Detecting Risk Gene and Pathogenic Brain Region in EMCI Using A Novel GERF Algorithm Based on Brain Imaging and Genetic Data

Xia-an Bi*, *Member, IEEE*, Wenyan Zhou, Lou Li, and Zhaoxu Xing

***Abstract*—Fusion analysis of disease-related multi-modal data is becoming increasingly important to illuminate the pathogenesis of complex brain diseases. However, owing to the small amount and high dimension of multi-modal data, current machine learning methods do not fully achieve the high veracity and reliability of fusion feature selection. In this paper, we propose a genetic-evolutionary random forest (GERF) algorithm to discover the risk genes and disease-related brain regions of early mild cognitive impairment (EMCI) based on the genetic data and resting-state functional magnetic resonance imaging (rs-fMRI) data. Classical correlation analysis method is used to explore the association between brain regions and genes, and fusion features are constructed. The genetic-evolutionary idea is introduced to enhance the classification performance, and to extract the optimal features effectively. The proposed GERF algorithm is evaluated by the public Alzheimer's Disease Neuroimaging Initiative (ADNI) database, and the results show that the algorithm achieves satisfactory classification accuracy in small sample learning. Moreover, we compare the GERF algorithm with other methods to prove its superiority. Furthermore, we propose the overall framework of detecting pathogenic factors, which can be accurately and efficiently applied to the multi-modal data analysis of EMCI and be able to extend to other diseases. This work provides a novel insight for early diagnosis and clinicopathologic analysis of EMCI, which facilitates clinical medicine to control further deterioration of diseases and is good for the accurate electric shock using transcranial magnetic stimulation.**

***Index Terms*—Early mild cognitive impairment, fMRI, gene, genetic-evolutionary random forest, imaging genetics.**

This work was supported by the National Natural Science Foundation of China (Grant No. 62072173), the Natural Science Foundation of Hunan Province, China (Grant No. 2020JJ4432), the Key Scientific Research Projects of Department of Education of Hunan Province (Grant No. 20A296), the Degree & Postgraduate Education Reform Project of Hunan Province (Grant No. 2019JGYB091), and the Hunan Provincial Science and Technology Project Foundation (Grant No. 2018TP1018).

Xia-an Bi (bixiaan@hnu.edu.cn), Wenyan Zhou, Lou Li and Zhaoxu Xing are with Hunan Provincial Key Laboratory of Intelligent Computing and Language Information Processing, Hunan Normal University, Changsha, P.R. China, College of Information Science and Engineering, Hunan Normal University, Changsha, P.R. China, and Hunan Xiangjiang Artificial Intelligence Academy, Changsha, P.R. China.

* indicates the corresponding author

I. Introduction

EARLY mild cognitive impairment (EMCI) is a clinical state between normal aging and Alzheimer's disease (AD). It is very dangerous to develop into AD, which is a high-risk state of dementia [1, 2]. Once it develops into AD, this process will be irreversible, adding heavy burden to families and medical institutions [3]. Therefore, early detection of EMCI is of great significance to prevent or delay the occurrence and development of dementia [4]. With the rapid development of complex brain disease detection technologies, the amount of biomedical data has increased significantly, such as functional magnetic resonance imaging (fMRI) data and genetic data [5, 6]. In the comprehensive and systematic studies of the brain diseases, the fusion research of gene and neuroimaging has attracted more and more attentions [7]. It is worthy to reveal the multifactorial pathogenesis, and further help with clinical diagnosis and precision medicine of complex brain diseases.

The fusion study of multi-modal data is an emerging field in brain science, but it is developing rapidly [8]. The fMRI is a widely used imaging technique that is often combined with other data including structural MRI (sMRI) and genes to detect the pathogenesis of brain diseases such as autism and schizophrenia [9, 10]. French *et al*. found out the link between CNR1 gene and schizophrenia in the cortical maturation process [11]. Romme *et al*. explored that the white matter disconnectivity was related to the cortical gene expression based on the findings of French [12]. Wang *et al*. proposed a multi-modality regression method to discover the relationships of risk genes and brain connectivities based on sMRI and fMRI [13]. These studies can be divided into two types. The one is the study of multi-modal neuroimaging technology to obtain multiple data of brain for information complementarity and cross verification [14]. The other is a combination of neuroimaging and genetics to illustrate that how the gene affects the structure and function of the brain, which is proposed by Hariri and Weinberger [15]. We pay attention to the latter in this paper.

The complex brain diseases such as EMCI often involve different omics of data, such as fMRI and gene. Gene expression is a complex process, and abnormal gene expression is also reflected in brain lesions. All these factors will jointly affect the development of brain diseases. Therefore, the integration of these multivariate data is expected to expose the

genetic basis of brain function, thereby further explaining the relationship between risk genes and brain diseases. At the same time, many studies show that the fusion of fMRI and gene will provide more comprehensive analysis of complex brain diseases. Yang *et al*. applied support vector machine (SVM) to combining fMRI data with genetic data for disease classification [16]. Based on MRI, clinical, and genetic data, Greenstein *et al*. used the random forest to conduct classification research by yielding an accuracy of 73.7% [17]. Lin *et al*. presented the sparse canonical correlation analysis approach to study the correlation between gene and high-dimensional fMRI data and explored the disease-related region of interests (ROIs) and genes for complex diseases [18]. These fusion methods recognized effectively the relationship between brain function and gene, and took advantage of the complementarity from two types of data, which made us deepen the understanding of biological mechanisms behind complex brain diseases [19-21].

In bioinformatics, there are several public, multi-modal, and reliable databases with small amount and high dimensionality of data [22, 23]. Some researchers have tried to improve the problem through small sample learning methods. In machine learning or deep learning, the small sample learning and dimensionality reduction methods are SVM [24], decision tree [25], neural network [26, 27], and many others [28-31]. However, the results are often unsatisfactory in practical problems. Therefore, the design of an efficient and robust method for small sample data is a motivation of this paper.

On the other hand, there are few overall research frameworks of feature fusion scheme, feature extraction, and sample classification in most multi-modal data studies. Hu *et al*. proposed a feature fusion approach of the distance canonical correlation analysis (DCCA) to study complex imaging-genetic associations, and the results revealed that the DCCA was a powerful method for analyzing the multi-modal data [32]. However, neither of them mentioned the classification task. It is very meaningful to construct a integral framework of fusion scheme, feature extraction, and sample classification for multi-modal data in small samples. Consequently, another underlying motivation of this paper is to design a more reasonable effective fusion method and further design subsequent methods of feature extraction and classification task in the context of rapid development of multi-modal data fusion analysis.

Specifically, we design the fusing method combining brain regions with genes, and construct fusion features via correlation analysis. The fusion features are called as brain region-gene (BR-gene) pairs. Subsequently, we propose a genetic-evolutionary random forest (GERF) method to overcome the limitations of few and high-dimensional data and incomplete research framework. Using the GERF model, we extract the most distinguishable features as the optimal features. Based on the optimal features, the classification accuracy of samples from Alzheimer's Disease Neuroimaging Initiative (ADNI) database is up to 86.21%, which fully shows that the method of GERF could effectively extract features. Moreover, we also find out disease-related brain regions and risk genes by the optimal features, such as some pathogenic brain regions like Rolandic operculum (ROL.L), Superior frontal gyrus, orbital part (ORBsup.R), Amygdala (AMYG.L), Angular gyrus (ANG.R), Middle frontal gyrus, orbital part (ORBmid.L), and Insula (INS.R), and some risk genes like CNTN5, GRM7, and FHIT. Therefore, our works will provide researchers with new insight to explore EMCI.

The remain of this paper is arranged as following. Section II introduce the method presented in this paper in full and detail. The experimental results and method performance demonstration are presented in Section III. Section VI and V are related discussions and conclusions, respectively.

II. Methodology

In this section, an overall framework including feature fusion, feature extraction, and sample classification is proposed to explore the etiologies of EMCI (e.g., abnormal genes and brain regions). Firstly, the framework applies the processed rs-fMRI and gene data to constructing BR-gene pairs by a correlation analysis method. Secondly, the proposed GERF method is used to find out the optimal fused features and classify EMCI patients and normal controls (NC) efficiently. Furthermore, some abnormal genes and brain regions are discovered through the optimal fused features. Fig. 1 shows the multi-task framework with GERF.

A. Data Acquisition and Preprocessing

The data are provided by the public database of ADNI (http://adni.loni.usc.edu/). The database provides neuroimaging, biomarkers, and gene data for studies of EMCI, AD, and other cognitive disorders. We consider data from a population of 73 samples, composing of 37 EMCI patients (age: 72.97 ± 7.40, 11 females) and 36 NC samples (age: 75.84 ± 6.27, 14 females). The data contain the rs-fMRI data and corresponding genetic data. ADNI has approved and authorized the use of data. We strictly screen data to ensure the homology of EMCI patients and NC. Subjects without MMSE and CDR scores are selected here because these scores cannot guarantee that the data are homologous and the robustness of the model cannot be assessed by adding these scores. Additionally, our study is conducted by suggestions of Federal Regulations, etc., which is also supported by Institutional Review Board of each participating site. All subjects have signed the informed consent. In order to make out the discrepancy of gender or age between EMCI and NC, two statistical tests are performed. Table I displays the information of participants and the results of tests. The results show that there are no differences in terms of gender and age between EMCI patients and NC (both $p > 0.05$).

TABLE I
BASIC INFORMATION OF EMCI AND NC

Variables (Mean ± SD)	EMCI (n = 37)	NC (n = 36)	p value
Gender (M/F)	26/11	22/14	0.410*
Age (years)	72.97±7.40	75.84±6.27	0.078**

* The p value is obtained via the chi-square test.
** The p value is obtained via the two-sample t-test.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2021.3067798, IEEE Journal of Biomedical and Health Informatics

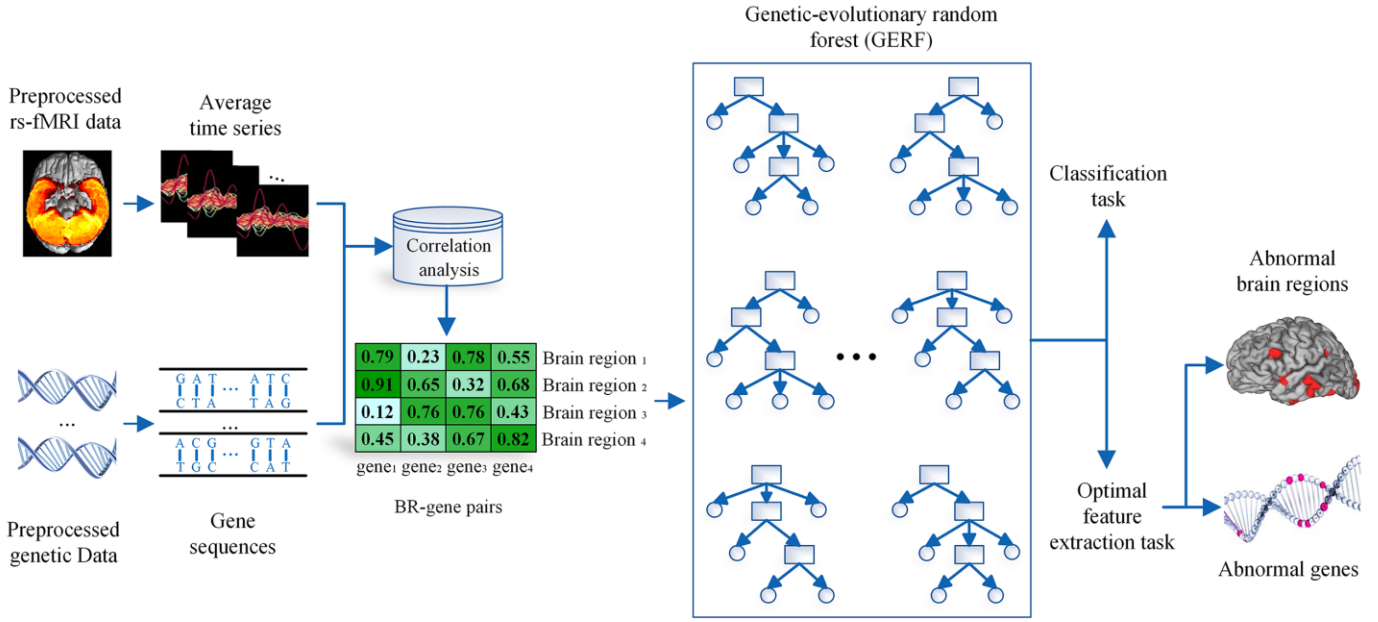


Fig. 1. An overall framework with the proposed GERF. The image of one subject is preprocessed to gain the average time series of brain regions, and some genes of the corresponding subject are preprocessed to gain the gene sequences. Through the correlation analysis method, BR-gene pairs are constructed. Then the GERF model is constructed to carry out the classification task and optimal feature extraction task. Based on the optimal features, the abnormal brain regions and genes are analyzed.

The rs-fMRI data are acquired by a magnetic resonance scanner. The image quality may be poor due to the noise in the data, which affects the experimental results. Therefore, based on the MATLAB 2014a platform, we use DPARSF software (http://rfmri.org/DPARSF) to preprocess the rs-fMRI data [33]. The preprocessing steps include transferring format, eliminating first 10 time volumes, correcting slice-time, correcting motion, normalizing space, smoothing, removing linear trend, filtering, and removing the covariates. The genetic data are stored on the Illumina Omni 2.5 M chip, including single nucleotide polymorphism (SNP) data. Owing to the subsistent noise in the data on the chip, we use the PLINK software to preprocess the genetic data. The preprocessing steps are as follows:

1) measuring the SNP call rate to detect the quality of data;
2) calculating the minimum allele frequency to get rid of the data that has less information;
3) calculating the genotyping rate to reduce errors;
4) carrying out the Hardy-Weinberg equilibrium to test whether the frequencies of alleles or genotypes remain stable;
5) transferring format to extract the required data from a large amount of data;
6) extracting the SNP data of subjects to reduce time complexity;
7) extracting genetic data from SNP data to facilitate subsequent experiments.

B. Construction of Multi-modal Fusion Features

As mentioned in the Introduction, factors such as brain appearance and genes can jointly affect the development of brain diseases. Here we will conduct a fusion study on brain regions and genes to study the relationship between them. Suppose we have $n \in N$ subjects. Each subject contains b ROIs from rs-fMRI data $B \in R^{n\times b}$ and g genes from genetic data $G \in R^{n\times g}$, where the average time series for ROIs are represented by w_b and the gene sequences for genes are represented by s_g. The fused approach is first designed for measuring the correlations between data B and G to construct fusion features. Firstly, w_b and s_g are clipped to ensure that the length of each ROI is equal to that of each gene sequence, resulting in $w_b{}'$ and $s_g{}'$. Secondly, gene sequences are encoded discretely by the way of using 1, 2, 3, and 4 to replace A, T, C, and G. Note that 1, 2, 3, and 4 are just markers. We have repeated experiments by changing their order or replacing them with different numbers, which does not make a difference to the result. Finally, the different correlation analysis methods are utilized to calculate the correlation coefficients between w_b' and s_g'. The Pearson correlation coefficient is defined as

$$\rho_{b,g} = \frac{\alpha \sum w_b' s_g' - \sum w_b' \sum s_g'}{\sqrt{\alpha \sum {w_b'}^2 - \left(\sum w_b'\right)^2}\sqrt{\alpha \sum {s_g'}^2 - \left(\sum s_g'\right)^2}} \quad (1)$$

where α represents the length of each gene or ROI. The canonical correlation coefficient is defined as

$$\theta_{b,g} = \frac{\beta^T \Sigma_{12} \gamma}{\sqrt{\beta^T \Sigma_{11} \beta}\sqrt{\gamma^T \Sigma_{22} \gamma}} \quad (2)$$

where β and γ represent two weight parameters that maximize the correlation between brain regions and genes. $\sum_{11}$ denotes a covariance matrix of w_b' and $\sum_{22}$ denotes a covariance matrix of s_g'. $\sum_{12}$ denotes a covariance matrix of w_b' and s_g'. The correlation distance coefficient is defined as

$$\mu_{b,g} = 1 - \frac{\alpha \sum w_b' s_g' - \sum w_b' \sum s_g'}{\sqrt{\alpha \sum {w_b'}^2 - \left(\sum w_b'\right)^2}\sqrt{\alpha \sum {s_g'}^2 - \left(\sum s_g'\right)^2}} \quad (3)$$

The best correlation analysis method is selected out through

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2021.3067798, IEEE Journal of Biomedical and Health Informatics

the comparison of the above three methods in part of Method Comparison and Validation. Then the best correlation coefficient is replaced by $c_{b,g}$. The coefficient matrix $\mathrm{C} = \{c_{1,1}, \cdots, c_{1,g}, \cdots, c_{b,1}, \cdots, c_{b,g}\}$ is used as fusion features and C_i represents the coefficient matrix of i^{th} subject.

C. Design Idea and Presentation of the GERF

The main work of this paper is the design of feature extraction and classification method, namely genetic-evolutionary random forest (GERF), to develop the fusion data analysis. In existing studies, conventional random forest methods have the characteristic of large randomness, reflecting in two aspects. One aspect is that the obtained numbers of sample subsets and feature subsets are generally determined based on previous experiences [34]. However, optimum values for the two parameters are discrepant in different experiments, which affects experimental results. On the other hand, the feature subsets are uncontrollable during constructing decision trees and may contain many irrelevant or redundant features, resulting in lower performance of random forest [35]. With the intention of making up for these deficiencies, we introduce an idea of genetic evolution on the basis of random forest and further propose the GERF method. The GERF has some advantages of automatic global optimization and automatic deletion of the irrelevant features. In a sense, the GERF method has carried out a transition from a black-box to a white-box, which has optimized the traditional random forest. Moreover, the GERF is designed to classify effectively and detect disease-related brain regions and genes through the persistent genetic evolutions.

Assuming that the original data set is $D = \{X, A\}$, where X represents the original sample set and A represents the original feature set. The original sample set is $X = \{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$, where x_n denotes the n^{th} sample and y_n denotes the classification label of this sample. The labels $y_n \in \{+1, -1\}$, where "$+1$" represents NCs and "-1" represents patients. There are $m \in M$ features in the feature set. The original feature set is $A = \{a_{11}, \cdots, a_{1m}, \cdots, a_{n1}, \cdots, a_{nm}\}$, where a_{nm} denotes the m^{th} feature of the n^{th} sample.

The original data set $D = \{X, A\}$ is divided into a training set $D_1 = \{X_1, A_1\}$ and a testing set $D_2 = \{X_2, A_2\}$ according to a certain ratio, where $D = D_1 + D_2$. D_1 is applied to training the model and D_2 is applied to testing the generalization performance of the model. The model adopts a binary form to characterize sample features for subsequent calculation. "0" represents irrelevant features that are unselected into the feature subset, and "1" represents relevant features that are selected into the feature subset. The i^{th} feature takes the following form:

$$w_i = \begin{cases} 1, & fusion\ feature\ is\ selected \\ 0, & fusion\ feature\ is\ unselected \end{cases}, i = 1, \cdots, m \quad (4)$$

We randomly choose samples and sample features from D_1 to establish a base classifier. It is worth noting that the base classifier is built based on the classification and regression tree (CART) algorithm. We repeat this process for $k \in K$ times to generate a premier random forest with k base classifiers. Each base classifier is encoded as a binary vector of length m. As a result, m features in k base classifiers constitute a binary matrix $w_{k,m}$ which is defined as

$$w_{k,m} = \begin{bmatrix} w_{1,1} & \cdots & w_{1,m} \\ \vdots & \ddots & \vdots \\ w_{k,1} & \cdots & w_{k,m} \end{bmatrix} \quad (5)$$

To better evaluate the classification ability of each feature subset, the model regards classification accuracy of base classifiers as a fitness function. If the classification accuracy is the lowest, the corresponding feature subset is replaced by that having the highest classification accuracy. Otherwise, the corresponding feature subset is preserved. This process is called as the selection. Thus, the formalization of the fitness function is expressed as

$$h_i = \frac{t_{right,i}}{T} \quad (6)$$

where h_i represents the value of the fitness function, $t_{right,i}$ represents the total of participants in the testing set correctly classified by the i^{th} base classifier. T represents the total quantity of samples in the testing set. Then we carry out the genetic evolution. Each base classifier in the random forest is evaluated by the fitness function, irrelevant or redundant sample features are gradually removed, and sample features with high classification ability are retained. The sample features in the random forest are further crossed and mutated to form the first-generation random forest. The process of selection, cross and mutation is repeated continually, which is referred to as the genetic evolution. However, the genetic-evolutionary process cannot continue indefinitely. We set up termination conditions to generate the ultima random forest. When the classification accuracy of random forest remains stable or the genetic-evolutionary times reach a threshold, the genetic-evolutionary process is stopped and the ultima random forest is formed. The Algorithm 1 gives the clear construction procedure.

D. Classification of the GERF

We adopt the GERF model for classification and use the majority voting to predict the classification label of each sample. The majority voting method is given as

$$MV(x_n) = \begin{cases} y_n^j, & if \sum_{i=1}^{K} mv_i^j(x_n) > 0.5 \sum_{z=1}^{Z} \sum_{i=1}^{K} mv_i^z(x_n) \\ reject, & otherwise \end{cases} \quad (7)$$

where $MV(x_n)$ represents predicted classification label of x_n, y_n^j represents the j^{th} classification label. The $\sum_{i=1}^{K} mv_i^j(x_n)$ represents the quantity of base classifiers labeled as y_n^j which is the predicted value of sample x_n, and $z \in Z$ represents the number of categories in the original data. The classification labels of all samples are predicted and the classification accuracy of GERF is computed by

$$ACC = \frac{E_{true}}{T} \quad (8)$$

where E_{true} represents the number of correctly classified samples in the testing set.

E. Extraction of Optimal Fusion Features

The extraction of optimal fusion features (i.e., optimal BR-gene pairs) is a crucial application of the GERF model. The optimal BR-gene pairs extracted can accurately distinguish between patients and NC. The specific extraction process is as follows. The first step is the extraction of important BR-gene pairs. Based on the random forest which has the optimal quantity of base classifiers, the classification accuracy tends to be stable. Each base classifier corresponds to a feature subset containing p features. The frequency of each feature included in all feature subsets is counted and sorted in descending order. Finally, the first q features are taken as important BR-gene pairs which are denoted as $O = \{o_1, o_2, \cdots, o_q\}$. The second step is the extraction of optimal BR-gene pairs. The first r ($r \leq q$) features are selected from O and their classification abilities are evaluated. Then the above procedure is repeated in a range of r to q with a step of 5. Eventually, the highest classification accuracy is found out from all classification accuracies. The corresponding feature subset is recorded as the optimal BR-gene pairs (i.e. optimal fusion features), which are denoted by $V = \{v_1, v_2, \cdots, v_r\}$. These optimal fusion features are automatically selected by GERF because they have more scientific value. The reason why it is called the optimal fusion feature is that it has strong classification ability for EMCI and normal people. That is to say, the differences in these features between EMCI and normal people are more pronounced, which indicates that brain regions involved in these features are more prone to lesions, and genes involved in these features are more prone to unusual expressions. By analyzing these features, there is a greater probability to identify risk genes and pathogenic brain regions of EMCI.

Algorithm 1 GERF learning process

Input: experimental data set D={X, A}
Output: the genetic-evolutionary random forest
1: Initialize D, D_1, D_2, k.
2: D={X, A} is experimental data set,
3: $D=D_1+D_2$,
4: k is the number of initial decision tree.
5: Partitioned the D={X, A} into $D_1=\{X_1, A_1\}_{training\_1}$, $D_2=\{X_2, A_2\}_{testing\_1}$, ⋯, $D_1=\{X_1, A_1\}_{training\_n}$, $D_2=\{X_2, A_2\}_{testing\_n}$
6: **for** z = 1 to k :
7: select $\{X_1, A_1\}_{training\_z}$
8: Randomly select a subset of features from D_1 as $\{Features\}_{training\_z}$
9: $\{X_1, A_1\}_{training\_z}$ and $\{Features\}_{training\_z}$ → Decision tree $_z$
10: $\{X_1, A_1\}_{testing\_z}$ → test the classification accuracy of Decision tree $_z$
11: **end for**
12: {Decision tree ensemble} = Ensemble of {Decision tree $_1$⋯⋯Decision tree $_k$}
13: **Do**
14: Retain the feature subset with the highest classification accuracy
15: Replace the feature subset with the lowest classification accuracy with the feature subset with the highest classification accuracy
16: Cross and mutate the sample features in the {Decision tree ensemble} to form the random forest
17: $\{Decision\ tree\ ensemble\}_{new}$ = Random forest with continuous genetic evolution
18: Calculate the classification ability of $\{Decision\ tree\ ensemble\}_{new}$
19: **Until** the classification accuracy remains stable or the genetic-evolutionary times reach the present threshold
20: GERF = {Decision tree ensemble} with the highest classification ability

F. Extraction of Disease-related Brain Regions and Genes

The optimal fused features with the highest classification accuracy are extracted from 3240 high-dimensional features, which have the best classification ability among all features. Consequently, the optimal fusion features are regarded as abnormal ones, in which two components (brain regions and genes) are most likely to be abnormal and cause brain diseases. Based on $V = \{v_1, v_2, \cdots, v_r\}$, the occurrence frequency of each brain region is counted and sorted in descending order. The higher the frequency is, the more abnormal the brain region is. Subsequently, the occurrence frequency of each gene is also computed and sorted in descending order. Similarly, the higher the frequency is, the more abnormal the gene is.

G. Parameters Optimization

In the proposed GERF model, there are two parameters that need to be optimized. The first parameter is the times of genetic evolutions, which affects the efficiency of the model. The times should satisfy the constraint:

$$\begin{aligned} &\min_{pr}\ (pr) \\ \text{s.t.}\quad &\Delta f_{RF} < \varepsilon \\ &pr \leq U \end{aligned} \tag{9}$$

where pr is the times of genetic evolutions and U is the largest value of genetic-evolutionary times. f_{RF} represents the classification accuracy of random forest. And the base classifiers quantity is the second parameter. Different quantities will influence the classification performance. In order to find out the optimal quantity of base classifiers, different values are selected to iterate the genetic-evolutionary process. When the random forest has the stabilized classification accuracy, the quantity corresponding to the least times of genetic evolutions is optimal. Consequently, the most befitting parameter is acquired.

III. Results

A. Results of Fusion Features Construction and GERF Construction

We combined brain regions with genes to construct BR-gene pairs. For the preprocessed rs-fMRI data, the brain image of each subject was matched utilizing the Anatomical Automatic Labeling template based on the voxel level to obtain 90 brain regions and corresponding 90 average time series. Based on the preprocessed genetic data, the quantity of SNPs in each gene was counted and sorted in descending order. We selected first 36 genes to ensure the precision of the experiment. Then we extracted first 30 SNPs in each gene and encoded discretely. Thus, each subject had 36 genes and recoded gene sequences. In order to match the gene sequences, first 60 time points were selected and treated as the final average time series. We took Pearson correlation analysis as an example to show the calculation process. Consequently, 3240-dimensional ($36 \times 90 = 3240$) BR-gene pairs were obtained and regarded as fusion features.

To better train the model, specifically, we divided the data

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2021.3067798, IEEE Journal of Biomedical and Health Informatics

set by a ratio of 6:4 and obtained 44 training samples and 29 testing samples. From the training data, 40 samples and 57 features were randomly extracted to construct a base classifier. The feature number of 57 was obtained through many practical experiments. It could play the advantages of cluster more effectively. We found that when the number of input features was small (<< 57), we needed to build many base classifiers to obtain a part of base classifiers with satisfactory performance, which would greatly increase the time complexity of cluster construction. However, when the number of input features was too large (>>57), the diversities among base classifiers would decrease, which would also increase the time cost of genetic evolution. Then, different base classifiers quantities were taken to build the GERF model. Eventually, we selected the optimal quantity of base classifiers to acquire the GERF model.

Here 100 base classifiers were taken as an example to illustrate the establishing process of the GERF model. Firstly, the quantity was set to 100 to construct a random forest with best performance. The classification accuracy of the random forest on the testing set was considered as the fitness function and the threshold of genetic-evolutionary times was set to 200. Secondly, sample features in the random forest were crossed and mutated to form the first-generation random forest. Thirdly, the genetic evolution was executed continuously until the classification accuracy of random forest remained stable. In Fig. 2, when the times of genetic evolutions were 84, the variation of classification accuracy tended to be stable and the value was 86.21%. As a consequence, when the quantity was 100, 84 was the optimal times of genetic evolutions. In order to find out the optimal quantity, we adjusted the quantity to build the GERF model. After repeated tests, the quantity of base classifiers was limited to a range of 100 to 300 with a step of 20 and the threshold times of genetic evolutions were set to 200. In Fig. 3, we summarized the changing situation of genetic-evolutionary times based on different quantities of base classifiers when classification accuracy of the GERF tended to be stable, hoping to find a combination with a less number of initial base classifiers and times of genetic-evolutionary to guarantee the effectiveness of model building. On the whole, a V-shaped

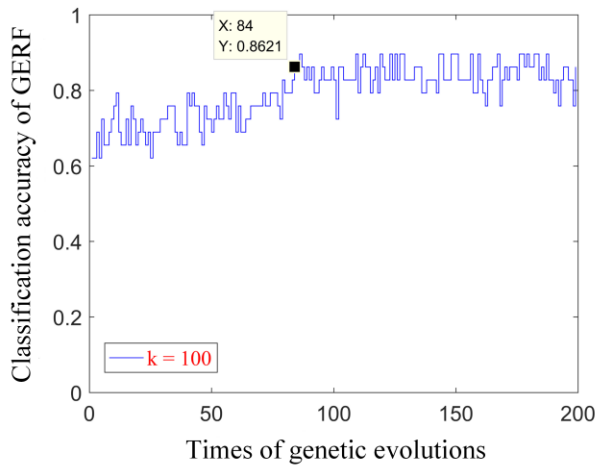


Fig. 2. The variation of classification accuracies for 100 base classifiers.

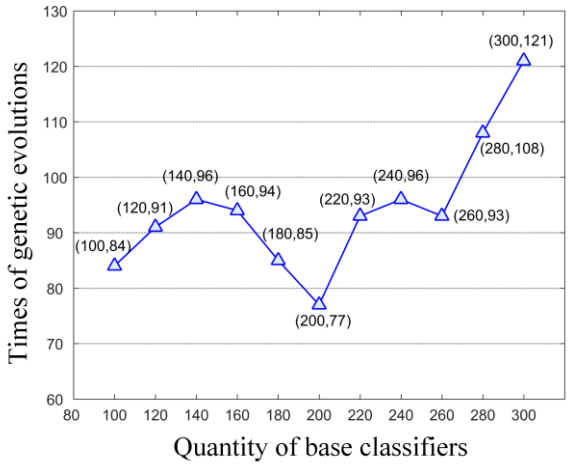


Fig. 3. Optimal quantity of the base classifiers. Based on different quantities of base classifiers, the lowest times of genetic evolutions were optimal.

curve was observed. When the quantity was 200, the genetic-evolutionary times were the least with the value 77, therefore the optimal quantity was 200. As a result, the GERF model with 200 base classifiers be constructed. The model did not have problem of overfitting, because in the process of model construction and assessment, samples and sample features were randomly extracted, which avoided overfitting to a certain level. Moreover, experimental results on ADNI dataset showed that the GERF model performed well, so it was impossible to overfit the model.

B. Results of Optimal Fusion Feature Extraction

As the genetic evolution of random forest continued, irrelevant or redundant features were gradually removed. Therefore, most of features in each base classifier had high classification ability. The frequencies of all features were calculated and ranked, and the first 400 features were extracted as important BR-gene pairs. We primarily extracted the first 70 features as a subset and applied the traditional random forest to classifying EMCI patients and NC. Then the range of extracted features was set to (70, 400), and the step length was 5. The starting point of range is set to 70 because the performance of the base classifier constructed with less than 70 features is too low, and such small number of features may not be enough to build a base classifier. Therefore, in order to ensure the performance of the ensemble learner, we need to use at least 70 features as feature subsets to build a base classifier, and then build the ensemble learner. The variational curve of the classification accuracies was described in Fig. 4.

The results showed that when the first 190 features were selected, the classification accuracy was the highest, reaching 86.21%. It was indicated that the subset including 190 features selected had the highest classification ability, eliminating irrelevant or redundant features. Therefore, the first 190 features were the optimal fused features, and the top 20 features were shown in Fig. 5.

C. Method Comparison and Validation

We further compared quantities of optimal fusion features

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2021.3067798, IEEE Journal of Biomedical and Health Informatics

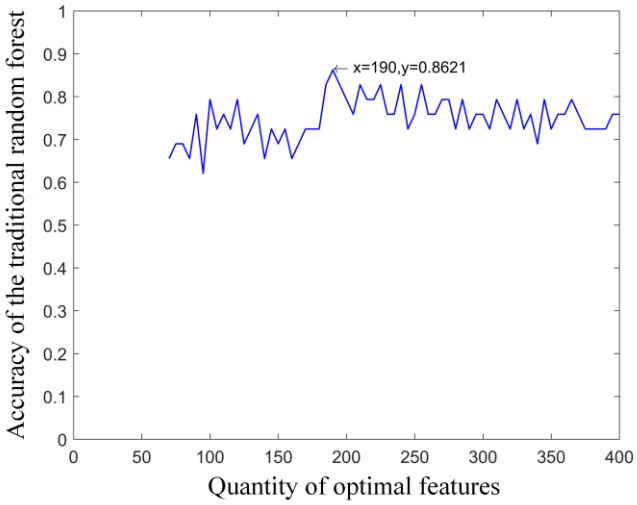


Fig. 4. The variational curve of the classification accuracies for different feature subsets.

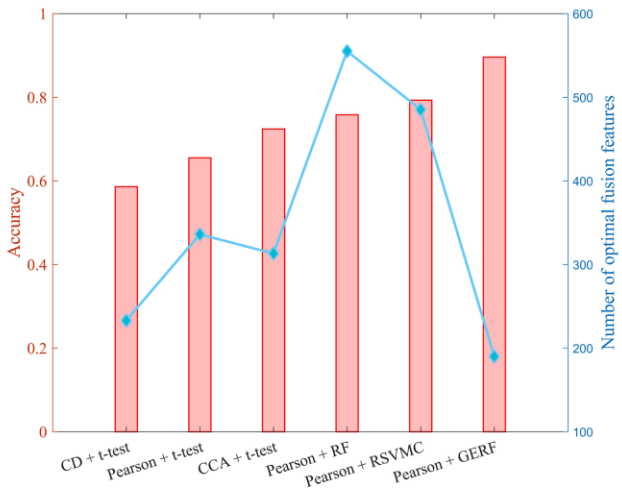


Fig. 6. The quantitative comparison of optimal fusion feature on multi-modal data.

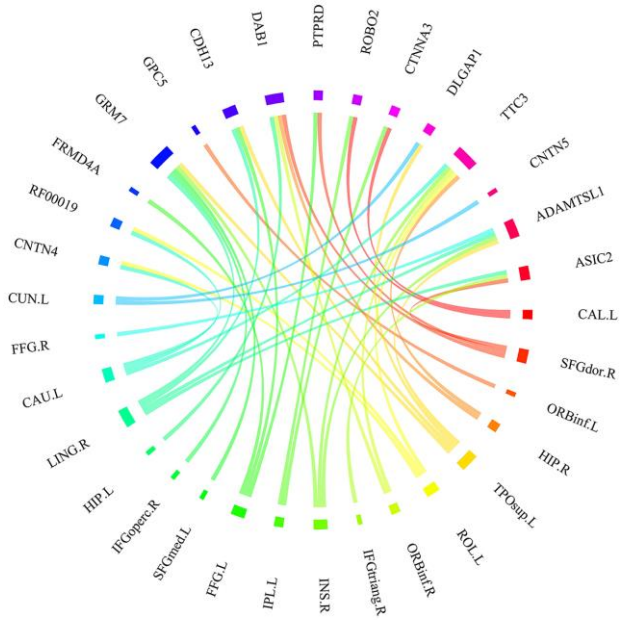


Fig. 5. The top 20 of optimal BR-gene pairs. The top represented genes and the bottom represented brain regions.

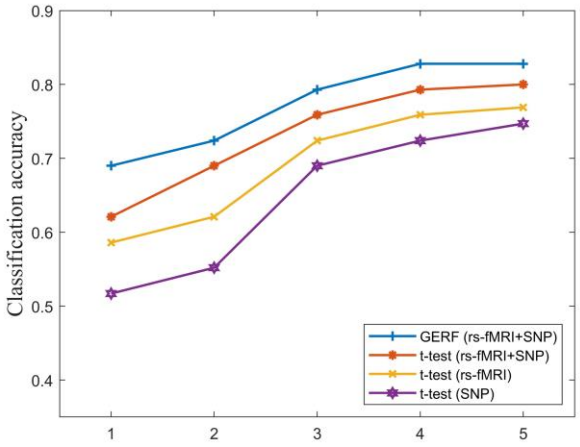


Fig. 7. Performance comparison. The comparison included the classification performance analysis of GERF and the comparison between GERF and two-sample t-test.

which were extracted by different methods on multi-modal data in Fig. 6. We utilized the different correlation analysis methods including Pearson correlation analysis, correlation distance (CD), and canonical correlation analysis (CCA) to construct feature matrixes, and extracted optimal features based on different machine learning methods involving the proposed GERF, two-sample t-test, random SVM cluster (RSVMC), and random forest (RF). Moreover, for the optimal features extracted by different methods, SVM classifier was used to test their classification performance. SVM is a two-class classifier, which has the advantages of excellent generalization performance, suitable for small sample learning, and can simplify common classification and regression problems, so it is widely used. As shown in Fig. 6, the quantity of optimal features selected by GERF was the minimum among all methods, while the classification accuracy was the highest, which indicated the features extracted by GERF were highly discriminative. In addition, to better compare the classification performances of GERF and two-sample t-test, we applied t-test to extracting features and then classified using an SVM classifier. Fig. 7 summarized the classification results of GERF and t-test based on different modal data for 5 times. The changing situation illuminated that the classification accuracy of GERF gradually tended to be stable and achieved optimal result with the increase of genetic-evolutionary times. It was observed that the performance of GERF appeared to be preferable than t-test in terms of the fusion features, and the stable value of classification accuracy of GERF was also superior to t-test based on the single modal data of rs-fMRI or SNP.

D. Results of Abnormal Brain Regions and Genes Extraction

In the optimal BR-gene pairs, we further analyzed the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2021.3067798, IEEE Journal of Biomedical and Health Informatics

occurrence frequencies of disease-related genes or brain regions. Brain regions with greater frequencies and the weights of those were provided in Fig. 8. The highest frequency was 6. The brain regions with greater frequencies were ROL.L, ORBsup.R, AMYG.L, ORBmid.L, ANG.R, INS.R, and other brain regions. Our discoveries are similar to many other studies. The ANG.R is involved in the processing of music performance, combinatorial semantics, and episodic memory [36]. The ANG.R accounts for a large proportion in all abnormal brain regions, indicating that the ANG.R may be connected with EMCI. The ORBmid.R plays a crucial role in the pathogenesis of EMCI. Xiang *et al.* identified the differences of ORBmid.R in the research on EMCI and NC [37]. Zhang *et al*. found that the decreased functional connectivity in MCI was between left dorsolateral superior frontal gyrus (SFGdor.L) and ORBmid.R [38]. The INS.R is a brain region connected with functions of cognition and affection [39]. Niu *et al.* employed one-way analysis of variance approach in the study of the conversion process from NC to EMCI to LMCI to AD and revealed significant differences of the INS.R on multiple time scales [40]. Zhao *et al.* suggested that some abnormal behaviors in patients with EMCI may reflect insular pathology [41].

Moreover, we also counted the frequencies of disease-related genes based on optimal BR-gene pairs. The frequencies of 36 genes with the number of SNPs greater than the set threshold were displayed in Fig. 9. Genes with greater frequencies were CNTN5, GERM7, CTNNA3, TTC3, FHIT, and other genes. Genes with greater frequencies suggested that these genes are likely to lead to EMCI. Our discoveries are also consistent with those of many scholars. The FHIT gene is involved in pathological characteristics of EMCI. Li *et al*. investigated the genetic interaction among cingulate amyloid-beta load in EMCI and found out the interaction between CLSTN2 and FHIT that associated with cingulate amyloid burden [42]. Similarly, Yan *et al*. employed the linear regression model to analyze the genetic interaction in EMCI and found out the interaction between FHIT and PRB1 [43].

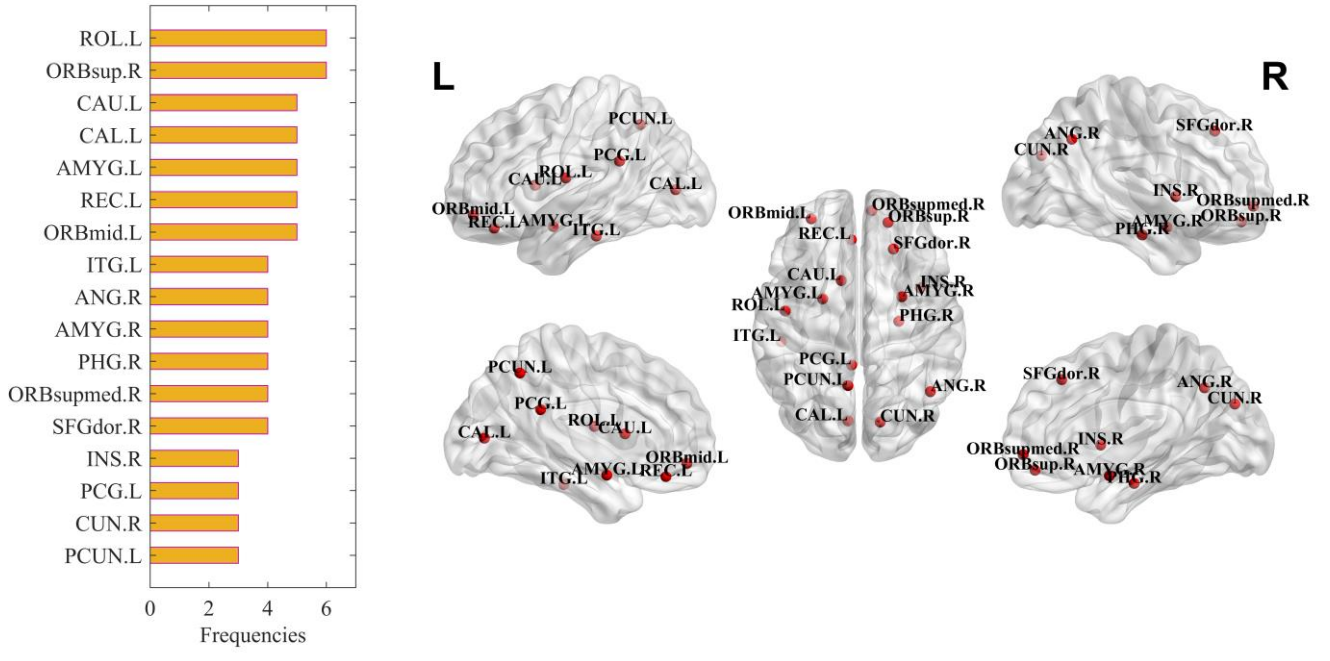
Moreover, there are some newfound brain regions (e.g., ROL.L and ORBsup.R) and genes (e.g., CNTN5 and GRM7), which are also likely associated with EMCI. Consequently, the GERF model not only provides a new method for the detection and diagnosis of EMCI, but also affords a novel mode for the pathological research of LMCI or AD.

IV. Discussions

A. Performance Analysis of Method

With the intensification of global aging process, the diagnosis study on senile diseases such as EMCI is becoming more and more essential. Some researchers focused on the classification study of EMCI by machine learning methods. Jie *et al*. adopted the multi-kernel SVM method to explore the dynamic connectivity networks of EMCI patients and NC, and the classification accuracy was up to 78.3% [44]. Wee *et al*. employed a sparse temporal network-based framework to classify EMCI patients and NC, and the accuracy reached 79.7% [45]. Peng *et al*. proposed a kernel-learning-based approach for multi-modal feature selection and used multiple kernel SVM to differentiate MCI from NC, resulting in an accuracy of 80.3% [46]. In this paper, we propose the GERF method to classify EMCI patients and NC, and the classification accuracy of the GERF method is up to 86.21%.

Compared with existing approaches, the outstanding performance of GERF is embodied in three aspects. Firstly, the



(a) (b)

Fig. 8. The frequencies and locations of abnormal brain regions. (a) Frequencies of a part of the brain regions. (b) Locations of the corresponding brain regions.

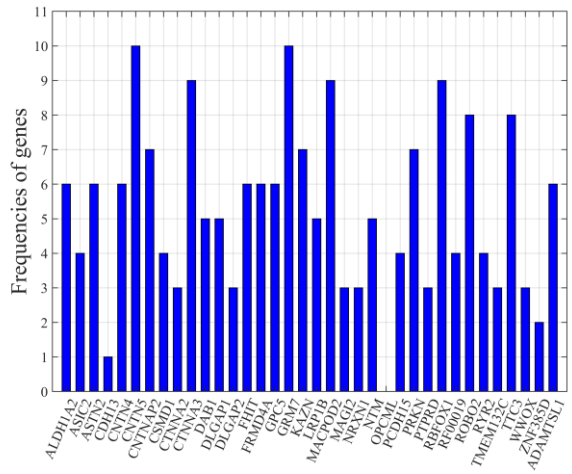


Fig. 9. The frequencies of 36 genes. The genes with larger weights were CNTN5, GERM7, CTNNA3, TTC3 and FHIT.

GERF model has good global optimization capabilities. The termination condition of genetic-evolutionary process is that the classification accuracy is stable or the times of genetic evolutions reach a certain threshold. Due to the stable accuracy, it shows that the most distinguishing features are preserved during the process of genetic evolutions, which improves the performance of the GERF model. The threshold of genetic-evolutionary times is set to simplify the model and ensure the learning efficiency. At the same time, the GERF model is optimized by changing the quantity of base classifiers. When the classification accuracy remains stable based on different quantities of base classifiers, the times of genetic evolutions corresponding to the optimal base classifier are the least, which means that the performance of GERF is optimal. Secondly, the model makes up for the deficiencies of random forest. Irrelevant or redundant features are gradually removed through continuous genetic evolutions, and most features in the ultima random forest have high classification ability. Thirdly, the overall framework consisting of the multi-modal data fusion method and the GERF model realizes the information complementation between fMRI data and gene data efficiently and effectively. The neuroimaging information and the genetic information are independent, both of which might affect the development of EMCI. We build a bridge between images and genes through correlation analysis, and the experimental results prove the good effect of information fusion.

B. Limitations and Future Efforts

Though achieving good feature extraction ability and classification performance, the proposed GERF method has some potential limitations. The first is that the selection of brain atlas may affect the generalization performance of the model. The constructed BR-gene pairs may be quite different because of the distinction of different atlases. In the future, we will consider different brain structures and use different brain atlas, like Broadman. The second is the selection of experimental data and fusion method. In this paper, Pearson correlation analysis method is used to fuse gene and fMRI data. In the future, we can also use other correlation analysis methods to fuse other modal data such as protein data [47, 48]. Lastly, for some less typical disease-related factors, due to the short of related research at present, we will find more data in subsequent research work, design novel algorithms for in-depth analysis, work with clinicians and better explain the role and rationality of these factors in the mechanism of EMCI.

V. Conclusions

In conclusion, we conduct a multi-modal data fusion study on rs-fMRI and gene data to detect brain diseases. The main achievements of this paper are listed. Primarily, we design a fused approach to carry out the feature fusion via combining rs-fMRI data with gene data. Secondly, we propose the GERF method to find out the optimal BR-gene pairs and classify EMCI patients and NC efficiently. Compared with other approaches, the classification accuracy of GERF is the highest. Finally, we provide an overall framework of feature fusion, feature extraction, and classification in small samples. The experimental results show that the overall framework had brilliant feature extraction ability and classification performance, which is supportive to explore the pathogenesis of EMCI. We discover many disease-related genes and brain regions, which sheds light on the pathogenic mechanism of EMCI.

References

[1] T. C. Pinto *et al.*, "Is the Montreal Cognitive Assessment (MoCA) screening superior to the Mini-Mental State Examination (MMSE) in the detection of mild cognitive impairment (MCI) and Alzheimer's Disease (AD) in the elderly?," *International psychogeriatrics*, vol. 31, no. 4, pp. 491-504, 2019.

[2] J. Ottoy *et al.*, "Association of short-term cognitive decline and MCI-to-AD dementia conversion with CSF, MRI, amyloid-and 18F-FDG-PET imaging," *NeuroImage: Clinical*, vol. 22, p. 101771, 2019.

[3] T.-E. Kam, H. Zhang, Z. Jiao, and D. Shen, "Deep learning of static and dynamic brain functional networks for early mci detection," *IEEE transactions on medical imaging*, vol. 39, no. 2, pp. 478-487, 2019.

[4] D. Shukla, P. K. Mandal, M. Tripathi, G. Vishwakarma, R. Mishra, and K. Sandal, "Quantitation of in vivo brain glutathione conformers in cingulate cortex among age - matched control, MCI, and AD patients using MEGA - PRESS," *Human Brain Mapping*, vol. 41, no. 1, pp. 194-217, 2020.

[5] X. Sun *et al.*, "ROBO1 polymorphisms, callosal connectivity, and reading skills," (in eng), *Human brain mapping*, vol. 38, no. 5, pp. 2616-2626, 2017.

[6] R. Han *et al.*, "AuTom-dualx: a toolkit for fully automatic fiducial marker-based alignment of dual-axis tilt series with simultaneous reconstruction," *Bioinformatics*, vol. 35, no. 2, pp. 319-328, 2018.

[7] J. Fang *et al.*, "Fast and Accurate Detection of Complex Imaging Genetics Associations Based on Greedy Projected Distance Correlation," *IEEE Transactions on Medical Imaging*, vol. 37, no. 4, pp. 860-870, 2018.

[8] L. Du *et al.*, "Associating Multi-modal Brain Imaging Phenotypes and Genetic Risk Factors via A Dirty Multi-task Learning Method," *IEEE Transactions on Medical Imaging*, vol. 39, no. 11, pp. 3416-3428, 2020.

[9] J. Liu, J. Wang, Z. Tang, B. Hu, F. Wu, and Y. Pan, "Improving Alzheimer's Disease Classification by Combining Multiple Measures," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 15, no. 5, pp. 1649-1659, 2018.

[10] D. Wang *et al.*, "KIBRA gene variants are associated with synchronization within the default-mode and executive control networks," (in eng), *NeuroImage*, vol. 69, pp. 213-222, 2013.

[11] L. French *et al.*, "Early Cannabis Use, Polygenic Risk Score for Schizophrenia and Brain Maturation in AdolescenceGenetic Risk for Schizophrenia and Cortical ThicknessGenetic Risk for Schizophrenia and Cortical Thickness," *JAMA Psychiatry*, vol. 72, no. 10, pp. 1002-1011, 2015.

[12] I. A. C. Romme, M. A. de Reus, R. A. Ophoff, R. S. Kahn, and M. P. van den Heuvel, "Connectome Disconnectivity and Cortical Gene Expression in Patients With Schizophrenia," *Biological Psychiatry*, vol. 81, no. 6, pp. 495-502, 2017.

[13] M. Wang, X. Hao, J. Huang, W. Shao, and D. Zhang, "Discovering network phenotype between genetic risk factors and disease status via diagnosis-aligned multi-modality regression method in Alzheimer's disease," *Bioinformatics*, vol. 35, no. 11, pp. 1948-1957, 2018.

[14] K. Uludağ and A. Roebroeck, "General overview on the merits of multimodal neuroimaging data fusion," *NeuroImage*, vol. 102, pp. 3-10, 2014.

[15] A. R. Hariri and D. R. Weinberger, "Imaging genomics," *British Medical Bulletin*, vol. 65, no. 1, pp. 259-270, 2003.

[16] H. Yang, J. Liu, J. Sui, G. Pearlson, and V. D. Calhoun, "A Hybrid Machine Learning Method for Fusing fMRI and Genetic Data: Combining both Improves Classification of Schizophrenia," *Frontiers in human neuroscience*, vol. 4, p. 192, 2010.

[17] D. Greenstein, J. D. Malley, B. Weisinger, L. Clasen, and N. Gogtay, "Using multivariate machine learning methods and structural MRI to classify childhood onset schizophrenia and healthy controls," *Frontiers in psychiatry*, vol. 3, p. 53, 2012.

[18] D. Lin, V. D. Calhoun, and Y.-P. Wang, "Correspondence between fMRI and SNP data by group sparse canonical correlation analysis," *Medical Image Analysis*, vol. 18, no. 6, pp. 891-902, 2014.

[19] L. Du *et al.*, "Detecting genetic associations with brain imaging phenotypes in Alzheimer's disease via a novel structured SCCA approach," *Medical Image Analysis*, vol. 61, p. 101656, 2020.

[20] B. Lei *et al.*, "Adaptive sparse learning using multi-template for neurodegenerative disease diagnosis," *Medical Image Analysis*, vol. 61, p. 101632, 2020.

[21] X. Hao *et al.*, "Multi-modal neuroimaging feature selection with consistent metric constraint for diagnosis of Alzheimer's disease," *Medical Image Analysis*, vol. 60, p. 101625, 2020.

[22] X. Wang, J. Yan, X. Yao, S. Kim, and H. Huang, "Longitudinal Genotype–Phenotype Association Study through Temporal Structure Auto-Learning Predictive Model," *Journal of Computational Biology A Journal of Computational Molecular Cell Biology*, 2018.

[23] J. Li *et al.*, "ncRNA-eQTL: a database to systematically evaluate the effects of SNPs on non-coding RNA expression across cancer types," (in eng), *Nucleic acids research*, vol. 48, no. D1, pp. D956-D963, 2020.

[24] M. Liu, D. Zhang, E. Adeli, and D. Shen, "Inherent Structure-Based Multiview Learning With Multitemplate Feature Representation for Alzheimer's Disease Diagnosis," (in eng), *IEEE transactions on bio-medical engineering*, vol. 63, no. 7, pp. 1473-1482, 2016.

[25] Z. Wang, Y. Zheng, D. C. Zhu, A. C. Bozoki, and T. Li, "Classification of Alzheimer's Disease, Mild Cognitive Impairment and Normal Control Subjects Using Resting-State fMRI Based Network Connectivity Analysis," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 6, pp. 1-9, 2018.

[26] L. Wang, Z.-H. You, Y.-A. Huang, D.-S. Huang, and K. C. Chan, "An efficient approach based on multi-sources information to predict circRNA–disease associations using deep convolutional neural network," *Bioinformatics*, vol. 36, no. 13, pp. 4038-4046, 2020.

[27] S. Wang *et al.*, "Iterative Label Denoising Network: Segmenting Male Pelvic Organs in CT from 3D Bounding Box Annotations," vol. 67, no. 10, pp. 2710-2720, 2020.

[28] L. Du *et al.*, "Identifying progressive imaging genetic patterns via multi-task sparse canonical correlation analysis: a longitudinal study of the ADNI cohort," *Bioinformatics*, vol. 35, no. 14, pp. i474-i483, 2019.

[29] S. Liu *et al.*, "MIR137 polygenic risk is associated with schizophrenia and affects functional connectivity of the dorsolateral prefrontal cortex," *Psychological medicine*, vol. 50, no. 9, pp. 1510-1518, 2020.

[30] B. Liu *et al.*, "Polygenic risk for schizophrenia influences cortical gyrification in 2 independent general populations," *Schizophrenia bulletin*, vol. 43, no. 3, pp. 673-680, 2017.

[31] X. Hao *et al.*, "Identifying candidate genetic associations with MRI-derived AD-related ROI via tree-guided sparse learning," *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 16, no. 6, pp. 1986-1996, 2018.

[32] W. Hu, A. Zhang, B. Cai, V. Calhoun, and Y.-P. Wang, "Distance canonical correlation analysis with application to an imaging-genetic study," *Journal of medical imaging (Bellingham, Wash.)*, vol. 6, no. 2, p. 026501, 2019.

[33] H. Yuan, X. Zhu, W. Tang, Y. Cai, S. Shi, and Q. Luo, "Connectivity between the anterior insula and dorsolateral prefrontal cortex links early symptom improvement to treatment response," *Journal of affective disorders*, vol. 260, pp. 490-497, 2020.

[34] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5-32, 2001.

[35] H. Ruan *et al.*, "Topographic diversity of structural connectivity in schizophrenia," *Schizophrenia Research*, vol. 215, pp. 181-189, 2020.

[36] S. Ramanan, O. Piguet, and M. Irish, "Rethinking the Role of the Angular Gyrus in Remembering the Past and Imagining the Future: The Contextual Integration Model," *The Neuroscientist*, vol. 24, no. 4, pp. 342-352, 2017.

[37] J. Xiang, H. Guo, R. Cao, H. Liang, and J. Chen, "An abnormal resting-state functional brain network indicates progression towards Alzheimer's disease," (in eng), *Neural regeneration research*, vol. 8, no. 30, pp. 2789-2799, 2013.

[38] X. Zhang, B. Hu*, X. Ma, and L. Xu, "Resting-State Whole-Brain Functional Connectivity Networks for MCI Classification Using L2-Regularized Logistic Regression," *IEEE Transactions on NanoBioscience*, vol. 14, no. 2, pp. 237-247, 2015.

[39] Y. Tian and A. Zalesky, "Characterizing the functional connectivity diversity of the insula cortex: Subregions, diversity curves and behavior," (in eng), *NeuroImage*, vol. 183, pp. 716-733, 2018.

[40] Y. Niu *et al.*, "Dynamic Complexity of Spontaneous BOLD Activity in Alzheimer's Disease and Mild Cognitive Impairment Using Multiscale Entropy Analysis," *Frontiers in neuroscience*, vol. 12, p. 677, 2018.

[41] X. Zhao *et al.*, "Investigating the Correspondence of Clinical Diagnostic Grouping With Underlying Neurobiological and Phenotypic Clusters Using Unsupervised Machine Learning," *Frontiers in applied mathematics and statistics*, vol. 4, p. 25, 2018.

[42] J. Li *et al.*, "Genetic Interactions Explain Variance in Cingulate Amyloid Burden: An AV-45 PET Genome-Wide Association and Interaction Study in the ADNI Cohort," *BioMed research international*, vol. 2015, p. 647389, 2015.

[43] J. Yan *et al.*, "Hippocampal transcriptome-guided genetic analysis of correlated episodic memory phenotypes in Alzheimer's disease," (in eng), *Frontiers in genetics*, vol. 6, p. 117, 2015.

[44] B. Jie, M. Liu, and D. Shen, "Integration of temporal and spatial properties of dynamic connectivity networks for automatic diagnosis of brain disease," *Medical Image Analysis*, vol. 47, pp. 81-94, 2018.

[45] C.-Y. Wee, S. Yang, P.-T. Yap, D. Shen, and I. Alzheimer's Disease Neuroimaging, "Sparse temporally dynamic resting-state functional connectivity networks for early MCI identification," *Brain imaging and behavior*, vol. 10, no. 2, pp. 342-356, 2016.

[46] J. Peng, X. Zhu, Y. Wang, L. An, and D. Shen, "Structured sparsity regularized multiple kernel learning for Alzheimer's disease diagnosis," *Pattern recognition*, vol. 88, pp. 370-382, 2019.

[47] X. Chen, D. Xie, Q. Zhao, and Z.-H. You, "MicroRNAs and complex diseases: from experimental results to computational models," (in eng), *Briefings in bioinformatics*, vol. 20, no. 2, pp. 515-539, 2019.

[48] L. Wei *et al.*, "Iterative feature representations improve N4-methylcytosine site prediction," *Bioinformatics*, vol. 35, no. 23, pp. 4930-4937, 2019.