ISSN: 1813-4890

Research on Alzheimer's Disease Diagnosis Model Based on Random Forest and XGBoost

Xiaohui Li, Ziheng Liu, Baixue Fan, Sheng Chu and Jitao He

School of Anhui University of Chinese Medicine, Anhui, China;

Abstract

In this paper, the significance of early and accurate diagnosis of Alzheimer's disease and mild cognitive impairment was studied, and a variety of Alzheimer's disease diagnostic models such as XGBoost model, random forest K nearest neighbor naïve Bayes, and decision tree were constructed, and the one with the highest accuracy was obtained, and participants with a long timeline and more than 15 times of information collection were selected to construct an overall evolution model of Alzheimer's disease and three different Alzheimer's disease evolution models. Conclusions: Early detection and treatment of Alzheimer's disease and mild cognitive impairment is of great significance. Finally, the advanced search function of the literature retrieval platform based on CNKI journal database was used to retrieve 44 relevant literatures, and the early intervention and diagnostic criteria of five types of patients: CN, SMC, EMCI, LMCI and AD were summarized and described. The results show that early and accurate diagnosis and treatment of Alzheimer's disease and mild cognitive impairment can effectively reduce the risk of disease, and summarize the criteria for early intervention and diagnosis of relevant patients, contributing to the research of Alzheimer's disease and mild cognitive impairment.

Keywords

Spearman;Random Forest;XGBoost Classification Model;Time Evolution Mode.

1. Introduction

1.1. Background

Since the proposal of the first meeting of the 13th National Committee of the Chinese People's Political Consultative Conference (CPPCC) in 2018, the proposal of the fourth meeting of the National Health and Wellness Committee of the People's Republic of China has clearly expressed that it attaches great importance to the health problems of the elderly with Alzheimer's disease, and proposes to strengthen the research and development and application of scientific and technological innovation in neurodegenerative diseases such as Alzheimer's disease (AD)[1]-[2].Alzheimer's disease usually occurs in the elderly, and it can be divided into the elderly with normal cognition (CN), mild cognitive impairment (MCI) and Alzheimer's disease (AD) according to its different clinical stages. Its clinical features are the decline of daily living activities and dementia, accompanied by various behavioral disorders and neuropsychiatric symptoms.

Because of the concealment of the onset of the disease, the unknown cause and the limitation of people's cognition, 67% of the patients were diagnosed as moderate or severe stage, usually died of complications 10-20 years after the onset, and missed the best intervention stage.Therefore, it is of great significance to develop an intelligent recognition model that can accurately diagnose Alzheimer's disease and mild cognitive impairment in the early stage according to the characteristics of brain structure and behavior cognition of different groups of people, to support the prevention and rescue services for disabled elderly people, and to

ISSN: 1813-4890

promote the organization and implementation of key projects such as the national key research and development plan "Active Health and Aging Technology Response"[2].

2. Symbol and Assumptions

2.1. Symbol Description

Numble	Instructions
P_s	Pearson grade correlation coefficient
d_i	Difference in rank
$\hat{y}_j$	Predicted value
L	Number of trees to adjust
f_l	The tree of l
F	The collection of decision trees
$Obj^{(t)}$	The target function for the tree of t
$loss(\ )$	Function of loss
$\Omega(f_t)$	The regular term of the tree of t

2.2. Fundamental assumptions

(1)It is assumed that the characteristics of brain structure and cognitive behavior in the appendix are enough to diagnose Alzheimer's disease, regardless of other influencing factors.

(2)It is assumed that both data extraction and feature extraction are random.

(3)Assume that there are no outliers in the processed data.

2.3. Indicators, data, and methods

The attached data sample size is large, there are too many characteristic indicators related to Alzheimer's disease, and some indicators have a large number of missing values, which will affect the solution of subsequent problems, so we first preprocess the data. The steps taken are as follows:

(1)All indexes are classified to facilitate the subsequent recognition and extraction of brain structural features and cognitive indexes.

(2)Missing value processing.The missing values of some data indicators exceed 80%, and there are many missing values, which are not representative, and affect the follow-up correlation analysis and modeling excellence, so we reject them. For the remaining missing values, we use the sequential mean method to interpolate.

(3)Abnormal value processing.The amount of data processed this time is large enough, and the proportion of outliers is very small. Therefore, we adopt the processing method of eliminating samples for outliers in the data.

ISSN: 1813-4890

3. Model

3.1. Model Ⅰ Models and Results

3.1.1. Thinking Analysis

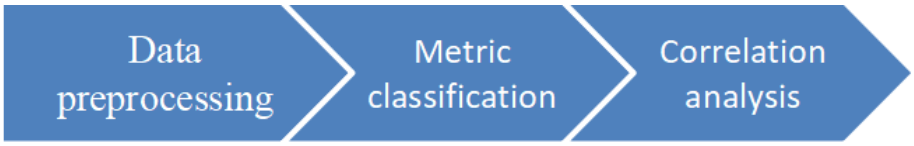

Figure 1 Analysis diagram

3.1.2. Data Preprocessing

(1)Data statistics

There are 16,223 samples in the annex, including 116 physical indicators, biological indicators, cognitive evaluation indicators and diagnosis results measured at different time points including patient information. We simply make descriptive statistics, and the results are shown in Table 1:

Table1 Descriptive Statistics

Numble	count	min	max	mean	std
RID	16222	2.00	7105.00	2866.05	2205.25
AGE	16213	50.40	91.40	73.25	7.01
PTEDUCAT	16222	4.00	20.00	16.10	2.77
APOE4	15907	0.00	2.00	0.52	0.65
FDG	3693	0.57	1.78	1.18	0.17
…	…	…	…	…	…
Years_bl	16222	0.00	16.54	2.70	2.89
Month_bl	16222	0.00	198.07	32.27	34.59
Month	16222	0.00	198.00	32.19	34.58
M	16222	0.00	198.00	32.06	34.50

(2)Treatment of missing values

First of all, the missing values of all indicators are listed in Table 2, and the top 7 indicators with missing values are listed. We found that the missing values of the following indicators exceed 80%, which has lost the significance of analysis, so we should eliminate them.

Table2 The top 7 indicators with missing values.

	count	mean	std	Missing count	proportion
PIB_bl	154	1.582256	0.3021138	16068	99.1
PIB	223	1.783161	0.4225111	15999	98.6
FBB	514	1.174792	0.2460774	15708	96.8
FBB_bl	1048	1.163628	0.2371746	15174	93.5
PTAU	2357	28.1355	14.29394	13865	85.5
TAU	2361	293.4798	129.30829	13861	85.4
AV45	3080	1.193030	0.2308664	13142	81.0

ISSN: 1813-4890

Secondly, the missing values of other indicators are filled up according to the sequential mean method.

(3)Abnormal value processing

Using Mahalanobis distance method, it is concluded that the outliers account for a small proportion, and we reject the samples where the outliers are located.

(4)Index classification

After searching the relevant literature of ADNI data set and searching the browser, and excluding the personal information of participants, we have established the classification of Alzheimer's disease-related indicators selected in the annex, which are divided into five categories: demographics, cognitive assessment, biomarkers, quantification of anatomical structure and others. Table 3 lists the classification of indicators.

Table3 Classification of indicators

Demographics	gender, years of education, marital status, marital status, whether or not engaged in a career in providing health information, race
Cognitive Assessment	CDRSB, FAQ, MMSE, ADAS-cog(ADAS11, ADASS13, ADASSQ4)RAVLT(RAVLT_immediate, RAVLT_learning, RAVLT_forgetting, RAVLT_perc_forgetting), LDELTOTAL, EcogPtTotal(EcogPtMem, EcogPtLang, EcogPtVisspat, EcogPtPlan, EcogPtOrgan, EcogPtDivatt), EcogSPTotal (EcogSPMem, EcogSPLang, EcogSPVisspa, EcogSPPlan, EcogSPOrgan, EcogSPDivatt), MOCA, DIGITSCOR, mPACCdigit, mPACCtrailsB, TRABSCOR, etc
Biomarkers	APOE4, TAU, ABEAT, FDG, PIB, AV45, FBB, PTAU
Anatomical Quantification Values	IMAGEUID, Ventricles, Hippocampus, WholeBrain, Entorhinal, Fusiform, MidTemp , ICV
Other	MRI(FLDSTRENG), M, Month, FSVERSION

3.1.3. Correlation analysis between characteristics and diagnosis of Alzheimer's disease

(1)Brief introduction of Spearman correlation coefficient

Spearman correlation coefficient is a method to calculate rank correlation coefficient. This method calculates product-moment correlation coefficient according to the rank of numerical values.

As for the diagnosis types of Alzheimer's disease that we study, we can regard them as a sort of ranking, with normal (CN) ranking 0, mild cognitive impairment (MCL) ranking 1, and Alzheimer's disease (AD) ranking 2, which can be regarded as a gradually deepening relationship among them. In this way, we can analyze and test the correlation between the characteristic indexes and the diagnosis of Alzheimer's disease according to Spearman's rank correlation coefficient theory.

(2)Spearman rank correlation coefficient principle

The product difference is calculated by using the rank of the variables studied by Spearman, and then the correlation coefficient is obtained. First, rank each index variable and the three diagnoses of Alzheimer's disease to get a new rank, then calculate the rank difference, and

ISSN: 1813-4890

calculate the Spearman correlation coefficient by using the following formula, where and respectively represent the rank of the reordered variable, and n represents the sample size and rank difference.

$$P_S = \frac{\sum_{i=1}^{n}(R_i-\bar{R})(S_i-\bar{S})}{\sqrt{\sum_{i=1}^{n}(R_i-\bar{R})^2 \sum_{i=1}^{n}(S_i-\bar{S})^2}} = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \quad (1)$$

(3)Correlation solution

First of all, All the processed data were imported into SPSS. Firstly, descriptive statistics were made on three diagnoses (DX) of Alzheimer's disease.

Secondly,Using the software's own Spearman correlation solving function, the correlation coefficients between all data indexes and Alzheimer's disease diagnosis are solved and tested. Table 4 lists the top 10 characteristic indicators of relevance.

Table 4 Top 10 Characteristic Indicators of Relevance

DX	ρ	Sig.
CDRSB	0.874	0.000
mPACCdigit	-0.761	0.000
mPACCtrailsB	-0.76	0.000
FAQ	0.756	0.000
ADASQ4	0.691	0.000
LDELTOTAL	-0.669	0.000
MMSE	-0.666	0.000
RAVLT_immediate	-0.654	0.000
RAVLT_learning	-0.523	0.000
EcogSPVisspat	0.502	0.000

It can be seen from the table that CDRSB, a cognitive evaluation index, has the strongest correlation with the diagnosis results of Alzheimer's disease, followed by MPACC Digit, MPACC Trailsb and FAQ, and their absolute correlation coefficients all exceed 0.7, which has a relatively great influence on the diagnosis results. Some biomarkers, such as APOE4, also have effects. In addition to the above cognitive evaluation indicators and biomarkers, there are also some brain structure indicators (quantified by anatomical structure) that also have an impact on the diagnosis of Alzheimer's disease, including Hippocampus, Entorhinal, MidTemp and other indicators.

Finally, we selected brain structure indicators and cognitive level indicators from the top 20 indicators of correlation, and used them to build the model of Question 2, as shown in Table 5.

Table 5 Top 20 indicators of relevance

cognitive behavioral characteristics	brain structural characteristics

ISSN: 1813-4890

CDRSB	
mPACCdigit	
mPACCtrailsB	
FAQ	
ADASQ4	
LDELTOTAL	
MMSE	
RAVLT_immediate	
RAVLT_learning	Hippocampus
EcogSPVisspat	Entorhinal
TRABSCOR	MidTemp
EcogSPDivatt	
MOCA	
EcogPtMem	
EcogPtTotal	
DIGITSCOR	
EcogPtPlan	
EcogPtVisspat	
FDG	

3.2. Model Ⅱ Models and Results

3.2.1. Random forest three classification principle

The forest algorithm includes three steps: decision tree construction, ensemble learning and voting decision. Several decision trees are combined to form a strong learner, thus improving the accuracy of the output results.The construction of random forest is mainly divided into three steps. Figure 2 demonstrates the formation process of random forest.

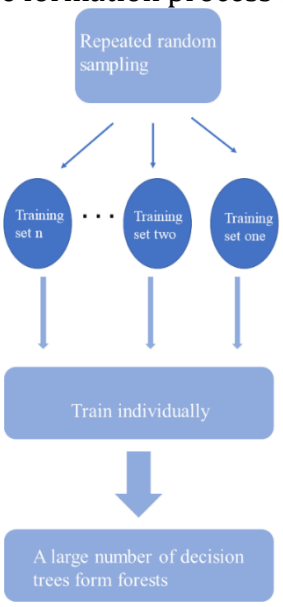

Figure 2 Random forest formation process

(1) Building a decision tree

A decision tree is constructed for each randomly selected training sample, and the decision tree randomly selects the characteristic indicators to split the nodes. Based on CART algorithm, the

ISSN: 1813-4890

data is processed, and Gini index is used as the evaluation index to evaluate the purity, and the better the purity, the better.

(2) Integrated learning

The sampling method based on Bagging is independent repeated sampling, and every sampling sample is selected with equal probability, so the prediction stability is higher. Therefore, we choose Bagging method to sample the sample set without weight.

(3) Voting decision

After repeated training of the training set, each decision tree will predict the original sample set. If a total of N sample sets are constructed, each sample will get N prediction results after each prediction. The majority voting method (more than half of the total votes) is adopted, and the prediction result of this sample is finally selected.

Construction of diagnostic model based on the principle of random forest classification

According to the indexes of brain structural characteristics and cognitive behavior characteristics selected in question one, we divide the diagnosis (DX) of Alzheimer's disease into three categories, namely, normal group (CN), mild cognitive impairment group (MCI) and Alzheimer's disease group (AD). Based on the principle of random forest three classification in 4.2.1, we construct the diagnosis model of Alzheimer's disease. The specific steps are as follows:

(1) Using the selected 22 feature indicators, build a data feature set. Assign the diagnosis result of Alzheimer's disease (DX), CN==1, MCI==2, AD==0, and import the data into python.

(2) Take the imported data as the initial data, and carry out independent repeated sampling on the data for N times with the method of Bagging, so as to obtain the training set sample and the test set sample.

(3) Continuing the second step, using the test machine samples selected in the second step to predict by using the built model to obtain the prediction result of the decision tree, and then obtaining the prediction of each sample according to the majority voting method.

(4) By comparing the predicted results of the test set with the real results, the accuracy of the established diagnostic model can be obtained.

(5) Through many iterations, the contingency of the results is reduced, and the accuracy and stability of the predicted results are improved.

3.2.2. Verification of model diagnosis results

(1)Model diagnosis results

After 20 iterative tests, the prediction accuracy tends to be stable. We choose to exclude outliers and take the average of the remaining results. Finally, it is concluded that the accuracy of this diagnostic model is 99.44%, and the diagnostic accuracy is high. The diagnostic model based on random forest can reach the diagnostic standard.

(2)Diagnostic accuracy

Taking 80% of the data set as the training group and 20% as the test group, the model was tested and evaluated. After 20 iterations, the average diagnostic accuracy of 16 times was taken, and the final diagnostic accuracy was 99.44%.

Figure3 is the confusion matrix heat diagram of one test result, where the horizontal axis represents the predicted value and the vertical axis represents the real value. It can be seen from the figure that all normal group (CN), mild cognitive impairment group (MCI) and Alzheimer's disease group (AD) can be diagnosed 100%, and there is no misdiagnosis.

ISSN: 1813-4890

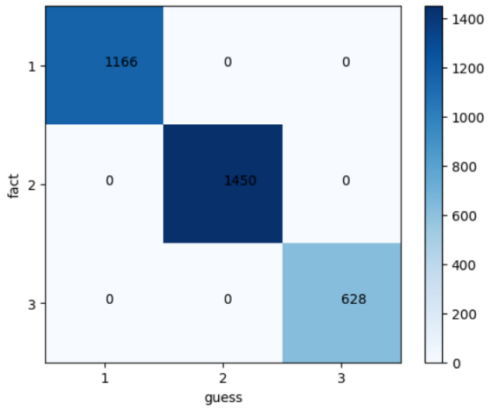

Figure 3 Confused matrix heat diagram

3.3. Model Ⅲ Models and Results

3.3.1. Establishment of XGBoost Classification Model

XGBoost (Extreme Gradient boosting) is a machine learning method based on the integration idea. It adopts the boosting idea and carries out high-precision modeling through multiple learners[3].XGBoost algorithm usually takes decision tree as the learner, and generates new trees by constantly splitting features. The newly generated tree is essentially the residual error between the predicted value and the real value of the previous tree. After N times of training, the learning results of multiple trees are accumulated to obtain the final prediction result[4].

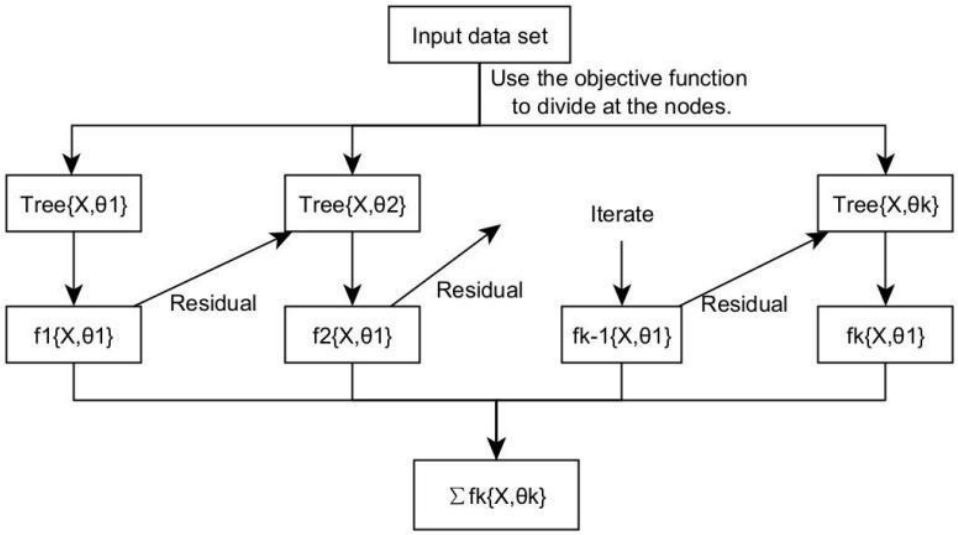

Figure 4 XGBoost algorithm flow chart

As shown in the principle formula (2):

$$\hat{y}_j = \sum_{l}^{L} f_l(x_j), f_l \epsilon F \tag{2}$$

Where $\hat{y}_j$ is the final predicted value of XGBoost model, L indicates the number of trees to be adjusted, f_l is the l tree, x_j is expressed as the jth input sample,F represents the set of all decision trees.The objective function and regularization term adopted by the model are shown in formulas (3) and (4):

ISSN: 1813-4890

$$Obj^{(t)} = \sum_{j=1}^{n} loss\left(y_j, \hat{y}_j^{(t-1)} + f_t(x_j)\right) + \Omega(f_t) + c \tag{3}$$

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{O=1}^{T} w_0^2 \tag{4}$$

Where $Obj^{(t)}$ represents the objective function when building the first tree; $loss(\ )$ represents the loss function,generally the mean square error; $\Omega(f_t)$ is expressed as the regular term of the t-th tree,Determines the depth of the tree to be adjusted (max _ depth);c is a constant term; γ and λ are coefficients of regular terms;T indicates the number of all leaf nodes of a tree; w_0 is the weight of the 0th leaf node in a tree.

Taylor expansion is carried out on formula (3):

$$Obj^{(t)} \approx \sum_{j=1}^{n} \left[loss\left(y_j, \hat{y}_j^{(t-1)}\right) + g_j f_t(x_j) + \frac{1}{2} h_j f_t^2(x_j)\right] + \Omega(f_t) + c \tag{5}$$

$$g_j = \partial_{\hat{y}_j^{(t-1)}} loss\left(y_j, \hat{y}_j^{(t-1)}\right), h_j = \partial^2_{\hat{y}_j^{(t-1)}} loss\left(y_j, \hat{y}_j^{(t-1)}\right) \tag{6}$$

Where the sum of all h_j's is the minimum sample weight sum of leaf nodes to be adjusted (min_child_weight).

Therefore, the selection of parameters such as eta, min_child_weight and max_depth in XGBoost classification model has an important influence on the effect of the model.

3.3.2. Evaluation Index of XGBoost Model

Considering the application of XGBoost model to multi-classification, confusion matrix can be used to evaluate the performance. The real class and the predicted class are divided into real class (TP), true negative class (TN), false positive class (FP) and false negative class (FN)[5].

Table6 Confusion Matrix

Confusion Matrix		Forecast label	
		Positive example	Negative example
Real label	Positive example	TP (True sample number)	FN (Number of false negative samples)
	Negative example	FP (Number of false positive samples)	TN (True negative sample number)

According to the actual situation, Precision, Recall, F1-score and Accuracy are adopted to evaluate the effect of the model:

(1)Precision describes the proportion of positive examples judged to be true to all true examples, as shown in formula (7):

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

ISSN: 1813-4890

(2)Recall describes the proportion of samples judged to be true to the total number of samples, as shown in formula (8):

$$\mathrm{Re}call = \frac{TP}{TP + FN} \tag{8}$$

(3)F1-score is composed of Precision and Recall, as shown in Formula (9):

$$F1\text{-}score = \frac{2 * Precision * \mathrm{Re}call}{Precision + \mathrm{Re}call} \tag{9}$$

(4)Accuracy describes the proportion of correctly classified samples, as shown in formula (10):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

Among them, the value range of each evaluation index is [0,1], and the closer it is to 1, the better the model effect will be.

3.3.3. XGBoost Classification Model Results

Using XGBoost-based machine learning method, the preprocessed data can be divided into three categories: CN, MCI and AD, among which MCI can be further divided into SMC, EMCI and LMCI. Through the comparison of several classification models, it can be found that the classification model based on XGBoost has the best effect.

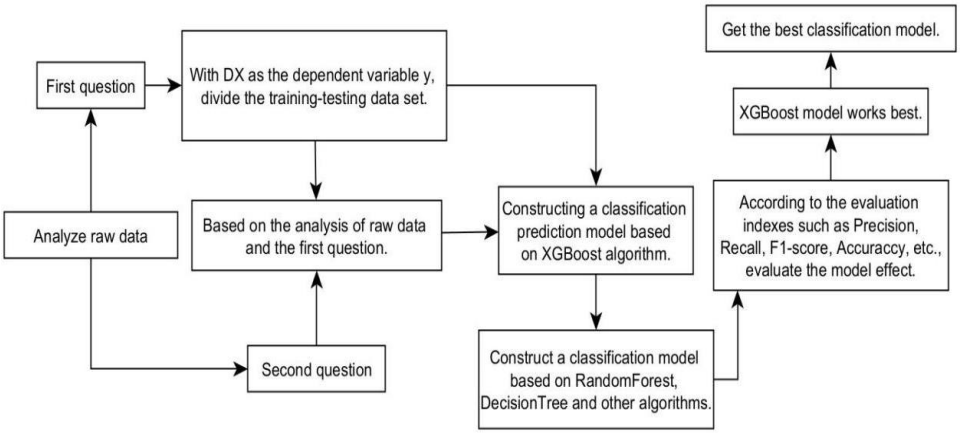

Figure 5 Solution process of question three

3.3.4. Feature Selection

In the feature selection of CN, MCI and AD, after a preliminary analysis of the original attachment data, DX features with three types of labels (CN, MCI and AD, respectively) and rich data are selected as dependent variable Y, and baseline data (initial data with 'bl' suffix) is deleted to avoid feature data redundancy, which is convenient for analysis. All the remaining features except DX are used as independent variables X. Among them:

(1) For the data with missing value of more than 50%, it is considered that it is not valuable for research and it is easy to interfere with the model effect, so all data will be discarded;

(2) The remaining few data are missing, qualitative variables are filled with the mode, and quantitative variables are filled with the mean;

(3) The existing outliers are deleted by 3σ method.

After standardization, there are 13119 sample data and 6059 feature columns (after coding) remaining.

ISSN: 1813-4890

As for the feature selection of MCI continued classification, based on the analysis of the original attachment data and the previous step, DX_bl features with five types of labels (CN, SCM, EMCI, LMCI, AD respectively) and excluding the sample data containing CN and AD are selected as dependent variables Y, and all the remaining features except DX and DX_bl are selected as independent variables X.

3.3.5. Results

In the three categories of CN, MCI and AD, the XGBoost model is divided into 9,183 pieces of training data for training, and the remaining 3,936 pieces of data are tested. It can be found that its effect is the best, with an accuracy rate of 95.12%.

When MCI continues to be subdivided into SCM, EMCI and LMCI, the XGBoost model is divided into 10,495 pieces of training data for training. After testing the remaining 2,660 pieces of data, it can be found that its effect is still the best, with an accuracy rate of 98.02%.

In addition, by comparing the models with precision, recall, f1-score and other evaluation indexes, it is not difficult to find that XGBoost classification model has the best effect no matter which classification.

Table 7 Comparison of models of CN, MCI and AD

Evaluating indicator	precision	recall	f1-score	accuracy
XGBoost	0.9508	0.9512	0.9507	0.9512
Random forest	0.9283	0.9281	0.9254	0.9281
KNN	0.6648	0.6839	0.6667	0.6839
Naive Bayes	0.7720	0.7551	0.7605	0.7551
DecisionTree	0.9196	0.9205	0.9199	0.9205

Table 8 Comparison of models for MCI fine classification

Evaluating indicator	precision	recall	f1-score	accuracy
XGBoost	0.9803	0.9802	0.9802	0.9802
Random forest	0.9337	0.9329	0.9331	0.9329
DecisionTree	0.9527	0.9527	0.9527	0.9527

3.4. Model Ⅳ Models and Results

3.4.1. Thinking analysis

The preprocessed data set contains the diagnosis results of a large number of participants at different time points. The data are jumbled, showing the evolution of all participants' diagnosis over time is very confusing, and it is difficult to get an effective evolution of Alzheimer's disease. Therefore, we first processed the data, and selected the data of participants who had a long time line and participated many times as samples to explore the evolution of Alzheimer's disease.

3.4.2. Data processing

Screening for the first time

(1) screening diagnosis times

Import the preprocessed data into python, and select the participants who have collected information more than 15 times. Through screening, we retained the data of 201 participants.

(2) Deal with the time of each participant.

ISSN: 1813-4890

Each participant takes the time of the first diagnosis as the benchmark, and the rest of the diagnosis time is expressed in years by rounding method relative to the benchmark time. For the data diagnosed in the same year, keep the latest diagnosis data and results, and delete the rest. For the diagnosis results of the missing year, the diagnosis results of the previous year are adopted.

Secondary screening

Using the data selected for the first time to establish the model of Alzheimer's disease with time is not effective, and it still shows the jumbled data. We process the data again.

(1) increase the number of diagnoses.

The accuracy of Alzheimer's disease evolution model increases with the increase of sample size. So we improve the selection criteria of participants. Select the diagnostic data of participants who collected information more than 22 times.

(2) Treatment of participants' diagnosis time

For the treatment of diagnosis time, we choose the diagnosis results from 0 to 16 years, and keep the latest data for the repeated years; For the exceeded years, we will eliminate them; For the missing years, we use the data of the previous year.

(3)Divide the participants into three categories, those who belong to normal (CN) or SMC, EMCI, LMCI or AD, and analyze the evolution of Alzheimer's disease with three initial different symptoms.

3.4.3. Alzheimer's disease evolution model

Alzheimer's Disease Evolution Model Based on the First Screening Data

Quantitative diagnosis results

Table9 Quantitative diagnosis results

Type	Assignment
CN	0
SMC	1
EMCI	2
LMCI	3
AD	4

Set the value of normal (CN) to 0, then take normal as the control, and assign the rest diagnosis results as SMC==1,EMCI==2,LMCI==3,AD==4 in order of severity.

(2) Model evolution diagram

The data were imported into SPSS, and the time series diagram of the evolution of the diagnosis results of 201 participants with time was drawn. It can be seen from the obtained diagram that the established Alzheimer's disease model is not ideal, and it is difficult to see the evolution trend from it. So we process the data again to build a new model.

Improved Alzheimer's disease evolution model

Use the diagnosis results of the selected representative 20 participants and the revised time to build the time series diagram again, as shown in the figure 6:

ISSN: 1813-4890

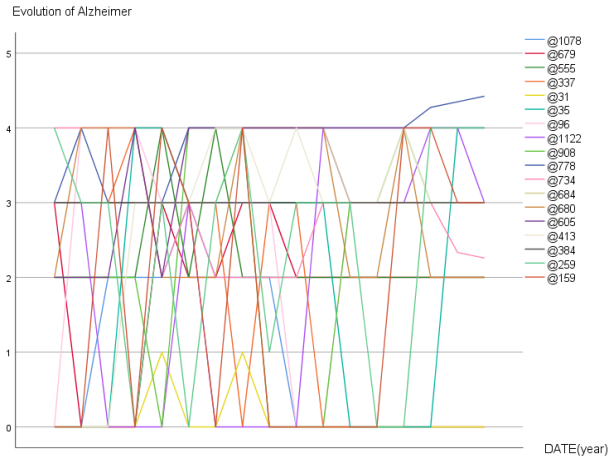

Figure 6 Evolution of Alzheimer's disease over time

A probe into the evolution of Alzheimer's disease

We classified the above participants according to their initial symptoms, and continued to explore the evolution of Alzheimer's disease.

(1)normal (CN) or SMC

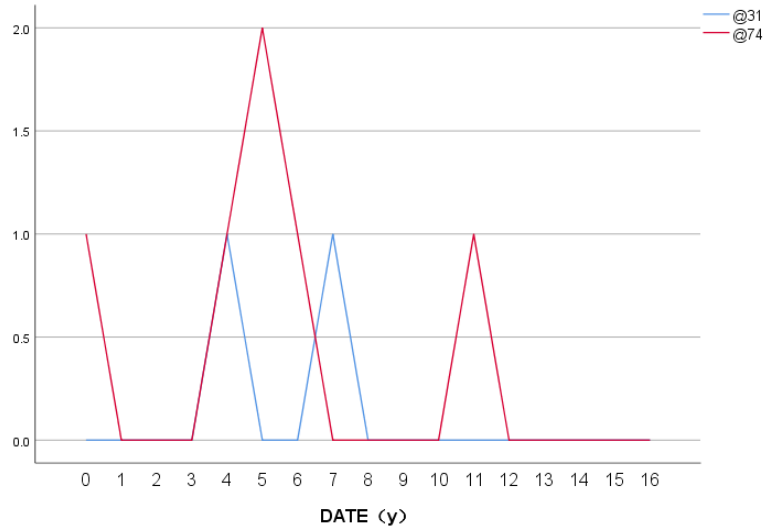

Figure7 CN or SMC

As can be seen from the figure, in CN and SMC groups, the diagnosis results of participants are basically hovering between normal (CN) and (SMC) every time, which indicates that the disease is easier to control when there are no symptoms of Alzheimer's disease or mild symptoms of Alzheimer's disease, and early intervention is more effective than active treatment.

(2)LMCI or AD

In the (LMCI) and (AD) groups, that is, when the number of participants' diagnoses is mostly in LMCI (Level 3) and AD (Level 4), although the diagnosis at different time points will fluctuate and even return to normal. But as time goes on, participants always return to the diagnosis results of AD or LMCI. This shows that Alzheimer's disease is not easy to recover when it develops deeply, and the disease is not easy to be effectively controlled. Therefore, early diagnosis and intervention of Alzheimer's disease are necessary.

ISSN: 1813-4890

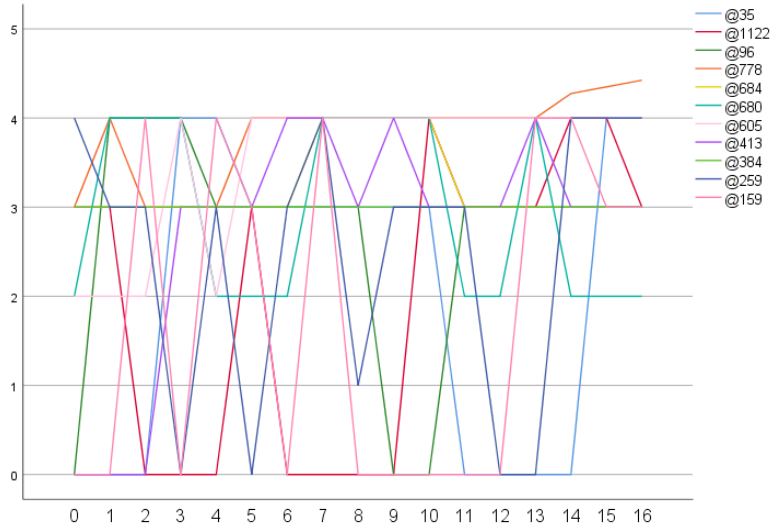

Figure8 LMCI or AD

(3)EMCI

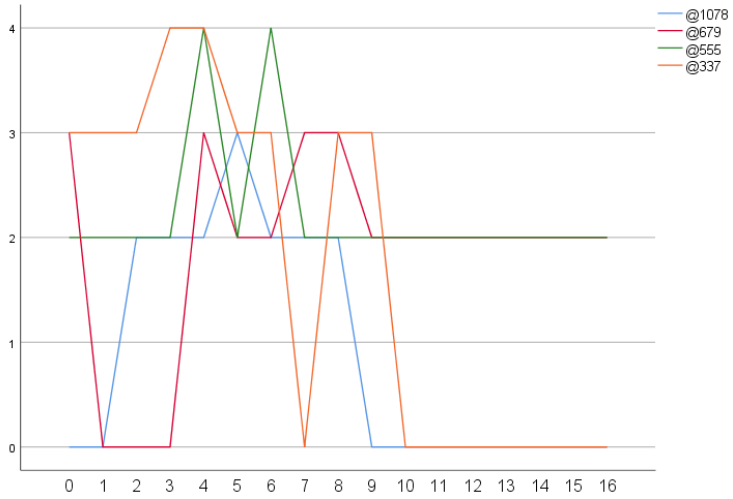

Figure9 EMCI

Compared with the (CN) and (SMC) groups, when the diagnosis times are mostly in EMCI, the evolution of Alzheimer's disease of participants is repetitive, that is, good or bad.

3.5. Model Ⅴ Models and Results

Alzheimer's disease(AD) commonly known as Alzheimer's disease, is a syndrome with acquired cognitive impairment as the core, which leads to a marked decline in patients' daily living ability, learning ability, working ability and social communication ability. Dementia is described as "neurocognitive disorder" in the 5th edition of Diagnostic and Statistical Manual of Mental Disorders (DSM-V) of American Psychiatric Association.

In order to describe the early intervention and diagnostic criteria of five types of patients: CN, SMC, EMCI, LMCI and AD, firstly, a literature retrieval platform based on China HowNet Journal Database (CNKI) was selected, and through the advanced retrieval function, the topics of mild cognitive impairment (MIC), Alzheimer's disease (AD), senile dementia, early intervention and diagnostic criteria were searched. Secondly, by manually eliminating those with weak correlation, 7 literatures can be consulted. Finally, by consulting the remaining literature, we summarized and described the early intervention and diagnostic criteria of five types of patients: CN, SMC, EMCI, LMCI and AD.

ISSN: 1813-4890

3.5.1. Diagnostic Criteria

(1)CN

CN(Normal Cognition) is the old people with normal aging, and there is no diagnostic standard.

(2)SMC

SMC(Subjective memory complaint) refers to the state in which an individual's subjective sensory memory or cognitive function declines or declines, but there is no obvious cognitive dysfunction in objective examination, that is, the middle-aged and elderly people's conscious memory declines or memory impairment without clear reasons[6].

(3)MCI

Mild Cognitive Impairment (MCI) refers to the progressive decline of memory or other cognitive functions, but it does not affect the ability of daily living and does not meet the diagnostic criteria of dementia[7]. According to the different time of this stage, it can be subdivided into Early Mild Cognitive Impairment (EMCI) and Late Mild Cognitive Impairment (LMCI).

Its diagnostic criteria mainly include the following four points [8]:

① Patients or insiders report, or experienced clinicians find cognitive impairment;

② There is objective evidence (from cognitive test) that one or more cognitive functional domains are damaged;

③ Complex instrumental daily living ability can be slightly damaged, but independent daily living ability can be maintained;

④ The diagnosis of dementia has not been reached.

However, there is no unified diagnostic standard for MCI at present. MCI caused by different causes should be diagnosed flexibly according to the actual situation.

(4)AD

For patients who had normal intelligence in the past, and then had acquired cognitive decline or abnormal mental behavior, which affected their working ability or daily life, and could not be explained by delirium or other mental diseases, they could be diagnosed as Alzheimer's disease[9]. Cognitive or psychobehavioral impairment can be objectively confirmed by medical history collection or neuropsychological evaluation, and it has at least two of the following five items:

① Impaired memory and learning ability;

② Impairment of executive functions such as reasoning, judging and handling complex tasks;

③ Impaired visual space ability;

④ impaired language function (listening, speaking, reading and writing);

⑤ Changes in personality, behavior or manners.

There are two international diagnostic criteria for dementia: the 10th edition of the International Classification of Diseases (10th edition) of the World Health Organization, ICD-10) and the Diagnostic and Statistical Manual of Mental Disorders (4th Edition, Revised, DSM-Ⅳ-R IV-R) of American Psychiatric Association.

3.5.2. Early Intervention

At present, a number of drug intervention studies aiming at the early stage of AD have been launched internationally. Although there are no drugs that can change the course of AD at present, it has been found that, besides the non-intervention factors such as age, sex, genetic factors and family history, the risk factors that can be intervened by human beings such as cardiovascular and cerebrovascular diseases, blood pressure, blood lipids, type 2 diabetes, smoking and drinking, diet, education level, physical activity and mental activity are also closely related to Alzheimer's disease[10]. Therefore, strengthening the control of these risk factors in

ISSN: 1813-4890

the early stage (SMC, MCI) will hopefully reduce the risk of AD and delay the development of AD[11].

Sensitivity Analysis

(1) The diagnosis model of Alzheimer's disease based on the principle of random forest three classifications was tested by bringing in different groups of data, and after 20 iterations, the stable accuracy rate was finally obtained, and the model effect was good.

(2) When initially clustering Alzheimer's disease based on XGBoot algorithm, after many tests, the final accuracy rate is steadily distributed at 95.12%; When MCI is subdivided into SCM, EMCI and LMCI clusters, the accuracy rate is steadily distributed at 98.02% after many tests, and both clusters achieve good results.

4. Strengths and Weakness

4.1. Strengths

(1) The correlation analysis with Spearman rank correlation coefficient does not need to consider the data form, that is, the data requirements are not strict.

(2) The diagnostic model based on random forest algorithm for classified data is very stable, and the diagnostic accuracy is very high.

(3) The regularization of 3)CART algorithm prevents over-fitting, and the loss is more accurate and the precision is high.

(4) When CART is used as the base classifier, XGBoost explicitly adds the regularization term to control the complexity of the model, which helps to prevent over-fitting, thus improving the generalization ability of the model.

(5)XGBoost adopts a strategy similar to random forest, which supports data sampling.

(6) Compared with the traditional GBDT, XG Boost can automatically learn the processing strategy of missing values.

4.2. Weakness

(1) The accuracy of Spearman rank correlation coefficient is lower than that obtained by product-difference relation.

(2) If the noise problem in the data can't be eliminated, the diagnostic accuracy of random forest will be affected.

(3) The model can be processed more accurately.

(4)XGBoosting adopts pre-sorting. Before iteration, the features of nodes are pre-sorted, and the optimal segmentation point is selected by traversal. When the data volume is large, the greedy algorithm is time-consuming and complex.

5. Conclusion

Firstly, this paper explores the correlation between brain structure characteristics, cognitive behavior characteristics and Alzheimer's disease, and then establishes a diagnosis model of Alzheimer's disease based on random forest algorithm. Then, the classification model is constructed by XGBoost algorithm, and then the evolution pattern of Alzheimer's disease with time is explored. Finally, the intervention table of different development stages of Alzheimer's disease is established by consulting relevant literature.

In question one, Spearman correlation coefficient explores the influence of brain structure and cognitive behavior characteristics on the diagnosis of Alzheimer's disease, and obtains the top 22 brain structure and cognitive behavior characteristics that have great influence on it.

ISSN: 1813-4890

Secondly, a diagnosis model of Alzheimer's disease based on brain structure characteristics and cognitive behavior characteristics is constructed by using the random forest three classification principle, and the diagnosis accuracy rate reaches 99.44%.

In the third, firstly, we build random forest, XGBoost, K-nearest neighbor, naive Bayes and decision tree classification models to classify Alzheimer's disease into three categories: CN, MCI and AD, and compare their accuracy rates. Finally, we conclude that XGBoost model has the best effect, and its accuracy rate reaches 95.12%. Secondly, based on the comparison of the previous model effects, XGBoost, random forest and decision tree classification models are constructed to further subdivide MCI, among which XGBoost model is still the best, with an accuracy rate of 98.02%.

In the fourth , we processed the data twice, and selected the diagnosis results and diagnosis time of the participants who participated in the data collection more times and participated in the long time line as samples. According to the proportion of CN, SMC, EMCI, LMCI and AD in the diagnosis results, they are clustered into three groups. On this basis, the overall evolution model of Aarmo's Alzheimer's disease and the evolution model of Aarmo's Alzheimer's disease with different participants were established respectively. According to the models, it is necessary to make early diagnosis and treatment.

In the fifth , we first selected the literature retrieval platform based on China HowNet Journal Database (CNKI), and searched the required literature with the keywords of mild cognitive impairment (MIC), Alzheimer's disease (AD), Alzheimer's disease, early intervention and diagnostic criteria through advanced retrieval functions. Secondly, by selecting the latest literature and manually rejecting the literature with weak relevance, 7 literatures are available for reference. Finally, by consulting the selected literature, we summarize and describe the early intervention and diagnostic criteria of five types of patients: CN, SMC, EMCI, LMCI and AD.

References

[1] People's Republic of China (PRC) National Health and Wellness Committee,Letter on Reply to Proposal No.4257 (Medical Sports No.422) of the First Session of the 13th National Committee of the Chinese People's Political Consultative Conference[EB/OL], http://www. nhc.gov.cn /wjw/tia/201812/ c176549541b043038dfec4567ac7dd14.shtml,(2022-11-18).

[2] People's Republic of China (PRC) National Health and Wellness Committee,Letter on Reply to Proposal No.4423 (Social Management No.362) of the Fourth Session of the 13th National Committee of the Chinese People's Political Consultative Conference[EB/OL], http://www.nhc. gov.cn/ wjw/tia/202112/fb2a72801872408c8b06a6ab429675bb.shtml,(2022-11-18).

[3] K Wang, S Jiang, H Zhang, Z Chao. Regression-classification-regression life prediction model based on XGBoost[J/OL].ChinaMeasurement:1-8[2022-11-19].http://kns.cnki.net/kcms/ detail/51.1714. TB.20220616.1537.002.html

[4] G Xu, Y Shen. Multi-classification detection method of malicious programs based on XGBoost and Stacking fusion model[J].Network Security,2021,21(06):52-62.

[5] W Zhang,D Liu,X Jia. Three-category coupon prediction method based on XGBoost[J].Journal of Nanjing University of Aeronautics and Astronautics,2019,51(05):643-651.DOI:10.16356/j.1005-2615.2019.05.009.

[6] China Writing Group of Diagnosis and Treatment Guidelines for Dementia and Cognitive Impairment, Cognitive Impairment Committee of Neurophysicians Branch of Chinese Medical Association.2018 Guidelines for Diagnosis and Treatment of Dementia and Cognitive Impairment in China (V): Diagnosis and Treatment of Mild Cognitive Impairment[J]. Journal of Internal Medicine,2018,98(17):1294-1301.

[7] Petersen R C. Mild cognitive impairment[J]. New England Journal of Medicine, 2011, 364(23): 2227-2234.

ISSN: 1813-4890

[8] Winblad B, Palmer K, Kivipelto M, et al. Mild cognitive impairment–beyond controversies, towards a consensus: report of the International Working Group on Mild Cognitive Impairment[J]. Journal of Internal Medicine, 2004, 256(3): 240-246.

[9] China Dementia and Cognitive Impairment Guide Writing Group, Cognitive Impairment Committee of Neurophysicians Branch of Chinese Medical Association.2018 Guidelines for Diagnosis and Treatment of Dementia and Cognitive Impairment in China (I): Dementia and its Classification Diagnostic Criteria[J].National Medical Journal of China,2018,98(13):965-970.

[10] China Writing Group of Diagnosis and Treatment Guidelines for Dementia and Cognitive Impairment, Cognitive Impairment Committee of Neurophysicians Branch of Chinese Medical Association. 2018 Guidelines for Diagnosis and Treatment of Dementia and Cognitive Impairment in China (VII): Risk Factors and Intervention of Alzheimer's Disease[J]. Journal of Internal Medicine, 2018, 98(1):1461-1466.

[11] China Writing Group of Diagnosis and Treatment Guidelines for Dementia and Cognitive Impairment, Cognitive Impairment Committee of Neurophysicians Branch of Chinese Medical Association. 2018 Guidelines for Diagnosis and Treatment of Dementia and Cognitive Impairment in China (VI): Pre-dementia Stage of Alzheimer's Disease[J]. Journal of Internal Medicine, 2018(1):1457-1460