fNIRS-Driven Depression Recognition Based on Cross-Modal Data Augmentation

Kai Shao, Yanjie Liu, Yijun Mo, *Member, IEEE*, Qin Yang,
Yixue Hao, *Member, IEEE*, and Min Chen, *Fellow, IEEE*

***Abstract*— Early diagnosis and intervention of depression promote complete recovery, with its traditional clinical assessments depending on the diagnostic scales, clinical experience of doctors and patient cooperation. Recent researches indicate that functional near-infrared spectroscopy (fNIRS) based on deep learning provides a promising approach to depression diagnosis. However, collecting large fNIRS datasets within a standard experimental paradigm remains challenging, limiting the applications of deep networks that require more data. To address these challenges, in this paper, we propose an fNIRS-driven depression recognition architecture based on cross-modal data augmentation (fCMDA), which converts fNIRS data into pseudo-sequence activation images. The approach incorporates a time-domain augmentation mechanism, including time warping and time masking, to generate diverse data. Additionally, we design a stimulation task-driven data pseudo-sequence method to map fNIRS data into pseudo-sequence activation images, facilitating the extraction of spatial-temporal, contextual and dynamic characteristics. Ultimately, we construct a depression recognition model based on deep classification networks using the imbalance loss function. Extensive experiments are performed on the two-class depression diagnosis and five-class depression severity recognition, which reveal impressive results with accuracy of 0.905 and 0.889, respectively. The fCMDA architecture provides a novel solution for effective depression recognition with limited data.**

***Index Terms*— Depression recognition, functional near-infrared spectroscopy (fNIRS), cross-modal, data augmentation, pseudo-sequence.**

I. Introduction

DEPRESSION, characterized by slow thinking, reduced volitional activities and an enduringly depressed mental state, is a typical mental disorder. Moderate or severe depression presents serious somatic symptoms, even leading to suicide, which is one of the leading causes of avoidable pain and premature death worldwide [1], [2]. A scientific brief from the World Health Organization (WHO) underscores the gravity of the situation, revealing a 28% increase in the global incidence of depression, becoming a challenge to global public health and medical communities [3], [4]. Traditional diagnostic methods for depression are primarily based on self-reports during clinical interviews, behavior reports from relatives or friends and responses to standardized questionnaires [5], such as the Patient Health Questionnaire-9 (PHQ9) [6], Hamilton Depression Scale (HAMD) [7] and Beck Depression Inventory (BDI-II) [8]. However, these assessment methods are susceptible to variability in subjective ratings, resulting in inconsistent outcomes across diverse temporal or environments. As the number of depression patients increases, the reassessment of early diagnosis and follow-up treatment effects becomes limited and time-consuming. Therefore, it is urgent to develop an effective auxiliary diagnosis approach to enhance the accuracy and efficiency of depression diagnosis [9].

Brain science-based research is rapidly advancing and provides important support for the early diagnosis of depression, through utilizing brain imaging techniques [10], [11]. Currently, functional brain imaging techniques employed for depression diagnosis mainly include functional magnetic resonance imaging (fMRI) [12], electroencephalography (EEG) [13], [14], [15] and functional near-infrared spectroscopy (fNIRS) [16]. Shen et al. [14] proposed a regularization parameter-based improved feature extraction method to explore the intrinsic characteristics of highly complex and nonstationary EEG signals. Notably, fNIRS records the reactions of oxygenated hemoglobin (HbO) and deoxyhemoglobin (HbR), offering high spatial resolution, comfortable equipment wearing, non-invasive measurement, and low sensitivity to head and limb shaking. These characteristics make fNIRS suit-

Manuscript received 28 February 2024; revised 21 June 2024; accepted 8 July 2024. Date of publication 16 July 2024; date of current version 29 July 2024. This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 62176101, Grant 62276109, and Grant 62322205; and in part by Guangdong Basic and Applied Basic Research Foundation under Grant 2024A1515030017 and Grant 2024A1515011153. *(Kai Shao and Yanjie Liu are co-first authors.) (Corresponding author: Qin Yang.)*

Kai Shao and Yixue Hao are with the School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, Hubei 430074, China (e-mail: sk_sk@hust.edu.cn; yixuehao@hust.edu.cn).

Yanjie Liu is with the School of Computer Science and Technology, South China University of Technology, Guangzhou, Guangdong 510641, China (e-mail: yanjieliu_001@163.com).

Yijun Mo is with the School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, Hubei 430074, China, and also with the Hubei Specialized Institute of Intelligent Edge Computing, Wuhan, Hubei 430074, China (e-mail: moyj@hust.edu.cn).

Qin Yang is with the Wuhan Second Ship Design and Research Institute, Wuhan, Hubei 430205, China (e-mail: qinyangnaoe@163.com).

Min Chen is with the School of Computer Science and Engineering, South China University of Technology, Guangzhou, Guangdong 510640, China, and also with the Pazhou Laboratory, Guangzhou, Guangdong 510640, China (e-mail: minchen@ieee.org).

Digital Object Identifier 10.1109/TNSRE.2024.3429337

able for a broader range of applications in auxiliary depression diagnosis. It is noteworthy that prior researches [17], [18] have underscored the significance of HbO as a crucial indicator of cerebral blood flow changes and depressive disorders. Zhu et al. [17] extracted ten features from HbO signals as inputs to classification algorithms. However, these approaches exhibit limitations that heavily rely on the experience of researchers. Furthermore, the extracted features primarily focus on a single statistical index, failing to capture the deeper physiological information contained within fNIRS data.

In recent years, deep learning technology has demonstrated remarkable achievements across various fields, including data analysis [19], representation learning [20] and medical-assisted diagnosis [15]. Automatic feature representation based on deep learning promotes the development of depression diagnosis research. Wang et al. [21] proposed a transformer-based fNIRS classification network to explore spatial-level and channel-level representations of fNIRS signals to improve data utilization and network representation. Liu et al. [22] focused on stimulation tasks to investigate the advantages of fNIRS in cognitive activation. fNIRS data has been shown to reliably reflect cognitive profiles on the brain in different stimulation tasks [23], [24], and presents signal differences under different stimulation task time points [25]. Notably, the fNIRS data during the different stimulation points varies, which is crucial to consider the different time points of the stimulation task. When realizing the depression recognition based on fNIRS, the temporal dynamics characteristics of fNIRS data should be fully utilized.

Moreover, collecting fNIRS data poses challenges given limited medical resources and the prevailing stigma associated with patients. The employment of data augmentation methods emerges as a viable strategy to expand fNIRS data. Several studies have made data augmentation for the fNIRS features [26], [27]. Nagasawa et al. [28] proposed generative adversarial networks for fNIRS data augmentation to generate artificial fNIRS data. Woo et al. [29] used deep convolutional generative adversarial networks to expand fNIRS data to improve classification accuracy and training stability. While this method reduces the level of subjectivity and domain knowledge required for manual feature extraction, the generated data still needs to be validated through medical studies.

In this paper, we present an **f**NIRS-driven depression recognition architecture based on **C**ross-**M**odal **D**ata **A**ugmentation (fCMDA), as shown in Fig. 1. The fNIRS data is collected to make testers receive verbal fluency task with the designated stimulation task. Following data collection, we conduct preprocessing and transform the raw sequence data into the HbO concentration change data. Then, we propose a time-domain augmentation (TDA) method acting on the time dimension to generate more HbO data to support the training of the deep networks, including time warping and time masking. Meanwhile, we design a stimulation task-driven data pseudo-sequence (DPS) cross-modal conversion method to transform the sequence modal to the pseudo-sequence activation image modal. The sequence image reflects the degree and the dynamic characteristics of brain activation during the stimulation task. Finally, considering the class imbalance in the real collecting situation, we establish a depression recognition model based on the focal loss function. Benefiting from these, the experimental results show the proposed fCMDA achieves high-precision depression recognition. Furthermore, the fCMDA model can be extended to other physiological data. Utilizing physiological data acquired under a stimulation task, features are extracted at key time points of the stimulation task and pseudo-sequence images are generated based on channel locations for depression diagnosis. In summary, the main contributions of this paper include:

- We propose an fNIRS-driven depression recognition architecture based on cross-modal data augmentation, converting the fNIRS sequence modal into the pseudo-sequence activation image modal. It offers a novel depression recognition approach based on fNIRS to provide an objective and rapid auxiliary diagnosis.
- We design a time-domain augmentation and stimulation task-driven pseudo-sequence method to enrich data to leverage the spatial-temporal and dynamic characteristics of fNIRS data. It generates diverse and reliable data to improve the accuracy and robustness of the model.
- Extensive experiments are performed to validate the effectiveness of the fCMDA on the two-class depression diagnosis and five-class depression severity recognition. The results show the superiority of the proposed method for the advancement of depression recognition.

The remainder of this paper is organized as follows. Section II gives a concise review of the pertinent literature. Section III describes the fNIRS data collection and data preprocessing. We present the depression recognition model in Section IV. Implementation details, results and discussion of the depression recognition task are given in Section V. Finally, Section VI concludes this paper.

II. Related Works

This section briefly overviews existing fNIRS-driven depression recognition research and data augmentation for fNIRS.

A. fNIRS-Based Depression Recognition

A wide variety of machine learning methods are devoted to improving the performance of depression recognition via feature learning. Traditional methods usually rely on prior knowledge to extract hand-crafted feature representation. Chao et al. [18] extracted four statistic-based features from HbO signals and four vector-based features extracted from HbO and HbR, and applied them to depression recognition. However, feature representation based on statistics is struggling to reflect deep semantic information. In contrast to the aforementioned research, Wang et al. [21] proposed a deep learning classification network to explore spatial-level and channel-level feature representations of fNIRS signals. Similarly, Zhang et al. [30] achieved mild cognitive impairment recognition by exploiting the multidimensional features of fNIRS data including channel, temporal, and spatial features. Wang et al. [31] transformed fNIRS signals into 2-D wavelet feature maps by using wavelet transform and parallel-CNN

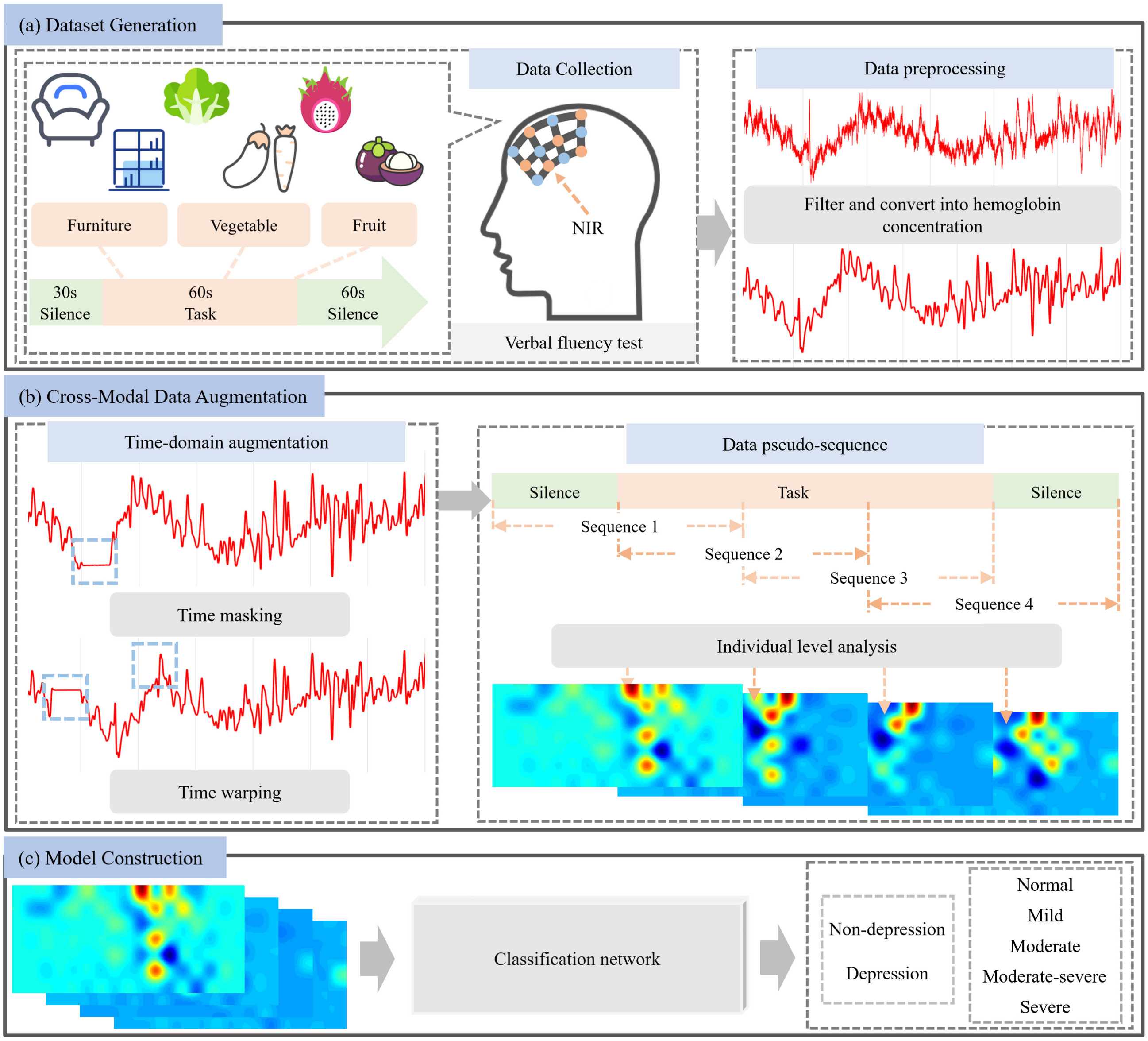


Fig. 1. The overview of the fCMDA architecture. (a) Data generation includes data collection and data preprocessing. (b) Introducing time-domain augmentation and data pseudo-sequence augmentation methods to generate rich data and map the hemoglobin concentration change data into sequence cognitive activation images. (c) Constructing the depression recognition architecture based on a classification network to realize the depression diagnosis and disease severity recognition.

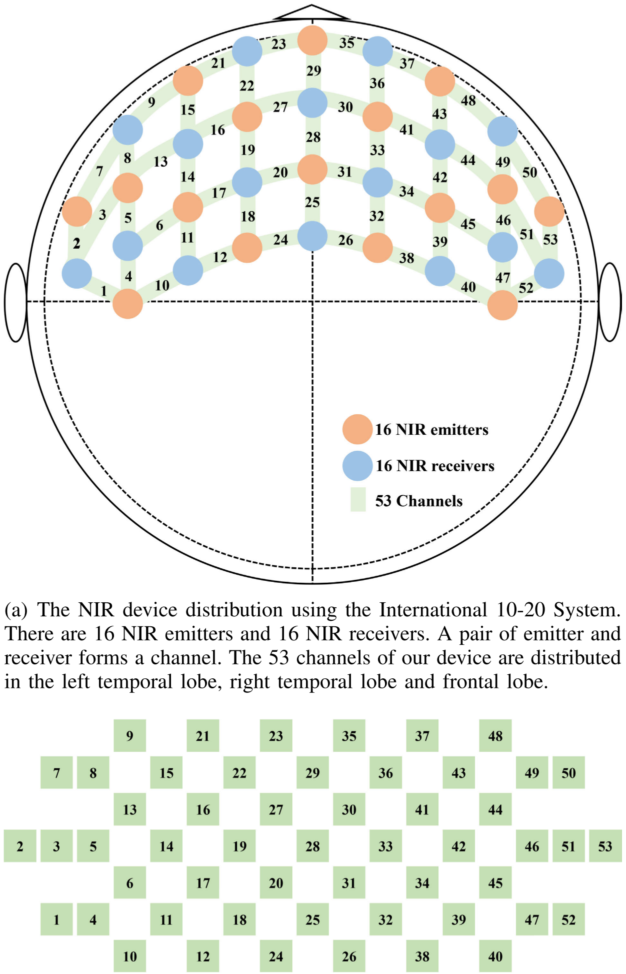
feature fusion to diagnose depressive disorder. Inspired by the work above, we transform fNIRS signals into activation images and utilize deep learning methods as the backbone network to implement depression recognition.

B. Data Augmentation for fNIRS

Data augmentation is an effective method to address the challenges of difficult access to fNIRS data and limited data. In recent years, research has mainly utilized generative adversarial networks to implement data augmentation. Nagasawa et al. [32] examined an fNIRS data augmentation method using Wasserstein generative adversarial networks. Wickramaratne et al. [33], [34] utilized conditional generative adversarial networks to generate artificial samples of a specific category to improve the classification accuracy when the sample size is insufficient. Zhang et al. [35] similarly employed a convolution-based conditional generative adversarial network for data augmentation. However, the expansion of fNIRS data based on the generative adversarial network ignores the data collection method and the dynamic change of the data with the stimulation task process. When implementing depression recognition using classification networks, the model may have difficulty focusing on the dynamic changes and deep semantic information of the data. Therefore, based on the fNIRS data collected with the stimulation task, we designed a cross-modal data enhancement method to mine the information of spatial-temporal, dynamic and brain activation semantic features.

III. Dataset Generation

The data used in this paper are from the Renmin Hospital of Wuhan University. The patients and volunteers participated in the fNIRS data collection under the guidance of doctors,



(a) The NIR device distribution using the International 10-20 System. There are 16 NIR emitters and 16 NIR receivers. A pair of emitter and receiver forms a channel. The 53 channels of our device are distributed in the left temporal lobe, right temporal lobe and frontal lobe.

(b) The 2D Channels distribution. The channels are used as a reference for generating the activation images.

Fig. 2. The fNIRS data collection device.

and they also received questionnaires evaluation and professional diagnosis to obtain the depression recognition results. This section describes in detail the data collection and data preprocessing, as shown in Fig. 1 (a).

A. Data Collection

fNIRS has become a widely utilized tool in both brain science research and clinical monitoring, facilitating the analysis of brain neural activity. Through the neurovascular coupling mechanism, when individuals engage in cognitive activities, the concentration of HbO in the blood of the active brain area increases, while the concentration of HbR decreases. fNIRS relies on the principle of the optical properties of biological tissues to enable the monitoring of changes in the hemoglobin concentration in the cerebral cortex [36]. These changes in hemoglobin concentration result in varying degrees of light intensity reduction. As hemoglobin concentration increases, the light intensity decreases more significantly. By examining changes in light intensity, we can acquire hemodynamic information related to neural activity [37]. The near-infrared (NIR) device utilized in this study comprises 16 NIR emitters and 16 NIR receivers, connecting into 53 channels, as shown in Fig. 2. These probes emit NIR light at wavelengths of 690 nm and 830 nm, enabling precise monitoring of neural activity.

During the data collection process, doctors assist participants in wearing a NIR device to ensure that the probe is tightly attached to the scalp until the channel pass rate exceeds 90%. As shown in Fig. 1 (a), the entire stimulation task spans a duration of 150s, including a pre-task silence period (30s), a task period (60s) and a post-task silence period (60s). During the silence periods, the participants need to sit up straight in front of the computer, remain calm and not shake their bodies. The task period involves participants responding to three questions displayed on the computer screen, prompting them to name the fruits, furniture and vegetables that they can associate. Each question has a 15s answer time followed by a 5s silence time. Throughout the entire test period, the NIR device continuously captures the intensity of the emitted light at two wavelengths at a sampling rate of 100hz. The data for each participant consists of $150 \times 100 \times 53 \times 2$, where 150 is the duration of the test, 100 is the data collection frequency, 53 is the number of channels and 2 is the number of wavelengths.

Participants comprise both male and female individuals, and they are enrolled using a non-randomized enrollment method under the guidance of a medical professional. The inclusion criteria for patients are: age 18-40 years; both first-onset and recurrence; previous cranial magnetic resonance imaging (MRI) within the past year, which ruled out organic brain lesions; no previous history of psychiatric illness; and right-handedness. The inclusion criteria for the control group are: age 18-40 years; cranial MRI with no organic brain lesions; no previous history of psychosis; and right-handedness. The common exclusion criteria for both patients and controls in the sample are: suffering from severe somatic diseases, including cardiovascular and cerebrovascular diseases, severe respiratory diseases, severe hepatic, renal, endocrine, and hematologic diseases, and malignant tumors; previous or existing psychotic disorders, alcohol or drug dependence, and definitively diagnosed cognitive impairments; and lactating and pregnant women. Under the guidance of professional doctors, combined with questionnaire evaluation and diagnostic results, the PHQ9 score serves as the class standard. For the depression diagnosis, participants with $0 \leq \text{PHQ9} \leq 9$ are considered as non-depression controls, while participants with $10 \leq \text{PHQ9} \leq 27$ are considered as depression controls. For the depression severity recognition, participants with $0 \leq \text{PHQ9} \leq 4$ are considered as normal controls, $5 \leq \text{PHQ9} \leq 9$ as mild, $10 \leq \text{PHQ9} \leq 14$ as moderate, $15 \leq \text{PHQ9} \leq 19$ as moderate-severe (mod-severe), and $20 \leq \text{PHQ9} \leq 27$ as severe depression. In total, fNIRS data are obtained for 96 participants as shown in Table I, including 17 with non-depression and 79 with depression. All personal information in the dataset is desensitized.

B. Data Preprocessing

During the collection of fNIRS data, noise is inevitable, including physiological information in different frequency bands such as heart rate, respiration, and typical motion noise. The band-pass filters are commonly utilized to remove physiological noise from cerebral oxygen signals and extract

TABLE I
DATA DISTRIBUTION OF FNIRS DATASET, WHERE NO. SAMPLE IS THE NUMBER OF SAMPLES

Depression diagnosis	Non-depression		Depression		
Depression severity recognition	Normal	Mild	Moderate	Mod-Severe	Severe
Number of subjects	12	5	25	25	29

hemoglobin low-frequency oscillatory signals. The preprocessing method of fNIRS has reached relative maturity and this study is based on near-infrared data analysis tools for data preprocessing [38]. The preprocessing steps begin with the elimination of motion artifacts unrelated to the raw data. Subsequently, the light intensity signal is converted into an optical density profile, which is then filtered using a band-pass filter to eliminate noise caused by physiological fluctuations such as pulse and respiration, as well as baseline drift caused by environmental and temperature changes. Finally, the optical density data are converted to concentration change of HbO and HbR using a modified Beer-Lambert method. Based on previous research [16], [17], [18], this study also deliberately focuses on the HbO concentration change data in subsequent method design, recognizing its relevance in capturing pertinent cognition variations associated with depression.

IV. AN FNIRS-DRIVEN DEPRESSION RECOGNITION ARCHITECTURE

The depression recognition architecture includes cross-modal data augmentation and recognition model construction, as shown in Fig. 1 (b) and (c). The cross-modal data augmentation methods include TDA and DPS, where TDA contains time masking and time warping.

A. Time-Domain Augmentation

Addressing the challenges stemming from limited sample size and class imbalance, data augmentation stands out as a viable solution. In particular, the temporal characteristics of fNIRS data and the spatial information related to detector placement in the generated activation image have inherent explanations. We aim to construct an augmentation policy directly impacting fNIRS data, thereby fostering the acquisition of more discriminative features by the network. Motivated by the goal that these features should be robust to deformations in the time direction, we draw inspiration from the augmentation methods in [39]. The time masking and time warping data augmentation methods act on the HbO concentration change data to enable the network to learn more meaningful and robust features. The different time points of stimulation tasks reflect the degree of brain activity. Therefore, according to the stimulation task of fNIRS data collection, we utilize the question start time of the task period as a segmentation point. The stimulation task consists of a pre-task silence period T_{pre}, a task period T, and a post-task silence period T_{aft}, where there will be N_q questions in the task period, with a total time of t_q for each question, and a rest time of t_r for each question. The two strategies are as follows.

1) Time Masking: The time step of the time masking method is $[t_0, t_0 + t_{tm}]$, where $t_0 \in [0, t_q)$ represents the start position of each problem period and the masking parameter $t_{tm} \in (0, \lambda]$, $\lambda \leq t_q$ denotes the time span, introducing an upper bound that the width of the time masking cannot be larger than the response time of each question, making the augmentation operation effective and facilitating the subsequent pseudo-sequence operation. From the change in hemoglobin concentration after time masking augmentation in Fig. 1 (b), it can be seen that the use of the time masking method is masking the data in the masking time step, which appears to be the unchanged value, and the before and after of the masked portion is discontinuous.

2) Time Warping: The time step of the time warping method is $[t_0, t_0 + t_{tw}]$, where $t_0 \in [0, t_q)$ represents the start position of each problem period and the warping parameter $t_{tw} \in (0, \lambda]$, $\lambda \leq t_q$ denotes the time span. From the change in hemoglobin concentration after time warping augmentation in Fig. 1 (b), it can be seen that the use of the time warping method is to lengthen as well as shorten the time scale, and the data before and after warping are continuous. To ensure that time masking and time warping processed data have the same latitude, so each time warping contains both times lengthening and shortening operations.

In this paper, the stimulation task of fNIRS data collection is verbal fluency test and its task period contains $N_q = 3$ questions, with a total time of $t_q = 15s$ per question. The time step parameter $t_0 = \{0s, 5s, 10s\}$, $t_{tm} = t_{tw} = 10s$ in the data augmentation method. The data augmentation methods all act on the time dimension. In particular, if the two augmentation methods act on the same data at the same time, it will cause a change in the data dimension or the failure of the methods. To avoid the data augmentation operation being utilized within the same problem period, time masking and time warping are applied to different question periods. Therefore, data augmentation includes three strategies: Time masking, Time warping, and Time masking with Time warping, and the three augmentation strategies are randomly applied to all sample data.

B. Data Pseudo-Sequence

In terms of data types, the generated HbO concentration change data emerges as multi-channel time series data. Traditional machine learning methods usually extract time-domain features such as kurtosis and skewness, or frequency-domain features represented by wavelet transform. Time-frequency features can only measure the overall characteristics of the data, making it challenging to extract deep information from physiological signals. Additionally, such methods lose the dynamic characteristics of fNIRS signals under task stimulation. Therefore, to exploit the feature learning capabilities of neural networks, we convert HbO concentration change data into cognitive activation images for depression recognition.

The size of the fNIRS data varies depending on the experimental paradigm of the stimulation task. To ensure uniformity in data analysis, we split the data into fixed-length time series

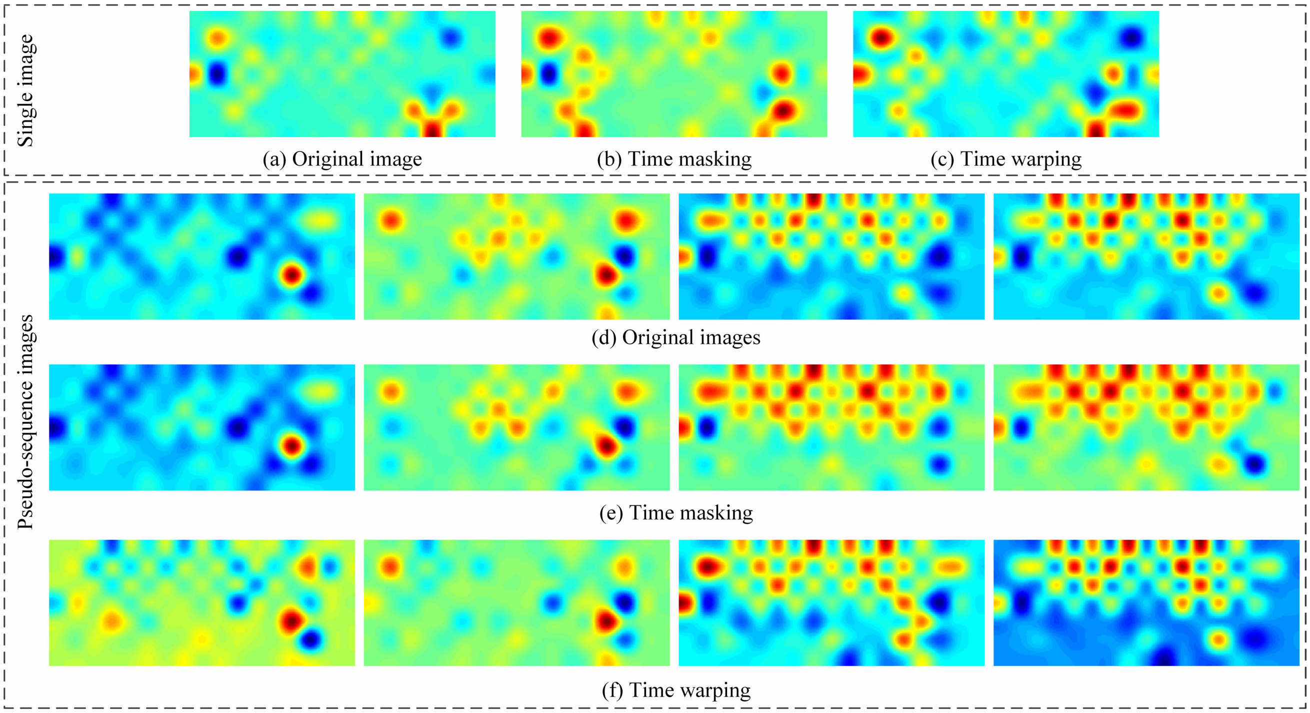


Fig. 3. Visualization of channel activation images based on the fCMDA. (a), (b) and (c) represent channel activation images generated without the pseudo-sequence method. (d), (e) and (f) show pseudo-sequence activation images, where (e) and (f) represent the application of time masking and time warping data augmentation methods, respectively. The activation image reflects the degree of activation of the brain region.

utilizing the time point of the stimulation task as the demarcation. This segmentation approach guarantees consistency in data analysis, with each time series encapsulating the pertinent aspects of the task. During the pre-task silence period and the post-task silence period, the continuous data of t_q duration at both ends of the task are selected to ensure that each segment of the segmented data has both temporal continuity and dynamic differences. The last t_q period of the pre-task silence period is denoted as t_{before}, and the first t_q period of the post-task silence period is denoted as t_{after}, then the first segment of the sequence data is the t_{before} duration data and the t_q duration data of the first question. The last segment of sequence data is the t_q duration data of the last problem and the t_{after} duration data, that is the data segments are $D_M = D_1(t_{before}, t_{q1}), D_2(t_{q1}, t_{q2}), \ldots, D_M(t_{qN}, t_{after})$, where $M = N_q + 1$ and the number of sequence data sets is one more than the number of stimulation task questions. There are three problems in our stimulation task experimental paradigm, which will generate four sequence data segments. Pseudo-sequence is realized by this segmentation method incorporating the overlapping selection strategy. This strategy serves to retain more time series information, enabling the model to capture more time series patterns and dynamic changes effectively. Moreover, the overlapping selection enhances the model's adaptability to noise and outliers in the data, ensuring a more robust representation of cognitive activities within the brain during stimulation tasks.

Based on the HbO concentration change data generated by the DPS method, the general linear model is utilized for individual-level analysis to test task-related neural activation:

$$X = G\beta + \varepsilon, \tag{1}$$

where X is the HbO concentration change data, G is the matrix generated by the convolution of the neural signal impulse and hemodynamic response functions function, and ε is the error matrix. β is the parameter to be estimated, indicating the activation value of each channel. Finally, based on the channel settings depicted in Fig. 2, the channel pseudo-sequence activation images are generated using the interpolation function.

To demonstrate the effectiveness of the fCMDA method, Fig. 3 displays the channel activation images. Fig. 3(a) shows the channel activation image without the pseudo-sequence method, and the red represents that the region has a large activation value. The exclusion of the influence of abnormal data indicates that the corresponding brain region has a large degree of activation. Notably, a red area in the lower right corner indicates a higher activation of the brain region corresponding to this channel. This channel is placed near the Brocas area of the brain which is a region of linguistic information processing and discourse production. Since the stimulation task we employed is a verbal fluency task in which participants are required to speak to answer, the activation value of this region is high. Fig. 3 (b) and (c) illustrate that the channel activation images employ the TDA method. The channel activation images exhibit greater differences among channels compared to the original image, which indicates that the TDA method does not change the semantic information of the data and expands the feature representations. Importantly, data augmentation methods influence activation values without changing the activation regions. Fig. 3 (d), (e) and (f) show the pseudo-sequence activation images utilizing the pseudo-sequence method. Each row displays four distinct activation images, highlighting varying activation levels for the same channels across different stimulation task problems.

Particularly, the middle and upper regions of the images are turning red in Fig. 3 (e), indicating an increase in activation values as the stimulation task progresses. This region of the image corresponds to the frontal lobe of the brain, which is the brain region that works with memory, attention, and emotional expression. In summary, the generated pseudo-sequence activation images contain both spatial-temporal and dynamic features of the channel, simultaneously reflecting differences in brain region activation levels corresponding to the channel under stimulation tasks.

C. Recognition Model Construction

The convolutional neural network (CNN) enables the capture of relationships of each channel at different time points and expands the receptive field with increasing convolutional layers. The learned features contain both the spatial-temporal and dynamic features of the entire sequence. In this paper, we establish the classification network based on CNN for depression recognition, with input comprising a set of activation images generated by the fCMDA method. The sequence images are mapped into a feature matrix through four 2D-CNN layers with a convolution kernel size of 3 and stride of 1. These feature matrices are then stitched together using a concatenation operation. Because different classification networks are constructed with different dimensions of their inputs, the transform reshape operation performs dimension adjustment. The resulting feature matrix is fed into the classification network for depression diagnosis and disease severity recognition. However, collecting medical data in a real environment often faces the problem of class imbalance, we utilize focal loss to construct the classification network. This strategy helps diminish the impact of class imbalance on classification accuracy and avoids undue bias towards a larger number of classes, ensuring a more robust model. The main idea of focal loss is to adjust the weights of the samples to pay more attention to the samples that are difficult to classify, thus mitigating the contribution of the easily classified samples to the loss, which is defined as follows:

$$FL(P) = -\alpha(1-P)^{\gamma}log(P), \tag{2}$$

where P denotes the predictive probability of the model, α is the weighting factor to balance the positive and negative samples, and γ is the adjustable parameter. The adjustment factor $(1-P)^{\gamma}$ can be adjusted adaptively according to the difficulty of the sample. In instances where samples are inherently easier to classify, the parameter P is larger, causing the adjustment factor tends to be zero. Consequently, this results in a reduced impact on the loss function, prompting the model to focus more on samples that are difficult to classify.

V. Experiments and Results

We conduct extensive experiments on depression diagnosis and disease severity recognition to validate our network. In this section, we first introduce the experiment details and then present the experimental results and comparisons with previous methods.

A. Experimental Setting

The data collection section comprehensively details the dataset and stimulation task settings. To evaluate the effectiveness of our method, we select ten baseline methods: Logistic Regression (LR), K-Nearest Neighbor (KNN), Support Vector Machine (SVM) [40], AlexNet [41], Residual Network (ResNet) [42], Random Forest (RF) [17], XGB [17], and the previous work of our group Corr-AlexNet [43], GCN [44] and Diffpool [44]. For the evaluation of depression diagnosis, the macro *Accuracy*, *Precision*, *Recall* and *F1-score* are used as evaluation indexes for the performance of the model. Then, we divided the data according to the depression severity and conducted experiments using the fCMDA method. For the evaluation of depression severity recognition, the *Accuracy*, *Macro-Precision*, *Macro-Recall*, *Macro-F1-score*and *Weighted-F1-score* are used as evaluation indexes for the performance of the model.

To evaluate the effectiveness of the fCMDA method, we select LeNet [45], ResNet18 [42], CNN-GRU and vision transformer (ViT) [46] as backbone networks to conduct ablation experiments. LeNet stands as a classic convolutional neural network with a total of seven layers, comprising convolutional, pooling and fully connected layers. ResNet18, a variant with 18 layers, employs 16 convolutional layers organized into four residual blocks. GRU, an improved model of the long short-term memory model, combines its forget gates and input gates into a single update gate. CNN-GRU incorporates two convolutional units which mainly consist of a convolution layer and pooling layer at the front of the two GRU basic units and can capture both spatial and temporal characteristics of data. Vision transformer (ViT) processes the input by dividing it into multiple patches, converting them into feature vectors using linear transformations, and adding position embedding vectors. The encoder block of ViT comprises multi-head self-attention (MSA) and MLP (two layers of fully connected neural network using GELU activation function). This paper adopts two transformer encoder blocks within the ViT architecture. For the training of deep network models, the Dropout layer, appropriate learning rate and focal loss function are adopted to ensure network convergence. The Adam optimizer is used to adaptively adjust different learning rates according to different parameters to complete the parameter updates.

B. Baseline Comparison

Extensive experiments are conducted to demonstrate the performance of the fCMDA method. We initially conduct a holdout cross-validation on a two-class depression diagnosis experiment, dividing the dataset into training, validation, and testing sets. Moreover, for different severity of depression, doctors will take different treatment options. We conduct a five-class depression severity recognition experiment to assist doctors in completing the diagnosis of the disease. However, since there are only 5 subjects in the Mild class, there is only 1 validation and 1 test data each if holdout cross-validation is utilized. Under such circumstances, the selection and evaluation of models would be subject to significant randomness,

TABLE II
EXPERIMENT RESULTS WITH THE BASELINE METHOD

Method	Accuracy	Precision	Recall	F1-score
LR	0.813	0.300	0.583	0.355
KNN	0.729	0.188	0.219	0.188
SVM [40]	0.823	0.000	0.000	0.000
AlexNet [41]	0.830	0.790	0.830	0.800
ResNet [42]	0.720	0.670	0.720	0.700
RF [17]	0.833	0.625	0.175	0.267
XGB [17]	0.833	0.525	0.413	0.446
Corr-AlexNet [43]	**0.900**	**0.910**	**0.900**	**0.880**
GCN [44]	0.854	0.700	0.488	0.563
Diffpool [44]	0.875	0.750	0.475	0.571
fCMDA	**0.905**	**0.889**	**0.929**	**0.899**

TABLE III
ABLATION EXPERIMENT RESULTS USING TIME-DOMAIN AUGMENTATION (TDA) AND DATA PSEUDO-SEQUENCE (DPS) METHODS

Method	TDA	DPS	Accuracy	Precision	Recall	F1-score
LeNet			0.762	0.637	0.662	0.646
	✓		0.810	0.691	0.691	0.691
		✓	0.810	0.767	0.767	0.767
	✓	✓	**0.857**	**0.769**	**0.816**	**0.788**
CNN-GRU			0.667	0.389	0.412	0.400
	✓		0.714	0.395	0.441	0.417
		✓	0.714	0.395	0.441	0.417
	✓	✓	0.762	0.583	0.566	0.571
ViT			0.714	0.395	0.441	0.417
	✓		0.714	0.556	0.537	0.537
		✓	0.762	0.583	0.566	0.571
	✓	✓	0.810	0.405	0.500	0.447
ResNet18			0.762	0.583	0.566	0.571
	✓		0.762	0.637	0.662	0.646
		✓	0.810	0.767	0.767	0.767
	✓	✓	**0.905**	**0.889**	**0.929**	**0.899**

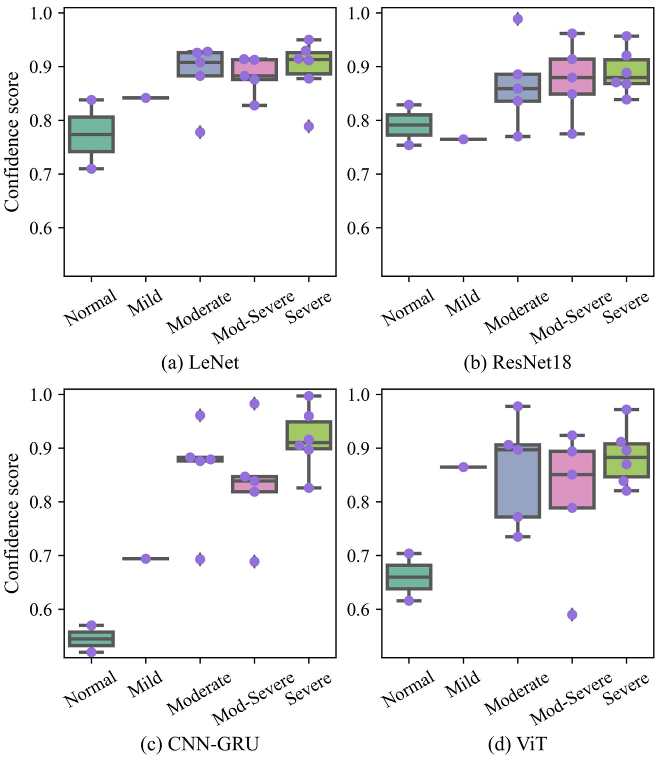
potentially leading to inconclusive and unconvincing results. Therefore, for the depression severity recognition experiment, we adopt a 5-fold cross-validation strategy. This method divides the data into five equal subsets, each serving as the testing set once, while the remaining four subsets are collectively utilized as the training set.

1) Experiment on Depression Diagnosis: Our experimental results utilize the fCMDA method with the ResNet18 classification network as the backbone network, incorporating the focal loss function. As shown in Table II, the color scheme emphasizes the outcomes, with red denoting the best result and blue indicating the second-best. Our proposed method achieves satisfactory results in terms of accuracy, precision, recall and F1-score. The accuracy of machine learning methods such as LR, KNN and SVM is relatively low, highlighting the superior performance of deep learning classification algorithms. While the Corr-AlexNet method achieves higher precision, it relies on hand-extracted features for network learning, lacking a profound exploration of dynamic features. Patients experiencing depression often manifest low mood and slow thinking. During performing stimulation tasks, the degree of brain activation in depressed individuals differs from that of non-depression controls. The activation image reflects the brain activation of the subjects in the stimulation task stage and the dynamic changes of activation, which can be used to distinguish patients with depression from the control group. According to the stimulation point of the stimulation task, we utilize the fCMDA method to generate pseudo-sequence activation images, enabling the network to learn spatial-temporal features and pay attention to the degree of activation at different stimulation time points. Moreover, our proposed fCMDA method proves effective even with small-scale datasets. Considering the challenge of class imbalance in the real diagnosis and treatment environment, we employ the focal loss function to construct the classification network, enhancing the robustness of the network to achieve higher recognition accuracy.

TABLE IV
EXPERIMENTAL RESULTS ON DEPRESSION SEVERITY RECOGNITION

Method	Accuracy	Macro-Precision	Macro-Recall	Macro-F1-score	Weighted-F1-score
LeNet	**0.889**	0.927	**0.920**	**0.920**	**0.888**
ResNet18	**0.889**	**0.943**	**0.920**	0.917	0.884
CNN-GRU	0.688	0.790	0.740	0.737	0.682
ViT	0.813	0.650	0.693	0.643	0.791

2) Experiment on Depression Severity Recognition: As shown in Table IV, the four networks all utilize fCMDA methods and focal loss, and the selection of LeNet and ResNet18 as backbone networks achieves an accuracy of 0.889. Evaluation indicators reveal that LeNet exhibits the best classification results. Similarly, the performance of LeNet and ResNet18 surpasses that of CNN-GRU and ViT. In the depression severity recognition task, given the limited data in the five classes, shallow networks such as LeNet and ResNet prove more adept at learning the differential characteristics of the data. Fig. 4 shows the confidence scores for one set of test data in the 5-fold cross-validation experiment of the four networks. The composition of the test data for depression severity recognition is {Normal, Mild, Moderate, Mod-Severe, Severe}={2, 1, 5, 5, 6}. The purple dots represent the confidence scores distribution, where the mild depression class is a horizontal line because the test data is only one. It can be seen that the models based on LeNet and ResNet18 show satisfactory results, with most confidence scores exceeding 0.8 and recognition accuracy surpassing 0.8 as well. We utilize the focal loss to avoid the network focusing on data-rich classes, addressing the challenge of class imbalance. It is noteworthy that the cognitive and brain activity of individuals with mild depression are not significantly different from those of the normal control group. Despite the low confidence scores of all networks for the normal and mild classes, the model is still capable of distinguishing between these two classes.



(a) LeNet

(b) ResNet18

(c) CNN-GRU

(d) ViT

Fig. 4. Confidence score distribution of the four networks on depression severity recognition experiment. The purple dots represent the test data.

Overall, depression severity recognition can assist doctors in completing accurate diagnoses.

C. Ablation Analysis

We validate the proposed fCMDA method as shown in Table III, where TDA means the utilization of the TDA method, DPS denotes the utilization of the DPS method and TDA with DPS represents the results using the fCMDA method. When applied to the ResNet18 network, our proposed method achieves the best scores, with accuracy, precision, recall and F1-score reaching 0.905, 0.889, 0.929 and 0.941, respectively. Initially, when a single activation image is fed into four networks, the accuracy of ResNet18 reaches 0.762, verifying the utility of the activation image in distinguishing between non-depression controls and patients with depression. Only relying on a single activation image for depression diagnosis, the recognition accuracy of the four networks is limited due to the small amount of data and fewer useful features. Using only the TDA method, a single activation image is also fed into the network. When using the TDA or DPS method, the classification performance of the network and the evaluation index exhibit improvement. The fCMDA method, incorporating both augmentation methods, demonstrates optimal performance across all four networks. Furthermore, the fCMDA method has different degrees of improvement in classification performance for the four networks. The classification accuracy of LeNet and ResNet18 is improved by 20%. For convolutional neural networks such as LeNet and ResNet18, the TDA method helps the network learn the channel and temporal features of fNIRS data. The implementation of the DPS method takes into account the differences in data at different time points and the dynamic changes on the time scale. Therefore, the fCMDA method enhances the network's ability to extract more effective features. Overall, complex networks such as CNN-GRU and ViT are not conducive to learning the features of the data for relatively simple and small-scale data compared to natural images, dispelling the notion that deeper and more complex networks inherently yield better results. In conclusion, our proposed fCMDA method can be applied to fNIRS-like data. Moreover, excellent algorithms from the visual field can be introduced into the study of depression recognition.

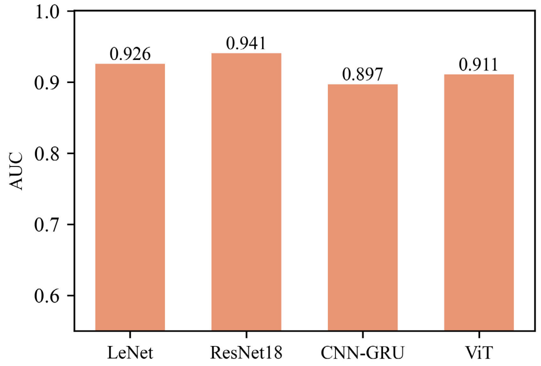
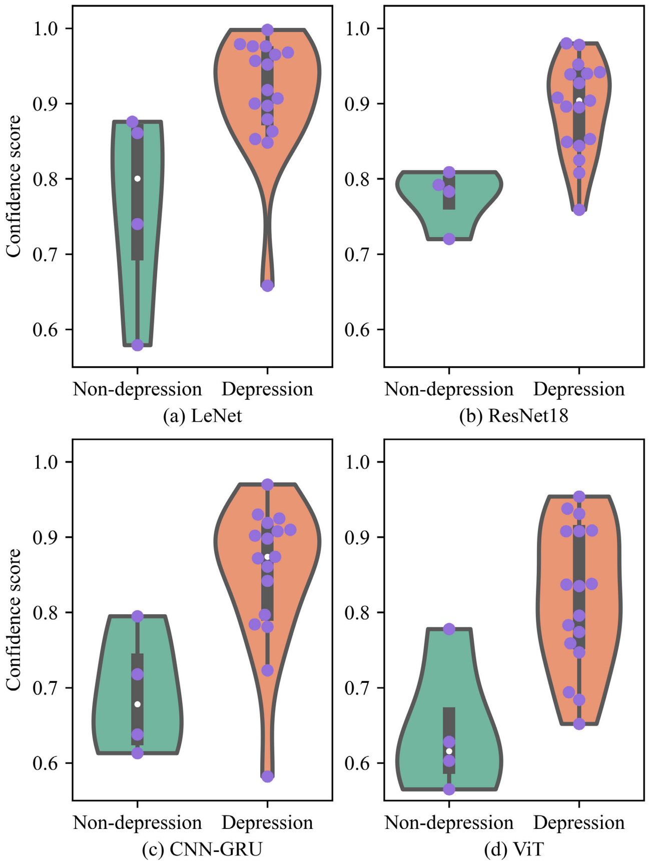


Fig. 5. The area under the curve results in a depression diagnosis experiment.

In this study, we calculate the area under the curve (AUC) based on the classification results of LeNet, ResNet18, CNN-GRU and ViT utilizing the fCMDA method, as shown in Fig. 5. All algorithms show satisfactory classification performance, among which ResNet18 performs the best experiment results with an AUC of 0.97. Notably, the AUC values for all methods surpass 0.8, which indicates that the fCMDA method combined with the classification network has excellent classification ability. The AUC value is influenced by the prediction results of the network on the test data. Fig. 6 shows the confidence scores distribution of test data of the four networks, where the purple dots represent the test data. It can be seen that ResNet18 has high confidence scores for all test data compared to other networks. However, all networks exhibit relatively low confidence scores for the non-depression class. Despite the incorporation of focal loss to mitigate the impact of class imbalance during training, the substantial difference in data volume between non-depression and depression classes affects the confidence scores. ViT, while not achieving high confidence scores, demonstrates a more concentrated score distribution, indicative of a more stable model. Overall, the data augmentation approach serves as an expansion of the data volume dimension, which facilitates the network to learn the differences between different data and the similarity of similar data. The pseudo-sequence method, on the other hand, fully utilizes the data feature, enabling the network to learn a characteristic particularly beneficial for ResNet18. In short, our proposed fCMDA method significantly improves the recognition ability of the network.

D. Parameter Analysis

To investigate the fCMDA model further, we analyze in detail the impact of the time step parameter of the data



(a) LeNet (b) ResNet18 (c) CNN-GRU (d) ViT

Fig. 6. Confidence score distribution of the four networks on depression diagnosis experiment. The purple dots represent the test data.

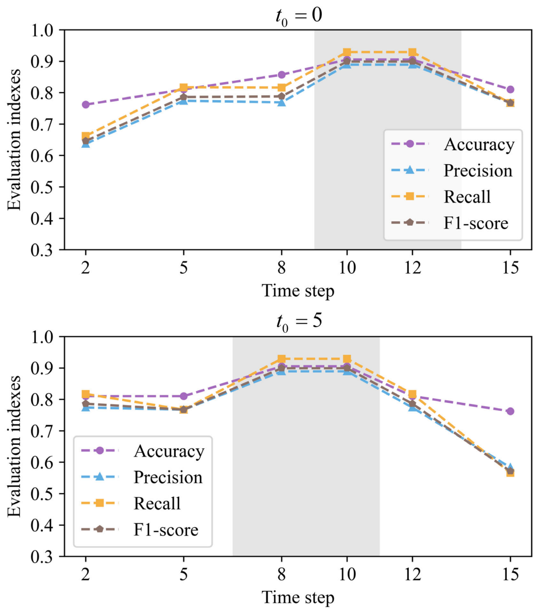


Fig. 7. Performance of fCMDA model with the different number of t_{tm}/t_{tw}. The shadow part represents the superior performance.

augmentation method on performance in this section. To avoid the data augmentation operation crossing the two problems period of the stimulation task, we set the time step parameter $t_0 = \{0, 5\}$, $t_{tm} = t_{tw} = \{2, 5, 8, 10, 12, 15\}$ and ResNet18 as the backbone network, and the experimental results are shown in Fig. 7. Setting $t_{tm} = t_{tw}$ ensures that the data dimension is consistent with the raw data when implementing TDA. With t_{tm}/t_{tw} set small, the classification ability of the fCMDA model is about the same as when utilizing only the DPS method. When the t_{tm}/t_{tw} is large or even across the problem period, it may cause the data to lose important information. When $t_0 = 5$ and $t_{tm} = t_{tw} = 15$, the TDA method operates on one problem period, while the entire stimulation task has only three problem periods. This results in a classification accuracy of 0.810. As shown in Fig. 7, the time step parameters represented by the shadow part will promote the model to achieve optimal performance.

VI. Conclusion

By observation of fNIRS data, it is found that depression patients exhibit diminished brain function and reduced brain function activation during cognitive stimulation tasks. In this paper, we propose a novel fNIRS-based depression recognition architecture fCMDA. We devise a cross-modal strategy that transforms fNIRS data into cognitive activation images corresponding to the stimulation task, thereby reflecting the degree of brain function activation in participants. First, we introduce time masking and time warping data augmentation methods to enrich the data. Subsequently, we develop a stimulation task-driven pseudo-sequence approach that focuses on the activation level of the brain at distinct stimulation points. The resulting pseudo-sequence data undergoes mapping to pseudo-sequence activation images through individual-level analysis, comprehensively considering spatial-temporal features. Finally, a neural network model for depression recognition is established, incorporating a class imbalance loss function to effectively address the challenges of imbalanced class distribution. Experimental results indicate that the depression diagnosis model based on ResNet18 demonstrates high accuracy. This research and the resultant recognition model hold significant implications for clinical aid in depression diagnosis.

References

[1] Y. Q. Lee, G. W. N. Tay, and C. S. H. Ho, “Clinical utility of functional near-infrared spectroscopy for assessment and prediction of suicidality: A systematic review,” *Frontiers Psychiatry*, vol. 12, Oct. 2021, Art. no. 716276.

[2] J. Shen, X. Zhang, G. Wang, Z. Ding, and B. Hu, “An improved empirical mode decomposition of electroencephalogram signals for depression detection,” *IEEE Trans. Affect. Comput.*, vol. 13, no. 1, pp. 262–271, Jan. 2022.

[3] *Wake-Up Call to All Countries to Step Up Mental Health Services and Support*, World health Org., Geneva, Switzerland, 2022.

[4] M. Taquet, E. A. Holmes, and P. J. Harrison, “Depression and anxiety disorders during the COVID-19 pandemic: Knowns and unknowns,” *Lancet*, vol. 398, no. 10312, pp. 1665–1666, Nov. 2021.

[5] M. Li, J. Zhang, J. Song, Z. Li, and S. Lu, “A clinical-oriented non-severe depression diagnosis method based on cognitive behavior of emotional conflict,” *IEEE Trans. Computat. Social Syst.*, vol. 10, no. 1, pp. 131–141, Feb. 2023.

[6] K. Kroenke and R. L. Spitzer, “The PHQ-9: A new depression diagnostic and severity measure,” *Psychiatric Ann.*, vol. 32, no. 9, pp. 509–515, Sep. 2002.

[7] M. Hamilton, “The Hamilton rating scale for depression,” in *Assessment of Depression*. Springer, 1986, pp. 143–152.

[8] L. He, C. Guo, P. Tiwari, H. M. Pandey, and W. Dang, "Intelligent system for depression scale estimation with facial expressions and case study in industrial intelligence," *Int. J. Intell. Syst.*, vol. 37, no. 12, pp. 10140–10156, Dec. 2022.

[9] G. S. Malhi and J. J. Mann, "Depression," *The Lancet*, vol. 392, no. 10161, pp. 0140–6736, 2018.

[10] B. Vai et al., "Predicting differential diagnosis between bipolar and unipolar depression with multiple kernel learning on multimodal structural neuroimaging," *Eur. Neuropsychopharmacology*, vol. 34, pp. 28–38, May 2020.

[11] Y. Wei et al., "Functional near-infrared spectroscopy (fNIRS) as a tool to assist the diagnosis of major psychiatric disorders in a Chinese population," *Eur. Arch. Psychiatry Clin. Neurosci.*, vol. 271, no. 4, pp. 745–757, Jun. 2021.

[12] Z. Zhu, Z. Zhen, X. Wu, and S. Li, "Estimating functional connectivity by integration of inherent brain function activity pattern priors," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 18, no. 6, pp. 2420–2430, Nov. 2021.

[13] J. Shen et al., "An optimal channel selection for EEG-based depression detection via kernel-target alignment," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 7, pp. 2545–2556, Jul. 2021.

[14] J. Shen et al., "Exploring the intrinsic features of EEG signals via empirical mode decomposition for depression recognition," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 356–365, 2022.

[15] J. Shen et al., "Depression recognition from EEG signals using an adaptive channel fusion method via improved focal loss," *IEEE J. Biomed. Health Informat.*, vol. 27, no. 7, pp. 3234–3245, Jul. 2023.

[16] J. Han, J. Lu, J. Lin, S. Zhang, and N. Yu, "A functional region decomposition method to enhance fNIRS classification of mental states," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 11, pp. 5674–5683, Nov. 2022.

[17] Y. Zhu et al., "Classifying major depressive disorder using fNIRS during motor rehabilitation," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, pp. 961–969, 2020, doi: 10.1109/TNSRE.2020.2972270.

[18] J. Chao et al., "FNIRS evidence for distinguishing patients with major depression and healthy controls," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 2211–2221, 2021.

[19] G. Li, C. H. Lee, J. J. Jung, Y. C. Youn, and D. Camacho, "Deep learning for EEG data analytics: A survey," *Concurrency Comput. Pract. Exper.*, vol. 32, no. 18, p. e5199, Sep. 2020.

[20] Y. Zhang, J. Shen, R. Zhang, and Z. Zhao, "Network representation learning via improved random walk with restart," *Knowl.-Based Syst.*, vol. 263, Mar. 2023, Art. no. 110255.

[21] Z. Wang, J. Zhang, X. Zhang, P. Chen, and B. Wang, "Transformer model for functional near-infrared spectroscopy classification," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 6, pp. 2559–2569, Jun. 2022.

[22] J. Liu, T. Song, Z. Shu, J. Han, and N. Yu, "FNIRS feature extraction and classification in grip-force tasks," in *Proc. IEEE Int. Conf. Robot. Biomimetics (ROBIO)*, Dec. 2021, pp. 1087–1091.

[23] S. Midha, H. A. Maior, M. L. Wilson, and S. Sharples, "Measuring mental workload variations in office work tasks using fNIRS," *Int. J. Hum.-Comput. Stud.*, vol. 147, Mar. 2021, Art. no. 102580.

[24] G. Rocco, J. Lebrun, O. Meste, and M.-N. Magnié-Mauro, "A chiral fNIRS spotlight on cerebellar activation in a finger tapping task," in *Proc. 43rd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Nov. 2021, pp. 1018–1021.

[25] C.-L. Yu, H.-C. Chen, Z.-Y. Yang, and T.-L. Chou, "Multi-time-point analysis: A time course analysis with functional near-infrared spectroscopy," *Behav. Res. Methods*, vol. 52, no. 4, pp. 1700–1713, Aug. 2020.

[26] S. Moon, S.-E. Moon, and J.-S. Lee, "Resting-state fNIRS classification using connectivity and convolutional neural networks," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Oct. 2022, pp. 1724–1729.

[27] D. Ma, M. Izzetoglu, R. Holtzer, and X. Jiao, "Deep learning based walking tasks classification in older adults using fNIRS," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 3437–3447, 2023.

[28] T. Nagasawa, T. Sato, I. Nambu, and Y. Wada, "Improving fNIRS-BCI accuracy using GAN-based data augmentation," in *Proc. 9th Int. IEEE/EMBS Conf. Neural Eng. (NER)*, Mar. 2019, pp. 1208–1211.

[29] S.-W. Woo, M.-K. Kang, and K.-S. Hong, "Classification of finger tapping tasks using convolutional neural network based on augmented data with deep convolutional generative adversarial network," in *Proc. 8th IEEE RAS/EMBS Int. Conf. Biomed. Robot. Biomechatronics*, Nov. 2020, pp. 328–333.

[30] C. Zhang et al., "Comparing multi-dimensional fNIRS features using Bayesian optimization-based neural networks for mild cognitive impairment (MCI) detection," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 1019–1029, 2023.

[31] G. Wang et al., "The diagnosis of major depressive disorder through wearable fNIRS by using wavelet transform and parallel-CNN feature fusion," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–11, 2023.

[32] T. Nagasawa, T. Sato, I. Nambu, and Y. Wada, "FNIRS-GANs: Data augmentation using generative adversarial networks for classifying motor tasks from functional near-infrared spectroscopy," *J. Neural Eng.*, vol. 17, no. 1, Feb. 2020, Art. no. 016068.

[33] S. D. Wickramaratne and M. S. Mahmud, "LSTM based GAN networks for enhancing ternary task classification using fNIRS data," in *Proc. 43rd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Nov. 2021, pp. 1043–1046.

[34] S. D. Wickramaratne and M. S. Mahmud, "Conditional-GAN based data augmentation for deep learning task classifier improvement using fNIRS data," *Frontiers Big Data*, vol. 4, Jul. 2021, Art. no. 659146.

[35] Y. Zhang, D. Liu, T. Li, P. Zhang, Z. Li, and F. Gao, "CGAN-RIRN: A data-augmented deep learning approach to accurate classification of mental tasks for a fNIRS-based brain–computer interface," *Biomed. Opt. Exp.*, vol. 14, no. 6, p. 2934, Jun. 2023.

[36] S. C. Wriessnegger, D. Kirchmeyr, G. Bauernfeind, and G. R. Muller-Putz, "Force related hemodynamic responses during execution and imagery of a hand grip task: A functional near infrared spectroscopy study," *Brain Cognition*, vol. 117, pp. 108–116, Oct. 2017.

[37] C. J. Hourdakis and A. Perris, "A Monte Carlo estimation of tissue optical properties for use in laser dosimetry," *Phys. Med. Biol.*, vol. 40, no. 3, pp. 351–364, Mar. 1995.

[38] X. Hou et al., "NIRS-KIT: A MATLAB toolbox for both resting-state and task fNIRS data analysis," *Neurophotonics*, vol. 8, no. 1, Jan. 2021, Art. no. 010802.

[39] D. S. Park et al., "SpecAugment: A simple data augmentation method for automatic speech recognition," 2019, *arXiv:1904.08779*.

[40] H. Song et al., "Automatic depression discrimination on FNIRS by using general linear model and SVM," in *Proc. 7th Int. Conf. Biomed. Eng. Informat.*, Oct. 2014, pp. 278–282.

[41] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Proces. Syst.*, vol. 25, 2012, pp. 1–23.

[42] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[43] R. Wang, Y. Hao, Q. Yu, M. Chen, I. Humar, and G. Fortino, "Depression analysis and recognition based on functional near-infrared spectroscopy," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 12, pp. 4289–4299, Dec. 2021.

[44] Q. Yu, R. Wang, J. Liu, L. Hu, M. Chen, and Z. Liu, "GNN-based depression recognition using spatio-temporal information: A fNIRS study," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 10, pp. 4925–4935, Oct. 2022.

[45] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[46] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.