Integrating Transient and Long-term Physical States for Depression Intelligent Diagnosis

Ke Wu#1
20373041@buaa.edu.cn
Han Jiang#*,1,2
vroffice@buaa.edu.cn
Li Kuang1
20373335@buaa.edu.cn
Yixuan Wang1
21373233@buaa.edu.cn
Huaiqian Ye1
21371161@buaa.edu.cn
Yuanbo He*,1,2
heluxixue@163.com

1 School of Computer Science and Engineering, Beihang University, Beijing, China
2 State Key Lab of Virtual Reality Technology and Systems, Beihang University, Beijing, China

Abstract

As social competition intensifies, the number of depression patients has rapidly increased. Many researchers have proposed diagnostic models for depression based on various physiological signals and behavioral information, such as Electroencephalogram (EEG) and facial expressions. However, it should be noted that these signals tend to reflect transient information, which may make them insufficient for accurately diagnosing depression characterized by persistent low mood over a prolonged period of time. Meanwhile, traditional Chinese medicine (TCM) believes that different parts of the tongue correspond to different Zang-fu and can indicate the long-term physical condition of the human body. Therefore, we use EEG and tongue images to reflect the subject’s instantaneous state and long-term physical condition respectively, and establish a multimodal model MMTV to assist doctors in diagnosing depression. Specifically, MMTV innovatively introduces the dual-stream input mechanism and self-attention mechanism to EEG-Net to better extract the spatio-temporal features of EEG. Meanwhile, to obtain higher quality tongue surface images, MMTV introduces a segmentation step before inputting tongue images into the ViT model. Meta-learning techniques are applied to gain better pretrained weights for ViT. Furthermore, we analyze the correlation between tongue images and EEG and subsequently fuse the output features of the brain-tongue branches in MMTV. The ability of MMTV to recognize patients with depression has been validated on multiple datasets, with the highest recognition accuracy reaching 98.18%. Our code is available at https://github.com/Clearlangw/Depression-diagnosis-based-on-EEG-and-Tongue.

© 2023. # The first two authors contributed equally to this work. *Corresponding author
1 Introduction

Accompanied by the rapid development of the global economy and increasing competition in various aspects of social life, the number of individuals suffering from depression has surged, making it a primary contributor to the global disease burden [26]. Depression is an emotional mental disorder characterized by persistent low mood [22], which can lead to functional impairments and an increased risk of suicide [11]. Current clinical diagnostic methods primarily rely on scale evaluations and physicians’ experience [23], which can result in issues like subjective bias, poor consistency, and high rates of misdiagnosis or missed diagnosis [32]. Consequently, there is an urgent need for objective and accurate approaches to assist doctors in screening and diagnosing depression. Many diagnostic models for depression based on behavioral information and physiological signals have been proposed to solve the problems [38] [10]. However, these signals typically reflect the subjects’ instantaneous states and lack assessment of their long-term physical conditions, which poses a potential conflict with the chronic nature of depression. Inspired by the use of tongue diagnosis in TCM to assess patients’ long-term visceral states, we proposed the idea of utilizing image recognition models to extract tongue features for depression diagnosis. Some TCM syndromes such as the liver qi stagnation are believed to be associated with the pathogenesis of depression and can also affect the brain [33]. Therefore, we have developed a model called MMTV that combines EEG and tongue images for depression diagnosis.

In terms of data acquisition, we initially collected open access separate datasets for EEG and tongue modalities. Inspired by Wang’s research [40] and Qian’s research [47], we selected VFT and the Chinese Facial Affective Picture System (CFAPS) as the event-related potential (ERP) triggers for EEG data collection. Prior to the EEG data collection, we captured tongue images of the subjects, which were then annotated with TCM labels by TCM practitioners.

In terms of model architecture, the MMTV model is composed of two branches: the Trans_EEGNet module for processing EEG and the ViT module for processing tongue images. Inspired by EEGNet [14], we introduced the dual-stream input mechanism and self-attention mechanism into Trans_EEGNet to better extract the spatio-temporal features of EEG data. We performed tongue surface segmentation on the tongue images and proposed a multi-step pre-training approach to enhance the ViT module’s ability to extract TCM features from the tongue surface. Additionally, we have completed the mapping work of EEG to tongue images, proving their correlation. MMTV has been validated on multiple datasets, achieving excellent results, demonstrating that it can assist physicians in diagnosing depression, reducing labor costs, and improving diagnostic accuracy.

In summary, the main contributions of this paper are as follows:

• We propose a multimodal intelligent diagnosis model for depression, MMTV, based on EEG and tongue images, which combines brain function information and the internal organ state of the body to reflect both the instantaneous state and long-term physical condition of the subject.

• In the EEG module, we design Trans_EEGNet featuring the dual-stream input and self-attention mechanism to effectively capture the spatio-temporal features of EEG.

• In the tongue module, we introduce tongue image segmentation before ViT. A multi-step pre-training method is proposed to better extract the TCM features from the tongue images.
2 Related works

2.1 Diagnosis Based on Transient State

In recent years, numerous researchers have been dedicated to leveraging machine learning techniques to analyze patients’ behavioral information (such as voice, video, and text) and physiological signals (such as EEG, eye movements (EM), and ECG) to identify individuals with depression. Xie et al. used CNN-LSTM to mine information from the videos focused on facial expressions, achieving an accuracy of 94.6% in diagnosing depression and anxiety. Compared to facial expression and speech signals, physiological signals have the advantage of being difficult to fake and can more subtly and objectively characterize emotions. Zang used TCN to analyze 5-second ECG segments, achieving a recognition accuracy of 97.1% in the depression diagnosis. Li et al. adopted a kernel extreme learning machine based on PCA features of EM for depression recognition, achieving a classification accuracy of up to 91%. Among the existing physiological signals used for diagnosing depression, EEG signals have the advantages of high time resolution and containing rich central nervous cognitive information, thus occupying a dominant position in the similar fields such as affective computing. To compensate for the limitations of single-modal physiological signals, the academic community has gradually focused on multimodal works that utilize the complementarity of signals, such as the combination of EEG and EM and the combination of speech and video.

However, the signals mentioned earlier mainly reflect transient states and still have limitations in characterizing the long-term emotional lows associated with depression.

2.2 Diagnosis Based on Long-term Physical Condition

Tongue diagnosis, as a key part of TCM examination, holds significant value for differential diagnosis, representing patients’ relatively long-term physical conditions, thus effectively compensating for the aforementioned limitations. According to TCM theory, the five parts of the tongue correspond to different Zang-fu: the tip to the heart and lungs, the edges to the liver and gallbladder, the center to the spleen and stomach, and the root to the kidneys. Tongue diagnosis infers patients’ internal Zang-fu states by observing the external manifestations of tongue texture and coating, integrating multiple attribute information of tongue images, thus serving as a diagnostic basis for diseases.

However, traditional tongue diagnosis relies on TCM practitioners’ experience and subjective judgment, necessitating quantification and objectification. Therefore, the integration of tongue images and machine learning has attracted the attention of many researchers. Li et al. used the GA_XGBT model to predict whether patients were in the early or late stages of diabetes, achieving an average accuracy of 82.1%; Ma et al. used complexity perception classification to identify TCM constitution using tongue images. Modern TCM categorizes depression into “Yuzheng” as an emotional disorder based on its clinical manifestations and characteristics. It is believed that the pathogenesis of depression is due to emotional injury, which leads to stagnation of Qi in the internal organs and imbalance of functions in Zang-fu. According to clinical observations and records, patients with depression exhibit distinctive tongue features, with 80.5% displaying abnormal tongue coating characterized by a greasy coating. To explain the relationship between tongue and depression better, the sample comparison is shown in Figure 1.

In addition, compared to EEG, tongue diagnosis has a lower sensitivity, while EEG has a higher temporal resolution, allowing it to detect subtle changes in brain activity. This makes
Figure 1: Notes on the tongue

EEG potentially more sensitive in identifying early or mild symptoms of depression.
Thus, the combination of physiological signals and tongue images may provide a more comprehensive assessment and improve diagnostic accuracy.

3 Methodology

3.1 Overview

The overall architecture of the Multimodal Model based on Trans_EEGNet and Vision Transformer (MMTV) is shown in Figure 2.

These modules have already been briefly introduced above. In the subsequent sections, we will provide a detailed description of each module.

3.2 EEG Module (Transient-term Physical State)

In the EEG module, we utilize Trans_EEGNet to analyze EEG data. Additionally, we replicate a CRNN architecture [29]. Besides, we analyze the feature extraction characteristics of these two models and verify the superiority of Trans_EEGNet in related issues.
Data preprocessing: The key steps in the preprocessing stage include: denoising, dividing the EEG signal into frequency bands, extracting differential entropy (the generalized form of Shannon entropy in continuous variables, which can be used as EEG features [8]) over fixed time intervals, and selecting a sequence of differential entropy values from consecutive time intervals.

Architecture: The CRNN architecture comes from a processing model for SEED dataset [48]. The key idea of preprocessing is to map the channel dimension of the data to a two-dimensional brain map based on the actual spatial distribution of the EEG channels. The characteristic of the CRNN model lies in its dependence on the actual spatial distribution, which inadvertently increases the requirements for data quality. Aiming at proposing a model with less stringent data requirements, we draw inspiration from EEGNet which consists of three convolutional layers: a conventional convolutional layer for extracting intra-channel features, a deep convolutional layer for extracting inter-channel features, and a separable convolutional layer for extracting spatio-temporal features [15]. Since the first convolution layer of original EEGNet only extracts features inside the EEG channel rather than considering the inter-channel features, we process input data in parallel and design a modified EEGNet that only extracts the inter-channel features in the first layer as a supplement. EEG data slices are simultaneously input into those two architectures to extract features. And the transpose mechanism is imposed on the intermediate results of the modified EEGNet. Therefore, the features of (sequence, channel) and (channel, sequence) are obtained respectively from EEGNet: $\mathbb{R}^{C \times S \times 1} \rightarrow \mathbb{R}^{S' \times C'}$ and the modified EEGNet: $\mathbb{R}^{C \times S \times 1} \rightarrow \mathbb{R}^{C'' \times S''}$. To better extract features between each channel and sequence features inside the channel, we introduce the self-attention mechanism to the model [37], which compensates for the difficulty of extracting global information in the convolutional layer. The core formula of the Attention mechanism is

$$\text{Attention}(Q, K, V) = \text{softmax} \left( \frac{QK^T}{\sqrt{d_k}} \right) V,$$

where $Q$ represents the query, $K$ represents the key, and $V$ represents the value. In the self-attention mechanism, $Q$, $K$, and $V$ are all the same, and thus the correlation between each element in a sequence and other elements can be extracted using the self-attention mechanism. After being extracted by self-attention(SA) mechanism $f_{\text{slef-attention}} : \mathbb{R}^{X \times Y} \rightarrow \mathbb{R}^{X \times Y}$, two branches of features are flattened and spliced. $f_{\text{eeg-fusion}} : (\mathbb{R}^{X \times Y}, \mathbb{R}^{Y \times X}) \rightarrow \mathbb{R}^{D}$. Finally, fully connected layer and SoftMax layer are used to output classification results or EEG features.

3.3 Tongue Module (Long-term Physical State)

In the tongue module, we choose Vision Transformer (ViT) [7] to process the tongue images.

Data preprocessing: The key steps in the preprocessing stage include: tongue surface segmentation using SAM [13](only in our private datasets) and image enhancement techniques such as image rotation and horizontal flipping.

Architecture: The ViT model mainly includes three crucial layers: Patch Embedding Layer $f_{\text{patch_embed}} : \mathbb{R}^{H \times W \times C} \rightarrow \mathbb{R}^{(1+B) \times F}$, Transformer Encoder Layer $f_{\text{encoder}} : \mathbb{R}^{(1+B) \times F} \rightarrow \mathbb{R}^{(1+B) \times F}$ and Classification Head $f_{\text{classification}} : \mathbb{R}^{(1+B) \times F} \rightarrow \mathbb{R}^{N}$. The model architecture follows the original paper [7].

Pretrained works: This study proposes a 3-step pre-training scheme for the tongue module. Due to the limited number of tongue images, we employ a few-shot related meta-
learning approach to solve the problem. In the first step, we load the weights pretrained on ImageNet and freeze the ViT weights. The second step involves adding a trainable linear layer for feature extraction behind ViT, constructing a Siamese network, selecting datasets from Oxford-flower102[24], forming training pairs of (anchor, positive, negative), and utilizing triplet loss to differentiate between different classes of objects. In the third step, we use the tongue data not paired with EEG as additional training data to fine-tune the model and improve its ability to distinguish different tongue features. In the scenario where only the tongue modality is used, we skip the third step and adopt a 1:1 ratio of the support set to the query set for training and prediction.

3.4 Multimodal Module (Transient & Long-term Physical State)

In the multimodal module, we first validate the correlation between EEG signals and tongue images (see Table 1), and then achieve more comprehensive depression prediction by fusing the EEG features extracted by Trans_EEGNet and the tongue features extracted by ViT (see Table 2). In the validation of the correlation, we use Trans_EEGNet to predict the corresponding tongue features.

Table 1: Validation of Correlation

<table>
<thead>
<tr>
<th>Task</th>
<th>Private Acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tongue’s Color</td>
<td>96.36% ± 3.40%</td>
</tr>
<tr>
<td>Coating’s Color</td>
<td>91.81% ± 6.03%</td>
</tr>
<tr>
<td>Coating’s Thickness</td>
<td>100.00% ± 0.00%</td>
</tr>
<tr>
<td>Coating’s Greasiness</td>
<td>95.45% ± 6.03%</td>
</tr>
<tr>
<td>TCM’s Syndromes</td>
<td>97.27% ± 3.63%</td>
</tr>
</tbody>
</table>

Table 2: Different modules for depression diagnosis

<table>
<thead>
<tr>
<th>Task</th>
<th>Private Acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>EEG</td>
<td>97.27% ± 3.63%</td>
</tr>
<tr>
<td>Tongue</td>
<td>90.00%</td>
</tr>
<tr>
<td>Multimodal</td>
<td>98.18% ± 3.63%</td>
</tr>
</tbody>
</table>

Using the EEG and tongue modules, we extract the EEG features $I_{EEG} \in \mathbb{R}^F$ and the tongue features $I_{Tongue} \in \mathbb{R}^{F'}$ respectively. To fuse these two types of features, we can employ various fusion methods, such as concatenation, converting them into parallel channels to utilize the SA mechanism, deep canonical correlation analysis [3], and bilinear pooling [18].

4 Experiments

In our experiments, we have validated the accuracy of MMTV and its submodules on several datasets. More details on experimental setting, datasets, training setting, and comparison and visual results are discussed in the following subsections.

4.1 Experimental Setting

We designed an EEG and tongue image acquisition experiment for the analysis of depression. Depressed patients and healthy participants aged between 18 and 55 were recruited for the study. After evaluating the depression scores using the 17-item Hamilton Depression Scale (HAMD-17), the experiment was conducted. Tongue images were collected by students of TCM using the phone camera, ensuring clear tongue surface features. EEG signals were collected using the NeuSenE 16-channel version (with an actual sampling rate of 1000 Hz), and the experimental procedures were designed using E-prime3, divided into a Verbal Fluency Test (VFT) and an emotional face picture stimulation experiment. The experiments were performed in a quiet environment. In the VFT experiment, participants are asked to form as many words as possible using three Chinese characters within 20 seconds for each character. In the face affective picture stimulation experiment, subjects are asked to view 30 sets of positive and negative pictures. Written informed consent was obtained from all
participants, and the experiment conforms to the World Medical Association Declaration of Helsinki.

4.2 Datasets

The following datasets are used in this paper to train and corroborate model effects: the SEED emotional EEG dataset [48], the MODMA depression EEG dataset [4], and an open source tongue dataset on Baidu Paddle platform. Furthermore, we collect available tongue and EEG data from 2 depressed and 9 normal control subjects to train multimodal model of EEG and tongue fusion. Another 29 tongue images collected in the same period are also used in the tongue module. In our experiment, we use EEG data slices shaped as (110,16,74,5) from private dataset. The SEED dataset contains EEG records of 15 subjects during their watching of 15 movie clips covering positive, negative and neutral emotional categories, which includes 62 channels of EEG channels. In our experiment, we use one of the subjects to obtain samples shaped as (5076,4,8,9,5). The MODMA dataset contains EEG records from 24 patients with depression and 29 healthy controls. In our experiment, we gain samples shaped as (318,128,50,5) from MODMA dataset (128 channel, resting-state). The tongue dataset on the Baidu Paddle platform is labeled by cooperating Chinese medicine practitioners and has four types of labels: the tongue color (5 classes: 'pale white':'bluish or purple': 'crimson': 'red': 'pale red' = 118 : 70 : 80 : 506 : 1314), the color of the tongue coating (3 classes: 'gray or black':'white':'yellow'=91 : 1466 : 531), the thickness of the tongue coating (2 classes), and whether the tongue coating is greasy(2 classes). Our private dataset has four types of labels same as the above-mentioned tongue dataset and two additional labels: whether the person is a depressed patient and the TCM syndromes of the patient.

4.3 Implement Details

In our experiments, we have used keras and the paddlepaddle framework for model building and training. We use keras to build the EEG and multimodal models, set the learning rate of Trans_EEGNet and the corresponding models to 1.5e-4, (tuning in 1e-2 to 1e-6), training epochs to 100 rounds, set the optimizer to Adam, and the rest of the control experiments are performed under the same conditions on the A100 GPU. The learning rate in tongue model training is set to 5e-5, the training epochs are 200 rounds in the single modality, and the optimizer is Adam. The size of the CRNN-related convolution kernel changes slightly depending on the size of the 2D brain map [29]. All datasets, except for those used in the meta-learning models, are divided into training and testing sets in a 4:1 ratio. The information of other hyperparameters can be accessed in the codes.

4.4 Comparsion Results

In the EEG module, for the CRNN architecture, we obtain \( I_{\text{CRNN}} \in \mathbb{R}^{\text{Sequence} \times \text{X} \times \text{Y} \times \text{Bands}} \), for the Trans_EEGNet architecture, the input data is shaped as \( I_{\text{EEGNet}} \in \mathbb{R}^{\text{Channel} \times \text{Sequence} \times \text{OneBand}} \). Experiments in the study have shown that the band of alpha performs well. The sequence length of data slices in SEED is significantly shorter than the rest of the dataset, while the spatialization of channels is more appropriate than the rest of the datasets.

We have used Trans_EEGNet, CRNN and traditional machine learning methods for classification on SEED, MODMA and our private dataset, achieving the following results. In the SEED dataset, the CRNN model which focuses on extracting spatial features, achieves the best classification results, while in both MODMA and private datasets, our Trans_EEGNet, which extracts spatio-temporal features in combination, achieves the best results in 5foldCV.
simulation. Trans_EEGNet also achieves competitive results in the SEED dataset. It also demonstrates that Trans_EEGNet is more suitable for data with longer time sequences but poor spatialization (i.e., fewer or more EEG channels). We also list below the accuracy of the relevant dataset in other models in recent years.

![Table 3: Results of EEG Module](image)

<table>
<thead>
<tr>
<th>Model</th>
<th>SEED Acc</th>
<th>MODMA Acc</th>
<th>Private Acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRNN</td>
<td>91.23% ± 0.97%</td>
<td>54.51% ± 0.51%</td>
<td>74.54% ± 17.86%</td>
</tr>
<tr>
<td>MLP</td>
<td>78.23% ± 0.05%</td>
<td>60.72% ± 7.10%</td>
<td>90.80% ± 7.06%</td>
</tr>
<tr>
<td>SVM</td>
<td>78.29% ± 0.78%</td>
<td>94.99% ± 4.97%</td>
<td>91.81% ± 5.30%</td>
</tr>
<tr>
<td>LogisticRegression</td>
<td>76.81% ± 1.53%</td>
<td>96.21% ± 4.09%</td>
<td>96.36% ± 1.81%</td>
</tr>
<tr>
<td>DecisionTree</td>
<td>71.47% ± 2.52%</td>
<td>82.07% ± 3.06%</td>
<td>95.45% ± 2.87%</td>
</tr>
<tr>
<td>RandomForest</td>
<td>76.96% ± 0.95%</td>
<td>90.25% ± 6.08%</td>
<td>92.72% ± 4.63%</td>
</tr>
<tr>
<td>Trans_EEGNet (Ours)</td>
<td>88.02% ± 1.55%</td>
<td>99.05% ± 1.26%</td>
<td>97.27% ± 3.63%</td>
</tr>
</tbody>
</table>

![Table 4: Comparision on MODMA](image)

<table>
<thead>
<tr>
<th>Model</th>
<th>MODMA Acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRNN</td>
<td>91.23% ± 0.97%</td>
</tr>
<tr>
<td>MLP</td>
<td>78.23% ± 0.05%</td>
</tr>
<tr>
<td>SVM</td>
<td>78.29% ± 0.78%</td>
</tr>
<tr>
<td>LogisticRegression</td>
<td>76.81% ± 1.53%</td>
</tr>
<tr>
<td>DecisionTree</td>
<td>71.47% ± 2.52%</td>
</tr>
<tr>
<td>RandomForest</td>
<td>76.96% ± 0.95%</td>
</tr>
<tr>
<td>Trans_EEGNet (Ours)</td>
<td>88.02% ± 1.55%</td>
</tr>
</tbody>
</table>

In the tongue module, in our private dataset, we have conducted the tongue surface segmentation using SAM to exclude the rest of the image from interfering with the classification. We have also conducted the multi-step pre-training method described above, which didn’t show significant differences in the tongue modality but improved the accuracy of each method by 1-10% when used in the multimodal context. The results of the model associated with the tongue are shown in Table 5.

![Table 5: Results of Tongue Module](image)

<table>
<thead>
<tr>
<th>Task</th>
<th>Paddle Acc</th>
<th>Private Acc</th>
<th>Private(Seg) Acc</th>
<th>Private(Seg,Meta-learning) Acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tongue's Color</td>
<td>70.09%</td>
<td>75.00%</td>
<td>75.00%</td>
<td>73.33%</td>
</tr>
<tr>
<td>Coating's Color</td>
<td>79.57%</td>
<td>83.33%</td>
<td>100.00%</td>
<td>73.33%</td>
</tr>
<tr>
<td>Coating's Thickness</td>
<td>79.43%</td>
<td>75.00%</td>
<td>50.00%</td>
<td>80.00%</td>
</tr>
<tr>
<td>Coating's Greasiness</td>
<td>78.12%</td>
<td>75.00%</td>
<td>75.00%</td>
<td>70.00%</td>
</tr>
<tr>
<td>TCM’s Syndromes</td>
<td>None</td>
<td>100.00%</td>
<td>87.50%</td>
<td>88.89%</td>
</tr>
<tr>
<td>Depression</td>
<td>None</td>
<td>100.00%</td>
<td>87.50%</td>
<td>90.00%</td>
</tr>
</tbody>
</table>

We have used some multimodal fusion methods for model fusion and achieved the following results. Table 6 shows that the direct concat method has surpassed other fusion methods. The method performs well and shows great prediction results of the model.

![Table 6: Results of Multi-modal Module](image)

<table>
<thead>
<tr>
<th>Fusion Method</th>
<th>Private Acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parallel Channel</td>
<td>92.72% ± 6.16%</td>
</tr>
<tr>
<td>DCCA</td>
<td>85.45% ± 1.81%</td>
</tr>
<tr>
<td>Concat</td>
<td>98.18% ± 3.63%</td>
</tr>
<tr>
<td>Bilinear Pooling</td>
<td>97.27% ± 3.63%</td>
</tr>
</tbody>
</table>

4.5 Ablation Study and Visualization

4.5.1 Impacts of the ERP tasks

In our experiment, we designed the above-mentioned ERP tasks as we did not achieve the desired results with each model on the resting state MODMA dataset at the beginning of the project. We compared the accuracy of the data from two tasks on the Trans_EEGNet model. The data from VFT achieved 97.27% ± 3.63% accuracy, while the data from CFAPS achieved 92.22% ± 7.53% accuracy. Therefore, we chose the data from VFT as our primary dataset.

4.5.2 Impacts of the components of models

We disassembled the CRNN model as well as the Trans_EEGNet model. The two components of the CRNN model, CNN and LSTM, are used separately for prediction, and an attempt is made to replace the LSTM component with the rest of the temporal correlation neural network. Table 7 initially demonstrates that the advantage of the CRNN model lies in the two-dimensionalization of the channels and better learning of spatial features. In addition, we split Trans_EEGNet and utilized EEGNet, Bi_input EEGNet (a variant of EEGNet with dual-stream input [15]), and Single_EEGNet (a variant of EEGNet with only attention
mechanism) for comparison. This comparison aims to demonstrate the advantages of the self-attention mechanism and dual-stream input mechanism.

Table 7: Impacts of the components of CRNN

<table>
<thead>
<tr>
<th>Model</th>
<th>SEED Acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNN</td>
<td>90.01% ± 0.69%</td>
</tr>
<tr>
<td>LSTM</td>
<td>84.12% ± 1.09%</td>
</tr>
<tr>
<td>CRNN</td>
<td>91.23% ± 0.97%</td>
</tr>
<tr>
<td>CRNN-BiLSTM</td>
<td>91.01% ± 1.61%</td>
</tr>
<tr>
<td>CRNN-TCN</td>
<td>91.06% ± 2.08%</td>
</tr>
</tbody>
</table>

Table 8: Impacts of the proposed module of Trans_EEGNet

<table>
<thead>
<tr>
<th>Model</th>
<th>MODMA Acc</th>
<th>Private Acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>EEGNet</td>
<td>95.35% ± 3.31%</td>
<td>82.72% ± 1.81%</td>
</tr>
<tr>
<td>Bi_input EEGNet</td>
<td>96.55% ± 2.83%</td>
<td>95.18% ± 1.53%</td>
</tr>
<tr>
<td>Single_EEGNet</td>
<td>96.85% ± 2.80%</td>
<td>95.45% ± 4.07%</td>
</tr>
<tr>
<td>Trans_EEGNet (ours)</td>
<td>99.05% ± 1.26%</td>
<td>97.27% ± 3.63%</td>
</tr>
</tbody>
</table>

Table 9: Impacts of the pretrained model

<table>
<thead>
<tr>
<th>Model</th>
<th>Paddle Tongue Color Acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>vit_small_patch16_224[7]</td>
<td>70.09%</td>
</tr>
<tr>
<td>vit_base_patch16_384[7]</td>
<td>68.03%</td>
</tr>
<tr>
<td>SwinTransformer_tiny_window7_224[19]</td>
<td>66.83%</td>
</tr>
<tr>
<td>ResNet50[12]</td>
<td>68.27%</td>
</tr>
<tr>
<td>DeiT_tiny_patch16_224[35]</td>
<td>65.14%</td>
</tr>
</tbody>
</table>

4.5.3 Impacts of the pretrained model and visualization of the Tongue module

For tongue image classification, we use different deep learning models for tongue color prediction, from which we select the optimal vit_small_patch16_224. We use t-SNE[36] and attention rollout[1] techniques to visualize the tongue color classification results. The 768-dimensional semantic features of all the images in ViT are reduced to two dimensions to visualize the model classification as shown in the Figure 2(a). There are five categories of tongue color in the figure: 'pale white': '0', 'bluish or purple': '1', 'crimson': '2', 'red': '3', and 'pale red': '4'. Among them, the two categories of pale red tongue (4) and red tongue (3) are the main clusters and can be separated better, but there is still less confusion, which we think may be related to the fact that the dataset itself is obtained by crawlers and thus the light standard is not uniform and the number of samples for each classification in the dataset is not the same.

Figure 3: Visualization of the Tongue module

Take the tongue coating color classification as an example, the heat map we obtained by attention rollout is shown in Figure 3(b). The heat map reveals the regions of interest in the image, and the decreasing heat from red to blue indicates the model’s level of attention to those regions. From Figure 3, we can observe that the model’s judgment of coating color focuses on the central area of the tongue and partially correlates with the tongue boundary, which aligns with the subjective assessment of tongue diagnosis in TCM.

5 Conclusion

In this paper, we proposed a multimodal model based on our original Trans_EEGNet and ViT for depression diagnosis. The model combines EEG representing instantaneous brain
function information and tongue images representing long-term organ state of the body for depression diagnosis. In the EEG module, we proposed the Trans_EEGNet to better capture the spatio-temporal features of EEG. In the tongue module, the segmentation of the tongue surface was conducted to exclude the interference of the rest of the image. The multi-step pre-training scheme helped in gaining better pretrained weights. We completed tests on several datasets and analyzed the role of core mechanisms in each model by ablation. In addition, we analyzed the tongue diagnosis using visualization techniques such as t-SNE and attention rollout to provide interpretive insights to validate the TCM diagnosis.

The study has achieved relatively satisfactory results in the diagnosis of depression, and future directions include the collection of more data, in-depth study of various multimodal approaches, and the strengthening of explanatory work related to TCM.

6 Acknowledgements

This work was supported by the National Natural Science Foundation of China(No.62272021), with equipment and manpower support from Danyang Huichuang Medical Equipment Co., Ltd, Beijing University of Chinese Medicine(BUCM) and Nanchang Institute of Science and Technology.

References


