IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 25, 2019, accepted August 1, 2019, date of publication August 7, 2019, date of current version August 21, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2933550

Schizophrenia Classification Using fMRI Data Based on a Multiple Feature Image Capsule Network Ensemble

BO YANG, YUAN CHEN, QUAN-MING SHAO, RUI YU, WEN-BIN LI, GUAN-QI GUO, JUN-QIANG JIANG, AND LI PAN

School of Information Science and Engineering, Hunan Institute of Science and Technology, Yueyang 414006, China

Corresponding authors: Jun-Qiang Jiang (jjq@hnist.edu.cn) and Li Pan (panli.hnist@gmail.com)

This work was supported in part by the Natural Science Foundation of Hunan Province under Grant 2019JJ40105, Grant 2018JJ2153, Grant 2017JJ2016, and Grant 2018JJ2152, in part by the Science and Technology Program of Hunan Province under Grant 2016TP1021, in part by the Scientific Research Fund of Hunan Provincial Education Department under Grant 15A079 and Grant 17A089, and in part by the Scientific Research Innovation Project for Postgraduate of Hunan Province under Grant CX2018B777.

ABSTRACT Automatic diagnosis and classification of schizophrenia based on functional magnetic resonance imaging (fMRI) data have attracted increasing attention in recent years. Most previous studies abstracted highly compressed functional features from the view of brain science and fed them into shallow classifiers for this purpose. However, their classification performance in practical applications is unstable and unsatisfactory. As an acute psychotic disorder, schizophrenia shows functional complexity in fMRI data. Therefore, additional features and deep classification methods are needed to improve classification performance. In this study, we propose a multiple feature image capsule network ensemble approach for schizophrenia classification. The proposed approach proceeds in three steps: 1) extracting multiple image features from the perspective of linear sparse representation, nonlinear multiple kernel representation, and function connection of brain areas respectively; 2) feeding these image features into three specially designed independent capsule networks for classification; 3) obtaining the final results by fusing the outputs of these three deep capsule network using a ensemble approach. To further improve the classification performance, we design a optimization model of maximizing the square of correlation coefficients and propose a weighted ensemble technology based on this model, which is mathematically proved to be solved as a eigenvalue decomposition problem in certain case. Finally, the proposed approach is implemented and evaluated on the schizophrenia fMRI dataset from COBRE, UCLA and WUSTL. From the experimental results, we conclude that the proposed method outperforms some current methods and further improves the accuracy of schizophrenia classification.

INDEX TERMS Schizophrenia classification, multiple features extraction, deep capsule network, classifier ensemble.

I. INTRODUCTION

A. BACKGROUND

Schizophrenia is a devastating mental disease with extraordinary complexity. Diagnosis of schizophrenia with high confidence is important in neurosciences and medical science [1], [2]. High-resolution brain imaging techniques, such as functional magnetic resonance imaging (fMRI) [3], structural magnetic resonance imaging(sMRI) [4], diffusion tensor imaging(DTI) [5], positron emission tomography(PET) [6], facilitate understanding of the structure and function of human brain. These techiques have contributed greatly to the improved analysis and diagnosis of schizophrenia in recent years [7]–[11].

As a complex psychiatric disorder, schizophrenia shows local abnormalities of brain activity and functional connectivity networks in the schizophrenic brain feature disrupted topological properties [12]. As a gold-standard functional imaging technique in neurosciences, fMRI has become the most widely used imaging technique among all the above

The associate editor coordinating the review of this manuscript and approving it for publication was Jeonghwan Gwak.

mentioned imaging technologies for the analysis and diagnosis of schizophrenia [13]. To free imaging specialists from the heavy task of interpreting fMRI images, methods to diagnose schizophrenia automatically with high reliability need to be developed to simplify analysis of fMRI images. Different machine learning methods, including classification and feature extraction, have been introduced in recent years to realize the automatic analysis and diagnosis of schizophrenia [14].

From the perspective of statistical machine learning, schizophrenia analysis and diagnosis can be regarded as a typical statistical classification task. Similar to other common statistical classification approaches, most of existing approaches for schizophrenia classification based on fMRI have the same processing steps. After some data preprocessing operations, these existing approaches can be divided into three main steps: 1) extracting features from original fMRI data based on functional region of interests(ROI) information; 2) transforming ROI features into compressed features using linear or nonlinear feature transformation techniques; 3) optimizing a classifier by training it on these compressed features. Recently proposed approaches using deep learning [19]–[21] involve concise operations in which the last two steps are merged. And automatic feature extraction is realized by end-to-end learning.

B. MOTIVATION

Despite the continuous advances in schizophrenia classification approaches based on fMRI data, existing approaches are still in the infancy stage, and their diagnosis levels are considerably lower than those of human experts. The research of schizophrenia classification based on fMRI data still faces several inevitable problems and challenges.

First, little is presently known about the structures and functions of human brain. In specific, the etiology of schizophrenia and the abnormal modes of the brains of patients with this condition remain unclear [14]. As mentioned above, existing approaches obtain ROI features from current knowledge about brain regions. The only available materials in schizophrenia classification are fMRI data and prior knowledge about this condition. Therefore, a method to extract deep information from fMRI data by using information procesing technologies has become increasingly important.

Second, most existing schizophrenia classification approaches can be classified as shallow learning. Intuitionally, deep learning approaches [15]–[18] that are suitable for complex tasks should be used for highly complex schizophrenia classification tasks. Besides, recently proposed approaches using deep learning [19]–[21] involve concise operations and automatic feature extraction is realized by end-to-end learning. However, mainstream deep learning methods are data hungry. In addition, these methods are easily over-fitting and show poor performance in tasks with a small sample size. Unfortunately, the number of samples in a schizophrenia classification research is usually small. As listed in the paper [48], the mean and median sample size of 21 researches on schizophrenia classification in the recent 5 years(2014-2018) are 208 and 147 respectively. Hence, how to choose appropriate deep learning methods and design network architecture suitable for schizophrenia classification with small sample size should also be determined.

C. OUR CONTRIBUTIONS

In view of these problems, this paper proposes a deep learning approach for schizophrenia classification from several perspectives, such as multiple features extraction, deep network architecture design and classifier ensemble, to improve the accuracy and stability of classification. The details of our contributions are listed as below:

(1)We present a novel whole schizophrenia classification approach which contains 3 steps: multiple features extraction, deep capsule network training and multiple classifiers weighed ensemble. In our experiments, the approach outperforms some current methods.

(2)We present multiple fMRI features extraction method and extract linear sparse feature image, nonlinear multiple kernel feature image and functional connectivity feature image from three different perspectives: statistical linear analysis, statistical nonlinear analysis and brain regions analysis. This multiple fMRI features extraction method can extract more useful information from the view of data science and brain science separately.

(3)We introduce capsule network into schizophrenia diagnosis and design new capsule network architecture more suitable for schizophrenia diagnosis. At the basis of original capsule network, we add more convolution layers to enlarge the local receptive field and cancel the RELU nonlinear mapping layers to control network capacity, which are effective for solving over-fitting.

(4)We present a weighted ensemble method to complete the final classifier ensemble of our schizophrenia classification approach. This weighted ensemble method based on an optimization model maximizes the square of correlation coefficients. We discuss and prove in theory that in certain case the presented optimization model is essentially a eigenvalue decomposition problem, which means we can obtain the best weights rapidly by the common eigenvalue decomposition algorithms.

The rest of the paper is organized as follows: In section II, we review the related literature. In section III, we present multiple feature image capsule networks ensemble approach and describe it in details. In Section IV, the experimental results on a multi-site schizophrenia fMRI dataset from the center for biomedical research excellence(COBRE, available at http://openfmri.org/), the university of California, Los Angles(UCLA, available at http://openfmri.org/) and the conte center for the neuroscience of mental disorders at Washington university school of medicine in St. Louis(WUSTL, available at http://openfmri.org/) are presented and discussed by comparing with some other representative methods. In Section V, we conclude the paper.

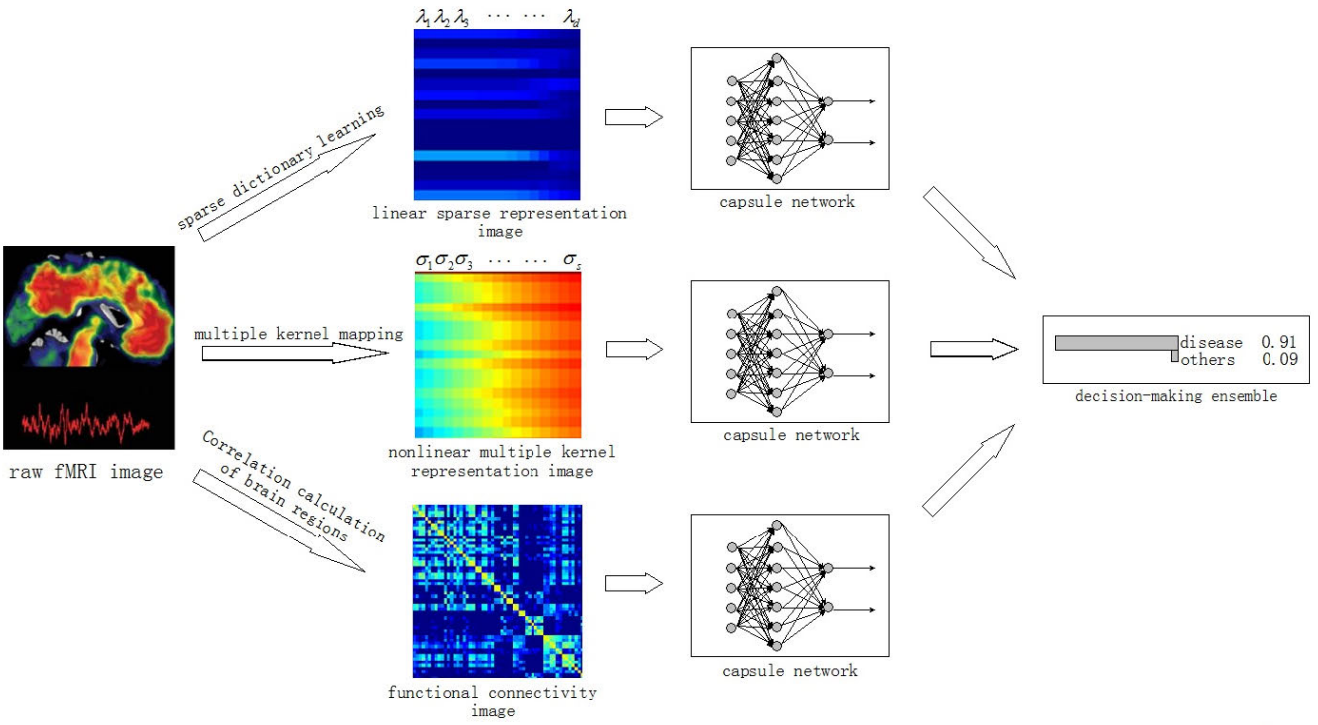


FIGURE 1. Illustration of multiple feature image capsule networks ensemble.

II. RELATED WORK

The researches of schizophrenia classification based on fMRI data have made great progress in recent years. AS a classification task, one of the most important things is choosing a good classifier. So far many linear and nonlinear classifiers have been applied in schizophrenia classification, which contain bayesian classifier [22], c-means classifier [23], k-nearest neighbor algorithm [24], artificial neural network [25], least square classifier [26], support vector machine(SVM) [27] and extreme learning machine(ELM) [28], [29].

The works mentioned above are generally based on ROIs information extracted from the original fMRI data. As pointed out in the paper [7], the dimension of functional connectivity is large even if ones only evaluate connectivity between defined ROIs. To avoid overfitting problem, many feature transformation technologies were also applied, which contain principle components analysis(PCA) [30], kernel PCA(KPCA) [31], fisher linear discriminant analysis (FLDA), kernel FLDA(KFLDA) [32], independent component analysis (ICA) [33], locally linear embedding (LLE) [23], canonical correlation analysis(CCA) [34], [35]. From the view of machine learning, all the above approaches can be classified as shallow machine learning. However, because there exists a gap between the highly complexity of brain function and the relatively poor representation power of shallow machine learning, further improving classification performance using shallow approaches becomes very difficult.

Recently, deep learning approaches were introduced in classification of brain disease [48], [49], which contain deep belief network [21], [36]–[38], Auto Encoder networks [19], [20], [39], [40], Convolution Neural Networks(CNN) [41]–[47]. Because of the powerful learning ability and automatic features extraction ability of deep learning, these approaches are hopeful to improve classification performance of brain disease. However, due to the relatively small number of samples [48], so far the deep neural networks widely used in classical computer vision tasks still have not shown distinguished classification power. As mentioned in the paper [48], there are still only 18 papers published in recent 5 years(2014–2018) using deep neural network for classification of various disorders(less than 4%, 18/209 papers) and SVM, a famous shallow learning method, is still the most popular methods(more than 55%, 117/209 papers). Hence, how to improve the generalization of deep learning in classification of brain disease becomes a key technology needed to be solved now.

III. OUR SCHIZOPHRENIA CLASSIFICATION APPROACH

A. ILLUSTRATION OF OUR APPROACH

Suppose we have a labeled fMRI image data set $S = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_i, y_i), \ldots, (\boldsymbol{x}_n, y_n)\}$, where $\boldsymbol{x}_i (\boldsymbol{x}_i \in R^{l_1 \times l_2 \times l_3 \times l_4}, 1 \leq i \leq n)$ is the ith 4D fMRI image data, $y_i (1 \leq i \leq n)$ is the category label of $\boldsymbol{x}_i$ and $y_i \in \{1, \ldots, C\}$ if there are C categories. Automatic diagnosis of brain diseases can be treated as a classification task in which unlabeled fMRI image are classified into certain categories by a classifier $y = f(\boldsymbol{x})$. In order to ensure the classification accuracy, the classifier $y = f(\boldsymbol{x})$ should be trained and optimized based on the minimization of empirical loss $\sum_{i=1}^{n} L(f(\boldsymbol{x}_i), y_i)$.

This paper proposes a multiple feature image capsule network ensemble approach for fMRI images classification. the proposed approach is illustrated in Fig.1. Different from the existing approaches, which mainly use functional connectivity feature set from ROI features to make decision, here we use classifier ensemble on multiple features to improve the accuracy and stability of brain disease classification. In our ensemble model, we first extract different 2D feature images based on the original high-dimensional 4D fMRI image data from three different perspectives: linear sparse representation, nonlinear kernel mapping and prior knowledge from brain science. Then the three 2D feature image are input into three capsule networks for individual decision-making. Finally, the classifier ensemble technology is used to integrate the three individual classification results and output the final ensemble decision-making results.

B. LINEAR SPARSE REPRESENTATION IMAGE

Linear sparse representation images are generated by sparse dictionary learning [50]. Because there is no need for temporal and spatial information here, first the 4D fMRI image $\boldsymbol{x}_i(\boldsymbol{x}_i \in R^{l_1\times l_2\times l_3\times l_4})$ is vectorized to a 1D vector $\boldsymbol{r}_i(\boldsymbol{r}_i \in R^{l\times 1}, l = l_1 \times l_2 \times l_3 \times l_4)$.

Suppose there exists a Dictionary matrix $\boldsymbol{D}(\boldsymbol{D} \in R^{l\times m})$ to be optimized, which contains m atoms $\{\boldsymbol{d}_i\}_{1\le i\le m}$ and $\boldsymbol{D} = (\boldsymbol{d}_1, \ldots, \boldsymbol{d}_i, \ldots, \boldsymbol{d}_m)$. The optimization model of sparse dictionary learning is defined as

$$\min_{\boldsymbol{D},\boldsymbol{S}} \quad Tr((\boldsymbol{R}-\boldsymbol{D}\boldsymbol{S})^T(\boldsymbol{R}-\boldsymbol{D}\boldsymbol{S})) + \lambda\Omega(\boldsymbol{S}), \tag{1}$$

where matrix $\boldsymbol{R}$ is the sample matrix and $\boldsymbol{R} \in R^{l\times n}, \boldsymbol{R} = (\boldsymbol{r}_1, \ldots, \boldsymbol{r}_i, \ldots, \boldsymbol{r}_n)$, matrix $\boldsymbol{S}(\boldsymbol{S} \in R^{m\times n}, \boldsymbol{S} = (\boldsymbol{s}_1, \ldots, \boldsymbol{s}_i, \ldots, \boldsymbol{s}_n))$ is the sparse presentation matrix using dictionary $\boldsymbol{D}$, $\Omega(\boldsymbol{S})$ is the sparsity regularization term for $\boldsymbol{S}$, $\lambda(\lambda \ge 0)$ is the regularization coefficient.

At present, the commonly used sparsity regularization terms are mainly based on L0 norm and L1 norm. Because L1 norm is of convexity in mathematics and brings more effective algorithms, here sparse dictionary learning based on L1 norm regularization is adopted. So model(1) can be rewritten as

$$\min_{\boldsymbol{D},\boldsymbol{S}} \quad Tr((\boldsymbol{R}-\boldsymbol{D}\boldsymbol{S})^T(\boldsymbol{R}-\boldsymbol{D}\boldsymbol{S})) + \lambda||\boldsymbol{S}||_1. \tag{2}$$

Eq.(2) can be optimized by alternating optimization of the below 2 sub problems

$$\min_{\boldsymbol{S}} Tr((\boldsymbol{R}-\boldsymbol{D}\boldsymbol{S})^T(\boldsymbol{R}-\boldsymbol{D}\boldsymbol{S})) + \lambda||\boldsymbol{S}||_1, \tag{3}$$

$$\min_{\boldsymbol{D}} Tr((\boldsymbol{R}-\boldsymbol{D}\boldsymbol{S})^T(\boldsymbol{R}-\boldsymbol{D}\boldsymbol{S})). \tag{4}$$

Details of the optimization algorithm is described in Algorithm 1.

Eq.(3) and (4) in Algorithm 1 are both convex optimization models. Eq.(3) can be solved quickly by some rapid algorithms such as the Least Absolute Shrinkage and Selection operator and Least Angle Regression, while Eq.(4) can be solved by K Singular Value Decomposition algorithm [50].

Algorithm 1 Optimization Algorithm for solving model(2)

Input: the sample matrix $\boldsymbol{R}$, the initialized Dictionary $\boldsymbol{D}(0)$, the regularization coefficient λ, stop threshold ε, $t = 0$

Output: the optimized Dictionary $\boldsymbol{D}^*$, the optimized sparse presentation matrix $\boldsymbol{S}^*$

1: Fix dictionary $\boldsymbol{D}(t)$ and optimize presentation matrix $\boldsymbol{S}(t)$ according to Eq.(3).
2: Fix sparse presentation matrix $\boldsymbol{S}(t)$ and optimize dictionary $\boldsymbol{D}(t+1)$ according to Eq.(4).
3: If $||\boldsymbol{S}(t-1)-\boldsymbol{S}(t)||_F^2 > \varepsilon \quad or \quad ||\boldsymbol{D}(t-1)-\boldsymbol{D}(t)||_F^2 > \varepsilon$, $t = t+1$, goto 1.
4: return the optimized Dictionary $\boldsymbol{D}^* = \boldsymbol{D}(t+1)$, the optimized sparse presentation matrix $\boldsymbol{S}^* = \boldsymbol{S}(t)$.

Using dictionary $\boldsymbol{D}$, a l-dimensional sample vector $\boldsymbol{r}_i$ is represented as a m-dimensional feature vector $\boldsymbol{s}_i$. Noting that the regularization coefficient λ has the effect of adjusting the sparsity of representation $\boldsymbol{S}$, here we set d different λ values $\lambda_1, \ldots, \lambda_d$ and obtain d different feature vectors of different degrees of sparsity. Finally, we combine these d different feature vectors into a sparsity representation matrix $\boldsymbol{S}_i$ which is called as linear sparsity representation image here

$$\boldsymbol{S}_i = (\boldsymbol{s}_i(\lambda_1), \ldots, \boldsymbol{s}_i(\lambda_d)). \tag{5}$$

C. MULTIPLE KERNEL REPRESENTATION IMAGE

Multiple kernel representation images are generated by using kernel mapping method [51]. For a l-dimensional 1D vector $\boldsymbol{r}_i$ in sample set $\{\boldsymbol{r}_i\}_{1\le i\le n}$, its nonlinear mapping can be described as

$$\boldsymbol{r}_i \to \phi(\boldsymbol{r}_i), \tag{6}$$

where $\phi(.)$ is a nonlinear mapping function.

Considering that $\phi(.)$ can hardly be defined in general, we cannot analyse it directly in this implicit nonlinear space. As an alternative, a well-defined kernel inner product function can be used. For two samples $\boldsymbol{r}_i, \boldsymbol{r}_j$, the kernel inner product about them is defined as

$$k(\boldsymbol{r}_i, \boldsymbol{r}_j) = \phi(\boldsymbol{r}_i)^T\phi(\boldsymbol{r}_j), \tag{7}$$

where $k(.,.)$ is the kernel inner product function.

Furthermore, the sample $\boldsymbol{r}_i$ can be mapped into n-dimensional kernel sampling space using kernel inner product function as a kernel sample $\varphi(\boldsymbol{r}_i)$, which can be described as

$$\boldsymbol{r}_i \to \varphi(\boldsymbol{r}_i) = (k(\boldsymbol{r}_1, \boldsymbol{r}_i), \ldots, k(\boldsymbol{r}_n, \boldsymbol{r}_i))^T. \tag{8}$$

Kernel sample $\varphi(\boldsymbol{r}_i)$ is related not only to the types of kernel inner product function but also to the values of super parameters. Here we set s different super parameter values $\sigma_1, \ldots, \sigma_s$ and obtain d different kernel samples for $\boldsymbol{r}_i$. Finally, we combine these s different kernel samples into a multiple kernel representation matrix $\boldsymbol{K}_i$ denoted as nonlinear

multiple kernel representation image

$$\boldsymbol{K}_i = \begin{bmatrix} k_{\sigma_1}(\boldsymbol{r}_1, \boldsymbol{r}_i) & \cdots & k_{\sigma_s}(\boldsymbol{r}_1, \boldsymbol{r}_i) \\ \vdots & \ddots & \vdots \\ k_{\sigma_1}(\boldsymbol{r}_n, \boldsymbol{r}_i) & \cdots & k_{\sigma_s}(\boldsymbol{r}_n, \boldsymbol{r}_i) \end{bmatrix}. \tag{9}$$

D. FUNCTIONAL CONNECTIVITY IMAGE

The above two presentation feature images are generated from the perspective of mathmatics.Unlike them, functional connectivity feature images are based on brain science. As we know, the first three dimensions of a 4D fMRI image $\boldsymbol{x}_i(\boldsymbol{x}_i \in R^{l_1 \times l_2 \times l_3 \times l_4})$ compose the spatial dimensions of brain voxels and the last one dimension is the temporal dimension. The time series of a brain voxel describes the variation in blood oxygen concentration with time at this brain voxel and evaluates the variation in activation and repression of this brain voxel. From the perspective of brain science, we can obtain the functional connectivity matrix about all ROIs by calculating the correlation coefficients of time series about every two ROIs. The calculation of functional connectivity matrix is described as follows:

(1) Select ROIs and obtain time series $\boldsymbol{v}_1^i, \ldots, \boldsymbol{v}_m^i$ in fMRI image $\boldsymbol{x}_i$ accrording to brain science;

(2) Calculate Pearson correlation coefficient w_{pq}^i of time series $\boldsymbol{v}_p^i$ and $\boldsymbol{v}_q^i$. The Pearson correlation coefficient w_{pq}^i is defined as

$$w_{pq}^i = \frac{(\boldsymbol{v}_p^i - \boldsymbol{m}_{\boldsymbol{v}_p^i})^T (\boldsymbol{v}_q^i - \boldsymbol{m}_{\boldsymbol{v}_q^i})}{||\boldsymbol{v}_p^i - \boldsymbol{m}_{\boldsymbol{v}_p^i}||_2 ||\boldsymbol{v}_q^i - \boldsymbol{m}_{\boldsymbol{v}_q^i}||_2}, \tag{10}$$

where $\boldsymbol{m}_{\boldsymbol{v}_p^i}$, $\boldsymbol{m}_{\boldsymbol{v}_q^i}$ are the mean values of time series $\boldsymbol{v}_p^i$, $\boldsymbol{v}_q^i$. Finally, we obtain a functional connectivity matrix $\boldsymbol{W}_i$ denoted as functional connectivity image here

$$\boldsymbol{W}_i = \begin{bmatrix} w_{11}^i & \cdots & w_{1m}^i \\ \vdots & \ddots & \vdots \\ w_{m1}^i & \cdots & w_{mm}^i \end{bmatrix}. \tag{11}$$

E. CAPSULE NETWORK DESIGN

Capsule network is a neural network by Hinton in 2017 [52]. Compared with CNN, the capsule network discovers the equivariance of features by introducing several capsule layers. It has been found experimentally that capsule network can solve some problems with a relatively small sample size effectively [52].

To improve the generalization performance of the capsule network for schizophrenia classification, here we adjusted the architecture of the original capsule network, which is illustrated in Fig.2.

In addition to the basic layers of the original capsule network, such as Convolution layer, PrimaryCaps layer, ClassCapes layer and L2 output layer, we make the following adjustments:

(1) The single convolution layer of the original capsule network is extended to multiple convolution layers to reduce the number of network parameters and expand the local receptive field.

(2) All convolution layers of our capsule network are designed not containing nonlinear activation, such as RELU, to avoid over fitting.

F. WEIGHTED CLASSIFIER ENSEMBLE

Suppose the score outputs of n training samples from m different classifiers are $\{\boldsymbol{o}_i\}_{1 \le i \le m}$, where $\boldsymbol{o}_i(\boldsymbol{o}_i = (o_i^1, \ldots, o_i^j, \ldots, o_i^n))$ is the output vector from the ith classifier and o_i^j is the output of the jth training sample from the ith classifier. To find the best score fusion results, here we propose a multiple classifier weighted ensemble method based on the maximization of the square of correlation coefficients. The optimization model is formulated as below

$$\begin{aligned} &max\ J(\boldsymbol{\alpha}) = \sum_{i=1}^{m} \frac{\boldsymbol{\alpha}^T \boldsymbol{O}^T \boldsymbol{o}_i \boldsymbol{o}_i^T \boldsymbol{O} \boldsymbol{\alpha}}{||\boldsymbol{O}\boldsymbol{\alpha}||_2^2 ||\boldsymbol{o}_i||_2^2} \\ &s.t.\ \boldsymbol{\alpha} \ge 0, \quad \boldsymbol{\alpha}^T \boldsymbol{e} = 1, \end{aligned} \tag{12}$$

where $\boldsymbol{O}(\boldsymbol{O} \in R^{n \times m})$ is the output matrix and $\boldsymbol{O} = (\boldsymbol{o}_1, \ldots, \boldsymbol{o}_m)$, $\boldsymbol{\alpha}(\boldsymbol{\alpha} \in R^{m \times 1})$ is the weight vector, and $\boldsymbol{e}(\boldsymbol{e} \in R^{m \times 1})$ is a vector whose elements are equal to 1.

In Eq.(12), we find the best weight vector to maximize the square of correlation coefficients between the fusion output vector $\boldsymbol{O}\boldsymbol{\alpha}$ and all outputs from the m single classifiers $\boldsymbol{o}_i$. We use the square of correlation coefficients instead of correlation coefficientsbecause the former is more convenient to calculate. Considering the mathematical significance of weights, we introduce the constraints $\boldsymbol{\alpha} \ge 0, \boldsymbol{\alpha}^T \boldsymbol{e} = 1$ into our model.

After defining inner product matrix of outputs $\boldsymbol{K} = \boldsymbol{O}^T \boldsymbol{O}$, Eq.(12) can be further reformulated as

$$\begin{aligned} &max\ J(\boldsymbol{\alpha}) = \frac{\boldsymbol{\alpha}^T \boldsymbol{K}\boldsymbol{S}\boldsymbol{K}\boldsymbol{\alpha}}{\boldsymbol{\alpha}^T \boldsymbol{K}\boldsymbol{\alpha}} \\ &s.t.\ \boldsymbol{\alpha} \ge 0, \quad \boldsymbol{\alpha}^T \boldsymbol{e} = 1, \end{aligned} \tag{13}$$

where $\boldsymbol{S}(\boldsymbol{S} \in R^{m \times m})$ is a diagonal matrix and $S_{ii} = \frac{1}{||\boldsymbol{o}_i||_2^2}$.

Observing the main optimization item of Eq.(13) separately, we can find that it is essentially a generalized eigenvalue decomposition model. Without considering the constraints, the best weight vector is exactly the first eigenvector corresponding to the first eigenvalue of the main optimization item. Unfortunately, after introducing the constraints, the conclusion is usually invalid.

After further analysis, we find that Eq.(13) can be treated as an unconstrained optimization model under certain conditions, which is described in Theorem 1.

Theorem 1: If outputs set $\{\boldsymbol{o}_i\}_{m \ge i \ge 1}$ are linearly independent set and every 2 outputs $\boldsymbol{o}_i, \boldsymbol{o}_j$ satisfy the condition $\boldsymbol{o}_i^T \boldsymbol{o}_j > 0$, then Eq.(13) is equivalent to the generalized eigenvalue decomposition model $max\ \frac{\boldsymbol{\alpha}^T \boldsymbol{K}\boldsymbol{S}\boldsymbol{K}\boldsymbol{\alpha}}{\boldsymbol{\alpha}^T \boldsymbol{K}\boldsymbol{\alpha}}$.

Proof: Suppose the best solution of weight vector for Eq.(13) is $\boldsymbol{\alpha}^*$, then we can deduce that using $\boldsymbol{\beta} = c\boldsymbol{\alpha}^*(c \in R, c \ne 0)$ can also obtain the best objective value and

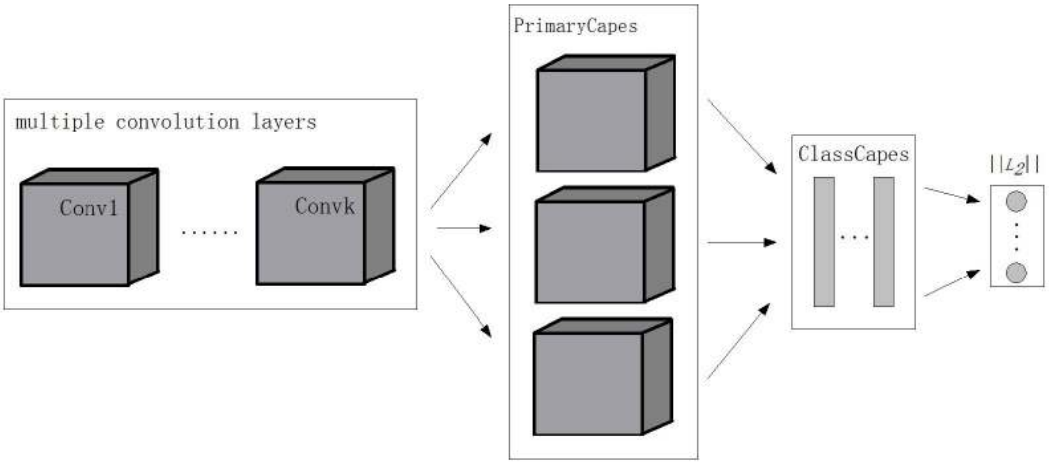


FIGURE 2. Basic framework of our capsule network.

$\frac{\boldsymbol{\alpha}^{*T}\boldsymbol{KSK}\boldsymbol{\alpha}^*}{\boldsymbol{\alpha}^{*T}\boldsymbol{K}\boldsymbol{\alpha}^*} = \frac{\boldsymbol{\beta}^T\boldsymbol{KSK}\boldsymbol{\beta}}{\boldsymbol{\beta}^T\boldsymbol{K}\boldsymbol{\beta}}$. It implies that the constraint $\boldsymbol{\alpha}^{*T}\boldsymbol{e} = 1$ is unnecessary in Eq.(13). Let $\boldsymbol{\beta}^*$ be the best solution of the similar model not containing the constraint $\boldsymbol{\alpha}^T\boldsymbol{e} = 1$. We can obtain the best vector $\boldsymbol{\alpha}^*$ for Eq (13) by normalizing $\boldsymbol{\beta}^*$

$$\boldsymbol{\alpha}^* = \frac{\boldsymbol{\beta}^*}{\boldsymbol{\beta}^{*T}\boldsymbol{e}}. \tag{14}$$

In this case, $J(\boldsymbol{\alpha}^*) = J(\boldsymbol{\beta}^*)$. Hence, the constraint $\boldsymbol{\alpha}^T\boldsymbol{e} = 1$ can be removed from Eq.(13). The main optimization item of Eq.(13) is a generalized eigenvalue decomposition model. According to Lagrangian Multiplier method, the best solution satisfies

$$\boldsymbol{KSK}\boldsymbol{\alpha} = \lambda\boldsymbol{K}\boldsymbol{\alpha}, \tag{15}$$

where $\lambda(\lambda > 0)$ is the Lagrangian multiplier. Because $\{\boldsymbol{o}_i\}_{m\geq i\geq 1}$ are linearly independent, matrix $\boldsymbol{K}$ is always invertible. Eq.(15) can be reformulated as

$$\boldsymbol{SK}\boldsymbol{\alpha} = \lambda\boldsymbol{\alpha}. \tag{16}$$

Eq.(16) shows that the generalized eigenvalue decomposition problem can be treated as a eigenvalue decomposition problem for matrix $\boldsymbol{SK}$ when matrix $\boldsymbol{K}$ is invertible.

Let $\boldsymbol{\alpha}^*$ be the first eigenvector corresponding to the first eigenvalue of matrix $\boldsymbol{P}(\boldsymbol{P} = \boldsymbol{SK})$. In this case, the best objective value is $\frac{\boldsymbol{\alpha}^{*T}\boldsymbol{P}\boldsymbol{\alpha}^*}{\boldsymbol{\alpha}^{*T}\boldsymbol{\alpha}^*}$. In general, $\boldsymbol{\alpha}^* \geq 0$ does not hold. However, when the condition, every 2 outputs $\boldsymbol{o}_i, \boldsymbol{o}_j$ satisfy the condition $\boldsymbol{o}_i^T\boldsymbol{o}_j > 0$, holds, we find it always holds.

In this case, $\boldsymbol{K} \geq 0$ and $\boldsymbol{P} \geq 0$ also hold, and we have

$$\begin{aligned}\frac{\boldsymbol{\alpha}^{*T}\boldsymbol{P}\boldsymbol{\alpha}^*}{\boldsymbol{\alpha}^{*T}\boldsymbol{\alpha}^*} &= \frac{\sum_{i=1}^m\sum_{j=1}^m \alpha_i^*\alpha_j^*P_{ij}}{\boldsymbol{\alpha}^{*T}\boldsymbol{\alpha}^*} \\ \leq \frac{\sum_{i=1}^m\sum_{j=1}^m |\alpha_i^*||\alpha_j^*|P_{ij}}{\boldsymbol{\alpha}^{*T}\boldsymbol{\alpha}^*} &= \frac{\sum_{i=1}^m\sum_{j=1}^m |\alpha_i^*||\alpha_j^*|P_{ij}}{|\boldsymbol{\alpha}^*|^T|\boldsymbol{\alpha}^*|}.\end{aligned}$$

The above inequality shows that $|\boldsymbol{\alpha}^*|$ is a better solution than the best vector $\boldsymbol{\alpha}^*$. Thus, when every two outputs $\boldsymbol{o}_i, \boldsymbol{o}_j$ satisfy the condition $\boldsymbol{o}_i^T\boldsymbol{o}_j > 0$, the best solution $\boldsymbol{\alpha}^* \geq 0$ always holds, which means that the constraint $\boldsymbol{\alpha}^* \geq 0$ can also be removed from Eq.(13). ■

In application, the condition $\boldsymbol{o}_i^T\boldsymbol{o}_j > 0$ in Theorem 1 can be achieved naturally. Taking a frequently-used classifier, softmax classifier, as an example, the outputs $\boldsymbol{o}_i$ of the ith softmax classifier are the probabilities of samples belonging to all categories. Clearly, $\boldsymbol{o}_i \geq 0$ always holds, which means that $\boldsymbol{o}_i^T\boldsymbol{o}_j > 0$ also holds for any 2 outputs $\boldsymbol{o}_i, \boldsymbol{o}_j$. SVM classifier, which is another frequently used classifier, is also used in this paper. The original outputs $\boldsymbol{o}_i(\boldsymbol{o}_i = \{o_{ij}\}, o_{ij} \in R)$ of the ith SVM classifier can also be transformed into the probability outputs $\frac{1}{1+exp(Ao_{ij}+B)}$(Refer to paper [53] for more details). Hence, here we find the best weight vector by solving eigenvalue decomposition (16) directly.

IV. EXPERIMENTS

A. DATASET

The dataset includes 385 subjects that are composed of 153 patients with schizophrenia and 232 healthy controls from three imaging resources.

The first sub-dataset is COBRE, which includes 72 patients with schizophrenia and 75 healthy controls. The original fMRI data were obtained from a 3 Tesla SIEMENS TIM scanner with the following parameters:time of repetition TR = 2000ms, echo time TE = 29ms, flip angle FA = 75°, field of view FOV = 192mm, 4mm thickness and 0mm gap, matrix size 64×64 and the number of axial slices is 32.

The second sub-dataset is UCLA, which includes 58 patients with schizophrenia and 134 healthy controls. The original fMRI data were obtained from a 3 Tesla SIEMENS TIM scanner with the following parameters:time of repetition TR = 2000ms, echo time TE = 30ms, flip angle FA = 90°, field of view FOV = 192mm, 4mm thickness and 0mm gap, matrix size 64×64 and the number of axial slices is 34.

The last sub-dataset is WUSTL, which includes 23 patients with schizophrenia and 41 healthy controls. All sub-datasets are available at the web site (https://openfmri.org/).

The original fMRI data were obtained from a 3 Tesla SIEMENS TIM scanner with the following parameters:time of repetition TR = 2500ms, echo time TE = 27ms, flip angle FA = 90°, field of view FOV = 256mm, 4mm thickness and 0mm gap, matrix size 64×64 and the number of axial slices is 33.

B. PREPROCESSING AND FEATURE GENERATION

All fMRI data were preprocessed as previously described [54], [55] by using a statistical parametric mapping software package(SPM8, which can be downloaded freely from the web site: http://www.fil.ion.ucl.ac.uk/spm). For each subject, the first 5 frames of the scanned data were discarded for magnetic saturation. The following preprocessing steps then proceeded in turn: a) slice timing correction; b) motion correction; c) normalization with an EPI template in the Montreal Neurological Institute atlas space (3mm isotropic voxels); d) spatial smoothing with a 6mm full-width half-maximum Gaussian kernel; e) linear detrend and band-pass temporal filtering (frequency range:0.01-0.08Hz); f) regression of nuisance variables, including the 6 parameters obtained by rigid body head motion correction, ventricular and white matter signals, and their first temporal derivatives, quadratic terms, and squares of derivatives (32P); and g) if frame-wise displacement at any point in time exceeded 0.3mm, then that time point was scrubbed.

After the above processing steps, data cleaning operations such as control of motion artifact, balancing for age and gender between the patient and control groups were then performed according to paper [19], [20]. Finally, we obtained 222 image samples composed of 102 patients with schizophrenia and 120 healthy controls, which were well matched in gender (patients vs. controls: 47/55 vs. 70/51 males/females) and age (patients vs. controls: 33.41±9.47 vs. 31.99±10.08 years). All samples were then adjusted to 4D images with the same size 53×63×52×140. The linear sparse representation image, nonlinear multiple kernel representation image, and functional connectivity image are all from the above preprocessed 4D images.

Linear sparse representation image. The dictionary contains 80 atoms. All the atoms are selected initially and randomly from another fMRI image dataset, Human Connectome Project and 20 values of the regularization coefficient $\lambda : \{2^{-10}, 2^{-9}, \ldots, 2^{8}, 2^{9}\}$ are set. Finally, we obtain a linear sparse representation feature image with size 80×20.

Nonlinear multiple kernel representation image. The Gaussian kernel function $k(r_i, r_j) = \exp(\frac{||r_i - r_j||_2^2}{-\sigma^2})$ is used and 20 values of the kernel parameter $\sigma : \{\sigma_s \times 1.2^0, \sigma_s \times 1.2^1, \ldots, \sigma_s \times 1.2^{48}, \sigma_s \times 1.2^{49}\}$ are set. σ_s is calculated as previously described [56]. In every training stage, all 100 samples are used as representation basis. Finally, we obtain a nonlinear multiple kernel representation feature image with size 100×50.

Functional connectivity image. All voxels are first divided into 116 brain regions according to the automated anatomical atlas(AAL). Then, we calculated the mean blood oxygen concentration of all brain regions, which are treated as the time series of all ROIs. We further calculate all the correlation coefficients between every 2 ROIs. Finally, we obtain a functional connectivity feature image with size 116×116.

C. APPROACHES AND EXPERIMENTAL SETTINGS

In our experiments, the 10 fold cross validation method is used to evaluate the proposed approach. Then, SVM, ELM, DAN, CNN, the original capsule network(capsule network-1) and our modified capsule network(capsule network-2) are employed as the final comparative classifiers for both single and multiple features.The details of settings are as follows:

SVM. We use the libSVM toolbox which can be downloaded freely from the web site(www.csie.ntu.edu.tw/cjlin/libsvm/). In our experiments, we use linear SVM and the penalty parameter C of SVM is selected from $\{10^{-5}, 10^{-4}, \ldots, 10^{4}, 10^{5}\}$ by grid search method.

ELM. The ELM is programmed on MATLAB platform. Our ELM is of single hidden layer and the number of nodes in the hidden layer is selected from $\{50, 100, \ldots, 450, 500\}$ by grid search method. Here the activation function of our ELM is sigmoid function $sigmoid(x) = \frac{1}{1+exp(-x)}$.

DAN. The DAN is also programmed on MATLAB platform. The number of hidden layers is chosen in {2, 3, 4, 5}. For simplicity, the numbers of nodes in hidden layers are all same. The number of nodes in a hidden layer is selected from {50, 100, 150}. Here the activation function is tanh function $tanh(x) = \frac{exp(x)-exp(-x)}{exp(x)+exp(-x)}$.

CNN. In our experiment, the CNN is designed and completed on the TensorFlow platform. The designed CNN contains K composite convolution layers and one fully connected classification layer. The output size of the fully connected classification layer is 2, which is the same as the number of classes in our experiment. For simplicity, every composite convolution layer is composed of a convolution layer and a max pooling layer. The RELU nonlinearly mapping layers are excluded from our composite convolution layers. In our CNN, the window size and stride size are 2 for all max pooling layers. In all convolution layers, the kernel size is 3, the stride size is 1, and padding is equal to "same". The number of output channels is chosen in {8,16,32}, and the number of composite convolution layers k is chosen in {1,2,3,4}. When training our CNN, batch size is 5, flop size is 200, initial learning rate is 0.01 and Adam optimizer is used for learning rate adjustment.

Capsule network-1. Here we use a publicly accessible edition of capsule network(https://github.com/naturomics/CapsNetTensorflow) on the TensorFlow platform.

TABLE 1. Classification results using linear sparse representation feature image(%).

$Classifiers$	SC	SS	ACC	PPV	NPV	$F1-Score$
SVM	85.12±14.40	55.84±14.92	73.74±11.94	76.12±24.38	75.07±7.58	62.55±15.28
ELM	75.21±8.22	74.03±14.02	74.75±7.19	66.04±8.99	82.62±8.70	69.08±9.06
DAN	74.38±20.65	71.43±14.74	73.23±12.06	70.44±17.59	73.97±18.46	68.48±13.36
CNN	80.17±13.37	58.44±20.66	71.72±8.03	67.88±15.87	76.07±7.56	59.9±15.95
$capsule\ network-1$	74.38±19.27	72.73±23.12	73.74±15.48	66.14±20.61	81.12±13.93	68.25±19.79
$capsule\ network-2$	80.99±12.50	68.83±15.41	76.26±8.99	71.84±14.39	80.85±8.04	69.23±11.52

TABLE 2. Classification results using nonlinear multiple kernel representation feature image(%).

$Classifiers$	SC	SS	ACC	PPV	NPV	$F1-Score$
SVM	90.91±8.13	42.86±18.07	72.22±9.94	75.00±20.22	71.77±7.60	53.33±18.64
ELM	85.95±8.49	48.05±21.45	71.21±11.34	67.14±19.76	72.84±9.22	54.99±20.53
DAN	76.86±17.08	68.83±14.59	73.74±10.35	71.00±15.94	78.55±15.46	67.64±12.66
CNN	87.60±10.18	44.16±25.12	70.71±11.14	70.38±22.34	72.19±10.07	51.15±22.84
$capsule\ network-1$	85.95±15.41	50.65±22.46	72.22±13.38	75.49±26.42	73.75±10.79	57.54±19.94
$capsule\ network-2$	80.17±17.52	72.73±14.29	77.27±10.10	75.89±16.31	80.40±16.59	71.93±12.64

TABLE 3. Classification results using functional connectivity feature image(%).

$Classifiers$	SC	SS	ACC	PPV	NPV	$F1-Score$
SVM	85.95±13.09	58.44±22.55	75.25±11.75	73.15±18.17	77.26±10.21	63.25±19.92
ELM	88.43±23.10	61.04±14.67	77.77±17.74	79.19±21.80	78.22±16.32	68.08±16.77
DAN	85.12±9.34	70.13±21.62	79.29±8.10	76.99±11.59	83.17±10.39	71.00±13.84
CNN	78.51±13.65	71.43±20.20	75.76±12.49	69.03±17.64	82.07±11.84	69.25±16.61
$capsule\ network-1$	82.64±13.15	71.43±23.04	78.28±9.45	76.40±16.24	83.47±10.89	70.43±14.88
$capsule\ network-2$	88.43±10.03	71.43±19.17	81.82±7.89	82.77±14.86	83.92±8.76	74.28±13.69

The capsule network-1 contains 1 RELU convolution layers, 1 primary capsule layer, 1 class capsule layer, and 1 L2 output layer. In the RELU convolution layer, the kernel size is 3, the stride size is 1, padding is equal to "valid". The number of output channels is chosen in {8,16,32}.In the primary capsule layer, the length of capsule is 4, the kernel size is 2, and the stride size is 2. In the class capsule layer, the length of capsule is 30 and the number of capsules is 2, which is equal to the number of classes in our experiment. When training capsule network-1, the batch size, flop size, initial learning rate, and optimizer are the same as those used in our CNN.

Capsule network-2. The capsule network-2 is designed and completed based on the above capsule network-1. Different from only 1 RELU convolution layer in the capsule network-1, the capsule network-2 contains k linear convolution layers. The number of convolution layers k is chosen in {3,4,5,6}, the number of output channels is also chosen in {8,16,32}, and the values of the other super parameters in the capsule network-2 are the same as those in the capsule network-1.

D. EXPERIMENTAL RESULTS AND ANALYSIS

1) CLASSIFICATION ACCURACY AND ROC CURVES

Tables 1-4 show the average specificity(SC), sensitivity(SS), classification accuracy(ACC), positive predictive value(PPV), negative predictive value(NPV) and F1-Score of all classifiers when using linear sparse representation feature image, nonlinear multiple kernel representation feature image, functional connectivity feature image and our score weighted fusion on the multi-site data set.

Tables 1-3 show that the capsule network-2 has better classification performance than the others when using each single features and achieves the best average classification accuracy of 81.82% when using functional connectivity feature image, with 2.53% higher than the second-best average classification accuracy of 79.29% from DAN. By comparison, the capsule network-1 shows no obvious improvement of classification performance. It proves that the proposed architecture adjustment for capsule network works indeed. Tables 4 shows that our weighted ensemble technology further improves the classification performance and that the final classification accuracy when using the capsule network-2 is up to 82.83%, with 2.02% higher than the second-best average classification accuracy of 80.81% from DAN. For the other classifiers except SVM and ELM, our weighted ensemble technology also improves the classification performance to some extent.

Furthermore, we calculate the AUC values and draw the ROC curves of all classifiers and illustrate them in Fig.3.

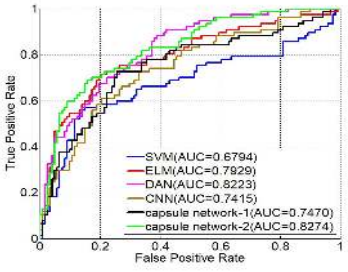
Fig.3 shows that the capsule network-2 obtains the best AUC values, followed by the DAN, finally the other classifiers. From the view of AUC values, functional connectivity feature image is the best, followed by linear sparse representation feature image, finally nonlinear multiple kernel representation feature image. Using our weighted ensemble technology, the best AUC value is up to 0.9141.

2) GRID SEARCH FOR PARAMETERS OF CNN AND CAPSULE NETWORK

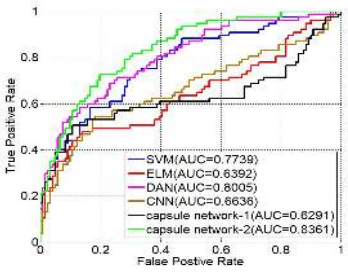
The above experimental results are obtained by using the best parameters, which are determined through grid search method on validation samples. Figs.4, 5, and 6 illustrate the

TABLE 4. Classification results using our weighted ensemble(%).

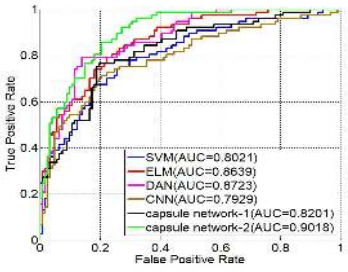
$Classifiers$	SC	SS	ACC	PPV	NPV	$F1-Score$
SVM	85.12±13.72	57.14±18.07	74.24±11.19	73.58±20.80	76.07±8.34	62.81±16.79
ELM	78.51±10.83	76.62±12.92	77.77±8.96	73.93±18.93	82.21±6.01	74.27±13.31
DAN	87.60±10.18	70.13±17.44	80.81±7.60	81.19±14.57	82.91±7.50	73.11±13.04
CNN	80.99±12.50	68.83±15.41	76.26±8.99	71.84±14.39	80.85±8.04	69.08±11.52
$capsule\ network-1$	80.99±14.56	75.32±13.48	78.79±9.33	74.74±10.08	84.74±7.95	73.17±9.49
$capsule\ network-2$	85.95±9.42	77.92±17.34	82.83±7.64	79.05±10.55	86.90±8.84	77.30±11.36



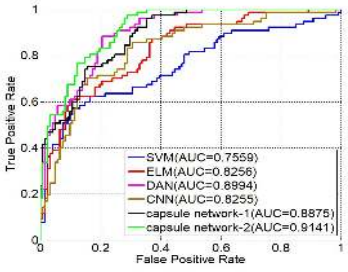
(a) linear sparse representation feature image



(b) nonlinear multiple kernel representation feature image



(c) functional connectivity feature image



(d) our weighted ensemble

FIGURE 3. ROC curves.

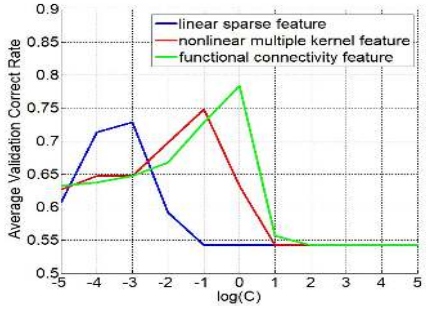


FIGURE 4. Average validation correct rates using SVM under different parameter settings.

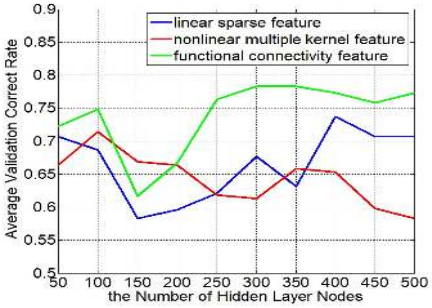


FIGURE 5. Average validation correct rates using ELM under different parameter settings.

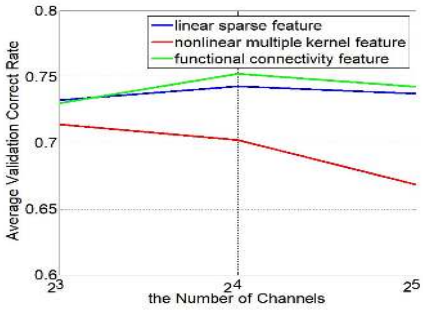
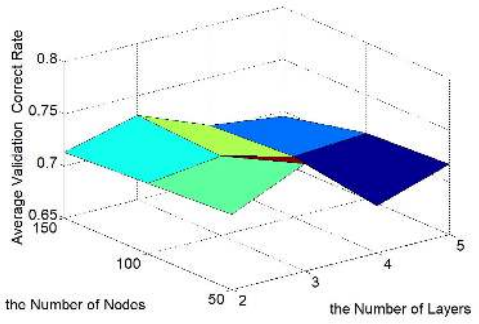


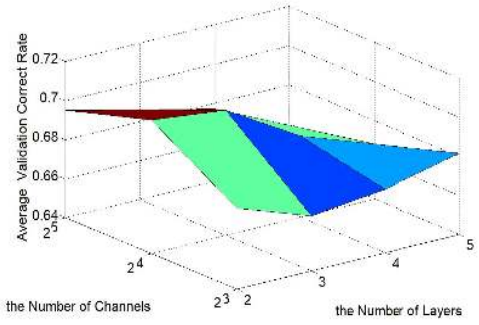
FIGURE 6. Average validation correct rates using capsule network-1 under different parameter settings.

average validation correct rates of SVM, ELM and capsule network-1 when using different single parameter settings. Figs.7, 8, and 9 illustrate the average validation correct rates of DAN, CNN and capsule network-2 when using different combinations of 2 parameters.

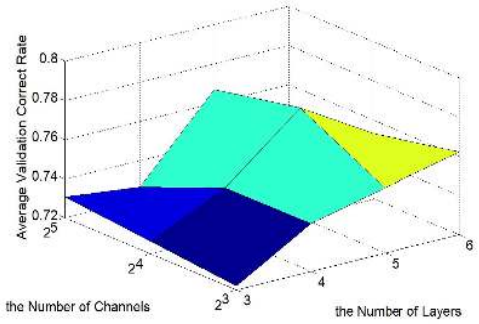
Using linear sparse representation feature image, we find that the best penalty parameter C for SVM is 10^{-3}, the best number of hidden layer nodes for ELM is 400, the best number of channels for capsule network-1 is 16, the best number of layers and the best number of nodes for DAN is 3 and 50 respectively, the best number of layers and the best number of channels for capsule network-2 are 5 and 16 respectively, the best number of layers and the best number of channels for



(a) DAN

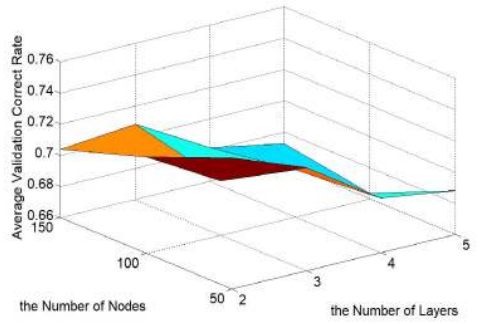


(b) CNN

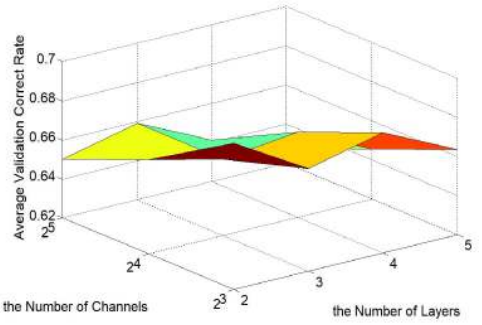


(c) Capsule Network-2

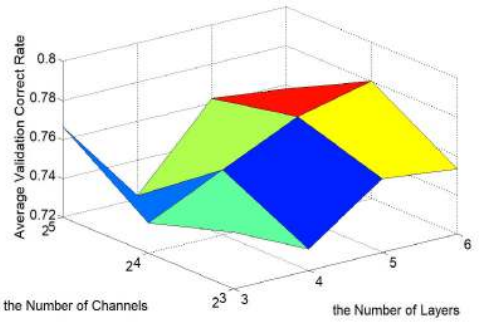
FIGURE 7. Average validation correct rates for linear sparse representation feature image using DAN, CNN and capsule network-2 under different parameter settings.



(a) DAN



(b) CNN

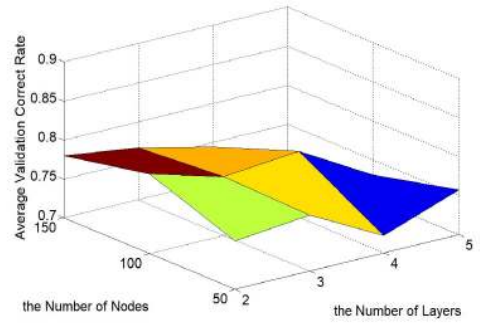


(c) Capsule Network-2

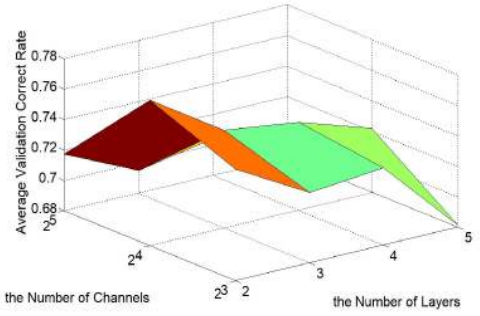
FIGURE 8. Average validation correct rates for nonlinear multiple kernel representation feature image using DAN, CNN and capsule network-2 under different parameter settings.

CNN are 2 and 16 respectively. Using nonlinear multiple kernel representation feature image, we find that the best penalty parameter C for SVM is 10^{-1}, the best number of hidden layer nodes for ELM is 100, the best number of channels for capsule network-1 is 8, the best number of layers and the best number of nodes for DAN is 2 and 50 respectively, the best number of layers and the best number of channels for our capsule network are 6 and 16 respectively, the best number of layers and the best number of channels for CNN are 2 and 8 respectively. Using functional connectivity feature image, we find that the best penalty parameter C for SVM is 10^{0}, the best number of hidden layer nodes for ELM is 300, the best number of channels for capsule network-1 is 16, the best number of layers and the best number of nodes for DAN is 2 and 100 respectively, the best number of layers and the best number of channels for our capsule network are 6 and 16 respectively, the best number of layers and the best number of channels for CNN are 2 and 16 respectively.

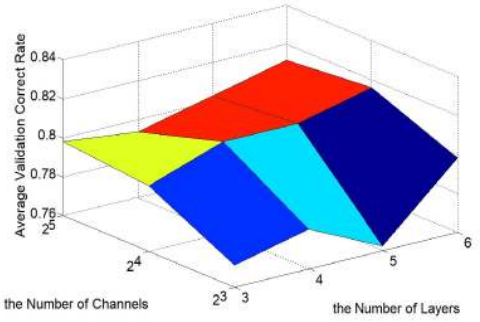
Figs.7, 8, and 9 illustrate that the average validation correct rates of DAN and CNN are reduced obviously as the number of layers is increased, which implies that over-fitting becomes more and more obvious with the increase of the number of nonlinear layers. By contrast, the variation of average validation correct rates of capsule network-2 with the number of layers are not the case, which shows that our introducing multiple linear layers into capsule network to increase the depth of network can avoid over-fitting effectively while improving the ability of covariant features extraction. Besides, the Figs.5,6,7,8 and 9 illustrate that the average validation correct rates of all neural networks including ELM,



(a) DAN



(b) CNN



(c) Capsule Network-2

FIGURE 9. Average validation correct rates for functional connectivity feature image using DAN, CNN and capsule network-2 under different parameter settings.

DAN, CNN, capsule network-1 and capsule network-2 are not very sensitive to the width of network.

V. CONCLUSION

To improve the effectiveness of schizophrenia classification, we propose a multiple feature image capsule networks ensemble method. In the proposed method, we introduce multiple features extraction, deep capsule network design, and a novel weighted classifier ensemble to increase the complementarity of features and improve the generality of classification. Finally, we conduct the comparative experiments on the schizophrenia fMRI dataset from COBRE, UCLA and WUSTL. Experimental results show that the proposed method performs better than the other comparative methods and and that the average correct rate of schizophrenia classification increases to 82.83%.

ACKNOWLEDGEMENT

(Bo Yang, Yuan Chen, and Quan-Ming Shao contributed equally to this work.) The authors would like to thank the anonymous reviewers for their valuable comments on improving the paper.

REFERENCES

[1] N. C. Andreasen and W. T. Carpenter, Jr., "Diagnosis and classification of schizophrenia," *Schizophrenia Bull.*, vol. 19, no. 2, pp. 199–214, Feb. 1993.

[2] N. C. Andreasen, "Diagnosis of schizophrenia," *Schizophrenia Bull.*, vol. 13, no. 1, pp. 9–22, Jan. 1987.

[3] N. K. Logothetis, J. Pauls, M. Augath, T. Trinath, and A. Oeltermann, "Neurophysiological investigation of the basis of the fMRI signal," *Nature*, vol. 412, pp. 150–157, Jul. 2001.

[4] Y.-D. Zhang, Z. Dong, L. Wu, and S. Wang, "A hybrid method for MRI brain image classification," *Expert Syst. Appl.*, vol. 38, pp. 10049–10053, Aug. 2011.

[5] Y. Assaf and O. Pasternak, "Diffusion tensor imaging (DTI)-based white matter mapping in brain research: A review," *J. Mol. Neurosci.*, vol. 34, no. 1, pp. 51–61, Jan. 2008.

[6] K. Vunckx, A. Atre, K. Baete, A. Reilhac, C. M. Deroose, K. Van Laere, and J. Nuyts, "Evaluation of three MRI-based anatomical priors for quantitative PET brain imaging," *IEEE Trans. Med. Imag.*, vol. 31, no. 3, pp. 599–612, Mar. 2012.

[7] P. Mcguire, O. D. Howes, J. Stone, and P. Fusar-Poli, "Functional neuroimaging in schizophrenia: Diagnosis and drug discovery," *Trends Pharmacol. Sci.*, vol. 29, no. 2, pp. 91–98, Feb. 2008.

[8] S. Ehrlich, S. Brauns, A. Yendiki, B.-C. Ho, V. Calhoun, S. C. Schulz, R. L. Gollub, and S. R. Sponheim, "Associations of cortical thickness and cognition in patients with schizophrenia and healthy controls," *Schizophrenia Bull.*, vol. 38, no. 5, pp. 1050–1062, Sep. 2012.

[9] J. Sui, H. He, Q. Yu, J. Rogers, G. Pearlson, A. R. Mayer, J. Bustillo, J. Canive, and V. D. Calhoun, "Combination of resting state fMRI, DTI, and sMRI data to discriminate schizophrenia by N-way MCCA+ jICA," *Frontiers Hum. Neurosci.*, vol. 7, p. 235, May 2013.

[10] D. F. Wong, H. Kuwabara, A. G. Horti, V. Raymont, J. Brasic, M. Guevara, W. Ye, R. F. Dannals, H. T. Ravert, A. Nandi, A. Rahmim, J. E. Minge, I. Grachev, C. Roy, and N. Cascella, "Quantification of cerebral cannabinoid receptors subtype 1 (CB1) in healthy subjects and schizophrenia by the novel PET radioligand [^{11}C] OMAR," *NeuroImage*, vol. 52, no. 4, pp. 1505–1513, Oct. 2010.

[11] P. Orban, C. Dansereau, L. Desbois, V. Mongeau-Pérusse, C.-É. Giguère, H. Nguyen, A. Mendrek, E. Stip, and P. Bellec, "Multisite generalizability of schizophrenia diagnosis classification based on functional brain connectivity," *Schizophrenia Res.*, vol. 192, pp. 167–171, Feb. 2017.

[12] I. Rish, G. Cecchi, B. Thyreau, B. Thirion, M. Plaze, M. L. Paillere-Martinot, C. Martelli, J.-L. Martinot, and J.-B. Poline, "Schizophrenia as a network disease: Disruption of emergent brain function in patients with auditory hallucinations," *PLoS ONE*, vol. 8, Jan. 2013, Art. no. e50625.

[13] G. S. Malhi, J. Lagopoulos, A. M. Owen, and L. N. Yatham, "Bipolaroids: Functional imaging in bipolar disorder," *Acta Psychiatrica Scandinavica*, vol. 110, no. s422, pp. 46–54, Feb. 2010.

[14] E. Veronese, U. Castellani, D. Peruzzo, M. Bellani, and P. Brambilla, "Machine learning approaches: From theory to application in schizophrenia," *Comput. Math. Methods Med.*, vol. 2013, pp. 1–13, Dec. 2013.

[15] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.

[16] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis, "Deep learning for computer vision: A brief review," *Comput. Intell. Neurosci.*, vol. 2018, pp. 1–13, Jul. 2018.

[17] K. Noda, Y. Yamaguchi, K. Nakadai, H. G. Okuno, and T. Ogata, "Audio-visual speech recognition using deep learning," *Appl. Intell.*, vol. 42, no. 4, pp. 722–737, Jun. 2015.

[18] S. P. Singh, A. Kumar, H. Darbari, L. Singh, A. Rastogi, and S. Jain, "Machine translation using deep learning: An overview," in *Proc. Int. Conf. Comput., Commun. Electron.*, New York, NY, USA, Jul. 2017, pp. 162–167.

[19] L.-L. Zeng, H. Wang, P. Hu, B. Yang, W. Pu, H. Shen, X. Chen, Z. Liu, H. Yin, Q. Tan, K. Wang, and D. Hu, "Multi-site diagnostic classification of schizophrenia using discriminant deep learning with functional connectivity MRI," *EBioMedicne*, vol. 30, pp. 74–85, Apr. 2018.

[20] J. Kim, V. D. Calhoun, E. Shim, and J.-H. Lee, "Deep neural network with weight sparsity control and pre-training extracts hierarchical features and enhances classification performance: Evidence from whole-brain resting-state functional connectivity patterns of schizophrenia," *NeuroImage*, vol. 124, pp. 127–146, Jan. 2016.

[21] W. H. Pinaya, A. Gadelha, O. M. Doyle, C. Noto, A. Zugman, Q. Cordeiro, A. P. Jackowski, R. A. Bressan, and J. R. Sato, "Using deep belief network modelling to characterize differences in brain morphometry in schizophrenia," *Sci. Rep.*, vol. 6, Dec. 2016, Art. no. 38897.

[22] J. I. Arribas, V. D. Calhoun, and T. Adali, "Automatic Bayesian classification of healthy controls, bipolar disorder, and schizophrenia using intrinsic connectivity maps from FMRI data," *IEEE Trans. Biomed. Eng.*, vol. 57, no. 12, pp. 2850–2860, Dec. 2010.

[23] H. Shen, L. Wang, Y. Liu, and D. Hu, "Discriminative analysis of resting-state functional connectivity patterns of schizophrenia using low dimensional embedding of fMRI," *NeuroImage*, vol. 49, no. 4, pp. 3110–3121, Feb. 2010.

[24] M. R. Arbabshirani, K. A. Kiehl, G. D. Pearlson, and V. D. Calhoun, "Classification of schizophrenia patients based on resting-state functional network connectivity," *Frontiers Neurosci.*, vol. 7, p. 133, Jul. 2013.

[25] M. J. Jafri and V. D. Calhoun, "Functional classification of schizophrenia using feed forward neural networks," in *Proc. Int. Conf. IEEE Eng. Med. Biol. Soc.*, New York, NY, USA, Aug./Sep. 2006, pp. 6631–6642.

[26] B. Yang, Q.-M. Shao, L. Pan, and W.-B. Li, "A study on regularized weighted least square support vector classifier," *Pattern Recognit. Lett.*, vol. 108, pp. 48–55, Jun. 2018.

[27] S. Laconte, S. Strother, V. Cherkassky, J. Anderson, and X. Hu, "Support vector machines for temporal classification of block design fMRI data," *NeuroImage*, vol. 26, no. 2, pp. 317–329, Jun. 2005.

[28] D. Chyzhyk, A. Savio, and M. Grana, "Computer aided diagnosis of schizophrenia on resting state fMRI data by ensembles of ELM," *Neural Netw.*, vol. 68, pp. 23–33, Aug. 2015.

[29] M. N. I. Qureshi, J. Oh, D. Cho, H. J. Jo, and B. Lee, "Multimodal discrimination of schizophrenia using hybrid weighted feature concatenation of brain functional connectivity and anatomical features with an extreme learning machine," *Frontiers Neuroinform.*, vol. 11, p. 59, Sep. 2017.

[30] J. Ford, "A combined structural-functional classification of schizophrenia using hippocampal volume plus fMRI activation," in *Proc. 24th Annu. Conf. Annu. Fall Meeting Biomed. Eng. Soc.*, New York, NY, USA, Oct. 2002, pp. 48–49.

[31] P. Khurd, S. Baloch, R. Gur, C. Davatzikos, and R. Verma, "Manifold learning techniques in image analysis of high-dimensional diffusion tensor magnetic resonance images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, New York, NY, USA, Jun. 2007, pp. 1–7.

[32] F. M. Vos, M. W. A. Caan, K. A. Vermeer, C. B. L. M. Majoie, G. J. Den Heeten, and L. J. Van Vliet, "Linear and kernel Fisher discriminant analysis for studying diffusion tensor images in schizophrenia," in *Proc. 4th IEEE Int. Symp. Biomed. Imag., From Nano Macro*, New York, VA, USA, Apr. 2007, pp. 764–767.

[33] S. A. Meda, M. C. Stevens, B. S. Folley, V. D. Calhoun, and G. D. Pearlson, "Evidence for anomalous network connectivity during working memory encoding in schizophrenia: An ICA based analysis," *PLoS ONE*, vol. 4, no. 11, Nov. 2009, Art. no. e7911.

[34] J. Sui, T. Adali, G. Pearlson, H. Yang, S. R. Sponheim, T. White, and V. D. Calhoun, "A CCA+ICA based model for multi-task brain imaging data fusion and its application to schizophrenia," *NeuroImage*, vol. 51, no. 1, pp. 123–134, May 2010.

[35] P. Guo, G. Xie, and R. Li, "Object detection using multiview CCA-based graph spectral learning," *J. Circuits, Syst. Comput.*, to be published. doi: 10.1142/S021812662050022X.

[36] H.-I. Suk, S.-W. Lee, and D. Shen, "Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis," *NeuroImage*, vol. 101, pp. 569–582, Nov. 2014.

[37] Y. Yoo, L. Y. W. Tang, T. Brosch, D. K. B. Li, S. Kolind, I. Vavasour, A. Rauscher, A. L. MacKay, A. Traboulsee, and R. C. Tam, "Deep learning of joint myelin and T1w MRI features in normal-appearing brain tissue to distinguish between multiple sclerosis patients and healthy controls," *NeuroImage Clin.*, vol. 17, pp. 169–178, Oct. 2018.

[38] M. A. Aghdam, A. Sharifi, and M. M. Pedram, "Combination of rs-fMRI and sMRI data to discriminate autism spectrum disorders in young children using deep belief network," *J. Digit. Imag.*, vol. 31, no. 6, pp. 895–903, Dec. 2018.

[39] H.-I. Suk, C.-Y. Wee, S.-W. Lee, and D. Shen, "State-space model with deep learning for functional dynamics estimation in resting-state fMRI," *NeuroImage*, vol. 129, pp. 292–307, Apr. 2016.

[40] A. S. Heinsfeld, A. R. Franco, R. C. Craddock, A. Buchweitz, and F. Meneguzzi, "Identification of autism spectrum disorder using deep learning and the ABIDE dataset," *NeuroImage, Clin.*, vol. 17, pp. 16–23, Jan. 2018.

[41] K. Aderghal, M. Boissenin, J. Benois-Pineau, G. Catheline, and K. Afdel, "Classification of sMRI for AD diagnosis with convolutional neuronal networks: A pilot 2-D+ ϵ study on ADNI," in *Proc. 23rd Int. Conf. Multimedia Modeling (MMM)*, 2017, pp. 690–701.

[42] H. Choi, K. H. Jin, and Alzheimer's Disease Neuroimaging Initiative, "Predicting cognitive decline with deep learning of brain metabolism and amyloid imaging," *Behav. Brain Res.*, vol. 344, pp. 103–109, May 2018.

[43] H.-I. Suk, S.-W. Lee, D. Shen, and Alzheimer's Disease Neuroimaging Initiative, "Deep ensemble learning of sparse regression models for brain disease diagnosis," *Med. Image Anal.*, vol. 37, pp. 101–113, Apr. 2017.

[44] M. Liu, D. Cheng, K. Wang, Y. Wang, and Alzheimer's Disease Neuroimaging Initiative, "Multi-modality cascaded convolutional neural networks for Alzheimer's disease diagnosis," *Neuroinformatics*, vol. 16, nos. 3–4, pp. 295–308, Oct. 2018.

[45] S.-H. Wang, P. Phillips, Y. Sui, B. Liu, M. Yang, and H. Cheng, "Classification of Alzheimer's disease based on eight-layer convolutional neural network with leaky rectified linear unit and max pooling," *J. Med. Syst.*, vol. 42, no. 5, pp. 295–308, May 2018.

[46] B. Jie, M. Liu, J. Liu, D. Zhang, and D. Shen, "Temporally constrained group sparse learning for longitudinal data analysis in Alzheimer's disease," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 1, pp. 238–249, Jan. 2017.

[47] Z. Akkus, I. Ali, J. Sedlár, J. P. Agrawal, I. F. Parney, C. Giannini, and B. J. Erickson, "Predicting deletion of chromosomal arms 1p/19q in low-grade gliomas from MR images using machine intelligence," *J. Digit. Imag.*, vol. 30, no. 4, pp. 469–476, Aug. 2017.

[48] K. Sakai and K. Yamada, "Machine learning studies on major brain diseases: 5-year trends of 2014–2018," *Jpn. J. Radiol.*, vol. 37, pp. 34–72, Jan. 2018.

[49] J. Ker, L. Wang, J. Rao, and T. Lim, "Deep learning applications in medical image analysis," *IEEE Access*, vol. 6, pp. 9375–9389, Dec. 2017.

[50] K. Kreutz-Delgado, J. F. Murray, B. D. Rao, and K. Engan, "Dictionary learning algorithms for sparse representation," *Neural Comput.*, vol. 15, no. 4, pp. 349–396, Feb. 2014.

[51] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*. New York, NY, USA: Cambridge Univ. Press, 2004, pp. 123–135.

[52] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 3856–3866.

[53] H. Lin, C. Lin, and R. Weng, "A note on Platt's probabilistic outputs for support vector machines," *Mach. Learn.*, vol. 68, no. 3, pp. 267–276, Oct. 2007.

[54] L. L. Zeng, H. Shen, L. Liu, and D. Hu, "Unsupervised classification of major depression using functional connectivity MRI," *Hum. Brain Mapping*, vol. 35, no. 4, pp. 1630–1641, Apr. 2014.

[55] L.-L. Zeng, D. Wang, M. D. Fox, M. Sabuncu, D. Hu, M. Ge, R. L. Buckner, and H. Liu, "Neurobiological basis of head motion in brain imaging," *Proc. Nat. Acad. Sci. USA*, vol. 111, no. 16, pp. 6058–6062, Apr. 2014.

[56] I. W. Tsang, A. Kocsor, and J. T. Kwok, "Efficient kernel feature extraction for massive data sets," in *Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, New York, NY, USA, 2006, pp. 724–729.

BO YANG received the B.Sc. degree in mechanical engineering from Zhengzhou University, China, in 1996, the M.Sc. degree in computer application technology from Xiangtan University, China, in 2004, and the Ph.D. degree in mechanical and electrical engineering from Central South University, China, in 2010. Since 2012, he has been an Associate Professor with the College of Information Science and Technology, Hunan Institute of Science and Technology. His research interests include MR brain image analysis, statistical pattern recognition, and machine learning.

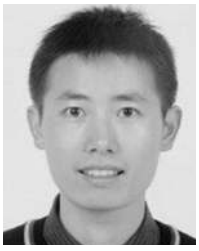
YUAN CHEN received the B.Sc. degree in information engineering from the Hunan Institute of Science and Technology, China, in 2017, where he is currently pursuing the M.Sc. degree with the College of Information Science and Technology. His research interests include deep learning and MR brain image analysis.

QUAN-MING SHAO received the M.Sc. degree in information engineering from the Hunan Institute of Science and Technology, China, in 2017. His research interests include deep learning, MR brain image analysis, and computer vision.

RUI YU received the B.Sc. degree in information engineering from the Hunan Institute of Science and Technology, China, in 2018, where he is currently pursuing the M.Sc. degree with the College of Information Science and Technology. His research interests include deep learning and multi-objective optimization.

WEN-BIN LI received the B.S. degree in computer science and technology from the Hunan Normal University, Changsha, China, in 2003, and the M.S. degree in computer applications technology from the Changsha University of Science and Technology, Changsha, in 2006. He is currently pursuing the Ph.D. degree with Central South University, Changsha. His research interests include deep learning, evolutionary computation, and multi-objective optimization.

GUAN-QI GUO received the B.Sc. degree in electronics from the Huazhong Institute of Science and Technology, Wuhan, China, in 1983, and the Ph.D. degree in control theory and control engineering from Central South University, Changsha, China, in 2003. He is currently a Professor with the School of Information Science and Engineering, Hunan Institute of Science and Technology. His research interests include evolutionary computation, multi-objective optimization, and machine learning.



JUN-QIANG JIANG received the Ph.D. degree in software engineering from Hunan University, Changsha, China, in 2017. He is currently a Master Supervisor with the School of Information Science and Engineering, Hunan Institute of Science and Technology, China. His research interests include cloud computing, parallel computing and workflow scheduling, and machine learning. He is a member of China Computer Federation.

LI PAN received the M.S. degree in computer software and theory from Sun Yat-sen University, Guangzhou, China, in 2004, and the Ph.D. degree in computer applied technology from Tongji University, Shanghai, China, in 2009. He is currently a Professor with the School of Information Science and Engineering, Hunan Institute of Science and Technology, China. He has published more than 20 papers in refereed journals and conference proceedings. His research interests include graph networks, Petri nets, and computational intelligence.