A Novel Convolutional Variation of Broad Learning System for Alzheimer’s Disease Diagnosis by Using MRI Images

RUIZHI HAN\textsuperscript{1}, C. L. PHILIP CHEN\textsuperscript{1,2,3}, (Fellow, IEEE), AND ZHULIN LIU\textsuperscript{4}, (Member, IEEE)

\textsuperscript{1}Faculty of Science and Technology, University of Macau, Zhuhai, Macau
\textsuperscript{2}School of Computer Science and Engineering, South China University of Technology, Guangzhou 510641, China
\textsuperscript{3}Pazhou Lab, Guangzhou 510355, China
\textsuperscript{4}School of Computer Science and Engineering, South China University of Technology, Guangzhou 510641, China

Corresponding author: Zhulin Liu (zhulinlu@gmail.com)

This work was supported in part by the National Key Research and Development Program of China under Grant 2019YFA0706200 and Grant 2019YFB1703600, in part by the National Natural Science Foundation of China under Grant 62006079, Grant 61702195, Grant 61751202, Grant U1813203, Grant U1801262, and Grant 61751205, in part by the Science and Technology Major Project of Guangzhou under Grant 202007030006, in part by the China Postdoctoral Science Foundation under Grant 2020TQ0105, in part by the Science and Technology Development Fund, Macau SAR under Grant 079/2017/A2 and Grant 0119/2018/A3, and in part by the Multiyear Research Grants of University of Macau.

ABSTRACT Alzheimer’s disease (AD) is a serious chronic health problem that causes great pain and loss to patients and their families. Its early and accurate diagnosis would achieve significant progress on the prevention and treatment of the disease. Magnetic Resonance Imaging (MRI) is a commonly used technique in nuclear medical diagnostics. However, it is still a challenging problem to diagnose AD, Control Normal (CN), and Mild Cognitive Impairment (MCI) because of the complex structures of MRI. In this paper, diagnosing models for MRI images are proposed to identify the various stages of AD based on the Broad Learning Systems (BLS), as well as its convolutional variants. To verify the validity of the proposed models, experiments on MRI images collected from the ADNI website are tested and evaluated. The results show that our algorithms outperform the other state-of-the-art algorithms for various tasks with better accuracy and less training times. Finally, the cross-domain learning ability of the proposed algorithms is verified on an independent AD dataset.

INDEX TERMS Alzheimer’s disease, broad learning system, convolutional neural network, image classification, magnetic resonance imaging.

I. INTRODUCTION Alzheimer’s disease (AD), the most common dementia, is a complex disease leading to memory impairment and other cognitive problems. Currently, there is no known treatment that could cure this disorder, and the average remaining life after diagnosed is about three to nine years [1]–[5]. According to the 2018 World Alzheimer report, nearly 50 million people worldwide were living with dementia in 2018. And this number will triple to 152 million by 2050. Around the world, there will be a new case every 3 seconds [6]. The state of AD can be divided into three categories: Alzheimer’s disease (AD), mild cognitive impairment (MCI), control normal (CN), of which MCI can be subdivided into stable mild cognitive impairment (sMCI) and progressive mild cognitive impairment (pMCI) [7], [8]. To diagnose Alzheimer’s disease, it is necessary to comprehensively analyze the patient’s medical history, behavioral assessment, cognitive test, brain imaging, and blood sampling. Among them, the patient’s medical history and behavioral assessment are the main clinical diagnostic pieces of evidence. However, both medical history and behavioral tests require evaluation by multiple visits of expert doctors for a long period of time. It is imperative to propose an efficient diagnostic method.

Recently, with the development of technology, high-end medical imaging technologies, such as magnetic resonance imaging (MRI) and positron computed tomography (PET), can provide more convincing information to diagnose AD [9]. Moreover, with the help of computer vision technology, some researchers used the generated method to construct the
corresponding required diagnostic images [10]–[12]. In comparison, MRI image is easier to obtain than PET because MRI requires a shorter time and is relatively less expensive for patients. At the same time, machine learning has grown by leaps and bounds. In terms of medical images, researchers have designed many effective algorithms for different data types to diagnose the state of AD.

On one hand, some of these algorithms are based on normal neural network and support vector machine (SVM) algorithms and contain a range of feature processing methods to preprocess MRI images. For example, Liu et al. [13] design a whole-brain hierarchical network (WBHN) to classify Alzheimer’s Diseases. WBHN is a special hierarchical neural network designed based on a number of regions of interest (ROI) in the brain. In addition, Jha et al. [14] take a series of feature processing methods, such as Principal component analysis (PCA) and linear discriminant analysis (LDA), to improve the effect of feature extraction on MRI images. The model designed by Salvatore et al. [15], [16] uses the SVM as a classifier, while the intermediate feature extraction and selection are processed by the PCA. In summary, such methods rely on a series of feature processing algorithms to classify accurately.

On the other hand, using convolutional networks to extract high dimension features from MRI or PET images is another approach. Lin et al. [17] propose a convolutional based model, which extracts features from three different axes of an MRI image, to diagnose pMCI and sMCI. Liu et al. [18] adopt a data-driven landmark discovery algorithm to locate the most informative image patches in MRI images. And a group of 3D CNN is designed to extract features and classify various states of AD [19]. Ding et al. [20] design a deep approach based on Inception Network version 3 [21], [22] which is a deep convolutional neural network to classify AD, MCI, or CN based on FDG-PET images. And the InceptionNet performs better than three radiologists on an independent dataset. The success of the above-mentioned methods inspires us to preprocess MRI images by convolutional networks. In addition, some methods based on multi-modal data and functional-MRI images are also proposed and applied [23]–[25], but they are beyond the scope of this paper.

Considering the above two types of methods comprehensively, the former traditional methods must be based on certain additional feature processing methods, and these preprocessing methods have a greater impact on the results and reduce the efficiency of the diagnostic algorithm. Among the many methods of the latter, some convolutional networks also depend on the prior knowledge of the data. Besides that, the InceptionNet model faces the problem of huge model size and unsatisfactory performance on MRI images although it performs well on PET data. Hence, in this paper, we first aim to accurately predict the various states of AD and the diagnosis of AD using MRI images. The second goal is to design an efficient algorithm that relies on feature processing as little as possible to achieve an accurate prediction of AD. Consequently, based on specific image processing, we first design a method named Convolution Feature-based Cascade of Enhancement Nodes BLS (CF-CEBLS) that combines the convolution variation of Broad Learning System (BLS) [26] algorithms to construct the effective diagnosis model of AD. After that, another variants based on the proposed CF-CEBLS and the direct solution BLS are introduced to solve the AD diagnosis problem. Fig. 1 presents the overall flow of AD diagnosis based on MRI images. The main contributions of this paper can be summarized as follows:

- BLS and its variation algorithms are introduced into medical image analysis for the first time, providing an effective tool for the diagnosis of AD by MRI images.
- Combined with the advantages of the direct solution method, a fast classification model of medical images is experimentally verified.
- Compared with other models, it provides a feature module with a smaller number of parameters, and the experimental results of our overall model show advantages in multiple data sets.

This paper is mainly composed of the following parts. Introduction and succinct related works are given in the first section. The second section offers the details of our dataset, data preprocessing and proposed models. In the third section, we evaluate our algorithm and other related algorithms with the experimental analysis. Finally, the summary of the full paper and future work are given in the fourth chapter.

II. MATERIALS AND METHODS

Here, we first introduce the datasets and MRI image processing pipeline used in this work and then present the proposed methods.
A. DATA DESCRIPTION

Data used in the preparation of this article were obtained from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) database [27]. The ADNI was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), and the Food and Drug Administration (FDA), as a 5-year public-private partnership, led by the principal investigator, Michael W. Weiner, MD. The primary goal of ADNI is to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessments subjected to participants could be combined to measure the progression of MCI and early Alzheimer’s disease (AD). For more details with the ADNI database, please check http://adni.loni.usc.edu.

In our experiments, T1-weighted MRI images are randomly selected from ADNI website to use as base training and test dataset. The dataset includes 768 MRI images and it contains 3 categories, each of which has 256 images. Specific details about the dataset are shown in Table 1. Mini-Mental State Examination (MMSE) [28], which is a questionnaire that has 30 questions used to assess cognitive impairment, is often used by the medical community to examine patients with dementia. The highest score for MMSE is 30 points and the higher the score, the better the ability. Usually, a person with an MMSE score of less than 24 is considered to be an AD patient. As shown in Table 1, the average age of the AD group is 75 years (range from 55 years to 92 years) with an average of 22.2 MMSE. The average age of the MCI group is 77 years (range from 58 years to 90 years). And the average age of the CN group is 78 years (range from 55 years to 92 years).

TABLE 1. Summary of participant demographics, mini-mental state examination (MMSE) considered in this study.

| Diagnostic Type | No. of images | Age [range] | Gender | MMSE | M/F |
|-----------------|---------------|-------------|--------|------|-----|
| AD              | 256           | 75.8 ± 7.7  | [55.2, 92.8] | 118/138 | 22.2 ± 3.6 |
| MCI             | 256           | 77.7 ± 6.4  | [58.0, 90.9] | 178/78  | 26.0 ± 3.3 |
| CN              | 256           | 78.7 ± 5.1  | [57.2, 92.8] | 137/119 | 28.9 ± 1.3 |
| ALL             | 768           | 77.4 ± 6.4  | [55.2, 92.8] | 433/335 | 26.5 ± 3.3 |

Values are presented as mean ± Std. M: Male, F: Female. AD: Alzheimer disease, MCI: mild cognitive impairment, CN: control normal.

Furthermore, an external independent dataset is also used to evaluate the performance of our algorithm according to [16]. The independent dataset is derived from MRI images obtained by each patient when their conditions are stable. It contains four categories, AD, CN, sMCI, and pMCI, for a total of 201 images. Table 2 shows specific information about the independent dataset. As shown in Table 2, the average age of the AD group is 77 years (range from 58 years to 89 years). The average age of the pMCI group is 76 years (range from 57 years to 90 years). The average age of the sMCI group is 76 years (range from 60 years to 88 years). And the average age of the CN group is 77 years (range from 62 years to 89 years).

Since the category number of this independent dataset is different from the training dataset, we need to adjust the categories of the validation set to fit the trained model when validating model generalization performance. The processing method of the independent dataset can be checked in Section 3. A. All images were downloaded in NIFTI format after being licensed on the ADNI website: http://ida.loni.usc.edu/.

B. DATA PROCESSING

According to the ADNI acquisition protocol [27], examinations were performed in 1.5T using a T1-weighted sequence. The MRI images are undergone with the following preprocessing steps: (1) 3D grad warp correction for geometry correction caused by gradient non-linearity [29], (2) B1 non-uniformity correction for intensity correction caused by non-uniformity [30]. These preprocessing steps help to improve the standardization among MRI images from different platforms. The whole process was performed by using the CAT12 [33] software package installed on the Matlab platform (Matlab R2016b, The MathWorks).

1) STEP 1
A sequence of processing procedure was performed on the images, consisting of: (1) image re-orientation; (2) cropping; (3) skull-stripping; (4) image normalization to the MNI standard space by means of co-registration to the MNI template [31]. After this phase, all MR images resulted to be with size 121 × 145 × 121 voxels. And the noise mitigation [32] is also considered by the above steps. The whole process was performed by using the CAT12 [33] software package installed on the Matlab platform (Matlab R2016b, The MathWorks).

2) STEP 2
Until now, all MRI images will retain the entire brain area, and the excess invalid area will be cut off by applying software Nilearn [34] and Nibabel [35]. And then, starting with the 20th slice of the axial view, a slice will be extracted for

TABLE 2. Summary of participant demographics, mini-mental state examination (MMSE) considered for independent dataset in this study.

| Diagnostic Type | No. of images | Age [range] | Gender | MMSE | M/F |
|-----------------|---------------|-------------|--------|------|-----|
| AD              | 50            | 77.2 ± 6.7  | [58.5, 89.9] | 28/22 | 19.0 ± 5.9 |
| pMCI            | 51            | 76.7 ± 7.0  | [57.4, 90.5] | 27/24 | 22.3 ± 4.1 |
| sMCI            | 50            | 76.7 ± 6.7  | [60.2, 88.7] | 28/22 | 27.1 ± 3.8 |
| CN              | 50            | 77.0 ± 6.0  | [62.1, 89.8] | 27/23 | 29.2 ± 0.9 |
| ALL             | 201           | 76.9 ± 6.6  | [57.4, 90.5] | 110/91 | 24.4 ± 5.7 |

Values are presented as mean ± Std. M: Male, F: Female. AD: Alzheimer disease, sMCI: stable mild cognitive impairment, pMCI: progression mild cognitive impairment, CN: control normal.
FIGURE 2. Illustration of data representation: (a) Original MRI image: a 3D NIfTI file. The sub-image from left to right is one slice extracted from Coronal View, Sagittal View, and Axial View respectively. The 3D raw data should be processed by image re-orientation, cropping, skull-stripping and image normalization (Processing step 1). (b) Processed image after Step 1. From left to right are: Coronal View, Sagittal View, and Axial View. (c) Final image feeds into the model.

Each of 5 slices from an MRI image. At the end of that, 16 slices are obtained and every slice will be scaled to the size of $125 \times 100$. For the purpose of fitting the convolutional neural network, the total 16 slices will be distributed in an image by $4 \times 4$ in the order of skull base to cranial top similar to [20]. After all the processing is completed, the original image is changed from 3D to a 2D image with a size of $500 \times 100$. Fig. 2.(c) is an example of the final image that feeds into the model.

From another perspective, medical workers often take a similar approach to achieve a 3D image reduction to 2D images. It is worth mentioning that although 3D computer vision technology has been widely used in the field of image and video processing, 2D convolution still plays an irreplaceable role. We visited physicians in medical imaging, learned from clinicians’ conventional processing methods of MRI images, and [20], and then designed relevant models in the hope that the models could be applied to practical applications. Therefore, from the perspective of practical application, this paper hopes to conduct image preprocessing in a way closer to the clinical practice. All processing steps were conducted on Python 3 and SciPy [36], [37]. MRI volumes were visually inspected for checking homogeneity and absence of artifacts both before and after the processing step. The image representation is visible in Figure 2.

C. METHOD PRELIMINARIES

The neural network is a machine learning tool widely used in disease diagnosis, among which Broad Learning System (BLS) is a special neural network structure. BLS is proposed by Chen et al. [26], [38] based on Functional-Link Neural Network (FLNN) and regularization techniques [39], [40]. It aims to offer an alternative way for classification problems instead of learning a deep structure. Prior to this, another graph CNN variant of BLS is prominent in processing electroencephalograms (EEG) [41]. The network structure of BLS is a three-layer network structure. Fig. 3 illustrates the basic model of BLS where $X \in \mathbb{R}^{N \times m}$ is the input matrix, $N$ is the number of instances, and $m$ is the dimension of every instance. $Y$ is the output layer and $W$ is the weights connecting the output layer to both the feature mapping layer and the enhancement layer.

After concatenating the output of the feature mapping layer and enhancement layer, we can think of it as the overall input of output layer, denoted as $A$, where $A = [\psi(XW_{e} + \beta_{e})]$, $\|e(ZW_{h} + \beta_{h})\|$. $W_{e}, \beta_{e}$ are the weights and bias from $X$ to the feature mapping layer and $W_{h}, \beta_{h}$ connect the feature mapping layer to enhancement layer. Furthermore, $Z$ is the output of the feature mapping layer, $W_{e,h}$ and $\beta_{e,h}$ are randomly generated and the dimensions of $W_{e}$ and $W_{h}$ are hyper-parameters. $\psi$ and $\xi$ are non-linear functions, such as $\tanh$ and $\tanh$. Finally, the problem is transformed into a linear equation system problem, which is

\[ AW = Y \]  

To obtain $W$, Eq. (1) can be solved according to the generalized inverse and ridge regression methods by optimizing:

\[ \min ||AW - Y||^{\rho} + \lambda||W||_{\rho}^{\rho} \]
where $\rho$ and $\sigma$ always equal to 2 and $\lambda$ is the regularization coefficient. We can deduce that:

$$W = A^{\dagger}Y$$

(3)

where

$$A^{\dagger} = \lim_{\lambda \to 0} (\lambda I + A^TA)A^T$$

(4)

1) VARIATION: CASCADE OF ENHANCEMENT NODES (CEBLS)

Through changing the network structure, BLS has derived many varieties, among which the most common one is to change the connection mode of nodes in the enhancement layer. Hence, one variation named Cascade of Enhancement Nodes (CEBLS) [38] is proposed. In this method, for the input data $X$, the first $n$ groups of feature nodes are generated by the following equations:

$$Z_i := \phi(XW_{e_1} + \beta_{e_1}), \quad i = 1, 2, \ldots, n$$

(5)

and $W_{e_1}$ and $\beta_{e_1}$ are sampled from the given distribution. Project the feature nodes $Z^n := [Z_1, Z_2, \ldots, Z_n]$ by function $\xi(\cdot)$, the first group of enhancement nodes is

$$H_1 := \xi(Z^nW_{h_1} + \beta_{h_1})$$

(6)

where the associated weights are randomly sampled. The enhancement nodes from the second group to $m$th group are compositely established as follows:

$$H_j = \xi(H_{j-1}W_{h_j} + \beta_{h_j}), \quad j = 2, 3, \ldots, m$$

(7)

Consequently, the nodes $Z^n$ and $H^m = [H_1, H_2, \ldots, H_m]$ are connected directly with the output $Y$.

D. METHOD PROPOSED

We design a multi-layer convolution module to serve as the feature extractor of the raw image. Meanwhile, in the enhancement layer of the BLS model, we draw the successful experience of CEBLS [38] structure and adopt the cascade network structure for our BLS enhancement layer. In a short, the proposed method could be regarded as two cascade network structure for our BLS enhancement layer. Hence, one variation named Cascade of Enhancement Nodes (CEBLS) [38] is proposed. In this method, for the input data $X$, the first $n$ groups of feature nodes are generated by the following equations:

$$Z_i := \phi(XW_{e_1} + \beta_{e_1}), \quad i = 1, 2, \ldots, n$$

(5)

and $W_{e_1}$ and $\beta_{e_1}$ are sampled from the given distribution. Project the feature nodes $Z^n := [Z_1, Z_2, \ldots, Z_n]$ by function $\xi(\cdot)$, the first group of enhancement nodes is

$$H_1 := \xi(Z^nW_{h_1} + \beta_{h_1})$$

(6)

where the associated weights are randomly sampled. Therefore, the enhancement nodes from the second group to $m$th group are compositely established as follows:

$$H_j = \xi(H_{j-1}W_{h_j} + \beta_{h_j}), \quad j = 2, 3, \ldots, m$$

(7)

Consequently, the nodes $Z^n$ and $H^m = [H_1, H_2, \ldots, H_m]$ are connected directly with the output $Y$.

As shown in Eq. (8), when $k$ is equal to $n$, we take a diverse operation. That is, the pooling layer is replaced by a special convolution layer which has $(1 \times 1)$ kernel size. Details about $B_1, B_2, \ldots, B_{n-1}$ and $B_n$ are illustrated in Fig. 4.

Second, the outputs of the last feature extractor block are fed into the first feature mapping node of CEBLS. The specific form can be formulated as Eq. (5), where $X$ is the output of the feature extraction module at this time. And the data flow through each enhancement node in turn. The representation of the above enhancement layer is formulated as follows:

$$H_j := \begin{cases} \xi(B^nW_{h_j} + \beta_{h_j}), & j = 1, \\ \xi(H_{j-1}W_{h_j} + \beta_{h_j}), & j = 2, 3, \ldots, m. \end{cases}$$

(10)

where $B^n := [B_1, B_2, \ldots, B_n]$ and $W_{h_j}, \beta_{h_j}$ are the weights and biases associated with the enhancement layer. And the nonlinear activation function $\xi(\cdot)$ is ReLU.

In addition, based on the convolution structure mentioned above, we also combine the direct solution BLS, as shown in Eq. (4), with the proposed convolution feature extractor to AD diagnosing in this paper.

In a word, the algorithms proposed in this paper can be summarized into the following two types according to the different top weight solutions. The first model is model 1, which integrates the convolution module and a CEBLS module as a whole, and realizes the simultaneous optimization of the weight of the two modules by gradient descent. The second model is model 2, which based on model 1. Once model 1 is trained, its convolution module will be regarded as the feature extractor, and a BLS module is stacked, which is solved by the pseudo inverse. Model 2 is only applied to the independent dataset.

1) MODEL SETTING

In our model, a basic (multiple convolutions)-pooling block is composed of 5 Convolution-ReLU layers and a pooling layer. The setting of the number of the Convolution-ReLU layers is based on the inspires of VGGNet and our trial experience in the experiment. The convolution kernel number within each CAP block is consistent, which are 32, 64,
FIGURE 5. The overall structure of CF-CEBLS. a) $B_1, B_2, \ldots, B_n$ illustrate the feature extract modules and the details are shown in Figure 4. According to the theory of BLS, the connection between the output of feature extract modules and $Y$ are optional. However, such kind of variants are not proposed and discussed here, and the connections are presented in dotted lines. b) $Z_1, Z_2, \ldots, Z_n$ are the feature mapping layers and $H_1, H_2, \ldots, H_m$ are the feature enhancement layers. The difference between $Z$ and $H$ is the type of inner connection. For $Z$, each layer has the same input and there are no connections between them. While for the feature enhancement layers $H$, every layer is connected to the previous layer except the first layer, and the first layer is connected with the concatenation of all the feature mapping layers. In order to reduce the parameter redundancy and the risk of model overfitting, we applied one feature mapping layer and two enhancement layers in the experiments, and detailed information of the structure are shown in Table 3.

TABLE 3. The structure of our model used in this paper.

| type               | input size | output size | conv layers | kernel size | kernel # | padding | parameters # |
|--------------------|------------|-------------|-------------|-------------|----------|----------|--------------|
| Batch Normalization| 500×400×1  | 500×400×1   | -           | -           | -        | -        | 0            |
| Block 1            | Conv layer | 500×400×1   | 490×390×32  | 5           | 3×3      | 32       | 37312        |
|                    | Max pooling| 490×390×32  | 245×195×32  | -           | 2×2      | -        | 0            |
| Block 2            | Conv layer | 245×195×32  | 235×185×64  | 5           | 3×3      | 64       | 166208       |
|                    | Max pooling| 235×185×64  | 117×92×64   | -           | 2×2      | -        | 0            |
| Block 3            | Conv layer | 117×92×64   | 107×82×128  | 5           | 3×3      | 128      | 664192       |
|                    | Max pooling| 107×82×128  | 53×41×128   | -           | 2×2      | -        | 0            |
| Block 4            | Conv layer | 53×41×128   | 43×31×256   | 5           | 3×3      | 256      | 2655488      |
|                    | Max pooling| 43×31×256   | 21×13×256   | -           | 2×2      | -        | 0            |
| Block 5            | Conv layer | 21×13×256   | 11×5×512    | 5           | 3×3      | 512      | 10619392     |
| Single Conv layer  | 11×5×512   | 11×5×32     | 1           | 1×1        | 32       | 16416     |
| Flatten layer      | 11×5×32    | 1×1×1760    | -           | -          | -        | 0         |
| Feature mapping layer (FML) | 1×1×1760 | 1×1×100     | -           | -          | -        | -         |
| Enhancement layer (EHL) 1 | 1×1×100  | 1×1×15      | -           | -          | -        | 1515      |
| Enhancement layer (EHL) 2 | 1×1×15   | 1×1×906     | -           | -          | -        | 49536      |
| Concatenate        | PML, EHL 1, EHL 2 | 1×1×3211 | -           | -          | -        | 0         |
| Output layer       | 1×1×3211  | 1×1×3       | -           | -          | -        | 9636      |

Note: Block 1,2,...,5 are defined as in Fig. 4.

128, 256, 512 respectively. Except for special instructions, all convolution kernel sizes in this study are $3 \times 3$. For the top-level network structure, the structural parameters are searched by the grid search method, and the resulted optimal structure could be checked in Table 3. For the BLS model in Model 2, similar searches through the searching space are implemented. The network structure parameters of the corresponding results are shown in Tables 5 and 7. Among them, the total nodes of the feature mapping layer and enhancement layer are $20 \times 19 + 2000$ and $5 \times 24 + 900$ in Tables 5 and 7, respectively.

2) LOSS FUNCTION

To solve the weights of the other layers to the output layer, that is, solving the objective function Eq. (1), analytic solutions could be obtained by pseudo inverse in BLS. Another alternative way to train output weights is the gradient descent method. Therefore, different from the direct solution method of Eq. (2), the cross-entropy loss function is used as the objective function of our model. Cross-entropy loss measures the performance of a classification model by probability value between 0 and 1. It increases if the predicted probability diverges from the actual label. A typical Categorical Cross-Entropy Loss can be formulated as follows:

$$\text{Loss} = - \sum_{c=1}^{M} y_{o,c} \log(p_{o,c})$$

where $M$ is the number of classes and $y$ is a binary indicator (0 or 1). When class label $c$ is correct for observation $o$, $y = 1$. $p$ is predicted probability that observation $o$ is of class $c$.

III. EXPERIMENTS AND DISCUSSION

In this section, experimental results are presented to verify the proposed methods. Firstly, we introduce the comparison
methods and detailed experimental settings. And the experimental results for both the AD state prediction task (AD vs. MCI vs. CN) and AD classification task (AD vs. CN) are shown in the following.

**A. METHODS FOR COMPARISON AND EXPERIMENTAL SETTINGS**

1) **METHODS FOR COMPARISON**
To prove the effectiveness of our model, we will compare the performance of our method with existing methods, including InceptionNet [20], which is a typical deep model, and a set of PCA-SVM models [16].

**Deep Model:** InceptionNet was firstly proposed by Christian Szegedy et al. [21] for normal computer vision problems like image classification. Ding et al. [20] apply the third version of InceptionNet [22] to settle the AD diagnosis problem using FDG-PET images and it achieves excellent results that outperform three clinicians. For InceptionNet [20], based on the code released by authors, we made extra improvements by reducing the learning rate($1e^{-3}$ → $1e^{-5}$) and adjusting the dropout rate. It is worth noting that InceptionNet is pre-trained on the ImageNet dataset.

**Traditional Model:** Traditional machine learning methods, such as PCA and SVM, and their variants often bring unexpected excellent results. Salvatore et al. propose various methods based on PCA and SVM to predict AD [16]. For PCA-SVM models, since the related codes are not released, the relevant results are obtained from the results claimed in their paper. The PCA-SVM models are available in multiple binary classification tasks, for instance, AD vs CN, CN vs MCI. The proposed models and the InceptionNet are multi-class classification models, we also divide the independent dataset into the same partition as PCA-SVM to achieve fairness in comparison.

2) **EXPERIMENTAL SETTINGS**
We validate our proposed methods on both AD states prediction (AD vs CN vs MCI) and AD classification (AD vs CN). For the AD states prediction, due to sMCI and pMCI both belong to MCI state, the prediction task is equivalent to classify AD vs CN vs MCI (sMCI + pMCI). sMCI means that patients with MCI will not convert to AD for at least 36 months, whereas pMCI means the patient would convert during 36 months. On the other hand, for patients with sMCI, it has much smaller risk to be an AD patient than pMCI. Therefore, we categorize the sMCI as CN and the pMCI as AD similarly. Consequently, the AD classification task is to divide AD (AD+pMCI) from CN(CN + sMCI)

Initialization method of **Truncated normal** is adopted in our model. For both models, the number of epochs is set at 200. The ADNI dataset is randomly divided into two parts: 80% is the training set and the other 20% is the testing set. Meanwhile, 5-fold cross-validation is applied to the independent dataset. Furthermore, in order to reduce the over-fitting, we adopt the early stopping technique.

**B. MODEL EVALUATION**
We take multiple assessment indicators to evaluate different models. Accuracy (ACC) is the overall accuracy of the test dataset. Sensitivity (SEN) is a measure of how well a model can correctly identify the number of positive samples in all positive samples. Similarly, specificity (SPE) is a measure of how well a model can correctly identify the number of negative samples in all negative samples. And precision refers to how many samples are predicted correctly in the positive samples. F1-scores is an indicator to check the balance between precision and sensitivity. The closer the value of F1-scores is to 1, the better the performance of the model.

\[
ACC = \frac{(TP + TN)}{(TP + TN + FP + FN)}\] (12)

\[
SEN = \frac{TP}{(TP + FN)}\] (13)

\[
SPE = \frac{TN}{(TN + FP)}\] (14)

\[
Precision = \frac{TP}{(TP + FP)}\] (15)

\[
F1 \text{ score} = \frac{2 \times \text{Precision} \times \text{SEN}}{(\text{Precision} + \text{SEN})}\] (16)

Here TP, TN, FP, FN is the number of true positives, true negatives, false positives, and false negatives, respectively.

**C. MODEL TRANSFER LEARNING**
As mentioned in section 2. A, two datasets, images collected from the ADNI website and the independent dataset, are used to validate the proposed models. In addition to the conventional model training, we also adopt the training method commonly used in transfer learning (fine-tuning): a). Fix the trained feature extractor module, b). Modify the number of final output nodes c). Re-train the top CEBSLS model. In detail, both the proposed model and the InceptionNet are trained in ADNI datasets for the classification of AD, CN, and MCI. After that, their feature extractor blocks are fixed while the left networks are retrained in the independent datasets for the classification task of AD and CN. Both models are also performed individually in the independent datasets to verify the cross-domain and cross-task abilities of our proposed model.

**D. RESULTS OF AD STATE PREDICTION: AD VS CN VS MCI**
1) **ADNI DATASET**
Table. 4 shows the performance of our model 1 and InceptionNet testing on the ADNI dataset. By comparing the different indicators of Table. 4, we can find that our model 1 has better performance than InceptionNet. The accuracy of our model can achieve 98.83% which is 2.6% higher than the accuracy of InceptionNet model. At the same time, indicators, such as SEN and SPE, indicate that our model is more balanced than that of InceptionNet model. It is worth mentioning that our model and the InceptionNet model need to learn about 14M and 23M parameters, respectively. And the InceptionNet has
TABLE 4. Performance of two models on ADNI dataset: AD vs CN vs MCI(\%).

| Model & Class | No. of Parameter | ACC | Sensitivity | Specificity | Precision | F1 Score |
|--------------|------------------|-----|-------------|-------------|-----------|----------|
| InceptionNet | 23.87 Million     |     |             |             |           |          |
| AD           | -                | 95.17 ± 3.96 | 97.26 ± 2.16 | 94.84 ± 3.85 | 94.89 ± 2.34 |
| CN           | -                | 97.34 ± 2.75 | 98.85 ± 0.92 | 97.64 ± 1.92 | 97.43 ± 0.99 |
| MCI          | -                | 96.05 ± 2.50 | 98.23 ± 0.98 | 96.51 ± 1.85 | 96.26 ± 1.69 |
|              |                  | 96.23 ± 1.50 |             |             |           |          |
| Our model 1  | 14.40 Million    |     |             |             |           |          |
| AD           | -                | 98.18 ± 1.08 | 99.58 ± 0.83 | 99.32 ± 1.36 | 98.74 ± 0.85 |
| CN           | -                | 99.27 ± 0.89 | 99.41 ± 0.48 | 98.87 ± 0.92 | 99.07 ± 0.62 |
| MCI          | -                | 99.11 ± 1.78 | 99.24 ± 0.72 | 98.39 ± 1.46 | 98.74 ± 1.28 |
|              |                  | 98.83 ± 0.76 |             |             |           |          |

Note: Values are presented as mean ± Std. The best results are marked in bold.

TABLE 5. Performance of multi models on Independent dataset: AD vs CN vs (sMCI + pMCI)(\%).

| Model & Class | No. of Parameter | ACC | Sensitivity | Specificity | Precision | F1 Score |
|--------------|------------------|-----|-------------|-------------|-----------|----------|
| InceptionNet | 2.10 Million     |     |             |             |           |          |
| AD           | -                | 73.33 ± 15.35 | 79.48 ± 7.76 | 55.18 ± 12.40 | 62.39 ± 12.23 |
| CN           | -                | 65.33 ± 19.62 | 94.93 ± 3.81 | 79.83 ± 15.23 | 71.03 ± 18.00 |
| MCI          | -                | 66.32 ± 11.61 | 74.00 ± 12.81 | 72.75 ± 11.67 | 69.00 ± 10.39 |
|              |                  | 67.86 ± 11.00 |             |             |           |          |
| Our model 1  | 0.24 Million     |     |             |             |           |          |
| AD           | -                | 58.71 ± 11.62 | 90.70 ± 5.16 | 68.99 ± 14.79 | 62.87 ± 11.46 |
| CN           | -                | 67.60 ± 14.50 | 91.05 ± 5.39 | 73.26 ± 12.74 | 69.14 ± 10.27 |
| MCI          | -                | 75.32 ± 9.17  | 65.74 ± 7.68 | 68.78 ± 6.00 | 71.72 ± 6.94 |
|              |                  | 69.19 ± 6.58  |             |             |           |          |
| Our model 2  | 2380             |     |             |             |           |          |
| AD           | -                | 74.00 ± 10.83 | 79.53 ± 6.38 | 55.24 ± 9.40 | 62.85 ± 8.77 |
| CN           | -                | 66.67 ± 15.35 | 93.84 ± 3.16 | 77.95 ± 10.70 | 71.27 ± 12.40 |
| MCI          | -                | 67.44 ± 9.72  | 77.67 ± 9.29 | 76.00 ± 9.05 | 71.05 ± 7.60 |
|              |                  | 68.91 ± 7.13  |             |             |           |          |

Note: Values are presented as mean ± Std. The best results are marked in bold. InceptionNet and our model are fine-tuned on the independent dataset based on the weights trained on ADNI dataset.

TABLE 6. Performance of two models on ADNI dataset: AD vs CN(\%).

| Model       | ACC       | Sensitivity | Specificity | Precision | F1 Score |
|-------------|-----------|-------------|-------------|-----------|----------|
| InceptionNet| 96.26 ± 1.67 | 95.17 ± 3.96 | 97.34 ± 2.75 | 98.13 ± 1.71 | 97.73 ± 2.20 |
| Our model 1 | 98.67 ± 0.45 | 98.18 ± 1.08 | 99.27 ± 0.89 | 99.66 ± 0.69 | 99.46 ± 0.73 |

Note: Values are presented as mean ± Std. The best results are marked in bold.

been pre-trained by the ImageNet dataset before apply to the AD diagnosis task.

2) INDEPENDENT DATASET
In order to verify the performance of models on different datasets, we apply the dataset of [16], which is the mentioned independent dataset in the previous section. For the purpose of matching the category settings of the independent dataset with the settings of the ADNI dataset, the classes of sMCI and pMCI are classified as MCI. To accomplish the above tasks on the independent set, we propose an additional model, model 2: combining our convolution feature extraction module and the BLS model with analytic solutions. Table 5 shows the relative results of the InceptionNet model, our model 1, and model 2 on the independent dataset. Among them, both our models and InceptionNet model are fine-tuned without structure changing on the independent dataset. In addition, the results of model 2 are obtained with 5-fold cross-validation on the convolution features of the independent dataset. It can be seen that our models have obvious superior compared with InceptionNet. On the other hand, our model 1 needs much fewer parameters to be trained.
TABLE 7. Performance of multi models on Independent dataset: (AD + pMCI) vs (CN + sMCI)(%).

| Model               | No. of Para | Train time(s) | ACC       | Sensitivity | Specificity | Precision | F1 Score |
|---------------------|-------------|---------------|-----------|-------------|-------------|-----------|----------|
| PCA-FDR-SVM\*       | -           | -             | 79.00 ± 8.00 | 83.00 ± 14.00 | 75.00 ± 1.00 | -         | -        |
| PCA-FDR-SVM+SNeD\*  | -           | -             | 92.00 ± 1.00 | 91.00 ± 4.00  | 93.00 ± 3.00 | -         | -        |
| PLS-SVM\*           | -           | -             | 82.00 ± 4.00 | 82.00 ± 7.00  | 81.00 ± 4.00 | -         | -        |
| PLS-SVM+SNeD\*      | -           | -             | 85.00 ± 5.00 | 87.00 ± 9.00  | 83.00 ± 4.00 | -         | -        |
| InceptionNet\*      | 23.87 Milloin | 185.6        | 82.55 ± 14.02 | 78.09 ± 29.78 | 87.00 ± 14.18 | 78.29 ± 28.98 | 77.58 ± 28.44 |
| Our model 1*         | 14.40 Million | 563.2       | 82.10 ± 9.17  | 82.18 ± 10.30 | 82.00 ± 10.33 | 82.44 ± 9.31  | 82.14 ± 9.10  |
| InceptionNet\#      | 2.10 Million | 71.8         | 80.60 ± 3.55  | 85.79 ± 5.43  | 75.33 ± 7.18  | 78.15 ± 4.76  | 81.61 ± 3.30  |
| Our model 2#        | 0.24 Million | 168.4        | \textbf{93.51 ± 8.63} | \textbf{94.18 ± 7.94} | 92.80 ± 11.44 | \textbf{93.68 ± 9.68} | \textbf{93.76 ± 8.08} |
| Our model 2\#       | 1020        | 0.23         | 86.57 ± 6.03  | 89.24 ± 8.55  | 83.73 ± 11.30 | 89.96 ± 8.06  | 87.07 ± 5.55  |

Note: Values are presented as mean ± Std. The best results are marked in bold.
\* Models are trained on the independent set.
\# Models are fine-tuned on the independent set based on models trained on ADNI dataset.
\f Models are trained (direct solution) on the independent set with the convolutional feature extracted by our model 1.

For our model 2, relatively excellent results could also be achieved with fixed convolution features.

E. RESULTS OF AD CLASSIFICATION: AD VS CN

1) ADNI DATASET

The results of the AD vs CN task on the ADNI dataset can be obtained by the AD vs CN vs MCI task, as the former is a simplification of the latter. Relevant experimental results are extracted and expressed in Table 6. Obviously, our model 1 performs better than InceptionNet in various indicators.

2) INDEPENDENT DATASET

For the purpose of evaluating the model, we first trained both our model 1 and InceptionNet model from scratch on the independent dataset, respectively. Results are shown in Table.7 with corner mark *.

Further, we fine-tuned the network structure of the model obtained from the ADNI dataset again and the results are marked corner mark # in Table.7. In this subsection, we still perform 5-fold cross-validation on the trained model.

As shown in Table. 7, our model 1 with corner mark # performs better under various evaluation indicators, for example, the ACC of our model is 93.51% which is 1.51% higher than the best PCA-SVM model; the parameters that our model needs are relatively minimal. Comparing with the algorithms denoted with corner mark *, we found that both our model 1 and InceptionNet would appear obvious overfitting due to the limitation of data size. However, when the experiments are conducted with fine-tuning, the performance of our model was improved, and the results are significantly better than that of InceptionNet. Therefore, we conclude that the top-level CEBLS structure of the proposed model will provide a more effective feature expression. For the same reason, although the experimental results of our model 2 are not the optimal results compared with other models, its structure is more flexible and the parameters and training time required for training are more concise. If some feature optimization methods are added, we believe that its performance will be improved to some extent.

In addition, we provide the training time and the number of parameters for a comprehensive description of the model performance. It can be found that our model 1 can achieve the best accuracy with only 0.24M parameters, which is about one-tenth of the InceptionNet model fine-tuning parameters, its training time is longer. The success of our model 1 that transfer learning to evaluate on the independent dataset proves the efficiency of our feature extractor module.

F. DISCUSSION

1) CLASS ACTIVATION MAPPING

Class activation mapping [42]–[44] is a typical method for visualization of the convolution features of the raw image. It can visualize the model’s attention location of the image’s underlying features. Similar to neural networks, such as [11], CNN also has excellent performance at data expression. In this paper, in order to demonstrate the correspondence between the class activation mappings and the original image, the class activation mappings of the first and second feature extraction block are shown in Fig. 6 and Fig. 7.
In addition, for the independent dataset, if only PCA or Partial Least Squares (PLS) methods are used for dimensionality reduction and feature extraction, and the SVM model is used for final classification, the results are obviously lower than other models. The main reason is that although the PCA method reduces the dimension of data, part of the useful information is lost in the meanwhile. With the different from it, our models apply the convolution modules as the feature extractor instead of PCA, which results that stronger features associated with various classes could be extracted.

IV. CONCLUSION

Once diagnosed, AD is hard to be cured and afflicts patients for the rest of their lives. Lately, the number of people suffering from AD continues to rise. The early diagnosis of AD is of great significance to delay the deterioration of the disease. In this paper, novel models combined with the BLS method is presented for the diagnosis of AD, CN, and MCI based on MRI images. The proposed models are tested in the selected ADNI dataset and the independent dataset, and the results are compared with InceptionNet and a series of PCA-SVM methods. Overall, the experimental results show that our model is superior to other models. On the other hand, the convolutional part of our model can be regarded as a better feature extraction module when the cross-domain tasks are needed.

As for the future work, we firstly aim to optimize the output of each feature extract block to achieve balance between network performance and parameters. Second, we would like to optimize the convolution module and improve our algorithm to achieve more realistic end-to-end AD diagnosis methods.

REFERENCES

[1] M. W. Weiner, “The Alzheimer’s disease neuroimaging initiative: A review of papers published since its inception,” Alzheimer’s Dementia, vol. 9, no. 5, pp. e111–e194, Sep. 2013.
[2] G. M. McKhann et al., “The diagnosis of dementia due to Alzheimer’s disease: Recommendations from the National Institute on Aging-Alzheimer’s Association workgroups on diagnostic guidelines for Alzheimer’s disease,” Alzheimer’s Dementia, vol. 7, no. 3, pp. 263–269, 2011.
[3] S. Todd, S. Barr, M. Roberts, and A. P. Passmore, “Survival in dementia and predictors of mortality: A review,” Int. J. geriatric psychiatry, vol. 28, no. 11, pp. 1109–1124, 2013.
[4] P. Rajarajeswari, S. V. Raja, A. S. Ashour, N. Dey, and V. E. Balas, “Active site cavities identification of amyloid beta precursor protein: Alzheimer disease study,” in Proc. IEEE 20th Jubilee Int. Conf. Intell. Eng. Syst. (INES), Jun. 2016, pp. 319–324.
[5] N. Dey, A. S. Ashour, S. Chakraborty, S. Samanta, D. Sifaki-Pistolla, A. S. Ashour, D.-N. Le, and G. N. Nguyen, “Healthy and unhealthy rat hippocampus cells classification: A neural based automated system for alzheimer disease classification,” J. Adv. Microsc. Res., vol. 11, no. 1, pp. 1–10, Jun. 2016.
[6] C. Patterson, “World alzheimer report 2018 the state of the art of dementia research: New frontiers,” in Proc. Alzheimer’s Disease Int. (ADI), London, UK, 2018, pp. 6–8.
[7] R. C. Petersen, R. Doody, A. Kurz, R. C. Mols, J. C. Morris, P. V. Rabins, K. Ritchie, M. Rossor, L. Thal, and B. Winblad, “Current concepts in mild cognitive impairment,” Arch. Neurol., vol. 58, no. 12, pp. 1985–1992, 2001.
[8] R. C. Petersen, G. E. Smith, S. C. Waring, R. J. Ivnik, E. G. Tangalos, and E. Kokmen, “Mild cognitive impairment: Clinical characterization and outcome,” Arch. Neurol., vol. 56, no. 3, pp. 303–308, 1999.
[9] M. L. Schreter, T. Stein, N. Maslowskij, and J. Neumann, “Neural correlates of Alzheimer’s disease and mild cognitive impairment: A system-
RUIZHI HAN received the M.S. degree in computer science and technology from the University of Jinan, Jinan, China, in 2017. He is currently pursuing the Ph.D. degree with the Faculty of Science and Technology, University of Macau, China. His current research interests include medical image analysis, broad learning systems, and machine learning.
C. L. PHILIP CHEN (Fellow, IEEE) graduated from the University of Michigan, Ann Arbor, MI, USA, in 1985. He is currently a Chair Professor and the Dean of the College of Computer Science and Engineering, South China University of Technology. Being a Program Evaluator of the Accreditation Board of Engineering and Technology Education (ABET) in USA, for computer engineering, electrical engineering, and software engineering programs, he successfully architects the University of Macau’s Engineering and Computer Science programs receiving accreditations from Washington/Seoul Accord through The Hong Kong Institution of Engineers (HKIE), of which is considered as his utmost contribution in engineering/computer science education for Macau as the former Dean of the Faculty of Science and Technology. His current research interests include cybernetics, systems, and computational intelligence. He is a Fellow of IEEE, AAAS, IAPR, CAA, and HKIE; a member of Academia Europaea (AE), European Academy of Sciences and Arts (EASA), and International Academy of Systems and Cybernetics Science (IASCYS). He received the IEEE Norbert Wiener Award, in 2018, for his contribution in systems and cybernetics and machine learning. He is also a highly cited researcher by Clarivate Analytics, in 2018 and 2019. He was a recipient of the 2016 Outstanding Electrical and Computer Engineers Award from his alma mater, Purdue University, in 1988. He was the Chair of TC 9.1 Economic and Business Systems of International Federation of Automatic Control, from 2015 to 2017, and is currently the Vice President of Chinese Association of Automation (CAA). He was the President of IEEE Systems, Man, and Cybernetics Society, from 2012 to 2013, the Editor-in-Chief of the IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS: SYSTEMS, from 2014 to 2019, and he is currently the Editor-in-Chief of the IEEE TRANSACTIONS ON CYBERNETICS, and an Associate Editor of the IEEE TRANSACTIONS ON ARTIFICIAL INTELLIGENCE, and IEEE TRANSACTIONS ON FUZZY SYSTEMS.

ZHULIN LIU (Member, IEEE) received the bachelor’s degree in mathematics from Shandong University, China, in 2005, and the M.S. degree in mathematics from the University of Macau, Macau, in 2009, and the Ph.D. degree from the Faculty of Science and Technology, University of Macau, Macau, in 2019. She currently holds a postdoctoral position with the School of Computer Science and Engineering, South China University of Technology, China. Her current research interests include machine learning, computational intelligence, and broad learning systems.