2021 27th International Conference on Mechatronics and Machine Vision in Practice (M2VIP) | 978-1-6654-3153-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/M2VIP49856.2021.9665129

# A Semi-automatic Feature Fusion Model for EEG-based Emotion Recognition

Gaotian Zhang[1], Shiqian Li[2], Jiabao Wang[1], Yun Zhou[2] and Tao Xu[1*]

***Abstract*— Electroencephalogram (EEG) is usually used to study cognitive activities, which have different temporal, frequency-domain features. Scientists attempted to find crucial features to improve recognition accuracy but challenging. This paper proposed a novel confused emotion recognition method based on EEG, which combine automatic feature extraction (deep learning) and knowledge-based feature extraction. To evaluate our method, we designed an experiment to collect data, the basic idea of which is to induce the confused emotion based on the English listening test. The results show that our method performs better in experiments than Convolution Neural Networks(CNN) and Support Vector Machine (SVM).**

## I. INTRODUCTION

The human brain produces spontaneous electrophysiological activities when carrying out cognitive activities. Through special equipment, the EEG signal can be detected in the cerebral cortex or inside. Different types or intensities of cognitive activities will produce different patterns of EEG signals.

EEG signals are very weak and easily interfered during collecting. Researchers attempt to find the efficient methods used to recognize the cognitive activities[1], [2], [3], [4]. These methods can be divided into two categories: one is the traditional method, extracts features based on prior knowledge, and then uses the machine learning method to classify the features extracted. Its performance mainly depends on whether to find a crucial feature. The other one is the deep learning method, which can automatically extract features and classify directly from original data. However, it requires big and high quantity data and a lack of interpretability.

The cost of collecting EEG data is relatively high, and the amount is relatively scarce. Since it is difficult to collect enough EEG data to train the model[5] The methods based on deep learning is not easy to play their advantage. Nowadays, researchers attempt to use traditional machine learning methods with fusion features to improve accuracy. The information of multiple modes can complement each other, this fusion can achieve a result superior to that of a single mode [6]. The EEG is time-series signal, the extracted features based on prior knowledge can be divided into time domain features, frequency domain features, time-frequency domain features and spatial domain features[7]. Inspired by this idea, Jia et.al [8] proposed a 3D Densenet based on the attention mechanism is proposed to classify the emotional EEG signals generated by multimedia stimulation. This model can extracts spatial, frequency and time features from data simultaneously under a unified framework.

We proposed a semi-automatic feature fusion model to take advantage of feature fusion and deep learning in both. The main contribution of this work is focused on the feature extraction layer. We attempt to combine automatic feature extraction (deep learning) and manual feature selection. It is used to solve the problem that deep learning fails to extract features effectively because of the small amount of data. Experimental results show that the proposed method is helpful to improve the classification accuracy in Emotion EEG classification.

## II. SEMI-AUTOMATIC FEATURE FUSION MODEL

We propose a semi-automatic feature extraction model for EEG-based emotion recognition, combined with fusing features extracted from deep learning and features selected based on prior knowledge. The overall structure of the model is shown in the Fig.1.

After raw data input, there is a feature extraction layer. It consists of two parallel feature extraction parts: The above one is the automatic feature extraction part based on deep learning. It is used EEGNet model[9] to extract features automatically. The below one is the knowledge-based feature extraction part. It extracts classic features from EEG data based on prior knowledge, like Power Spectral Density(PSD), Differential Entropy(DE), and asymmetric feature. The features extracted from the two parts are concatenated in the feature fusion layer and then input to the fully connected network in the classification layer. Finally, the fully connected network is adopted to obtain the final result of emotion classification after the Softmax function. The details of the feature extraction layer will be introduced as follows.

### *A. The Automatic Feature Extraction*

Deep learning, that is the convolution network-based architecture, has shown impressive performance in time series data classification[10], [11]. In the field of EEG, EEGNet network[9] based on convolution network has achieved good results in the task of EEG classification. EEGNet is based on a convolution neural network, which consists of three convolution layers and a fully connected layer.

*This research was supported by the National Natural Science Foundation of China(62077036), the National Key Research and Development Program of China(2018AAA0100500)

[1]Gaotian Zhang, Jiabao Wang and Tao Xu (Corresponding author) are with the School of Software, Northwestern Polytechnical University, 127 West Youyi Road, Xi'an, 710072, P.R.China. zhanggaotian@mail.nwpu.edu.cn, wjiabao@mail.nwpu.edu.cn, xutao@nwpu.edu.cn

[2]Shiqian Li and Yun Zhou are with the School of Education, Shaanxi Normal University, Xi'an, 710062, P.R.China. devount@qq.com, zhouyun@snnu.edu.cn

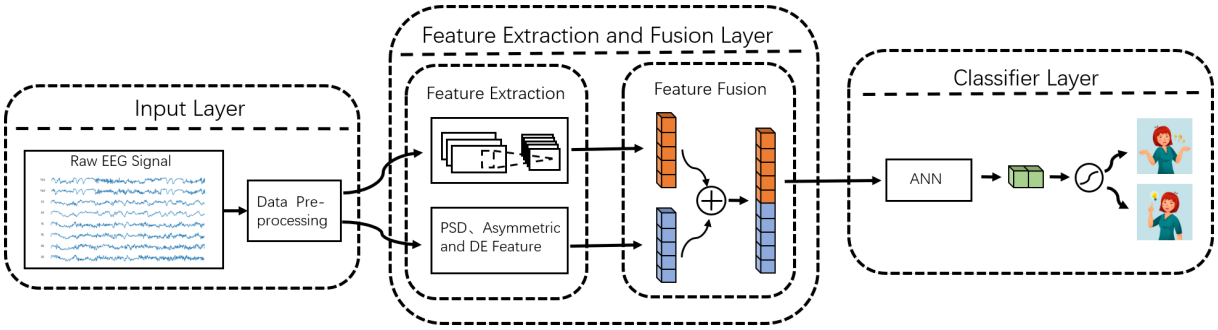


Fig. 1. Semi-automatic feature fusion model

The three-level convolution of EEGNet consists of a common convolution layer, a depthwise convolution layer and a separable convolution layer. In the concrete implementation, the ordinary convolution layer only performs convolution operations in the time dimension. The parameter kernLength of the convolution kernel is set to 64, and the parameter F1 of the number of convolution kernels is set to 8. Depthwise convolution is a channel-by-channel operation at the convolution layer. That is, the convolution operation is performed on each channel separately. The number of channels in the output feature graph of depthwise convolution is the same as the original input. In addition, this layer only performs convolution operations on the dimension of EEG channel, the size of convolution kernel is equal to EEG channel, and the number of convolution kernel F2 is set to 16. Separable convolution is composed of one depthwise convolution layer and one pointwise convolution layer. The convolution kernel size of pointwise convolution is fixed as $1 \times 1$. It is a point-by-point convolution layer operation on the same position of different channels. In each convolution layer, two-dimensional BatchNorm and ELU activation functions were used. In order to prevent overfitting, dropout mechanism was used during training and the dropout ratio set to 0.5.

We only use the three convolution layers for feature extraction. The automatic feature is get from raw EEG data, shown as follows:

$$Automatic\ Feature_n \in R^{1 \times f_1}$$

where $f_1$ is the number of features extracted by EEGNet, which is determined by the length of the original data $t$ and the size and the number of convolution kernels. 112 automatic features are extracted by this part.

### B. The Knowledge-based Feature Extraction

In the knowledge-based feature extraction part, classic feature extraction methods can be used. We choose PSD, DE and asymmetry feature of EEG signals from the frequency and spatial domain. Before feature extraction, we use the Butterworth filter to decompose the EEG signal of each channel into five different frequency bands, namely Theta (4-8 Hz), Slow Alpha (8-10 Hz), Alpha (8-12 Hz) , Beta (12-30 Hz), and Gamma (30-50 Hz). The PSD, DE and asymmetry are applied to extract features.

The definition of power spectrum feature is as follows:

$$h_p(X) = E\left[x^2\right]$$

where $x$ represents the signal obtained in the specific frequency band of each channel.

Differential entropy is the generalized form of Shannon entropy on continuous variables, which can distinguish EEG patterns between low-frequency and high-frequency energy. It is defined as:

$$h_D(X) = -\int_X f(x)\log(f(x))dx$$

where $f(x)$ is the probability density function of $x$.

Asymmetry feature is used to indicate the energy imbalance between the channel pairs. Take the EEG channels FP1, FP2, C3, C4, O1, O2, T5, T6 used in this article as an example, the asymmetry features need to be separately (FP1, FP2) , (C3,C4), (O1,O2), and (T5,T6) channel pairs are calculated, the calculation method is:

$$Asymmetry\ Feature = \log(PSD(Chan1)) - \log(PSD(Chan2))$$

Among them, Chan1 and Chan2 are a channel pair.

Using the above feature extraction methods, 96 features were extracted from each segment of trial. One the one hand, The 96 features extracted based on knowledge are directly input into SVM for classification, which is used to compare the performance of SVM classifier. After repeated experiments, the kernel function of the SVM classifier is set as Gaussian kernel, the kernel coefficient Gamma is set between 0.1 and 10, and the regularization parameter C of the classifier is set as 2. We use a feature selection method, every time in the training, according to specified rules divided into the training set and validation set, we will using the one-way ANOVA method on the training set, help training set to obtain the best classification results K features as optimal features, Only the optimal feature is selected for classification on the validation set.

One the other hand, We employ the early fusion method of multi-modal features as a feature fusion layer. We use the feature map obtained by EEGNet three-layer convolution to expand into one-dimensional feature vector, which is used for the fusion of automatically extracted features and manual features extracted by traditional methods. The fusion method adopts the concat method, that is, connecting the two

features to obtain new features. In the feature fusion layer, the automatic features and the knowledge-based features are concatenated together as follows:

$$Fusion\ \ feature_n \in R^{1\times f_3}$$

where $f_3$ is the number of fusion features, and $f_3 = f_1 + f_2$.

The classifier layer, a fully connected network, is applied based on the fusion feature for emotion recognition.The fusion features are directly input to the fully connected network. The number of neurons in the input layer of the fully connected network is equal to the number of fusion features, and the number of neurons in the output layer is equal to the number of categories. No additional hidden layer is set in the middle. To evaluate this model, we conduct a emotion EEG database. The detail of database will be introduced in the next chapter.

## III. EEG EMOTION DATABASE INDUCED BY AUDIO

Audio EEG Emotion Database is an emotional EEG Database that induces subjects to produce confused and non-confused emotions through different types of Audios, and collects corresponding EEG signals.

14 subjects' EEG data were recorded in the database, but one of them had a problem with the data label, so there were only 13 subjects is valid data. Their ages ranged from 21 to 35 years old (Mean=24.5, Standard Deviation=3.84), including 5 males and 8 females. All the subjects have bachelor's degree or above, including 3 undergraduate students, 7 master's degree students and 3 doctor's degree students. All the subjects are right-handed, without intellectual or hearing impairment, and all of them are Chinese.

The experimental paradigm used in this database is shown in the Fig.2:

Subjects in each experiment need to listen to some audios, each section of the audio recording for a trial, before the start of each trial first present the fixation point, it reminds the subjects that an audio is about to be played. The audio starts playing 1 second after the fixation point appears. Then the subjects began to answer questions, answer this question time is set to 5 second, After the subjects answered the questions or lasted more than 5 seconds, the next trial began. According to the content of audio, it can be divided into word experiment, sentence experiment and paragraph experiment. The details are as follows:

- **Word Experiment**
  In the word experiment, there were 120 audio clips of different words in total, and the audio length of each word was about 1s. During the word experiment, the fixation point would appear to remind the subject audio will be played, and the word audio would be played after 1s. After the audio was played, the subjects would have four different word definitions for choice. Subjects were asked to choose what they thought was the correct interpretation of the word they heard, as well as an additional option of "I didn't hear" if they didn't understand the audio content. Each question will take 5s to answer. When the subject finishes the choice or takes more than 5s to answer, a new round of fixation will appear. In the specific recording, if the subject chooses, the index of the corresponding option will be recorded as the subject's answer; if the subject does not chooses, the empty value will be recorded. Each participant was asked to listen to 120 audio clips and give their answer to each question. But the order in which each subject heard the words was completely random. The experiment uses an objective standard as an evaluation standard of the participants were confused or not, namely, the subjects answered correctly or not, in other words, if the subject answered correctly, the subject was considered unconfused; otherwise, the subject was considered confused.
- **Sentence Experiment**
  The sentence experiment was similar to the word experiment, except that in the sentence experiment, only 20 audio sentences were prepared and played to the subjects in sequence, each of which was about 4s long. Subjects no longer choose the Chinese definition of the sentence. After listening to each sentence, four options are prepared. The first three are possible English answers to the sentence, and the subjects need to choose a reasonable answer to the sentence question or content. When the subjects did not understand the sentence, they had to choose the fourth option, "I did not understand". For example, subjects might hear the sentence "Did you order more copy paper yesterday?". At the end of the sentence audio, the screen will give subjects the different options to choose. Taking this question as an example, the following options appear on the screen: "Option1: It was paid by card. Option2: Yes, after you left. Option3: Not for a while. Option4: I did not understand". Subjects make choices according to their own understanding and press the corresponding number keys on the keyboard. The numbers 1-4 correspond to four options.
- **Paragraph Experiment**
  In the paragraph experiment, each subject is required to listen to 10 different English paragraphs with the length of each paragraph being around 30s. After listening to each paragraph, questions related to the content of the paragraph will appear. The subject is required to give the answer to the given question according to their own understanding of the paragraph.

In the experiment, OpenBCI was used as the EEG data acquisition device. During data collection, the subjects wore eight-lead electrode caps and recorded the EEG signals of eight channels, FP1, FP2, C3, C4, T5, T6, O1 and O2, according to the international 10-20 lead standard. Eight EEG channels were located in the frontal pole, central, posterior temporal and occipital regions of the brain. The sampling rate is set to 250 Hz when recording, which means 250 data points are recorded per second. In order to ensure

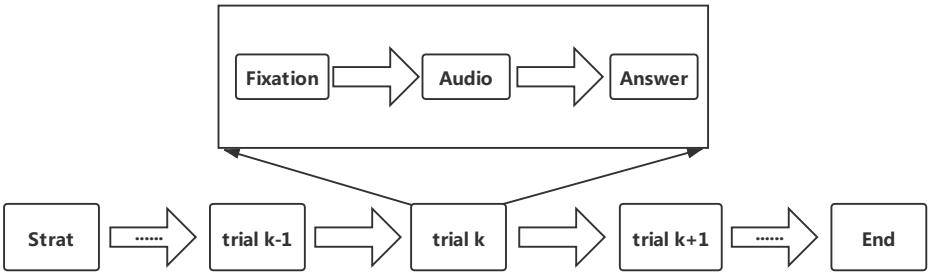


Fig. 2. Audio EEG database experimental paradigm

the consistency of data records, each trial data record is not recorded in a single file, but all trials of each subject and each experiment are recorded in the same file. Therefore, in order to distinguish trials from the overall data, The time points were marked at the beginning of the experiment, when the fixation point appeared, when the audio appeared, when the question appeared, and when the experiment ended. Each trial data can be extracted from these triggers during subsequent processing.

## IV. EVALUATION AND RESULTS ANALYSIS

### A. Data Preprocessing

In this paper, some common preprocessing methods of EEG data are used to process the EEG data in the audio emotion database. There are the following processing procedures:

- **Data Segmentation**
  The collected EEG data were divided into different trials according to the trigger marked at the time of collection. Taking subject 1 as an example, the recorded word EEG data could be divided into 120 different trials, sentence EEG data could be divided into 20 trials, and paragraph EEG data could be divided into 10 trials. The number of trials was related to the number of audio played. Subsequent processing is conducted on the trial after subsection.
- **Band Filters and Notch Filter**
  The valid frequency band of EEG data ranges from 1-50Hz, which can be further divided into five frequency bands: Delta(1-4Hz), Theta(4-8Hz), Alpha(8-12Hz), Beta(12-30Hz) and Gamma(30-45Hz). Therefore, high-pass filtering and low-pass filtering are used to remove high-frequency noise and low-frequency baseline drift. In order to remove power-frequency interference, notch wave is used to process the collected EEG data at 50Hz.
- **Independent Component Analysis (ICA)**
  During the collection of EEG signals, the subjects' behaviors such as swallowing, blinking, eye movement, heartbeat, and muscle activity during answering questions will cause certain interference to the collected EEG data. For EMG and ECG components in EEG data, due to their inconsistent frequency range, they can be removed by filtering method. But for electrical signals, the frequency of the electric eye and brain electrical signal is extremely close, direct filtering can't very useful, so using independent component analysis, the original EEG is decomposed into several independent components, remove independent components of the suspected eye electrical components, using the mixed matrix can recover clean EEG signals from independent component.
- **Data Augmentation**
  Since the length of each segment of audio is not fixed, and the answering time used by the subjects for each question is also not fixed, the sliding window method is used to make the length of each segment of EEG data consistent. The practice is to use a window with a specified length to slide overlaps across the entire trial, and the overlap rate is generally set at 0.5. New data with the length of the window will be obtained. Meanwhile, because the window overlaps during the sliding, the amount of EEG data used for training will be increased to a certain extent.
- **The Normalized Procedure**
  The data values recorded in each channel of each trial are different in order of magnitude, so the min-max method is used to normalize each channel of each trial. Make all data numerical maps to the [0,1] interval. Another advantage of using a normalization approach is to reduce the differences between individuals.
- **Visualization**
  In order to verify that the data is balanced and to find evidence that there is indeed the difference between the EEG corresponding to confused and non-confused emotions, we conducted visualization on the EEG data of experiment, and the distribution of three kinds of experimental emotion labels are shown in Fig.3.

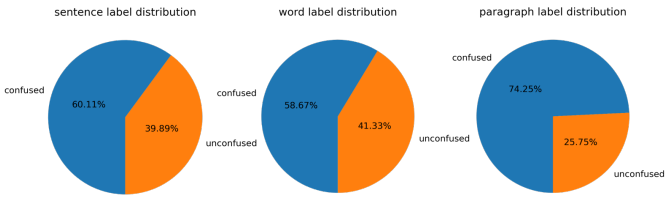


Fig. 3. The distribution of three kinds of experimental emotion labels

For the word and sentence experiment, the number of confused and non-confused EEG was relatively balanced, but with the increase of difficulty, that is, when listening to the paragraph, more confused emotions were generated, which was consistent with our cognition. There is no extra work done on the balance of paragraph emotional labels.

## V. RESULTS AND DISCUSSIONS

### A. Word Level Result

The classification effect of the three models and the validity of the semi-automatic feature fusion method were verified by mixing all subjects' data for 5 fold cross-validation and leaving one subject for cross-validation respectively. The results of the 5 fold cross-validation for all subjects' data were shown in the Table.I.

TABLE I
RESULT OF MIX DATA AND 5 FOLD CROSS-VALIDATION

| Model \ Fold | 1 | 2 | 3 | 4 | 5 | Avg |
|---|---|---|---|---|---|---|
| CNN | 0.641 | 0.537 | 0.514 | 0.523 | 0.668 | 0.576 |
| SVM | 0.637 | 0.531 | 0.508 | 0.548 | 0.707 | 0.586 |
| Fusion | 0.640 | 0.540 | 0.521 | 0.563 | 0.706 | 0.594 |

The header number is the fold number of cross-validation, and Avg Acc represents the average classification accuracy of the 5 fold cross-validation. As can be seen from the table, the feature fusion model proposed by us not only shows the same classification ability as the other two models in each fold but also is significantly better than the simple use of convolution network and the use of SVM classifier after feature extraction in the average classification accuracy of 5 fold.

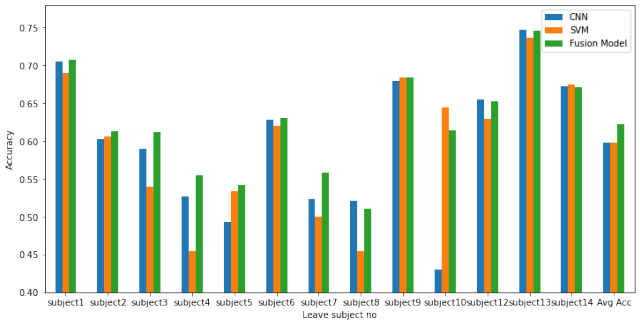


Fig. 4. Results of the leave one subject cross-validation

The number on the abscess axis in Fig.4 represents the serial number of the subjects left in the cross-validation of keeping one subject, and Avg Acc represents the average classification accuracy. It can be seen from the Fig.4 that, similar to the result of 5 fold cross validation, the semi-automatic feature fusion method proposed and used by us achieves significantly better results than the other two models in terms of average classification accuracy.

### B. Sentence Level Result

The classification effect of the three models was verified by mixing the data of all subjects and using 5 fold cross-validation, and leaving one person cross-validation on the total data. The accuracy was showed in Table.II and Fig.5:

TABLE II
RESULT OF MIX DATA AND 5 FOLD CROSS-VALIDATION

| Model \ Fold | 1 | 2 | 3 | 4 | 5 | Avg |
|---|---|---|---|---|---|---|
| CNN | 0.587 | 0.666 | 0.542 | 0.510 | 0.661 | 0.594 |
| SVM | 0.582 | 0.714 | 0.495 | 0.467 | 0.745 | 0.601 |
| Fusion | 0.611 | 0.679 | 0.517 | 0.535 | 0.713 | 0.612 |

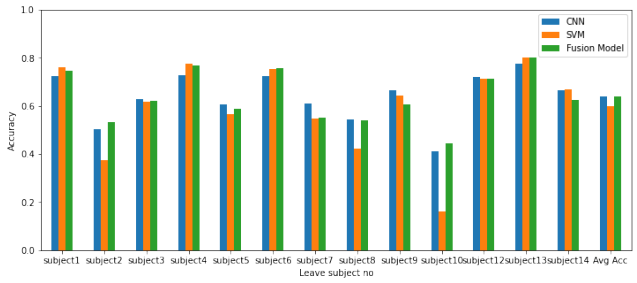


Fig. 5. Results of the leave one subject cross-validation

In the sentence experiment, the average accuracy of 5-fold cross validation by mixing all the subjects' data reached about 0.6. However, in the cross-subject experiment, the classification effect of the convolution model and the mixed model is relatively stable, while the robustness of the SVM classification effect is poor. However, in general, most of the classification accuracy can reach above 0.6.

### C. Paragraph Level Result

The classification effect of the three models was verified by mixing all subjects' data and using 5-fold cross-validation, and leaving one person cross-validation on the total data. The accuracy was showed in Table.III and Fig.6:

TABLE III
RESULT OF MIX DATA AND 5 FOLD CROSS-VALIDATION

| Model \ Fold | 1 | 2 | 3 | 4 | 5 | Avg |
|---|---|---|---|---|---|---|
| CNN | 0.871 | 0.651 | 0.856 | 0.530 | 0.810 | 0.744 |
| SVM | 0.871 | 0.651 | 0.855 | 0.523 | 0.810 | 0.742 |
| Fusion | 0.871 | 0.656 | 0.846 | 0.539 | 0.809 | 0.744 |

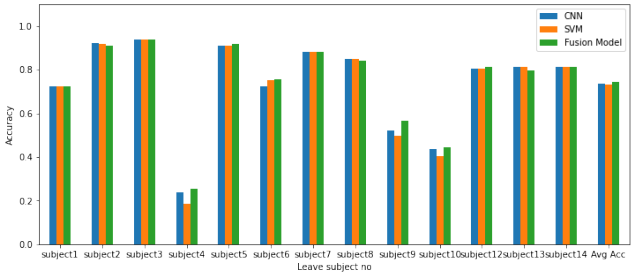


Fig. 6. Results of the leave one subject cross-validation

For paragraph classification, the average classification accuracy of all data mixed with 5-fold cross validation is better than that of word and sentence data classification. However, this result may be affected by the relatively small amount of paragraph data and the imbalance of paragraph labels, so further verification is needed. Similarly, in the cross-subject classification, the classification effect of the three models is similar, and the average classification result also reaches more than 0.6.

## VI. CONCLUSIONS

In this paper, one-dimensional convolution, EEGNet network model, SVM model and feature fusion model are used to classify emotional EEG induced by English audio, aiming at effectively classifying confused and non-confused emotions. The average classification accuracy of 0.594, 0.6387 and 0.7446 were obtained at the word, sentence and paragraph levels respectively.

In addition, we verify the availability of the semi-automatic feature fusion method, namely the convolution and fusion network classify extracted features and manual extraction is superior to the direct use of convolution network classification and manual extraction and feature classification.

## REFERENCES

[1] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," vol. 3, no. 1, pp. 42–55. [Online]. Available: http://ieeexplore.ieee.org/document/5975141/

[2] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "DEAP: A database for emotion analysis ;using physiological signals," vol. 3, no. 1, pp. 18–31, conference Name: IEEE Transactions on Affective Computing.

[3] R.-N. Duan, J.-Y. Zhu, and B.-L. Lu, "Differential entropy feature for EEG-based emotion classification," in *2013 6th International IEEE/EMBS Conference on Neural Engineering (NER)*, pp. 81–84, ISSN: 1948-3554.

[4] W.-L. Zheng and B.-L. Lu, "Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks," vol. 7, no. 3, pp. 162–175, conference Name: IEEE Transactions on Autonomous Mental Development.

[5] F. Lotte, L. Bougrain, A. Cichocki, M. Clerc, M. Congedo, A. Rakotomamonjy, and F. Yger, "A review of classification algorithms for EEG-based brain–computer interfaces: a 10 year update," vol. 15, no. 3, p. 031005. [Online]. Available: https://iopscience.iop.org/article/10.1088/1741-2552/aab2f2

[6] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," vol. 41, no. 2, pp. 423–443, conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

[7] M. Yu, D. Zhang, G. Zhang, G. Zhao, Y.-J. Liu, Y. Han, and G. Chen, "A review of EEG features for emotion recognition," vol. 49, no. 9, pp. 1097–1118. [Online]. Available: http://engine.scichina.com/doi/10.1360/N112018-00337

[8] Z. Jia, Y. Lin, X. Cai, H. Chen, H. Gou, and J. Wang, "SST-EmotionNet: Spatial-spectral-temporal based attention 3d dense network for EEG emotion recognition," in *Proceedings of the 28th ACM International Conference on Multimedia*. ACM, pp. 2909–2917. [Online]. Available: https://dl.acm.org/doi/10.1145/3394171.3413724

[9] 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," vol. 15, no. 5, p. 056013. [Online]. Available: https://iopscience.iop.org/article/10.1088/1741-2552/aace8c

[10] H. Ismail Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P.-A. Muller, "Deep learning for time series classification: a review," vol. 33, no. 4, pp. 917–963. [Online]. Available: http://link.springer.com/10.1007/s10618-019-00619-1

[11] H. I. Fawaz, B. Lucas, G. Forestier, C. Pelletier, D. F. Schmidt, J. Weber, G. I. Webb, L. Idoumghar, P.-A. Muller, and F. Petitjean, "InceptionTime: Finding AlexNet for time series classification," vol. 34, no. 6, pp. 1936–1962. [Online]. Available: http://arxiv.org/abs/1909.04939