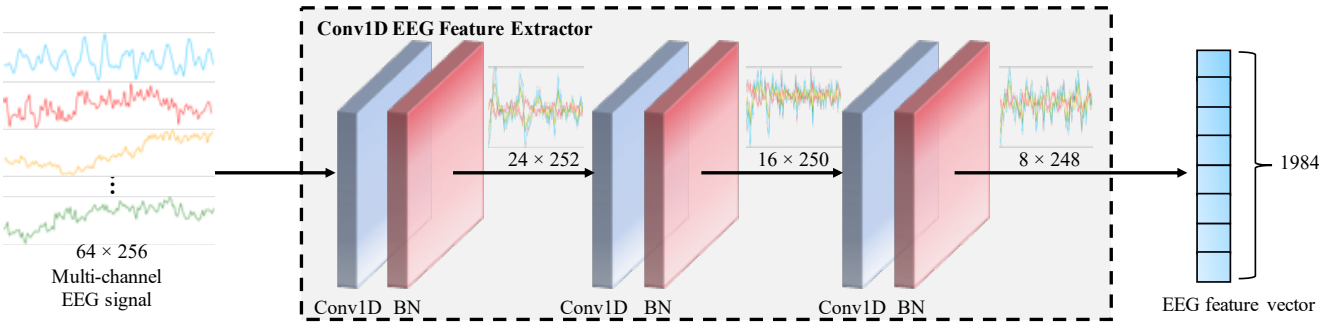


Fig. 4. The architecture of 1D convolutional neural network EEG feature extractor for alcoholism classification

$$\hat{\mathbf{y}} = \begin{cases} Alcoholism, & s > 0.5 \\ Control, & s \leq 0.5 \end{cases} \tag{3}$$

2.4 Validation Strategy

In this study, there are two hyper-parameters in the models, namely the hidden layer size λ and learn rate γ. Here hidden layer size λ refers to the number of neurons in the hidden layer of the classifier. In order to get improve the model generalisation ability, I run validation experiments for fine-tuning. I choose {8, 16, 32, 64, 128 } and {10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}} as candidates for λ and γ, respectively, so there are 25 hyper-parameter combination. A grid search is applied to find the optimal hyper-parameter combination with the highest mean accuracy which are $\lambda = 16, \gamma = 10^{-4}$.

2.5 EEG Feature Significance Analysis

In this study, I apply Gedeon Method [5] to analyse the significance of the EEG features by magnitude and functional measures. Firstly, we use a magnitude measure P_{ij} to represent the contribution of a last layer neuron i to a next layer neuron j as Equation 4, where N denotes the total neuron number in the last layer.

$$P_{ij} = \frac{|W_{ij}|}{\sum_{p=1}^{N} |W_{pj}|} \tag{4}$$

Our model is a 3-layer network, and the output layer only contains 1 neuron, so the contribution of an input neuron i to the output neuron o can be calculated by Equation 5, where h and M denote a hidden neuron and the total number of hidden neurons. The contribution magnitude measure Q_{io} determines the significance of a certain EEG feature in our study. When the value is larger, the corresponding EEG feature is more important; vice versa.

$$Q_{io} = \sum_{h=1}^{M} (P_{ih} \times P_{ho}) \tag{5}$$

Secondly, I use the weight vector angle between input neurons as a functional measure to represent the distinctiveness of the input neuron. It is extended from the hidden neuron distinctiveness analysis technique proposed in [6]. The vector angle of the neurons i and j can represent how different between two neurons. If the angle is small, one of these neurons may be redundant and can be removed. In this study, I calculate the average angle of one input neuron to other 191 neurons as the functional measure, and the angle is calculated by Equation 6, where $sact(p,h) = norm(weight(h)) - 0.5$.

$$angle(i,j) = tan^{-1}\left(\sqrt{\frac{\sum_{p}^{pats} sact(p,i)^2 * \sum_{p}^{pats} sact(p,j)^2}{\sum_{p}^{pats} (sact(p,i) * sact(p,j))^2}} - 1\right) \tag{6}$$

Since deep learning-based EEG features are extracted by our deep extractors automatically, the features do not have the actual meaning. However, the spectral power EEG features are more explainable, and it is important to analyse which frequency band or which electrode is more significant to EEG data. So, we use both magnitude and functional measures introduced above to analyse our spectral power EEG features in the alcoholism classification experiments.

2.6 Implementation Details

For model training phase, I utilise binary cross-entropy loss function and Adam optimiser. The epoch number and batch size are 50 and 8 respectively. The code implementation is based on Pytorch deep learning framework, and it runs on Google Colab.

3 Results & Discussions

This section reports the classification accuracy using different EEG features extracted from UCI dataset. We compare the performances of these features and also with previous studies. Besides, we analyse the significance of hand-crafted EEG features, namely spectral power of different electrodes using Gedeon method [5].

3.1 Classification Results

I use classification accuracy as performance metric to evaluate our methods. Table 1 reports the cross-subject experiment results of our experiments using different hand-crafted and deep learning EEG features, and also a comparison with previous studies.

We use the performance of hand-crafted EEG features as a baseline which is 57.83% mean accuracy. The image-based spatial features surpasses the baseline and achieves the highest classification result with 67.31% accuracy, demonstrating the effectiveness of our 2D CNN based feature extractor. I believe that image-based spatial features perform better because: ***(a)*** EEG images remain the spatial information of EEG signals from different electrodes and makes the input EEG data more meaningful; ***(b)*** our CNN-based feature extractor can obtain the salient spatial EEG features successfully contributing to the higher classification accuracy. Besides, the higher result of image-based method also demonstrates the better generalisation ability of our CNN-based model. For temporal features, when we use LSTM as feature extractor, the result is only 51.67% and even worse than the baseline. However, when we use Conv1D to extract temporal features, the classification accuracy improves a lot and achieves 64.06%. This demonstrates the EEG temporal information are also meaningful and contributes to the alcoholism identification. We also notice that the train accuracy of LSTM model is much higher than the test accuracy, thus the model is overfitting and it is difficult to solve it by changing hyper-parameter. We believe this is because EEG signal has a high inter-subject variability, While Conv1D model alleviates this overfitting because we utilise batch normalisation after each convolutional layer. Overall, the EEG deep learning features outperforms than hand-crafted features, and spatial features are better than temporal features in alcoholism classification,

Compared to the previous studies, our image-based method surpasses some earlier works such as EEGNet [8], DE & PSD [16] and rEEG [12], but fails to beat some recent studies such as Image-wise Autoencoders [15]. I believe this is because the encoder-decoder-based model proposed in [15] performs better in high-level feature extraction.

To conclude, I have the following finds for alcoholism classification results: ***(a)*** The proposed methods successfully surpass some previous studies and achieve a satisfied results for alcoholism classification; ***(b)*** EEG deep learning features perform better than hand-crafted features and has a better generalisation ability in alcoholism identification; ***(c)*** Image-based spatial features performs better than temporal features and achieve the highest accuracy, which demonstrates the importance of the location information of EEG electrodes.

Method	Accuracy
Hand-crafted features	**57.83%**
Image-based spatial features	**67.31%**
Temporal features (LSTM)	**51.67%**
Temporal features (Conv1D)	**64.06%**
Channel-wise Autoencoders (Yao et al. [15])	73.1%
Image-wise Autoencoders (Yao et al. [15])	75.6%
EEGNet (Lawhern et al. [8])	67.2%
SyncNet (Li et al. [10])	72.3%
DE (Zheng and Lu [16])	62.2%
PSD (Zheng and Lu [16])	60.5%
rEEG (O'Reilly et al. [12])	61.4%

Table 1. Mean classification accuracy for cross-subject experiments

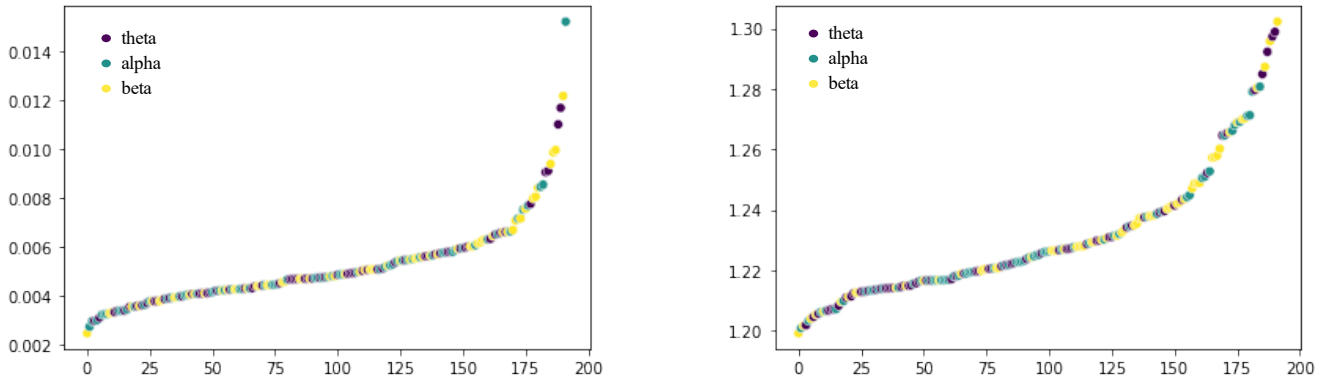
3.2 EEG Feature Significance Analysis

Although EEG deep learning features perform better in alcoholism classification, hand-crafted features are more meaningful and explainable. This means we explicitly know the meaning of each value in the feature vector (the spectral power of a frequency band of a electrode). So, we only analyse the hand-crafted features in this study.

We utilise magnitude and functional measures proposed in [5] to evaluate the EEG hand-crafted features significance based on our trained model (i.e. 2-linear-layer neural network classifier). Table 2 reports 5 most significant features and also 5 least significant features using both two measures. It it notable that, beta32, theta45, and theta33 appear in 5 most significant features for both two measures demonstrating the consistency of two measures. The significance of all 192 features are distributed in Figure 5. The average magnitude/functional measures of theta, alpha and beta frequency bands are 5.10e-3/1.23, 5.03e-3/1.23 and 5.49e-3/1.24, respectively. So, we find that beta frequency band features are most significant and contribute the most to the alcoholism classification output.

Measure	Most significant					...	Least significant				
Magnitude	alpha47	beta32	theta45	theta33	beta64	...	alpha30	theta22	alpha52	alpha61	beta9
Functional	beta34	theta9	theta33	beta32	theta45	...	alpha1	theta20	beta5	alpha45	beta57

Table 2. EEG features distinctiveness analysis using magnitude measure and functionality; alpha47 means the mean value of theta frequency band from 47th channel.



(a) EEG feature significance using magnitude measure

(b) EEG feature significance using functional measure

Fig. 5. Feature significance

4 Conclusion & Future Work

This study proposed three deep learning-based EEG feature extractors, namely Conv2D spatial feature extractor, LSTM temporal feature extractor and Conv1D temporal feature extractor, for alcoholism classification. We use hand-crafted spectral power EEG features as the baseline and find that the deep learning features perform better because they successfully capture the spatial and temporal information of EEG data. The highest classification accuracy is 67.31% and it is achieved by CNN spatial features. While the best performance of deep temporal features is 64.06% achieved by Conv1D. So, we conclude that spatial information is more significant than temporal information in short-term EEG alcoholism classification task. Besides, two input distinctiveness measures, namely magnitude and functional measures, are used to evaluate the spectral power EEG features. The result shows the features from beta frequency band are the most significant features.

However, the proposed methods fail to beat some recent works and there is much space for improvement. EEG is a kind of sequential data which also contain spatial information, however this study only extract the spatial and temporal feature separately but fails to combine them. In the future, we plan to explore more advanced

techniques which can extract both spatio-temporal information from EEG data, such as 3D CNNs, convolutional long short-term memory (ConvLSTM) to extract more salient spatio-temporal EEG features and try to improve the classification accuracy. Besides, I will prune some units in the model according to the distinctiveness obtained in this study to facilitate the efficiency, and compare the accuracy between before and after pruning.

References

1. Abhang, P.A., Gawali, B.W.: Correlation of eeg images and speech signals for emotion analysis. Current Journal of Applied Science and Technology pp. 1–13 (2015)
2. Alfeld, P.: A trivariate clough—tocher scheme for tetrahedral data. Computer Aided Geometric Design **1**(2), 169–181 (1984)
3. Bashivan, P., Rish, I., Yeasin, M., Codella, N.: Learning representations from eeg with deep recurrent-convolutional neural networks. arXiv preprint arXiv:1511.06448 (2015)
4. Garson, D.G.: Interpreting neural network connection weights. AI Expert pp. 47–51 (1991)
5. Gedeon, T.D.: Data mining of inputs: analysing magnitude and functional measures. International Journal of Neural Systems **8**(02), 209–218 (1997)
6. Gedeon, T., Harris, D.: Network reduction techniques. In: Proceedings International Conference on Neural Networks Methodologies and Applications. pp. 119–126. AMSE (1991)
7. Koelstra, S., Muhl, C., Soleymani, M., Lee, J.S., Yazdani, A., Ebrahimi, T., Pun, T., Nijholt, A., Patras, L.: Deap: A database for emotion analysis using physiological signals. IEEE Transactions on Affective Computing **3**(1), 18–31 (2012)
8. Lawhern, V.J., Solon, A.J., Waytowich, N.R., Gordon, S.M., Hung, C.P., Lance, B.J.: Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces. Journal of neural engineering **15**(5), 056013 (2018)
9. Lebedev, M.A., Nicolelis, M.A.: Brain–machine interfaces: past, present and future. TRENDS in Neurosciences **29**(9), 536–546 (2006)
10. Li, Y., Murias, M., Major, S., Dawson, G., Dzirasa, K., Carin, L., Carlson, D.E.: Targeting eeg/lfp synchrony with neural nets. In: NIPS. pp. 4620–4630 (2017)
11. Lotte, F., Congedo, M., Lécuyer, A., Lamarche, F., Arnaldi, B.: A review of classification algorithms for eeg-based brain–computer interfaces. Journal of neural engineering **4**(2), R1 (2007)
12. O'Reilly, D., Navakatikyan, M.A., Filip, M., Greene, D., Van Marter, L.J.: Peak-to-peak amplitude in neonatal brain monitoring of premature infants. Clinical neurophysiology **123**(11), 2139–2153 (2012)
13. Ullah, I., Hussain, M., Aboalsamh, H., et al.: An automated system for epilepsy detection using eeg brain signals based on deep learning approach. Expert Systems with Applications **107**, 61–71 (2018)
14. Wolpaw, J.R., Birbaumer, N., McFarland, D.J., Pfurtscheller, G., Vaughan, T.M.: Brain–computer interfaces for communication and control. Clinical neurophysiology **113**(6), 767–791 (2002)
15. Yao, Y., Plested, J., Gedeon, T.: Deep feature learning and visualization for eeg recording using autoencoders. In: International Conference on Neural Information Processing. pp. 554–566. Springer (2018)
16. Zheng, W.L., Lu, B.L.: Investigating critical frequency bands and channels for eeg-based emotion recognition with deep neural networks. IEEE Transactions on Autonomous Mental Development **7**(3), 162–175 (2015)