Bioinformatics, 36(8), 2020, 2561–2568
doi: 10.1093/bioinformatics/btz967
Advance Access Publication Date: 23 January 2020
Original Paper

Bioimage informatics

Morbigenous brain region and gene detection with a genetically evolved random neural network cluster approach in late mild cognitive impairment

Xia-an Bi [1,2,*], Yingchao Liu[1,2], Yiming Xie[1,2], Xi Hu[1,2], Qinghua Jiang[3,*] and for the Alzheimer's Disease Neuroimaging Initiative

[1]Hunan Provincial Key Laboratory of Intelligent Computing and Language Information Processing, [2]College of Information Science and Engineering, Hunan Normal University, Changsha, China and [3]Center for Bioinformatics, School of Life Science and Technology, Harbin Institute of Technology, Harbin, China

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on September 26, 2019; revised on December 12, 2019; editorial decision on December 25, 2019; accepted on January 18, 2020

Abstract

Motivation: The multimodal data fusion analysis becomes another important field for brain disease detection and increasing researches concentrate on using neural network algorithms to solve a range of problems. However, most current neural network optimizing strategies focus on internal nodes or hidden layer numbers, while ignoring the advantages of external optimization. Additionally, in the multimodal data fusion analysis of brain science, the problems of small sample size and high-dimensional data are often encountered due to the difficulty of data collection and the specialization of brain science data, which may result in the lower generalization performance of neural network.

Results: We propose a genetically evolved random neural network cluster (GERNNC) model. Specifically, the fusion characteristics are first constructed to be taken as the input and the best type of neural network is selected as the base classifier to form the initial random neural network cluster. Second, the cluster is adaptively genetically evolved. Based on the GERNNC model, we further construct a multi-tasking framework for the classification of patients with brain disease and the extraction of significant characteristics. In a study of genetic data and functional magnetic resonance imaging data from the Alzheimer's Disease Neuroimaging Initiative, the framework exhibits great classification performance and strong morbigenous factor detection ability. This work demonstrates that how to effectively detect pathogenic components of the brain disease on the high-dimensional medical data and small samples.

Availability and implementation: The Matlab code is available at https://github.com/lizi1234560/GERNNC.git.

Contact: bixiaan@hnu.edu.cn or qhjiang@hit.edu.cn

1 Introduction

Late mild cognitive impairment (LMCI) is a late stage of mild cognitive impairment (MCI). LMCI commonly causes slight cognitive dysfunction and easily converts to dementia involving Alzheimer's disease (AD), frontotemporal dementia or Lewy body dementia, which is an irreversible process. Early diagnosis of LMCI is an important step toward preventing dementia and has a great significance to patients or medical development. At present, the multimodal data fusion analysis is an emerging area for exploring the multiple pathogenic factors of LMCI. It can extract the most significant characteristics and further find out the morbigenous factors (e.g. abnormal brain regions and genes) via the multimodal fusion characteristics, while how to choose the appropriate methods for analyzing the multimodal data is still an increasing challenge.

Neural network technology has been an essential research direction in the field of machine learning. Owing to the variable activation of the internal nodes and the adjustability of the connection weights among the nodes, the neural network can adapt to the complex data analysis tasks in many research fields, especially life sciences (Lutnick *et al.*, 2019). In recent years, increasing neuroscientists have noticed that many machine learning methods including neural network possess unique advantages in sequence learning, medical image processing and prediction (Cui *et al.*, 2016; Erickson *et al.*,

2017; Liu *et al.*, 2019; Wei *et al.*, 2018), which may provide an attractive approach to the emerging multimodal study of brain science.

On account of the heterogeneousness and complexity of the brain data, a great deal of effort is invested to bring potential of neural network in brain science into full play. For example, the neural network can successfully assist the medical personnels to study on brain tumor detection (Havaei *et al.*, 2017), brain image automatic segmentation (Moeskops *et al.*, 2016) and brain image reconstruction (Zhu *et al.*, 2018) through learning the existing large-scale neural imaging. At present, many researchers use the neural network technology to diagnose brain diseases and explore the pathological mechanism, and the results reveal that the technology is effective in diagnosis of these diseases (Deshpande *et al.*, 2015). Moreover, the researchers detect the activation patterns of brain in different states of motion by modifying the number of hidden layers in deep neural network. With the convolutional neural network technology, the medical staffs can also make more reliable survival prediction for people with brain diseases (Cole *et al.*, 2017). These researches show that the neural network is a powerful means in classification, and exhibits a great classification effect (Naseer *et al.*, 2016). The technological advances of neural network provide a vital approach to explore the function of the brain and the pathogenesis of the brain disease.

With the deepening development of the brain science research, new tasks are frequently put forward in the aspects of multidimensional medical data handling, dimensional reduction and precision medicine (Du *et al.*, 2019; Wang *et al.*, 2019). For example, the cost of data acquisition is expensive and time-consuming in the emerging study of multimodal data fusion in brain science (Dosenbach *et al.*, 2017), which limits the accumulation of public data, especially genetic and functional magnetic resonance imaging (fMRI) data. Owing to the specialization of medical datasets, there are a few public and multimodal databases that possess a small amount of brain disease data. Furthermore, the brain science data generally has high-dimensional characteristics and the direct input of high-dimensional data will lead to the computational difficulty or 'dimension disaster', which undoubtedly increases the complexity of the data analysis task. In the traditional research paradigms, the classical methods of processing high-dimensional data include independent component analysis (Smith *et al.*, 2015) and principal component analysis (Artoni *et al.*, 2018), which are likely to cause the loss of raw information. A dimension reduction methods based on artificial bee colony algorithm and clustering are further proposed and make up for the shortcomings of conventional methods in solving high-dimensional data problems (Li *et al.*, 2001; Rao *et al.*, 2019). Similarly, multiple unsupervised feature selection method of dimensional reduction based on sparse linear regression is formulated in recent research (Zheng *et al.*, 2018). These studies clarify that the characteristic selection based on machine learning can eliminate redundant characteristics to achieve the purpose of dimensional reduction (Xu *et al.*, 2019). Additionally, it is well known that the neural network is a typical 'data-hungry' technique (Qu *et al.*, 2019). Therefore, it is a valuable and challenging task for how to maintain the generalization performance of neural network and extract the potential correlation information from the high-dimensional multimodal brain science data under the condition of small samples.

In addition, most current researchers concentrate on the single brain science research task. For example, Hao *et al.* (2017) and Du *et al.* (2016) both improved canonical correlation analysis (CCA) to fuse MRI and single nucleotide polymorphisms (SNPs) data, and found out the significant characteristics associated with AD or MCI. Deshpande *et al.* (2015) applied the reformative artificial neural network to classifying attention deficit hyperactivity disorder and normal people, and the result suggested that the method exhibited the satisfactory classification performance. More importantly, the design of a multi-tasking framework based on neural network for brain diseases could be a significant but easily overlooked assignment which integrates characteristic learning, sample classification and significant characteristic extraction.

In order to meet the above challenges, this article expands the fusion study of multimodal brain scientific data including genetic and fMRI data under the condition of small samples based on the neural network technology. Different from the general optimization strategies, we propose a novel genetically evolved random neural network cluster (GERNNC) model to accomplish the outside optimization of neural network. A random neural network cluster (RNNC) model is first constructed through random selection of samples and characteristics and then the continuous evolutions are carried out to form the GERNNC model. This method integrates a large number of randomly constructed neural networks and introduces the genetically evolved idea to conduct adaptive iterative optimization, which ensures the generalization performance of the cluster in small samples. Additionally, this study applies the GERNNC model to the sample identification and morbigenous factor extraction of brain diseases, and forms a multi-tasking brain science data analysis framework. Finally, we verify the validity of the method in the multimodal data of LMCI. Our work provides a reference for the extensive application of neural network in brain science research and presents a propagable and rapid analytical framework for many brain disease studies.

The remainder of this article is arranged in the following fashion. In Section 2, we describe the proposed GERNNC model and the overall framework of multimodal data analysis. Section 3 shows the experimental results and some discussions. Sections 4 and 5, respectively, give the limitations and conclusions.

2 Materials and methods

In this section, we fuse the fMRI and genetic data to perform the multimodal data analysis and further introduce the GERNNC model. The framework of data analysis and diseases detection with GERNNC is summarized in Figure 1.

2.1 Multimodal fusion characteristics

First, the average time series of the brain regions are got. According to the anatomical automatic labeling (AAL) template, each subject's brain is divided into 90 brain regions, and the first 60 time points for each brain region are extracted as an average time sequence. Each participant is given the average time series of 90 brain regions. Second, the gene sequences are got. The gene corresponding to each SNP after pre-processing is calculated, and the frequency of each gene is calculated and sorted in descending order. For experimental accuracy, we extract the first 36 genes, and further extract the first 30 SNPs from each gene to match the average time series of each brain region, using 1, 2, 3 and 4 to encode the discrete values of base A, T, C and G. Each subject is given 36 gene sequences. Finally, the fusion characteristics are constructed. For each subject, the Pearson correlation coefficient of the brain region and the gene is calculated, and the 3240-dimensional fusion characteristics are eventually obtained. We treat a brain region as a region of interest (ROI); therefore, each dimensional fusion characteristic is called a 'region of interest-gene' (ROI-G) pair. The ROI-G pairs are denoted as $C = \{c_1, \ldots, c_m, \ldots, c_M\}$ and we see the ROI-G pairs as the input to the neural network.

2.2 GERNNC design idea and algorithm

2.2.1 Neural network

It is well known that neural network is a strong and powerful tool. On account of differences in connection nodes, activation functions and training methods for nodes in the neural network (Schmidhuber, 2015), a variety of neural networks are created. To better find out the most appropriate neural network for establishing the GERNNC model, we first choose five general neural networks as base classifiers of RNNC model.

The backpropagation neural network (BPNN) is one of the most outstanding algorithms, which can realize the advantage of the nonlinear mapping (Ma and Meng, 2019). A single BPNN is constructed

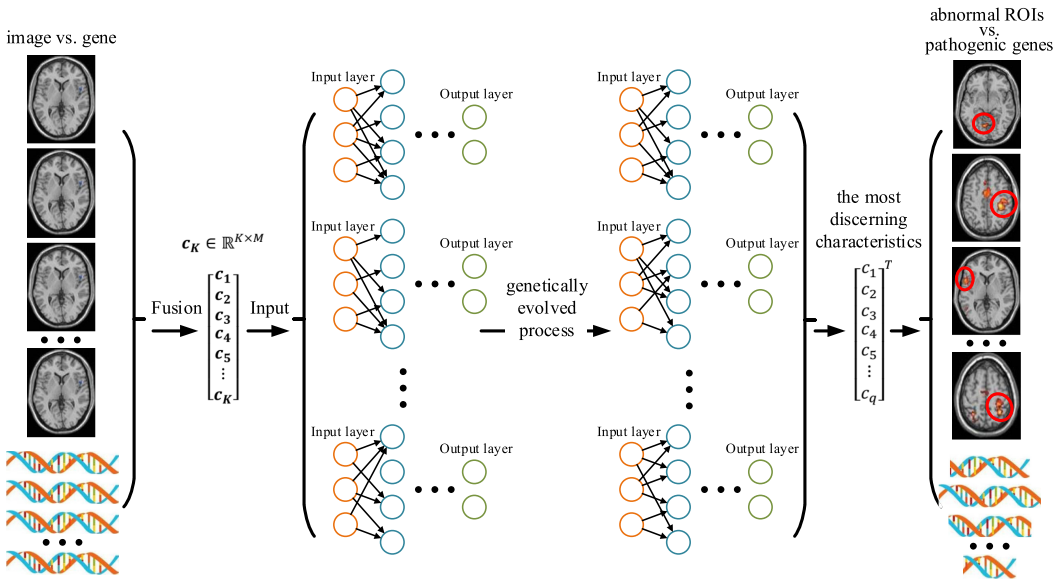

Fig. 1. Multimodal data analysis and diseases detection framework with GERNNC. We fuse gene and neuroimaging data to form characteristic matrix, where c_K represents characteristic vector in a base classifier. A total of K base classifiers are constructed to integrate a cluster that is genetically evolved and q characteristics with the most discerning abilities are extracted to further find out abnormal ROIs and pathogenic genes

and trained using the training set and the fusion characteristics, the training error of which is denoted as:

$$E_{\mathrm{BPNN}} = \frac{1}{2}\sum_{d=1}^{D}(\hat{h}_d - h_d)^2 \tag{1}$$

where D represents the quantity of output layer, $\hat{h}_d$ represents the output of BPNN and h_d represents the target output.

The probabilistic neural network (PNN) is a simple neural network on the basis of Bayesian decision theory (Specht, 1990), which possesses the properties of short training time and good expansion performance. The input–output relationship of output layer is denoted as:

$$\varphi_{ij}(s) = \frac{1}{(2\pi)^{\frac{1}{2}}\sigma^d} e^{-\frac{(s-s_{ij})(s-s_{ij})^T}{\sigma^2}} \tag{2}$$

$$T = \mathrm{argmax}\left(\frac{\sum_{j=1}^{L}\varphi_{ij}(s)}{L}\right) \tag{3}$$

where $\varphi_{ij}(s)$ and T, respectively, represent the input–output relationships of hidden layer and output layer. s_{ij} represents the jth sample of ith class, σ represents the smoothing factor and L represents the number of neurons in ith class.

The Elman neural network (ENN) has the better stability and functions of local memory and feedback, which can process time-varying data (Kremer, 1995). The training error of ENN is denoted as:

$$E_{\mathrm{ENN}} = \frac{1}{2}\left(h(t) - h(\hat{t})\right)^T\left(h(t) - h(\hat{t})\right) \tag{4}$$

where $h(\hat{t})$ represents the output of ENN at t time and $h(t)$ represents the target output at t time.

The competitive neural network (CNN) has the merits of simple structure and simple learning algorithm. The output vector is denoted as:

$$Z = Vw \tag{5}$$

where V represents the input vector of CNN and w represents the connection weight.

The learning vector quantization neural network (LVQNN) is a simple and powerful neural network classification method (Chen *et al.*, 2017). The output class of LVQNN is determined by the weights of input layer versus competition layer and competition layer versus output layer. The five neural networks are, respectively, integrated by the ensemble learning idea.

2.2.2 GERNNC algorithm

We develop the traditional neural network by combining the ensemble idea with the genetically evolved idea to design the GERNNC model. The model is accomplished in two main steps. First, the ensemble idea is introduced to effectively integrate multiple neural networks for building an RNNC model, where the randomness is reflected in the random selections of samples and characteristics from the entire sample set and characteristic set. Second, the genetically evolved idea is introduced into the constructed RNNC to eliminate the neural network with poor classification performance and increase the explanatoryness of the cluster.

We aim to design a neural network model capable of distinguishing patients from normal controls (NC) and extracting the most discerning characteristics. The sample set is denoted as $S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ and divided into training set S_{train} and testing set S_{test}. x_n represents nth sample that has M fusion characteristics and y_n represents the label of nth sample. We severally exploit the five neural networks to construct five RNNCs by randomly selecting a samples and b characteristics. In each RNNC, we construct K base classifiers and integrate the multiple neural networks to circumvent the negative influences of superabundant parameters in a single neural network. The classification performances of five clusters at stable state in the genetically evolved process are evaluated and contradistinguished to find out the optimal neural network.

Based on the cluster with optimal base classifier, the binary system that possesses the superiorities of easier implementation and comprehensibility is employed to initialize characteristics, where '1' represents the randomly selected characteristic and '0' represents the

unpicked characteristic. Therefore, the cluster with K base classifiers is constituted by a binary matrix $C \in \mathbb{R}^{K\times M}$, which is expressed as:

$$\begin{bmatrix} c_{1,1} & c_{1,2} & c_{1,3} & \cdots & c_{1,M} \\ c_{2,1} & c_{2,2} & c_{2,3} & \cdots & c_{2,M} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_{K,1} & c_{K,2} & c_{K,3} & \cdots & c_{K,M} \end{bmatrix} \tag{6}$$

where $c_{K,M}$ represents mth characteristic of kth base classifier in the cluster. Since the characteristics are randomly selected, one of the expressions for the cluster is likely to be (7).

$$\begin{bmatrix} 0 & 1 & 1 & \cdots & 0 \\ 1 & 0 & 1 & \cdots & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 1 & \cdots & 1 \end{bmatrix} \tag{7}$$

The quantity of '1' is b in each base classifier, which is regarded as the input of each neural network. We consider the cluster with the binary system as the initial RNNC model.

In order to enhance the stability of RNNC model, we adopt genetically evolved idea for cluster optimization. Through the ongoing procedure of genetic evolution, a large number of high-performance base classifiers are preserved to allow the evolved cluster to cover the most useless information of the pre-evolutionary cluster, and lots of superior base classifiers in the cluster are generated every time to make the performance of the cluster continuously close to optimal. The genetically evolved process is executed by three steps of selecting, crossing and variation. The selecting procedure is based on the fitness evaluation of the base classifier. The classification accuracy of each neural network in the cluster is computed and defined as the fitness function to appraise the performance of neural network. The fitness function is formulated as:

$$fin_k = \frac{g_{\text{true},k}}{G} \tag{8}$$

where $g_{\text{true},k}$ represents the quantity of truly classified samples in the kth neural network and G represents total amount of samples in the testing samples. It is noted that the neural network with highest fitness function value has strongest classification ability and will be selected to form all base classifiers. On the contrary, the network with the lowest fitness function value is picked out and replaced by that with the highest value. The crossing procedure is performed to acquire the recombination of characteristics in partial base classifiers, which keeps most base classifiers with strong classification performance and generates superior base classifiers. Since some base classifiers are not selected in the process of crossing, the variation procedure is carried to escalate the diversity and performance of base classifiers by changing the binary characteristic from '1' to '0' or '0' to '1' in some classifiers. The genetically evolved process is continuously go on until the constraints are reached as follows:

$$\begin{aligned} &\min(p) \\ &\text{s. t. } \Delta\text{ACC} < \varepsilon \\ &p \le P \end{aligned} \tag{9}$$

where ACC represents the accuracy of cluster. p is the genetically evolved times and P is the largest value of that. Equation (9) means that when the accuracy of cluster tends to be stable within the range of P, the genetically evolved process is terminated, resulting in the optimized RNNC model. The construction procedure of GERNNC is shown in Algorithm 1.

Algorithm 1 The GERNNC Algorithm

Input: Original dataset $S=\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, Original characteristic set C
Output: The lowest genetically evolved times
1: Initialize S, C, k, w, where k is the quantity of base classifiers, w is the genetically evolved termination condition
2: Partitioned the S into S_{train} and S_{test}
3: **for** $i=1$ to k **do**
4: Randomly select a subset of samples and characteristics from S_{train} and C
5: Generate {base_classifier$_i$}
6: Update binary matrix C using (6)
7: **end for**
8: **repeat**
9: Calculate the fitness function using (8)
10: Selecting
11: Crossing
12: Variation
13: **until** w
14: Select the lowest genetically evolved times when the classification accuracy of RNNC is stable

2.3 Hyperparameter tuning

To better optimize the model, we apply the optimal neural network to running the cluster with different quantities of base classifiers for many experiments in the genetically evolved process to confirm the optimal combination of base classifiers' number and genetically evolved times. The quantity of neural networks having the highest stable classification accuracy is observed at each experiment and the corresponding minimum number of genetic evolutions is considered as the optimal times, which ensures the lowest space–time complexity and avoids the overfitting. Through the grid search approach, all results are marked to pick out the optimal combination. Consequently, we find out the optimal hyperparameter of the most suitable number of neural networks via using the tuning methods.

2.4 Extraction of morbigenous brain regions and genes

The GERNNC model is primarily used for characteristic selection, and the 3240-dimensional ROI-G pairs that have been built are the classification characteristics. Through the genetically evolved process, the base classifiers in the final cluster have strong identification ability and the characteristics in each base classifier have powerful capability of identification. With the aim of selecting the optimal characteristics, the frequencies of the selected characteristics in each base classifier are first summed to get the frequencies of occurrence for all characteristics based on the final cluster. According to frequencies, these characteristics are sorted in descending order. With the decreasing of the frequency, the distinguishing ability of the ROI-G pair corresponding to each frequency becomes weaker and weaker; therefore, the first t characteristics are extracted from 3240-dimensional ROI-G pairs to further find out the optimal characteristics. We choose the first r ($r \le t$) characteristics as the input of the reconstructed RNNC and increase the number of characteristics with the step of 5 to test the classification performance of each characteristic set. The characteristics having the highest accuracy are seen as the optimal characteristics, which are considered to be the most conspicuous characteristics that distinguish patients from normal people in the whole characteristic set.

There are two elements of ROI and gene in each characteristic, the abnormalities of which further incarnate significance of optimal characteristics set. For estimating the contribution of each ROI or gene to identify patients and normal people, we split the optimal characteristics into ROIs and genes, and count the frequencies of ROIs or genes. The frequencies are used to measure the extent of the anomaly for an ROI or a gene, the larger frequency means that the element is more abnormal. In the context of statistical analysis for optimal characteristics, it can be used to find out the morbigenous genes and brain regions.

Table 1. Demographic information of LMCI and NC

Variables	LMCI ($n=26$)	NC ($n=36$)	P-value
Gender (male/female)	14/12	14/22	0.243[a]
Age (mean ± SD)	72.45±7.47	75.84±6.27	0.057[b]

[a]The P-value was acquired via the chi-square test.
[b]The P-value was acquired via the two-sample t-test.

3 Results

3.1 Data acquisition and preprocessing

The genetic and fMRI data were acquired from the public database of Alzheimer's Disease Neuroimaging Initiative (ADNI, adni.loni.usc.edu/). Since 2003, the ADNI was dedicated to collecting neuroimaging data, genetic data and biomarkers of various subjects including AD, LMCI and NC, making great contributions to the research of AD and other cognitive disorders. We collected 26 LMCI patients and 36 NC from the ADNI database, containing fMRI and SNP data. The selected dataset required the participant to remain at rest with closed eyes, no thinking and no task. The research was supported by the Banner Alzheimer's Institute, etc. and all participants have signed the informed consent. Table 1 summarized the basic information of the participants.

For the fMRI image, the Philips Medical System was employed to scan brains, with a field strength of 3.0 T, a flip angle of 80.0°, a slice thickness of 3.3 mm, a TR (Time of Repeatation) of 3000.0 ms and a TE (Time of Echo) of 30.0 ms. The images were preprocessed by DPARSF toolbox within MATLAB (MathWorks, Inc., version 2014b). The image processing steps included the format conversion, removing the first 10 volumes, correcting time slices and head action parameters, space standardization, Gaussian smoothing (FWHM, Full Width at Half Maximum = 6 mm), removing covariate and temporal filtering. Using the AAL template, the fMRI image of each sample was divided into 90 brain regions, containing corresponding time series.

For the SNP data preprocessing, the PLINK was utilized for the quality control. The sample recall threshold was set to 95% to evaluate overall quality of data. Moreover, the thresholds for Hardy–Weinberg equilibrium test, minimum allele frequencies and genotyping rates were further set to 1e−4, 99.9% and 5%, respectively, resulting in 82 400 SNPs.

3.2 Constructions of characteristics and GERNNC model

The proposed method was validated using LMCI-related data in this study. All experimental sample data were provided by ADNI. For the quality control, these data need a series of preprocessing. For each participant, 90 brain regions and 82 400 SNPs were retained to establish the multimodal fusion characteristics. Through the designed fusion approach in the section of methods, we extracted the sequence information from 90 brain regions and 36 genes to carry out the correlation analysis, and finally constructed 3240 fusion characteristics which are also called the ROI-G pairs.

In order to select the best type of neural network, we evaluated the integrated performances of the five different types of neural network including BPNN, PNN, ENN, CNN and LVQNN. The 57 fusion characteristics and 36 samples were randomly selected as the training set of a neural network at each time. The number of neural networks in each type of RNNC was set to 300, and the genetically evolved process was carried out to stabilize classification performances of clusters. The integrated performances of five neural networks clusters were summarized in Figure 2. We found that with the increase of genetically evolved times, the classification accuracies of five clusters were gradually stable. The variation trend of the random BPNN cluster's classification accuracy was ascend in first and descend at last. The variation trend of the random PNN cluster's classification accuracy was gradually ascend and leveled off, and the accuracy was 84%. The variation trend of the random ENN cluster's classification accuracy was relatively smooth, while the accuracy was ∼72%. The variation trend of the random CNN cluster's

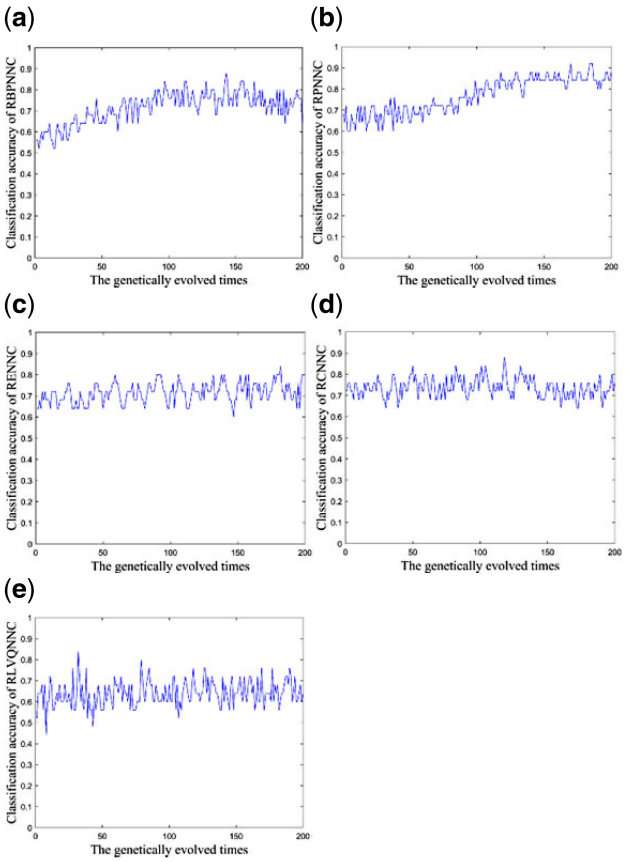

Fig. 2. Integrated performance comparison of five neural networks. The LMCI-related dataset is applied to five random neural network ensembles to assess the classification performances. The genetically evolved times are set to a range of (0, 200) and the different variation tendencies are observed, which is convenient for finding out the optimal neural networks. (**a**) random BPNN cluster (RBPNNC), (**b**) random PNN cluster (RPNNC), (**c**) random ENN cluster (RENNC), (**d**) random CNN cluster (RCNNC), (**e**) random LVQNN cluster (RLVQNNC)

classification accuracy floated up or down and the accuracy was ∼70%. The variation trend of the random LVQNN cluster's classification accuracy had large trend fluctuation. In all clusters, the accuracy of the random PNN cluster achieved the highest, when the accuracy rate tends to a stable value. Consequently, the PNN was affirmed as the base classifier to construct the genetically evolved random PNN cluster (GERPNNC) model in the subsequent experiments.

In this study, we further carried out plentiful training and debugging of the GERPNNC to optimize the classification performance. Specifically, through adjusting the number of neural networks in the cluster, we found out the genetically evolved times when the cluster classification performance tended to be stable. Figure 3 summarized the relationships between the different numbers of base classifiers and the corresponding genetically evolved times. At the coordinate of (160, 70), the number of neural networks and the genetically evolved times were both less and the cluster was regarded as the final cluster for further analysis, which also helped to reduce the consumption of system resources in practical application.

3.3 Selection of the most discerning characteristics

The final cluster composed of 160 PNNs was obtained through 76 genetically evolved times. In the genetically evolved process, the accuracy rate gradually increased from 60% to 84% and eventually remained stable. It is indicated that the genetically evolved process realized the filtering of irrelevant or redundant characteristics, which achieved the purpose of preserving characteristics with strong recognition capabilities. Since these characteristics were randomly

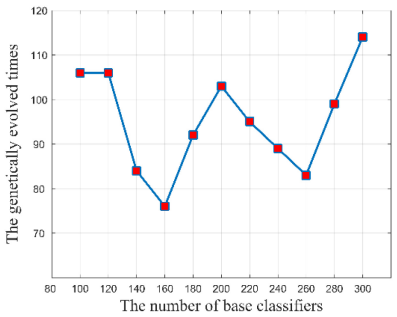

Fig. 3. Construction of the GERPNNC model. We choose the different quantities of base classifiers to construct the multiple random PNN clusters. When the accuracies of clusters tended to be stable, the different clusters correspond to diverse genetically evolved times. In order to reduce the time complexity and space complexity, the 160 base classifiers and corresponding 76 genetically evolved times are considered the optimal combination for establishing the GERPNNC model

selected by base classifiers, we counted the frequencies of different characteristics selected by each base classifier in the final cluster, and took the first 400 characteristics with higher frequencies as important characteristics. The first t ($t = 70, 75, \ldots, 400$) characteristics of important characteristic set were extracted and their recognition abilities were assessed via a random PNN cluster, as shown in Figure 4a. The identification ability of the first 205 characteristics was the best and retained as the most discerning characteristic set, which further filtered out unrelated or redundant characteristics. The 20 most significant fusion characteristics were exhibited in Figure 4b. We observe that some ROIs or genes have multiple connections, which illuminates those are likely to make important contributions on the classification of brain diseases.

3.4 Performance evaluation and comparison

As a benchmark to compare with our proposed characteristic fusion method and the GERNNC model, we trained the random support vector machine cluster (RSVMC), random forest (RF) and two-sample t-test using the same sample set and fusion characteristics constructed based on the Pearson correlation analysis. We further computed the overlaps of the optimal characteristics extracted between the GERNNC model and other methods. We also applied the classical CCA and correlation distance (CD) methods to establishing the fusion characteristics and used the traditional two-sample t-test to extract the optimal characteristics. The comparison attempted to combine some classical fusion characteristic construction approaches and characteristic extraction methods to form the overall frameworks. The overlaps and the performance difference are summarized, as shown in Table 2. The numbers of optimal ROI-G pairs extracted by different methods were analyzed statistically, and the SVM was used to test the identification abilities of the optimal characteristic sets.

From Table 2, it is learned that the number of optimal characteristics extracted by our method is the least among all frameworks, while the identification ability is the best. In addition, the overlaps of optimal characteristics between the GERPNNC and other methods are observed, the non-contingency of which are proved by the hypergeometric test. More interestingly, the larger the overlaps are, the better the classification performance is. Therefore, it is believed that the GERPNNC method can be used as a novel sample classification and pathogenic factors detection method.

To better assess the effectivities of the genetically evolved idea and multimodal data fusion scheme, we further calculated the accuracies of random PNN clusters with different genetic evolution times, and carried out comparative experiments with the two-sample t-test underlying the multimodal data (fMRI and SNP) and unimodal data (fMRI or SNP), which are summarized in Figure 5. With the increasing of evolved times, the accuracy rates of GERPNNC are first raised and then stabilized at around 84%, showing that genetically evolved idea can effectively improve the performance of cluster. The comparison methods are accomplished

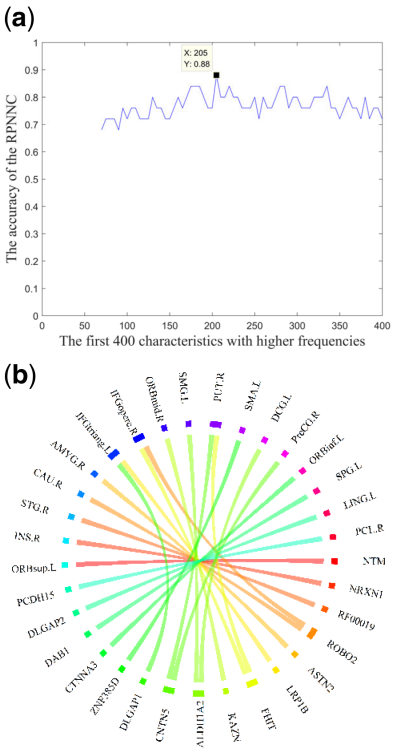

Fig. 4. The most discerning characteristics. Different quantities of characteristics constitute multiple characteristic sets to construct the random PNN clusters, the classification accuracies of which are computed. (**a**) The peak value represents that the characteristic set has the best distinguishing ability to make the performance of the cluster optimized. (**b**) The first 20 most discerning characteristics are extracted to embody the links between ROIs and genes

on the multimodal data of combining fMRI with gene and the unimodal data of fMRI or gene. The accuracy rates are below 80% and lower those of GERPNNC model. It is noted that there is a higher classification result on fMRI data in the t-test experiment, which may be due to the randomness of sample and characteristic selection during model construction. These results not only reflect that the GERPNNC model has better adaptability to fusing multimodal data, but also confirm the capacity for multimodal data information complementarity. The advantages of machine learning could be more obvious than that of the conventional methods.

From the results of comparison with other methods, it can be found that the classification performance of the GERPNNC is better than that of other models, and the high-dimensional characteristics can be selected to achieve the purpose of dimensional reduction. The performance advantages of this model are mainly reflected in the following three levels. First, the genetic evolution is used as an optimization strategy to improve the learning ability of the cluster. In the genetically evolved process, the redundant or invalid characteristics are removed, which makes the performance of clusters gradually increase and remain stable. At the same time, by setting the genetically evolved terminal condition, learning efficiency is guaranteed and the overfitting is avoided. Second, via searching for the optimal number of neural networks in the model, the genetically evolved times of the cluster are as small as possible to ensure the efficiency of the cluster. The above two levels make the GERPNNC have a good global optimization ability. Finally, the multimodal data fusion method and the GERPNNC approach are effectively integrated into the overall framework, which can make better use of the complementarity between genetic data and fMRI data to improve the performance.

3.5 Analysis of abnormal ROIs and pathogenic genes

Figure 4b visually displays the fusion characteristics constituted by the correlations between the ROIs and the genes. The ROI-G pairs

Table 2. Comparison with classical characteristic extraction methods

Methods	Discoveries	Accuracy (%)	Overlaps
Pearson versus GERPNNC	205	87.5	—
Pearson versus RF	620	75.0	131 ($P = 1.670202e{-}24$)
Pearson versus RSVMC	705	70.8	104 ($P = 1.288661e{-}53$)
Pearson versus *t*-test	335	66.7	91 ($P = 2.435895e{-}06$)
CCA versus *t*-test	294	58.3	56 ($P = 2.918568e{-}14$)
CD versus *t*-test	323	66.7	73 ($P = 8.994176e{-}11$)

Note: The *P*-value was acquired by the hypergeometric test.
t-test, two-sample *t*-test.

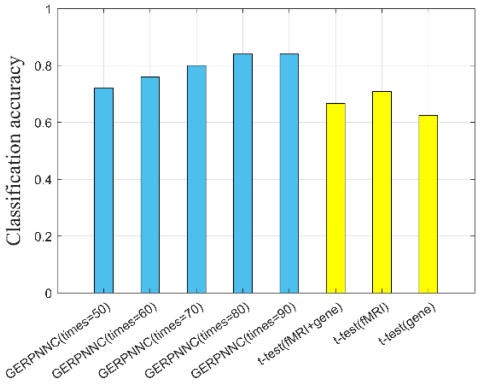

Fig. 5. Comparison with two-sample *t*-test. Based on the training results of GERPNNC model, we set the genetically evolved times to 50, 60, 70, 80 and 90 for observing the performance variation during genetically evolved progression. Moreover, the *t*-test is compared to our model underlying unimodal and multimodal data

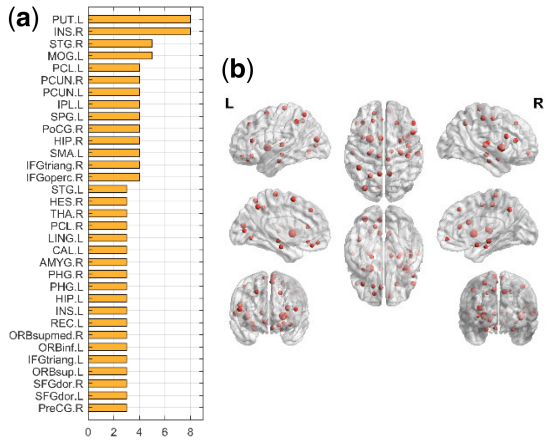

Fig. 6. Locations, frequencies and sizes of segmental ROIs. The ROIs with higher frequencies are extracted as morbigenous ROIs. (**a**) The highest frequency is 8 and a small percentage of the 90 ROIs possesses higher frequencies. It is noted that there are not ROIs with frequencies of 7 and 6. (**b**) The locations and sizes of morbigenous ROIs are shown in coronal, sagittal and axial maps of the brain

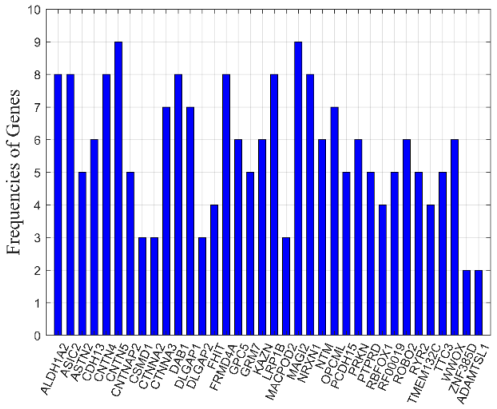

Fig. 7. Frequencies of all genes. By the ROI-G pairs with strong identification abilities, the frequency of each gene occurred in optimal characteristic set is calculated to find out the morbigenous genes, such as CNTN5, MAGI2 and ALDH1A2

in the most discerning characteristics mean that they have significantly strong identification abilities in terms of classification between normal people and patients with brain disease. These ROI-G pairs are regarded as the abnormal characteristics which can provide evidences for us to explore abnormal ROIs and pathogenic genes.

Therefore, we isolated the ROIs and genes from the optimal characteristic set, and counted the frequencies of different ROIs or genes. The greater the frequency of ROI or gene is, the more likely it is to be associated with the brain disease. Figures 6 and 7 depicted the abnormal ROIs and disease-causing genes connected with LMCI found in this study, and these pathogenic factors have been verified by existing studies. For example, Liang *et al.* (2014) found that the four groups of CN (control normal)/early MCI (EMCI)/LMCI/AD had significant differences in gray matter in bilateral hippocampus, bilateral insula, bilateral postcentral gyrus and right angular gyrus. Zhu *et al.* (2019) found that during the development of EMCI to LMCI, the APOEε4 gene carriers showed increased functional connectivities in the precuneus and hippocampus, and decreased functional connectivities in the insula, while the APOEε4 gene non-carriers showed an entirely opposite pattern. Li *et al.* (2015) revealed that the MAGI2 gene was possibly associated with LMCI in 2015, and Li *et al.* (2017) found out abnormalities in the CDH13 gene in genome-wide association analysis of LMCI patients in 2017.

4 Results and discussion

Although we demonstrated the efficiency of the GERNNC algorithm that distinguished people with LMCI from NC and provided a reliable basis for predicting brain diseases, there are some limitations that need to be further discussed. First, this model is mainly based on the characteristics of ROI-G pairs to study brain diseases, therefore we can study brain diseases based on the characteristics of voxel-gene pairs in the follow-up work. Second, the AAL template is used to match the brain. We can also use other templates to match the brain.

5 Conclusions

In this article, we efficaciously fused the fMRI and gene using the complementary information, and proposed a framework for the exploration of potential pathogenic factors and the early diagnosis of LMCI based on the machine learning method. In the framework, the GERNNC model was presented for the first time to effectively deal

with the classification challenge under the condition of small samples. Compared with the frameworks consisting of general correlational analysis methods and classical machine learning or statistical approaches, we took advantage of the neural networks to achieve a high classification accuracy and the effects of characteristic extraction achieved the most outstanding, demonstrating that the GERNNC was a powerful tool for identifying brain diseases. The rapid and scalable approach was easy to deploy and could have a significant impact on clinical decision-making and understanding disease mechanisms. Furthermore, the proposed framework can be extended to other brain diseases, such as AD.

Acknowledgements

ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

Funding

This work was supported in part by the Hunan Provincial Science and Technology Project Foundation [2018TP1018] and the National Science Foundation of China [61502167]. Data collection and sharing for this project was funded by ADNI [National Institutes of Health Grant U01 AG024904]. ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego.

Conflict of Interest: none declared.

References

Artoni,F. *et al.* (2018) Applying dimension reduction to EEG data by principal component analysis reduces the quality of its subsequent independent component decomposition. *Neuroimage*, **175**, 176–187.

Chen,Y. *et al.* (2017) A feature-free 30-disease pathological brain detection system by linear regression classifier. *CNS Neurol. Disord. Drug Targets*, **16**, 5–10.

Cole,J.H. *et al.* (2017) Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker. *Neuroimage*, **163**, 115–124.

Cui,Y. *et al.* (2016) Continuous online sequence learning with an unsupervised neural network model. *Neural. Comput.*, **28**, 2474–2504.

Deshpande,G. *et al.* (2015) Fully connected cascade artificial neural network architecture for attention deficit hyperactivity disorder classification from functional magnetic resonance imaging data. *IEEE Trans. Cybern.*, **45**, 2668–2679.

Dosenbach,N.U.F. *et al.* (2017) Real-time motion analytics during brain MRI improve data quality and reduce costs. *Neuroimage*, **161**, 80–93.

Du,L. *et al.*; for the Alzheimer's Disease Neuroimaging Initiative (2016) Structured sparse canonical correlation analysis for brain imaging genetics: an improved GraphNet method. *Bioinformatics*, **32**, 1544–1551.

Du,L. *et al.*; Alzheimer's Disease Neuroimaging Initiative (2019) Identifying progressive imaging genetic patterns via multi-task sparse canonical correlation analysis: a longitudinal study of the ADNI cohort. *Bioinformatics*, **35**, i474–i483.

Erickson,B.J. *et al.* (2017) Machine learning for medical imaging. *RadioGraphics*, **37**, 505–515.

Hao,X. *et al.*; for the Alzheimer's Disease Neuroimaging Initiative (2017) Identification of associations between genotypes and longitudinal phenotypes via temporally-constrained group sparse canonical correlation analysis. *Bioinformatics*, **33**, i341–i349.

Havaei,M. *et al.* (2017) Brain tumor segmentation with deep neural networks. *Med. Image Anal.*, **35**, 18–31.

Kremer,S.C. (1995) On the computational power of Elman-style recurrent networks. *IEEE Trans. Neural Netw.*, **6**, 1000–1004.

Li,W. *et al.* (2001) Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics*, **17**, 282–283.

Li,J. *et al.* (2015) Genetic interactions explain variance in cingulate amyloid burden: an AV-45 PET genome-wide association and interaction study in the ADNI cohort. *Biomed. Res. Int.*, **2015**, 647389.

Li,J. *et al.* (2017) Genome-wide association and interaction studies of CSF T-tau/Aβ42 ratio in ADNI cohort. *Neurobiol. Aging*, **57**, 247.e241–247.e248.

Liang,P. *et al.* (2014) Altered amplitude of low-frequency fluctuations in early and late mild cognitive impairment and Alzheimer's disease. *Curr. Alzheimer Res.*, **11**, 389–398.

Liu,B. *et al.* (2019) BioSeq-Analysis2.0: an updated platform for analyzing DNA, RNA and protein sequences at sequence level and residue level based on machine learning approaches. *Nucleic Acids Res.*, **47**, e127.

Lutnick,B. *et al.* (2019) An integrated iterative annotation technique for easing neural network training in medical image analysis. *Nat. Mach. Intell.*, **1**, 112–119.

Ma,Y. and Meng,Y. (2019) Research on enterprise innovation persistence patterns recognition and selection based on BP neural network. *Am. J. Ind. Bus. Manage.*, **9**, 658–679.

Moeskops,P. *et al.* (2016) Automatic segmentation of MR brain images with a convolutional neural network. *IEEE Trans. Med. Imaging*, **35**, 1252–1261.

Naseer,N. *et al.* (2016) Analysis of different classification techniques for two-class functional near-infrared spectroscopy-based brain-computer interface. *Comput. Intell. Neurosci.*, **2016**, 1–11.

Qu,Y. *et al.* (2019) Migrating knowledge between physical scenarios based on artificial neural networks. *ACS Photonics*, **6**, 1168–1174.

Rao,H. *et al.* (2019) Feature selection based on artificial bee colony and gradient boosting decision tree. *Appl. Soft Comput.*, **74**, 634–642.

Schmidhuber,J. (2015) Deep learning in neural networks: an overview. *Neural Netw.*, **61**, 85–117.

Smith,S.M. *et al.* (2015) A positive-negative mode of population covariation links brain connectivity, demographics and behavior. *Nat. Neurosci.*, **18**, 1565–1567.

Specht,D.F. (1990) Probabilistic neural networks. *Neural Netw.*, **3**, 109–118.

Wang,M. *et al.* (2019) Discovering network phenotype between genetic risk factors and disease status via diagnosis-aligned multi-modality regression method in Alzheimer's disease. *Bioinformatics*, **35**, 1948–1957.

Wei,L. *et al.* (2018) ACPred-FL: a sequence-based predictor using effective feature representation to improve the prediction of anti-cancer peptides. *Bioinformatics*, **34**, 4007–4016.

Xu,X. *et al.* (2019) Review of classical dimensionality reduction and sample selection methods for large-scale data processing. *Neurocomputing*, **328**, 5–15.

Zheng,W. *et al.* (2018) Identification of Alzheimer's disease and mild cognitive impairment using networks constructed based on multiple morphological brain features. *Biol. Psychiatry Cogn. Neurosci. Neuroimaging*, **3**, 887–897.

Zhu,B. *et al.* (2018) Image reconstruction by domain-transform manifold learning. *Nature*, **555**, 487–492.

Zhu,Y. *et al.*; on behalf of Alzheimer's Disease Neuroimaging Initiative. (2019) Default mode network connectivity moderates the relationship between the APOE genotype and cognition and individualizes identification across the Alzheimer's disease spectrum. *J. Alzheimers Dis.*, **70**, 843–860.