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 Kuang[1]
20373335@buaa.edu.cn
Yixuan Wang[1]
21373233@buaa.edu.cn
Huaiqian Ye[1]
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 EEGNet 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.

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 symdromes 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) [45] [44] and physiological signals (such as EEG, eye movements (EM), and ECG) [8] [9] to identify individuals with depression. Xie *et al.* [42] 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 [46] used TCN to analyze 5-second ECG segments, achieving a recognition accuracy of 97.1% in the depression diagnosis. Li *et al.* [17] 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 [28]. 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 [20] and the combination of speech and video [25].

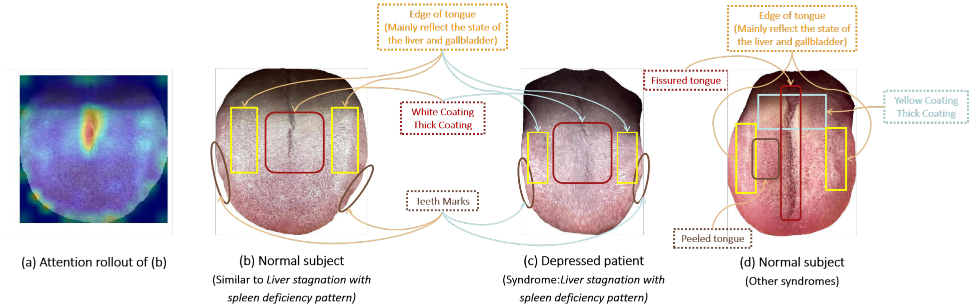
However, the signals mentioned earlier mainly reflect transient states and still have limitations in characterizing the long-term emotional lows associated with depression.

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 [2]. 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.* [16] 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.* [21] 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 [39]. 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 [43].To explain the relationship between tongue and depression better, the sample comparsion 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



(a) Attention rollout of (b)

(b) Normal subject (Similar to *Liver stagnation with spleen deficiency pattern)*

(c) Depressed patient (Syndrome:*Liver stagnation with spleen deficiency pattern)*

(d) Normal subject (Other syndromes)

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.

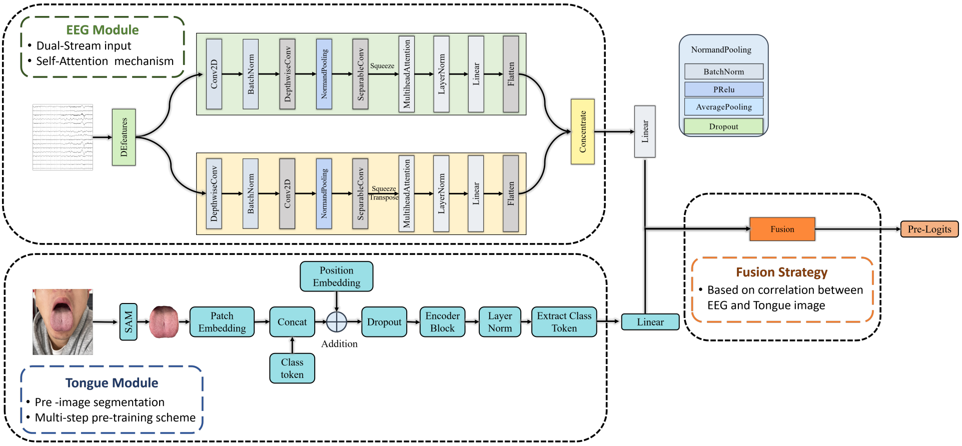


Figure 2: Overview of the MMTV

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 $f_{EEGNet} : \mathbb{R}^{C\times S\times 1} \to \mathbb{R}^{Classes}$ 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 $f_{EEGNet} : \mathbb{R}^{C\times S\times 1} \to \mathbb{R}^{S'\times C'}$ and the modified EEGNet $f_{modifiedEEGNet} : \mathbb{R}^{C\times S\times 1} \to \mathbb{R}^{C''\times S''}$. 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 $f_{self-attention} : \mathbb{R}^{X\times Y} \to \mathbb{R}^{X\times Y}$, two branches of features are flattened and spliced. $f_{eeg-fusion} : (\mathbb{R}^{X\times Y}, \mathbb{R}^{Y\times X}) \to \mathbb{R}^{D}$ 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 $f_{patch\_embed} : \mathbb{R}^{H\times W\times C} \to \mathbb{R}^{(1+B)\times F}$, Transformer Encoder Layer $f_{encoder} : \mathbb{R}^{(1+B)\times F} \to \mathbb{R}^{(1+B)\times F}$ and Classification Head $f_{classification} : \mathbb{R}^{(1+B)\times F} \to \mathbb{R}^{N}$. 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

Task	Private Acc
Tongue's Color	96.36% ± 3.40%
Coating's Color	91.81% ± 6.03%
Coating's Thickness	100.00% ± 0.00%
Coating's Greasiness	95.45% ± 6.03%
TCM's Syndromes	97.27% ± 3.63%

Table 2: Different modules for depression diagnosis

Task	Private Acc
EEG	97.27% ± 3.63%
Tongue	90.00%
Multimodal	98.18% ± 3.63%

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 $I_{CRNN} \in \mathbb{R}^{Sequence \times X \times Y \times Bands}$, for the Trans_EEGNet architecture, the input data is shaped as $I_{EEGNet} \in \mathbb{R}^{Channel \times Sequence \times OneBand}$. 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

Model	SEED Acc	MODMA Acc	Private Acc
CRNN [18]	**91.23% ± 0.97%**	54.71% ± 0.51%	74.54% ± 17.86%
MLP	78.23% ± 4.05%	60.72% ± 7.10%	90.90% ± 7.60%
SVM	78.29% ± 0.78%	94.99% ± 4.97%	91.81% ± 5.30%
LogisticRegression	76.81% ± 1.53%	96.21% ± 4.09%	96.36% ± 1.81%
DecisionTree	71.47% ± 2.52%	82.07% ± 3.06%	95.45% ± 2.87%
RandomForest	76.96% ± 0.95%	90.25% ± 6.08%	92.72% ± 4.63%
Trans_EEGNet(Ours)	88.02% ± 1.55%	**99.05% ± 1.26%**	**97.27% ± 3.63%**

Table 4: Comparision on MODMA

Model	MODMA Acc
Trans_EEGNet	**99.05% ± 1.26%**
MPA[27]	92.73%(LOSO)
mKTAChSel[30]	89.97%(LOSO)
GRL[31]	88.88%(10fold)
CNN-GRU-ATTN[11]	99.33%(9:1split)
TPTLP[34]	83.96%(LOSO)100%(10fold)
SparNet[1]	94.37%(LOSO)

In the tongue module, in our private dataset, we have conducted the tongue surface segmentation using SAM[13] 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

Task	Paddle Acc	Private Acc	Private(Seg) Acc	Private(Seg,Meta-learning) Acc
Tongue's Color	70.09%	75.00%	75.00%	73.33%
Coating's Color	79.57%	83.33%	100.00%	73.33%
Coating's Thickness	79.43%	75.00%	50.00%	80.00%
Coating's Greasiness	78.12%	75.00%	75.00%	70.00%
TCM's Syndromes	None	100.00%	87.50%	88.89%
Depression	None	100.00%	87.50%	90.00%

Table 6: Results of Multimodal Module

Fusion Method	Private Acc
Parallel Channel	92.72% ± 6.16%
DCCA	85.45% ± 1.81%
Concat	**98.18% ± 3.63%**
Bilinear Pooling	97.27% ± 3.63%

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.

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

Model	SEED Acc
CNN	90.01% ± 0.67%
LSTM	84.12% ± 1.10%
CRNN [[illegible]]	**91.23% ± 0.97%**
CNN-BiLSTM	91.01% ± 1.61%
CNN-TCN	91.06% ± 2.05%

Table 8: Impacts of the proposed module of Trans_EEGNet

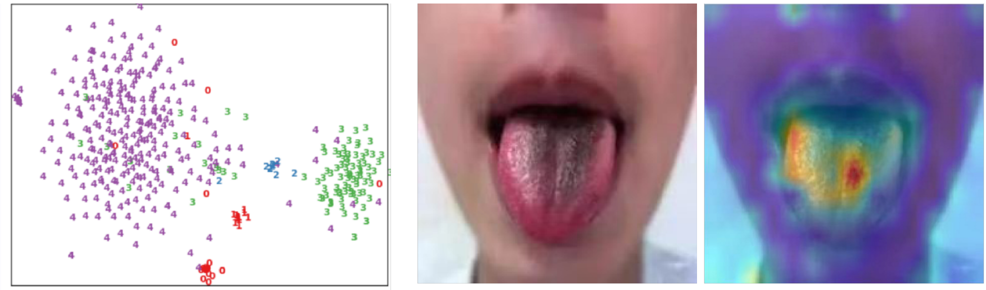
Model	MODMA Acc	Private Acc
EEGNet	95.59% ± 3.81%	82.72% ± 1.81%
Bi_input EEGNet	96.85% ± 2.83%	88.18% ± 1.81%
Single_EEGNet	96.85% ± 2.80%	95.45% ± 4.07%
Trans_EEGNet (ours)	**99.05% ± 1.26%**	**97.27% ± 3.63%**

Table 9: Impacts of the pretrained model

Model	Paddle Tongue Color Acc
Vit_small_patch16_224[[illegible]]	**70.09%**
Vit_base_patch16_384[[illegible]]	68.03%
SwinTransformer_tiny_window7_224[[illegible]]	66.83%
ResNet50[[illegible]]	68.27%
DeiT_tiny_patch16_224[[illegible]]	65.14%

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[[illegible]] 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.



(a) t-SNE embedding of the digits

(b) attention_rollout

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

[1] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4190–4197. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.385.

[2] Zhenyu An, Emiri Nonaka, Ren Wu, Mitsuru Nakata, and Qi-Wei Ge. Automatic diagnosis of tongue using mask-rcnn. In *2021 36th International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC)*, pages 1–4, 2021. doi: 10.1109/ITC-CSCC52171.2021.9501265.

[3] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *International conference on machine learning*, pages 1247–1255. PMLR, 2013.

[4] Hanshu Cai, Yiwen Gao, Shuting Sun, Na Li, Fuze Tian, Han Xiao, Jianxiu Li, Zhengwu Yang, Xiaowei Li, Qinglin Zhao, et al. Modma dataset: a multi-modal open dataset for mental-disorder analysis. *arXiv preprint arXiv:2002.09283*, 2020.

[5] En-Jui Chang, Abbas Rahimi, Luca Benini, and An-Yeu Andy Wu. Hyperdimensional computing-based multimodality emotion recognition with physiological signals. In *2019 IEEE International Conference on Artificial Intelligence Circuits and Systems (AICAS)*, pages 137–141. IEEE, 2019.

[6] Xin Deng, Xufeng Fan, Xiangwei Lv, and Kaiwei Sun. Sparnet: A convolutional neural network for eeg space-frequency feature learning and depression discrimination. *Frontiers in Neuroinformatics*, 16, 2022. doi: 10.3389/fninf.2022.914823.

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[8] Ruo-Nan Duan, Jia-Yi Zhu, and Bao-Liang Lu. Differential entropy feature for eeg-based emotion classification. In *2013 6th International IEEE/EMBS Conference on Neural Engineering (NER)*, pages 81–84. IEEE, 2013.

[9] Turker Tekin Erguzel, Serhat Ozekes, Oguz Tan, and Selahattin Gultekin. Feature selection and classification of electroencephalographic signals: an artificial neural network and genetic algorithm based approach. *Clinical EEG and neuroscience*, 46(4): 321–326, 2015.

[10] Jiang-Jian Guo, Rong Zhou, Li-Ming Zhao, and Bao-Liang Lu. Multimodal emotion recognition from eye image, eye movement and eeg using deep neural networks. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 3071–3074. IEEE, 2019.

[11] Keith Hawton, Carolina Casañas i Comabella, Camilla Haw, and Kate Saunders. Risk factors for suicide in individuals with depression: a systematic review. *Journal of affective disorders*, 147(1-3):17–28, 2013.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. doi: 10.1109/CVPR.2016.90.

[13] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

[14] Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung, and Brent J Lance. Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces. *Journal of neural engineering*, 15(5):056013, 2018.

[15] Chunbin Li. *EEG-based Emotion Recognition and Depression Diagnosis*. PhD thesis, Hefei University of Technology.

[16] Jun Li, Pei Yuan, Xiaojuan Hu, Jingbin Huang, Longtao Cui, Ji Cui, Xuxiang Ma, Tao Jiang, Xinghua Yao, Jiacai Li, et al. A tongue features fusion approach to predicting prediabetes and diabetes with machine learning. *Journal of biomedical informatics*, 115:103693, 2021.

[17] Mi Li, Lei Cao, Qian Zhai, Peng Li, Sa Liu, Richeng Li, Lei Feng, Gang Wang, Bin Hu, and Shengfu Lu. Method of depression classification based on behavioral and physiological signals of eye movement. *Complexity*, 2020:1–9, 2020.

[18] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. Bilinear cnn models for fine-grained visual recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 1449–1457, 2015.

[19] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[20] Yifei Lu, Wei-Long Zheng, Binbin Li, and Bao-Liang Lu. Combining eye movements and eeg to enhance emotion recognition. In *IJCAI*, volume 15, pages 1170–1176. Buenos Aires, 2015.

[21] Jiajiong Ma, Guihua Wen, Changjun Wang, and Lijun Jiang. Complexity perception classification method for tongue constitution recognition. *Artificial Intelligence in Medicine*, 96:123–133, 2019.

[22] Katja Machmutow, Ramona Meister, Alessa Jansen, Levente Kriston, Birgit Watzke, Martin Christian Härter, and Sarah Liebherz. Comparative effectiveness of continuation and maintenance treatments for persistent depressive disorder in adults. *Cochrane Database of Systematic Reviews*, (5), 2019.

[23] Mario Maj, Dan J Stein, Gordon Parker, Mark Zimmerman, Giovanni A Fava, Marc De Hert, Koen Demyttenaere, Roger S McIntyre, Thomas Widiger, and Hans-Ulrich Wittchen. The clinical characterization of the adult patient with depression aimed at personalization of management. *World Psychiatry*, 19(3):269–293, 2020.

[24] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729. IEEE, 2008.

[25] Mingyue Niu, Jianhua Tao, Bin Liu, Jian Huang, and Zheng Lian. Multimodal spatiotemporal representation for automatic depression level detection. *IEEE Transactions on Affective Computing*, 14(1):294–307, 2023. doi: 10.1109/TAFFC.2020.3031345.

[26] World Health Organization et al. Depression and other common mental disorders: global health estimates. Technical report, World Health Organization, 2017.

[27] Hong Peng, Chen Xia, Zihan Wang, Jing Zhu, Xin Zhang, Shuting Sun, Jianxiu Li, Xiaoning Huo, and Xiaowei Li. Multivariate pattern analysis of eeg-based functional connectivity: A study on the identification of depression. *IEEE Access*, 7:92630–92641, 2019.

[28] XueLiang Quan, ZhiGang Zeng, JianHua Jiang, YaQian Zhang, B Lu, and D Wu. Physiological signals based affective computing: A systematic review. *Acta Automatica Sinica*, 47(8):1769–1784, 2021.

[29] Fangyao Shen, Guojun Dai, Guang Lin, Jianhai Zhang, Wanzeng Kong, and Hong Zeng. Eeg-based emotion recognition using 4d convolutional recurrent neural network. *Cognitive Neurodynamics*, 14:815–828, 2020.

[30] Jian Shen, Xiaowei Zhang, Xiao Huang, Manxi Wu, Jin Gao, Dawei Lu, Zhijie Ding, and Bin Hu. An optimal channel selection for eeg-based depression detection via kernel-target alignment. *IEEE Journal of Biomedical and Health Informatics*, 25(7): 2545–2556, 2021. doi: 10.1109/JBHI.2020.3045718.

[31] Surbhi Soni, Ayan Seal, Anis Yazidi, and Ondrej Krejcar. Graphical representation learning-based approach for automatic classification of electroencephalogram signals in depression. *Computers in Biology and Medicine*, 145:105420, 2022. ISSN 0010-4825. doi: https://doi.org/10.1016/j.compbiomed.2022.105420.

[32] Michael Sung, Carl Marci, and Alex Pentland. Objective physiological and behavioral measures for identifying and tracking depression state in clinically depressed patients. *Massachusetts Institute of Technology Media Laboratory, Cambridge, MA, Tech. Rep. TR*, 595, 2005.

[33] Yulei Tao, Qian Ding, and Jianqiang Mei. Based on the cheng-zhi tiao-ping theory: Strategies for preventing and treating depression. *Hebei journal of traditional China medicine*, 44(1906-1910), 2022.

[34] Gulay Tasci, Hui Wen Loh, Prabal Datta Barua, Mehmet Baygin, Burak Tasci, Sengul Dogan, Turker Tuncer, Elizabeth Emma Palmer, Ru-San Tan, and U Rajendra Acharya. Automated accurate detection of depression using twin pascal's triangles lattice pattern with eeg signals. *Knowledge-Based Systems*, 260:110190, 2023.

[35] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021.

[36] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[38] Adrián Vázquez-Romero and Ascensión Gallardo-Antolín. Automatic detection of depression in speech using ensemble convolutional neural networks. *Entropy*, 22(6):688, 2020.

[39] Meng Wang and Yongxue Zhou. Origin and development of depression theory in traditional chinese medicine. *China journal of traditional Chinese medicine and pharmacy*, 37:1878–1881, 2022.

[40] Shoujing Wang. Investigation of gamma wave activity induced by emotional images in patients with depression and healthy individuals. Master's thesis, University of electronic science and technology of China, 2012.

[41] Zhuozheng Wang, Zhuo Ma, Wei Liu, Zhefeng An, and Fubiao Huang. A depression diagnosis method based on the hybrid neural network and attention mechanism. *Brain Sciences*, 12(7):834, 2022.

[42] Wanqing Xie, Chen Wang, Zhixiong Lin, Xudong Luo, Wenqian Chen, Manzhu Xu, Lizhong Liang, Xiaofeng Liu, Yanzhong Wang, Hui Luo, and Mingmei Cheng. Multimodal fusion diagnosis of depression and anxiety based on cnn-lstm model. *Computerized medical imaging and graphics : the official journal of the Computerized Medical Imaging Society*, 102:102128, 2022.

[43] Ying Xu, Yan Liu, and Lifang Yu. Variation regulation in anxiety depression patient's tongue. *Journal of Liaoning university of TCM*, 13:23–24, 2011.

[44] Le Yang, Dongmei Jiang, Xiaohan Xia, Ercheng Pei, Meshia Cédric Oveneke, and Hichem Sahli. Multimodal measurement of depression using deep learning models. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, pages 53–59, 2017.

[45] Le Yang, Hichem Sahli, Xiaohan Xia, Ercheng Pei, Meshia Cédric Oveneke, and Dongmei Jiang. Hybrid depression classification and estimation from audio video and text information. In *Proceedings of the 7th annual workshop on audio/visual emotion challenge*, pages 45–51, 2017.

[46] Xiaohan Zang. Depression recognition based on deep learning and ecg signal. Master's thesis, Shandong university, 2022.

[47] Qian Zhai, Lei Feng, and Guofu Zhang. Study of cognitive dysfunction in patients with depression. *Beijing medical journal*, 42(597-601), 2020.

[48] Wei-Long Zheng and Bao-Liang Lu. 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.