Problem Chosen
C

2022
ShuWei Cup
Summary Sheet

Team Control Number
2022110513494

summary

Alzheimer's disease (AD) is one of the leading causes of death in older adults. Its early stage is mild cognitive impairment (MCI), which is mainly manifested as memory loss, judgment decline, etc. When it develops to the AD stage, it is manifested as severe memory impairment, emotional irritability, etc., and even daily life cannot take care of itself. The disease is serious, not only bringing great suffering to patients, but also bringing great burden to family and social medical care. At present, there is no cure for AD, and prevention and early intervention of the disease have become the main means and goals to overcome the disease. Therefore, if patients with MCI can be identified at an early stage and whether MCI patients will be further transformed into the AD stage, it will help patients receive treatment as soon as possible and improve the diagnosis and treatment effect.

The main contents of this article are as follows:

For questions 1, data preprocessing is carried out, and missing values and outliers are processed; Establish feature equations and extract feature vectors; High-dimensional feature vector dimensionality reduction processing; 21 feature vectors were extracted by combining Spearman correlation analysis, random forest and ANOVA.

For questions 2, combined with the first problem, an AD intelligent diagnosis algorithm based on the KSVM model and LSTM model of KPCA-FDA is established, and the model is well equipped.

For questions 3, Using the preprocessed data, the CN, AD, MCI; SMC, EMCI, DMCI, these two groups of data are classified in three categories, and the traditional machine learning and deep learning-based classification methods are used to solve the third problem.

For questions 4 and 5, the data were divided into five categories: CN, AD, SMC, EMCI and LMCI based on the first visit to study the relationship between disease and time. 6 of the 21 feature vectors in question 1 were selected to analyze the relationship between features and diseases. Literature was reviewed to give diagnostic criteria and early interventions.

Key words: AD; Intelligent diagnosis; KSVM; Machine learning; LSTM

Content

1. Introduction

1.1 Background

Alzheimer's Disease (AD) is a kind of disease with hidden onset and slow deterioration of brain function, which is related to some neuropathy and neurochemical changes in the brain. The brain cells of patients with this disease will be rapidly aging and degenerated in a pathological way, and the brain function will gradually decline, leading to cognitive function decline and memory impairment, and various mental symptoms[1]. When the disease is serious, memory will be seriously lost, and daily life will not be able to take care of itself, leading to coma. AD is the main cause of senile dementia. With the aging of the world population, the incidence rate of senile dementia is also rising rapidly. According to the survey report, there are about 50 million people suffering from Alzheimer's disease in the world, and the number of patients is estimated to double every 20 years. It is estimated that AD patients will reach 152 million by 2050, that is, about one person in every 64 elderly people suffers from Alzheimer's disease on average, and almost one person changes to Alzheimer's disease every 3 seconds.

Alzheimer's disease not only brings great pain to patients themselves, but also brings great economic burden to families and society. A patient with Alzheimer's disease will spend no less than 400000 yuan in 10 years[2]. The annual medical diagnosis cost of AD in the world is up to 604 billion US dollars, which is huge and has exceeded the cost of heart disease, cancer and other diseases. According to expert research, the total social and economic burden caused by Alzheimer's disease in China reached US $167.74 billion in 2015 and is expected to reach US $2.54 trillion by 2030. Unfortunately, there is no specific drug that can cure Alzheimer's disease or effectively reverse the disease process. However, for patients at the early stage, the combination of drug, non drug treatment and careful care can reduce symptoms and delay the development of the disease, so prevention becomes the most important measure. In view of this, finding a method that can improve the early diagnosis effect of Alzheimer's disease, to find early and potential AD patients, and to specifically control their disease is of great significance for clinical, patients, patients' families, and social and economic development[3].

Mild cognitive impairment (MCI) is a state between normal aging and Alzheimer's disease, and is also considered as the early stage of AD. Compared with the normal elderly of the same age, the patients at this stage have a certain degree of cognitive decline, but have not reached the ability to seriously affect daily life. The development of cognitive dysfunction is a continuous process, which may have begun decades before the clinical symptoms appear. Patients diagnosed with MCI have a very high probability of conversion to AD. The research shows that AD conversion rate is 6%~25% every year, and 30%~50% of patients will be converted to AD within 5 years[4]. The conversion time varies from 6 months to 36 months, usually about 18 months. In medicine, those MCI patients who convert to AD within a specified time range are called progressive MCI (pMCI), and those who do not convert to AD are called stable MCI (sMCI). If we can diagnose pMCI patients and sMCI patients more efficiently, we can take different treatment measures or interventions as soon as possible to improve the diagnosis and treatment effect.

Structural magnetic resonance imaging (MRI) provides an effective method to improve our understanding and evaluation of AD brain changes (for example, regional or global brain atrophy caused by axonal degeneration and cell death), and plays an important role in routine clinical practice. It is also considered as an important biomarker of AD progress.

1.2 Work

Question 1: Preprocess the characteristic indicators of the data in the annex to study the correlation between the data characteristics and the diagnosis of Alzheimer's disease.

Question 2: Design an intelligent diagnosis of Alzheimer's disease using the brain structure and cognitive behavior characteristics in the appendix.

Question 3: First cluster CN, MCI and AD into three categories, and then subdivide the three subcategories (SMC, EMCI and LMCI) contained in MCI into three subcategories.

Question 4: The same sample in the attachment contains characteristics at different time points, and its relationship with time points is analyzed to reveal the evolution law of different types of diseases over time.

Question 5: Describe the early intervention and diagnostic criteria of CN, SMC, EMCI, LMCI and AD.

2. Problem analysis

2.1 Analysis of question one

Question 1 requires studying the correlation between data characteristics and AD diagnosis. First, the attachment data is cleaned, which is mainly divided into three steps: First, remove the variables with a large number of missing values; Secondly, feature engineering is established as an alternative for feature extraction and feature selection; Third, the random forest algorithm is used to fill the missing values, and the 3σ criteria and median interpolation are used to deal with outliers. Then, the high-dimensional feature vectors are reduced, and spearman correlation analysis, random forest algorithm and variance analysis are proposed to extract more important feature indicators.

2.2 Analysis of question two

Question 2 requires the design of intelligent diagnosis algorithm for AD. In this paper, we will use the feature indicators proposed in question 1 as the model input. And use the KPCA-FDA-KSVM model and the LSTM model to design intelligent diagnosis algorithm to effectively diagnose Alzheimer's disease and mild cognitive impairment.

2.3 Analysis of question three

For question 3, the data sets of Alzheimer's disease classification adopted in this chapter are: (1) CN, AD, MCI; (2) SMC, EMCI and DMCI classify the data of these two groups of categories into three categories. In CN, AD and MCI, the category label of CN is 1, that of AD is 2, and that of MCI is 3; In SMC, EMCI and DMCI, the category label of SMC is 3, that of EMCI is 4, and that of LMCI is 5. This chapter will adopt the classification method of Alzheimer's disease based on traditional machine learning and depth learning.

2.4 Analysis of question four

For question 4, the law of disease change can be seen in two ways: changes in th e same disease and changes in disease characteristics (variables) over the same period. Therefore, the data set is classified: according to the first method, the data is divided into five categories: CN, SMC, EMCI, LMCI, and AD, taking the initial diagnosis as the standard; According to the second method, the same variable of the five types of diseases is used as a data set, the relationship between the variables and the change of the disease is analyzed, and the evolution of the disease over time is sought.

3. Fundamental assumptions

1. This paper assumes that the data given by the title is true and effective, and has research value

2.The model selects the correct variables.

3. The model selects the correct functional form.

4.The data we have collected are completely true.

4. Data preprocessing

4.1 Data observation and preliminary analysis

AD intelligent diagnosis and classification is to learn rules from data through various machine learning and deep learning algorithms to judge whether the subject is ill. Therefore, data is crucial for AD intelligent diagnosis and classification. ADNI (Alzheimer's Disease Neuroimaging Initiative) [5] was founded by Dr. Michael W. Weiner in 2004. ADNI provides a large number of horizontal and vertical tracking research samples of AD related diseases, which provides important data support for the research of early diagnosis methods of AD. At present, the ADNI dataset is divided into four subsets: ADNI1, ADNIGO, ADNI2, and ADNI3.

The experimental data in this paper are derived from the ADNIMERGE_New data published in Question C of the 2022 "ShuWei Cup" International College Student Modeling Challenge, and the data shape is (16222, 116). The attached data contain specific information characteristics of 4850 cognitive normal elderly (CN), 1416 patients with subjective memory complaint (SMC), 2968 patients with early mild cognitive impairment (EMCI), 5236 patients with late mild cognitive impairment (LMCI) and 1738 patients with Alzheimer's disease (AD) collected at different time points (one time point is a quantity). Table 4.1 gives the specific information of the MRI images of the ADNI dataset given in this paper.

Table 4.1 MRI Information Statistics of ADNI Dataset

	ADNI1		ADNIGO		ADNI2		ADNI3
Collection year	2005-2022		2006-2022		2005-2022		2007-2022
MR protocol	1.5T	3T	1.5T	3T	1.5T	3T	-
Data Set Composition	1049CN 999AD 1478MCI	2CN 2MCI	65CN 44AD 48MCI	7CN 2CN 204MCI	99CN 66AD 74MCI	878CN 508AD 1477MCI	985CN 311AD 1113MCI

MRI brain scan is divided into 1.5T MRI scan and 3.0T MRI scan. Due to the different configuration environments of different MRI equipment manufacturers, the image preprocessing process provided by ADNI is also different. Therefore, it is necessary to select the characteristic data collected by the same MRI equipment, and select the data set with less missing information.

It can be seen from Table 4.1 that ADNI1's MRI data almost all use 1.5T MR protocol, and ADNI2's MRI data mostly use 3.0T MR protocol. Combined with the missing values in the given data set, this paper mainly uses COLPROT (Study protocol of data collection), ORIGPROT (Original study protocol), all of which are ADNI2 and the protocol is 3.0TMR to study the intelligent diagnosis and classification of AD, namely, the attachment data ADNI2_ 3.0T is taken as sample data.

In this paper, 2400 subject characteristic data were extracted from the subject data set. The data shape is (2400, 56). The distribution is shown in Table 4.2.

Table 4.2 Demographic information of subjects

Type	Number	Age	gender(M/F)
CN	781	55-89	380/401
MCI	1154	55-91.4	621/533
AD	465	55-90.3	260/205

4.2 Characteristic engineering

（1） Feature extraction

There are a large number of missing values in the data given by the title, which will seriously affect the accuracy of the model, so it is necessary to conduct preliminary screening and elimination. The data set with a data shape of (2400, 49) is obtained after removing duplicate rows and columns and characteristic indicators with missing values exceeding 50%.

At present, the research methods of AD early diagnosis are mainly divided into two categories. The first is the traditional machine learning methods, which are typically represented by support vector machines (SVM) and random forests; The second is based on deep learning methods, such as RNN, CNN, etc. Both of these methods depend on feature engineering, and good feature tuning can achieve high accuracy. Therefore, this paper constructs a high-dimensional feature collection based on the following principles to provide sufficient alternatives for subsequent feature extraction and feature selection, as follows.

(1) Brain structure: construction includes IMAGEUID, Ventures, Hippocampus, WholeBrain, Entrhinal, Fusiform, MidTemp, ICV and other brain structure features.

(2) Cognitive behavior: construction includes RAVLT_ learning 、RAVLT_ forgetting 、RAVLT_ perc_ Cognitive behavior characteristics such as targeting, DIGITSCOR, TRABSCOR, FAQ, etc.

The 46 dimensional feature vectors constructed in this paper, including brain structure features and cognitive behavior features, are shown in Table 4.3, providing feature sets for subsequent feature screening.

Table 4.3 Partial feature set

Influencing factors	Features
Brain Structural Features	IMAGEUID Ventricles Hippocampus WholeBrain Entorhinal …
Cognitive Behavioral	RAVLT_learning RAVLT_forgetting RAVLT_perc_forgetting DIGITSCOR TRABSCOR …

（2） Data quantification processing

The data provided by the title of this paper is of text nature and is not suitable for model input. Therefore, it is necessary to quantify the character indicators in text form. The quantization rules are shown in Table 4.4 - Table 4.9, and the results are shown in the appendix table Quantified data.csv.

Table4.4 PTGENDER

Male	Female
1	2

Table4.5 PTETHCAT

Unknown	Hisp/Latino	Not Hisp/Latino
0	1	2

Table4.6 PTMARRY

Unknown	Married	Divorced	Widowed	Never married
0	1	2	3	4

Table4.7 DX

CN	AD	SMC	EMCI	LMCI
1	2	3	4	5

Table4.8 VISCODE

bl	m06	m12	m24	m36	m48	m60
0	6	12	24	36	48	60

Table4.9 PTRACCAT

Unknown	White	Black	Asian	Am Indian/Alaskan	Hawaiian/Other PI	More than one
0	1	2	3	4	5	6

4.3 Missing value processing

There are three ways to deal with missing data: direct elimination, data interpolation and no processing. Directly removing data is to directly use software to delete data in the process of data analysis. However, this method is applicable to cases where the sample data is large and the missing value accounts for a small proportion. When the data sample is small or the missing value accounts for a large proportion, the accuracy of the prediction results will be seriously affected. Due to the large proportion of missing values in this paper, the random forest algorithm is used to preprocess the missing values (see the appendix Table Missing_value processed data. csv for the results). The algorithm process is as follows:

Step1. The columns with missing values are changed from small to large according to the number of missing values;

Step2. Enter the for loop to fill in the blank value;

Step3. Iterate by the number of null values, sorted from small to large;

Step4. Take the column with null value and store it in fillc as the y variable, except for the column as the x variable;

Step5. The non zero in the fillc column is taken as y_ train, 0 is taken as y_ test；

Step6. Rows not 0 in the fillc column are treated as x_ train, the line of 0 is used as x_ test；

Step7. Apply the trained model to y_ test.

4.4 Analysis and handling of abnormal values

During data information collection, it may be affected by unexpected factors, which may cause some data to deviate from the true value and have abnormal valley peak fluctuations. Such data points are called abnormal points. If they are not properly corrected, they will affect the model's learning of AD laws. Therefore, this paper adopts the 3σ criterion (Formula 4.1)

$$|x-\bar{x}|>3\sigma \tag{4.1}$$

The data set is processed with outliers. The values that meet the 3σ criteria account for less than 1% of the total sample points, which can be judged as outliers, and filled with the median (see the appendix table Outlier processed data. csv for the processing results).

4.5 Data normalization

In the research of AD intelligent diagnosis, its input data usually contains different types of data, representing different feature vectors that have an impact on the diagnosis results. For example, the ICV value in this paper fluctuates between 1072880 and 2108790, while the MSSE range is between 4 and 30. In order to avoid the problem of different degrees of influence caused by different dimensions, it is necessary to normalize the feature vectors of each dimension to obtain the feature vectors of different dimensions of the same dimension. To ensure that data with a large numerical range will not submerge data with a small numerical range.

Through normalization, data sets are standardized, that is, data of different dimensions are transformed into data sets within a unified range. Generally, the normalized data range is between [0, 1] and [— 1, 1]. In this paper, we uniformly normalize the data to the range of [0,1]. The calculation formula is shown in Formula (4.2).

$$x_i^* = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \tag{4.2}$$

Where, x_i^* and x_i are the values before and after normalization, $x_{\min}$ is the minimum value of each characteristic sample data, and $x_{\max}$ is the maximum value of each characteristic sample data.

It should be pointed out here that to ensure the consistency of neural network training and prediction, all data, including sample set and target set, adopt the same normalization standard.

5. Model establishment and solution of question 1

5.1 Feature selection and feature extraction

In AD related CAD systems, we often encounter the problem of "dimension disaster". At this time, it is very important to reduce feature dimensions and improve the efficiency of data analysis. Both feature extraction and feature selection can filter redundant and irrelevant features, which can not only reduce computational complexity, but also select features sensitive to AD based on different modes. Feature selection refers to selecting the optimal feature subset from a set of original feature data sets to achieve the purpose of dimension reduction. It is an inclusive relationship and does not change the original feature space.

5.1.1 Spearman correlation analysis

Spearman correlation coefficient is proposed by Charles Edward Spearman, a British statistician, and is a non-parametric indicator to measure the dependence between two variables.

Spearman correlation coefficient is defined as Pearson correlation coefficient between hierarchical variables. For samples with a sample size of n, n raw data are converted into grade data, and the correlation coefficient ρ For:

$$\rho_{X,Y}=\frac{\sum_i (x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_i (x_i-\overline{x})^2 \sum_i (y_i-\overline{y})^2}} \tag{5.1}$$

In practical application, the connection between variables is insignificant, so it can be calculated by simple steps ρ. The difference between the grades of the two observed variables, then ρ For:

$$\rho=1-\frac{6\sum d_i^2}{n(n^2-1)} \tag{5.2}$$

The correlation strength of variables is represented by the absolute value of the correlation coefficient. The larger the absolute value is, the stronger the correlation is. On the contrary, the smaller the absolute value is, the weaker the correlation is. The specific coefficient range and correlation degree are shown in Table 5.1.

Table 5.1 Range of correlation coefficient

Coefficient range	0.0~0.1	0.1~0.3	0.3~0.8	0.8~1.0
Degree of relevance	No correlation	Weak correlation	Moderate Correlation	Strong correlation

According to the above spearman correlation analysis method, the correlation analysis between 46 characteristic indicators and AD is carried out, and the results are shown in Figure 5.1.

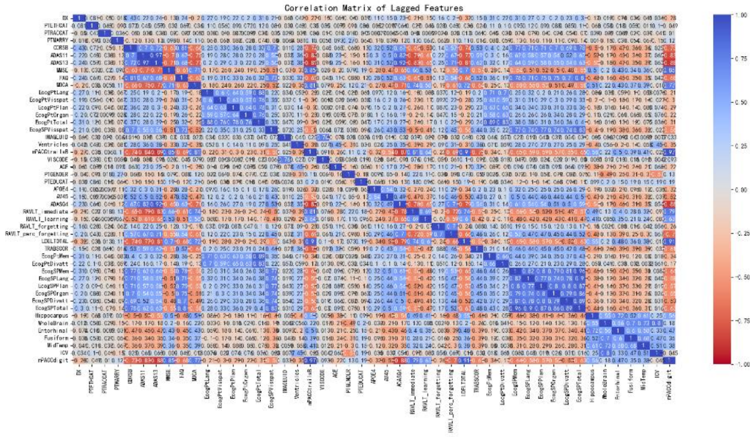

Figure 5.1 Thermodynamic Correlation Diagram of Characteristic Indexes

5.1.2 Variable importance measurement based on random forest

Stochastic forest algorithm inherits and improves the traditional decision tree, can analyze complex and interactive features, and has fast learning speed and high robustness when dealing with data with missing values. In addition, as an important feature of the random forest algorithm, the variable importance measure can be used to select features of high-dimensional data. In essence, the random forest algorithm used in this paper is a combined classifier consisting of multiple classification regression trees. Several decision trees are constructed using bootstrap technology and node random splitting technology. In the process of random resampling, some unselected samples are called out of bag (OOB) data. Using OOB data to evaluate the random forest model can get the OOB error, so as to get the importance of each feature vector. The principle can be understood as: when the independent variable of the OOB data is slightly disturbed, the larger the increase of the OOB error, the more important the variable is. Therefore, OOB error can be used to quantitatively evaluate the importance of feature vectors, and then select high-dimensional feature data.

The process (pseudo code) of random forest algorithm to select variables is as follows:

(1) Suppose there are k trees in the random forest algorithm, and each eigenvector x1, x2,..., xn, for i=1: k

① For each tree, a certain size of data is randomly sampled from dataset N by random resampling to form a sample training subset Ni, and b OOB data are composed of unexampled data.

② Repeat step a) - c) in Ni, and make the decision tree grow to the maximum extent in each cycle. Do not prune it to get the decision tree Ti.

a) Suppose there are M characteristic attributes input, and m attributes are randomly selected as the attribute set of the current decision tree split.

b) Select the best variable j and segmentation point s from m eigenvectors θ i (j, s)。

c) Set the node according to θ I (j, s) is divided into 2 sub nodes.

end for

(2) When generating k decision trees to form a random forest, vote on the b OOB data corresponding to Ti of each decision tree, so that the voting score of each sample in the OOB data is

$$s_1, s_2, \cdots, s_b \tag{5.4}$$

(3) Randomly change the value of each characteristic vector xi in the OOB data sample to generate a new OOB data test sample, and vote on the new OOB data through the random forest to obtain

$$\begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1b} \\ s_{21} & s_{22} & \cdots & s_{2b} \\ \vdots & \vdots & \ddots & \vdots \\ s_{k1} & s_{k2} & \cdots & s_{kb} \end{bmatrix} \tag{5.5}$$

(4) Calculate the importance score of feature vector xi

$$c_i = \sum_{j=1}^{b} \frac{s_j - s_{ij}}{b} \tag{5.6}$$

In Formula (5.6), sj and sij represent the OOB error rate of the ith tree before and after the change of variables respectively; Ci represents the contribution of each feature vector to the classification process, which can measure the independent classification ability of each feature, so it helps to determine the importance of each feature in the classification process and provides a basis for feature selection.

The results of importance evaluation of random forest algorithm are shown in Table 5.2, and the training accuracy of the model is 0.9206.

Table 5.2 Partial Analysis Results of Random Forest Importance

Importance ranking	Features	Importance ranking	Features
1	CDRSB	5	mPACCdigit
2	IMAGEUID	6	AGE

3	VISCODE	7	AV45
4	LDELTOTAL	…	…

5.2 Multivariate ANOVA

ANOVA, also known as "F test", was proposed by Sir Ronald Fisher to test the significance of mean difference between two or more samples[6]. The original assumption was that the factor was not significant.

Multivariate ANOVA is used to study whether two or more control variables have significant effects on the observed variables. Multi factor ANOVA can not only analyze the independent influence of multiple factors on the observed variables, but also analyze whether the interaction of multiple control factors can have a significant impact on the distribution of the observed variables, and finally find the optimal combination of the observed variables.

The results of ANOVA are shown in Table 5.4. If the p value is less than 0.05, the original hypothesis will be rejected, indicating that the characteristics have a significant correlation with AD diagnosis. Table 5.5 shows 21 characteristic indexes that have significant correlation with AD diagnosis.

Table 5.4 Results of ANOVA

	df	sum_sq	mean_sq	F	PR(>F)
PTETHCAT	1	13.9559	13.9559	24.49504	7.98E-07
PTRACCAT	1	1.536474	1.536474	2.696781	0.100685
PTMARRY	1	0.444006	0.444006	0.779309	0.377443
CDRSB	1	58.93523	58.93523	103.4416	8.27E-24
ADAS11	1	7.78E-07	7.78E-07	1.37E-06	0.999068
…	…	…	…	…	…
Entorhinal	1	1.188761	1.188761	2.086484	0.148741
Fusiform	1	0.051316	0.051316	0.090069	0.764116
MidTemp	1	0.274324	0.274324	0.481486	0.487819
ICV	1	0.227634	0.227634	0.399537	0.52739
mPACCdigit	1	2.399376	2.399376	4.211324	0.040265
Residual	2353	1340.607	0.569744	-	-

Table 5.5 Characteristic indicators with significant correlation in AD diagnosis

Features	PR(>F)	Features	PR(>F)	Features	PR(>F)
PTETHCAT	7.98E-07	EcogPtOrgan	0.020134	RAVLT_immediate	1.97E-05
CDRSB	8.27E-24	EcogPtTotal	1.17E-13	RAVLT_forgetting	0.003114
ADAS13	6.20E-23	IMAGEUID	2.47E-05	RAVLT_perc_forgetting	0.026585
MMSE	3.84E-13	mPACCtrailsB	1.55E-18	LDELTOTAL	2.57E-25
FAQ	2.04E-20	VISCODE	2.36E-07	EcogPtDivatt	0.026268
MOCA	0.002178	AGE	1.22E-07	EcogSPMem	2.06E-13
EcogPtLang	2.79E-24	ADASQ4	0.000275	mPACCdigit	0.040265

6. Model establishment and solution of question 2

6.1 KSVM intelligent diagnosis algorithm based on KPCA-FDA

This chapter mainly combines the kernel principal component analysis (KPCA) and Fisher discriminant analysis (FDA) to extract features, and uses the kernel support vector machine (KSVM) to classify the extracted MRI data features. The overall framework of this chapter is shown in Figure 6.1. First, the feature subset is placed in the KPCA module, and then the data is projected onto the higher dimensional kernel space to reduce the principal component coefficients to increase the linear separability; Then, the KPCA coefficients are projected into a more effective FDA to select the optimal feature subset; Finally, the new projection data is used to classify AD, MCI and NC data using KSVM.

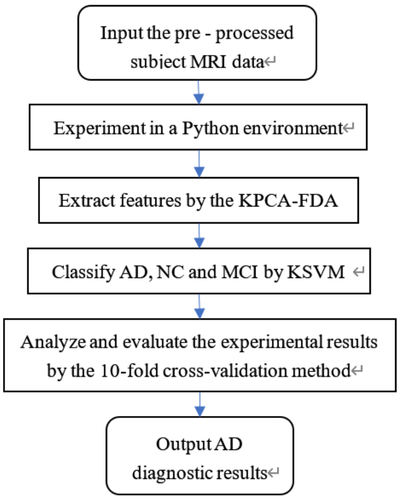
Input the pre - processed subject MRI data↵
Experiment in a Python environment↵
Extract features by the KPCA-FDA
Classify AD, NC and MCI by KSVM ↵
Analyze and evaluate the experimental results by the 10-fold cross-validation method↵
Output AD diagnostic results↵

Figure6.1 Schematic diagram of the KPCA-FDA -KSVM classification framework

6.1.1 Nuclear principal component analysis

Principal component analysis (PCA) is a commonly used multivariable method. Its principle is: first, remove the correlation linear transformation between elements in the original sample data set; Then, find out the most "main" ingredients; Finally, the original complex data is reduced to the feature subspace to obtain a set of projection coefficients for early diagnosis of AD. However, this method only considers second-order correlation information and does not use high-order correlation information in the image, which often contains information more useful for early diagnosis of AD[7].

Therefore, KPCA algorithm is proposed by combining kernel function with PCA. It uses nonlinear methods to extract the principal components, that is, it maps the original vector $X(X \in \mathbb{R}^N)$ to a high-dimensional feature space F , $F=\left\{\phi(X): X \in \mathbb{R}^N\right\}$ through a nonlinear function ϕ , and then performs PCA analysis on the F . It can map data that can not be linearly separable in the input space to the feature space to achieve linear separability. Compared with PCA, KPCA can not only achieve the goal of dimensionality reduction on the basis of maintaining the original data information, but also extract nonlinear features with better recognition performance.

Suppose there are N sample $x_k(k=1,2,\cdots,N)$, $x_k \in \mathbb{R}^N$ in the input space, where $\sum_N x_k = 0$ (satisfying the centralization condition), the covariance matrix is:

$$C=\frac{1}{N}\sum_{j=1}^{N} X_j X_j^T \qquad (6.1)$$

By solving the characteristic equation (6.2), large eigenvalue λ and corresponding eigenvector v can be obtained:

$$\lambda_v = C_v \qquad (6.2)$$

The nonlinear mapping function ϕ is introduced to convert the sample points $x_1, x_2, \cdots, x_N$ in the original space into the sample points $\phi(x_1), \phi(x_2), \cdots, \phi(x_N)$ in the feature space and meet the centralization conditions, namely:

$$\sum_{k=1}^{N} \phi(x_k) = 0 \tag{6.3}$$

The covariance matrix in the feature space F is:

$$\overline{C} = \frac{1}{N} \sum_{j=1}^{N} \phi(x_j) \phi(x_j)^T \tag{6.4}$$

In the feature space, PCA is used to solve the eigenvalue λ and eigenvector v in equation (6.2). Further, there are:

$$\lambda(\phi(x_k) \cdot v = \phi(x_k) \overline{C} v \, (k = 1, 2, \cdots, N) \tag{6.5}$$

Where v is represented linearly by $\phi(x_i)(i = 1, 2, \cdots, N)$, namely:

$$v = \sum_{i=1}^{N} \alpha_i \phi(x_i) \tag{6.6}$$

According to Formula (6.4) - (6.6):

$$\lambda \sum_{i=1}^{N} \alpha_i (\phi(x_k) \cdot \phi(x_i)) = \frac{1}{N} \sum_{i=1}^{N} \alpha_i \left(\phi(x_k) \cdot \sum_{j=1}^{N} \phi(x_j) \right) (\phi(x_j) \cdot \phi(x_i)) (k = 1, 2, \cdots, N) \tag{6.7}$$

Defined the K matrix of $N * N$:

$$K_{ij} = \phi(x_i) \cdot \phi(x_j) \tag{6.8}$$

Formula (6.7) is simplified as:

$$M \lambda K \alpha = K^2 \alpha \tag{6.9}$$

Obviously satisfied:

$$M \lambda \alpha = K \alpha \tag{6.10}$$

The calculated eigenvalues λ and eigenvectors v can be obtained by formula (6.10).

The projection of the test sample in space F is:

$$Y = \left(V^k \cdot \phi(x) \right) = \sum_{i=1}^{N} \alpha_i^k (\phi(x_i) \cdot \phi(x)) \tag{6.11}$$

Because the distance between the selected data points is close and the characteristics are similar, the local kernel function is selected in this paper. Radial Basis Function (RBF), namely Gaussian kernel Function, is used for nonlinear mapping as follows:

$$k(x, y) = \exp\left(-\frac{\| x - y \|^2}{2\sigma^2} \right) \tag{6.12}$$

KPCA algorithm process is as follows:

Step 1. Write the original data into the $m \times n$ dimensional data matrix A;

Step 2. Calculate the kernel matrix K by formula (6.8);

Step 3. Calculate the eigenvalue $\lambda_1, \lambda_2, \cdots, \lambda_n$ and its corresponding eigenvector $v_1, v_2, \cdots, v_n$ through formula (6.10);

Step 4. Arrange the obtained $\lambda_1, \lambda_2, \cdots, \lambda_n$ in descending order to obtain $\lambda_1 > \lambda_2 > \cdots > \lambda_n$;

Step 5. Orthogonalize the units $v_1, v_2, \cdots, v_n$ to obtain $\alpha_1, \alpha_2, \cdots, \alpha_n$;

Step 6. Calculate the cumulative contribution rate $B_1, B_2, \cdots, B_n$ of $\lambda_1, \lambda_2, \cdots, \lambda_n$.

According to the designed extraction rate ρ , select t principal components $\alpha_1,\alpha_2,\cdots,\alpha_t$ when $B_i>\rho$;

Step 7. Calculate the projection Y of the sample $\phi(x_1),\phi(x_2),\cdots,\phi(x_N)$ on $\alpha_1,\alpha_2,\cdots,\alpha_t$ by formula (6.11);

Step 8. Use the new projection data to further analyze with FDA.

6.1.2 Fisher discriminant analysis

FDA considers how to determine a set of projection vectors so that data sets can be distinguished from each other as much as possible in projection. After Fisher transformation, the pattern vectors of the same class are "closer" to each other, while those of different classes are "farther" to each other, so that they can be more easily distinguished, as shown in Figure 6.2.

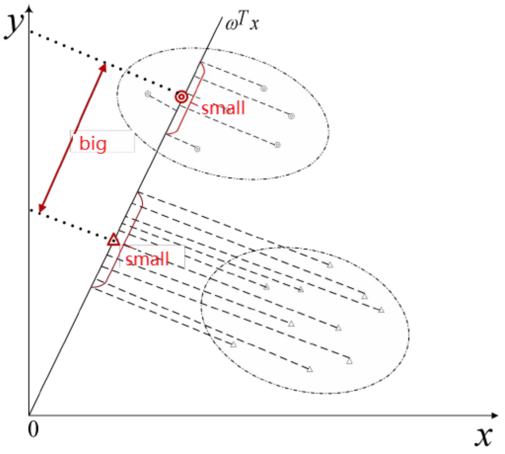

Figure6.2 FDA schematic

In the original space, the class mean is:

$$m_i=\frac{1}{N}\sum_{X\in\omega_i}X,i=1,2 \qquad (6.13)$$

Where N_i is the number of samples of the class ω_i, $X=(x_1,x_2,\cdots,x_n)$ is the training sample set, that is, the coefficient of KPCA is projected on the FDA axis as the training sample; m_i is a matrix of $d*1$, assuming that each dimension is a variable value, and each dimension in m_i is the mean value of these variable values.

In order to determine the category divergence projection axis, we must determine the intra class divergence matrix D_i, the total intra class divergence matrix D_w and the inter class divergence matrix D_b, respectively. Where the intra class dispersion matrix is the quasi covariance matrix.

$$D_i=\sum_{X\in\omega_i}(X-m_i)(X-m_i)^T,i=1,2$$

$$D_w=D_1+D_2 \qquad (6.14)$$

$$D_b=(m_1-m_2)(m_1-m_2)^T$$

Further, define Fisher criterion function:

$$G_F(V)=\frac{\omega^T D_b\omega}{\omega^T D_w\omega} \qquad (6.15)$$

The generalized eigenvalue problem is described as:

$$D_b V = \lambda D_w V \qquad (6.16)$$

The above equation is simplified as:

$$D_w^{-1} D_b V = \lambda V \qquad (6.17)$$

The eigenvector matrix of $D_w^{-1} D_b$ is V :

$$V = [V_1 V_2 V_3 \cdots V_k] \qquad (6.18)$$

The maximum value of $G_F(V)$ obtained by Lagrangian algorithm, at this time, the projection direction ω^* is:

$$\omega^* = D_w^{-1}(m_1 - m_2) \qquad (6.19)$$

Then, project all samples in the training sample set:

$$X^* = \omega^* X \qquad (6.20)$$

Next, KSVM is used to classify the selected optimal feature subset.

FDA algorithm process is as follows:

Step 1. Standardize and calculate the mean value of each category by Formula (6.13);

Step 2. Calculate the intra class divergence matrix D_i, total intra class divergence matrix D_w and inter class divergence matrix D_b according to formula (6.14);

Step 3. Define Fisher criterion function;

Step4 4. Calculate the eigenvalue and eigenvector of matrix D_w and D_b;

Step 5. Select the first k features and their corresponding eigenvectors to construct a $d \times k$ dimensional matrix V, where the eigenvectors are arranged in columns;

Step 6. Map the training samples to the new feature space through the matrix V.

6.1.3 Kernel support vector machine

SVM is an effective tool to deal with small sample learning problems[8]. It has good advantages in dealing with "model selection", "over learning", "dimension disaster" and "local minima". It finds a hyperplane in the feature space and separates positive and negative samples with the minimum error rate. When the samples are linearly indivisible in the original feature space, that is, when a good enough hyperplane cannot be found, the kernel function $\phi: R^d \to R^{d'} (d' > d)$ can be used to map the original data to a higher dimensional space. After mapping, the samples are linearly separable in the new feature space, as shown in Figure 6.3.

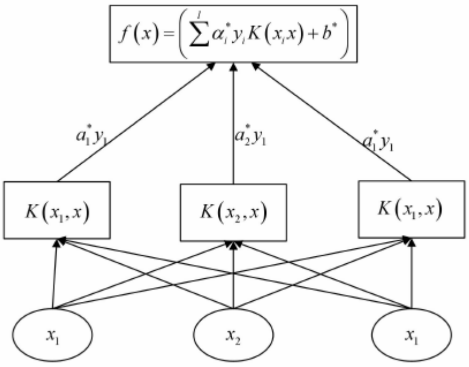

Figure6.3 Schematic diagram of SVM structure

Given the labeled training set $\{(x_i, y_i)\}^n = 1$, where $x_i \in \mathbb{R}^d$ is the training sample and $y_i \in \{1,-1\}$ represents the corresponding category label, the main optimization problems of traditional KSVM are:

$$\begin{aligned} &\min_{g,b,\xi_i} \frac{1}{2} \| \mu \|^2 + \kappa \sum_{i=1}^{n} \xi_i , s.t. \\ &y_i \left(g^T \phi(x_i) + b \right) \geq 1 - \xi_i , \xi_i \geq 0, i = 1,2,\cdots,n \end{aligned} \quad (6.21)$$

Among them, ξ_i represents the non-negative relaxation variable of the error classification degree of the measurement data, κ represents the penalty parameter controlling the constraint violation amount introduced by ξ_i , b represents the deviation, μ represents the normal vector of the hyperplane, and ϕ represents the mapping function triggered by the kernel.

Use Lagrange and kernel techniques to solve optimization problems:

$$\min_{\omega,b,\xi} \max_{\alpha,\beta} \left\{ \frac{1}{2} \| \mu \|^2 + \varepsilon \sum_{k=1}^{N} \xi_k - \sum_{k=1}^{N} \alpha_k \left[y_k \left(\mu x_k - b \right) - 1 \right] + \xi_k - \sum_{k=1}^{N} \beta_k \xi_k \right\} \quad (6.22)$$

The dual form is as follows:

$$\begin{aligned} &\max_{\alpha} \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j k(x, x_j), s.t. \\ &0 \leq \alpha_i \leq \kappa, \sum_i \alpha_i y_i = 0, i = 1,2,\cdots,n \end{aligned} \quad (6.23)$$

Where, k is the kernel matrix, α_i and α_j is the Lagrange multiplier.

Given a new test sample $U = \{u^1, \cdots, u^j, \cdots u^m\}$, the decision function of KSVM is:

$$F(u) = sign\left(\sum_{i=1}^{n} \alpha_i y_i k(x_i, u) + b \right) \quad (6.24)$$

The Radial Basis Function (RBF) kernel is one of the most widely used kernels at present, in the form of:

$$K(x_m, x_n) = \exp\left(-\frac{\| x_m - x_n \|^2}{2\sigma^2} \right) \quad (6.25)$$

σ is the scale factor in the RBF kernel. We introduce formula (6.22) into formula (6.20) to obtain the final KSVM training function:

$$\begin{aligned} &\max_{\alpha} \sum_{i=1}^{n} \alpha_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} \alpha_n \alpha_m y_n y_m \exp\left(-\frac{\| x_m - x_n \|^2}{2\sigma^2} \right), \\ &s.t. \begin{cases} 0 \leq \alpha_n \leq \varepsilon \\ \sum_{n=1}^{N} \alpha_n y_n = 0 \end{cases}, n = 1,2,\cdots,N \end{aligned} \quad (6.26)$$

6.2 Intelligent diagnosis algorithm based on LSTM model

6.2.1 Analysis of model

LSTM[9] is a special recurrent neural network, which is called long short term memory network in full. Because LSTM algorithm is improved based on RNN algorithm, a special RNN network avoids the problem of gradient explosion and gradient disappearance in standard RNN, and the network is designed to solve the problem of long-term dependence. This network was introduced by Hochreiter & Schmidhuber (1997), and has been improved and popularized by many people. Their work has been used to solve a variety of problems and has been widely used until now.

LSTM consists of four parts: input gate, output gate, forget gate and memory unit;

Input gate: determines whether the outside world can write data to the memory unit. Data can be written only when the input gate is open;

Output gate: determines whether the outside world can read data from the memory unit. Only when the output gate is opened can data be read;

Forget gate: determines when to clear the data in the memory cell. It is not cleared when it is opened, but will be cleared when it is closed;

Memory cell: the basic cell of LSTM. The core is the cell state, which is the line to the right as shown in the figure above.

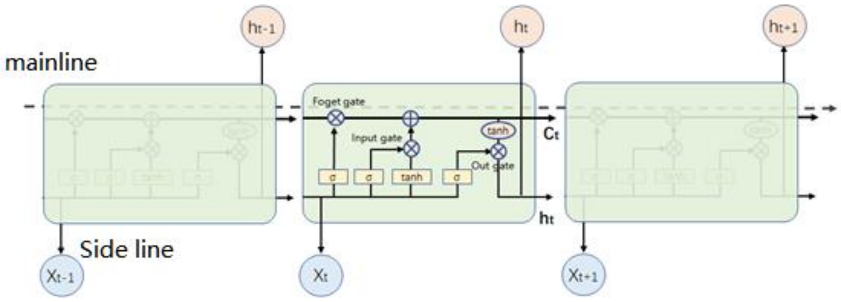

Figure6.4 LSTM schematic

6.2.2 Establish of model

Based on the principle of LSTM model, in the Python 3.9 environment, the data processed in Section 4 is sent to the constructed model for training. In this paper, the qualitative cycle epoch is set to 50, the batchsize is set to 2400, and the activation function is adam. After 50 rounds of training, a trained model is obtained.

The ratio of training set to test set in this section is 0.80: 0.20. The first n points in the training set correspond to n+1 points, that is, $[x_1, x_{2,} \cdots, x_n]$ correspond x_{n+1}, $[x_2, x_{3,} \cdots, x_{n+1}]$ correspond x_{n+2}. The training set is divided in turn, and the test set is treated in the same way. 2400 pieces of data are processed and fed into the trained model for AD diagnosis. In order to reduce the impact on prediction caused by large values and characteristics of different dimensions, the whole training process uses normalized data. Therefore, it is necessary to reverse normalize the prediction results to obtain the required prediction value. The reverse normalization formula (6.27) is:

$$x_i = x_i^*(x_{\max} - x_{\min}) + x_{\min} \tag{6.27}$$

Where, x_i^* is the prediction result output by the model, $x_{\max}$ and $x_{\min}$ are the maximum and minimum loads in the sample data respectively, and xi is the final value of load prediction.

6.2.3 Solution of model

After training the neural network model, input the characteristic indexes of the test set into the model for AD diagnosis, and the diagnosis results are shown in Table 6.1.

Table6.1 AD diagnostic results

Test	Predict	Test	Predict	Test	Predict
1	1	1	1	1	1
2	2	3	3	3	3

1	1	3	1	3	3
1	1	3	2	3	2
2	2	2	2	1	1
3	1	3	3	2	2
2	2	3	3	…	…

The model evaluation indicators mainly include the precision, recall and F1, as shown in Table 6.2. It can be seen that the model has a good effect.

Table6.2 Model evaluation

Metrics	Score
Precision score	0.7958
Recall score	0.8628
F1	0.7432

7. Model establishment and solution of question 3

7.1 Analysis of model

Before the advent of deep learning, the feature extraction steps of Alzheimer's disease image data were very cumbersome. Researchers usually used manual methods to extract features from MRI images, and then used support vector machine (SVM) or random forest and other classification algorithms in machine learning for classification. Figure 7.1 shows the flow chart of Alzheimer's disease classification method based on traditional active learning. The method is still alive today, with two key steps: feature extraction and feature selection.

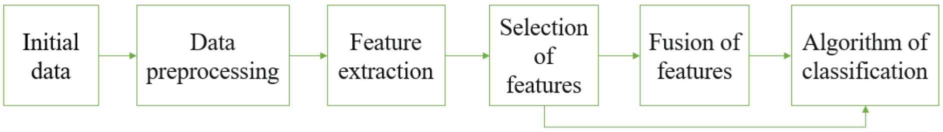

Figure 7.1 Flow of AD classification method based on traditional machine learning method

7.2 Establish of model

7.2.1 AD classification algorithm based on deep features and nonlinear dimension reduction

（1） Deep feature extraction

In this chapter, we selected VGG16[10], which is pre trained and widely used, to extract deep features. VGG16 is a 16 layer network built by Oxford Visual Geometry Group (VGG) and participated in the ImageNet competition of ILSVRC-2014. VGG16 consists of an input layer, a hidden layer, a fully connected layer and an output layer. The hidden layer is composed of five blocks in series, each of which includes a different number of convolution layers and pooling layers.

Like other CNN, this network architecture is superior to traditional machine learning algorithms in computer vision tasks. VGG16 has strong universality. It can use additional task specific images to fine tune, and then transfer the learned image features (such as edges, lines, circles, etc.) to other image classification tasks.

（2） Nonlinear dimensionality reduction method

LargeVis algorithm[11] is an improvement of t-SNE. LargeVis algorithm is mainly divided into three steps, namely, reconstruction of neighborhood graph, construction of conditional probability distribution in high-dimensional and low-dimensional space, and optimization of loss function by steepest descent method. In high-dimensional space, the t-SNE algorithm can calculate the nearest neighbor probability only by calculating the entire dataset, which leads to the reduction of the efficiency of the t-SNE algorithm. At the same time, the universality of t-SNE is poor, and the adjusted parameters in one dataset cannot achieve the same effect in another dataset, so retraining is required in different datasets to find the optimal parameters. LargeVis algorithm's improvement on t-SNE is mainly reflected in the reconstruction of neighborhood graph. Unlike t-SNE, which needs to calculate the probability of the entire dataset, LargeVis algorithm only needs the probability of k nearest neighbors of the computer[12], which greatly reduces the calculation cost. In the reconstruction of neighborhood graph, LargeVis does not seek to reach the goal in one step, but first calculates an approximate k-Nearest Neighbor (kNN) graph, and then calculates a kNN graph with higher accuracy. First, a space partition is obtained by using the random projection tree, and the k-nearest neighbor of each point is found to obtain a kNN graph that does not require complete accuracy. Then, according to the idea of direct access to the nearest neighbor, use the neighbor search algorithm to find potential neighbors, calculate the distance between neighbors, neighbors and the current point, take the nearest k points as kNN, and finally obtain an accurate kNN graph. The pseudo code of the algorithm of the LargeVis algorithm to reconstruct the neighborhood graph is shown in Table 7.1.

Table 7.1 Pseudo code of algorithm of LargeVis algorithm to reconstruct neighborhood graph

```
(1)data: ${x_i}$ where i=1,2,…,N, number of trees NT, number of neighboors K,number of iterations NI.
(2)Result:Approximate K-nearest neighbor graph G.
(3)Build NT random projection trees on ${x_i}$ where i=1,2…,N.
(4)Search neared neighbors:
    for each node I in parallel do
  Search the random projection trees for i’s K neighbors,store the result in knn(i)
    end
(5)Neighbor exploring:
while t<T do
        Set old_knn()=knn(),clear knn()
        for each node I in parallel do
                Create max heap Hi;
                for j ∈ old_knn(i) do
                        for I ∈ old_knn(j) do
                                calculate dist (I,1)= $\|x_i - x_I\|$
                                pull I with dist(I,1) into Hi
                        end
                end
                put nodes in Hi into knn(i);
        end
        T++;
end
for each node I and each j ∈ knn(i) do
        Add edge(i,j) into graph G;
end
(6)calculate the weight of the edges accordin $W_{ij} = \frac{p_{j|i} + p_{i|j}}{2N}$
```

The whole framework of the above methods is shown in Figure 7.2.

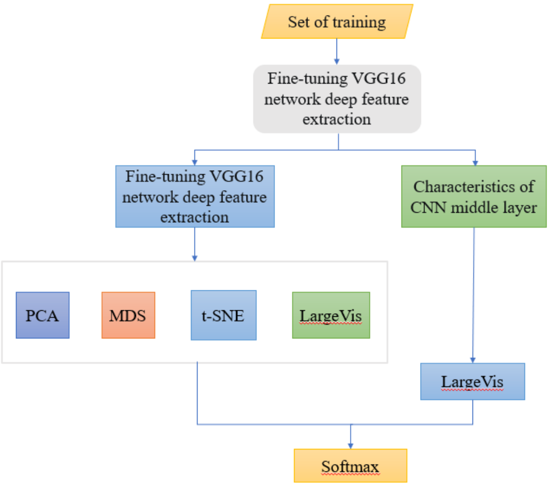
Set of training
Fine-tuning VGG16 network deep feature extraction
Fine-tuning VGG16 network deep feature extraction
Characteristics of CNN middle layer
PCA
MDS
t-SNE
LargeVis
LargeVis
Softmax

Figure 7.2 Flow Chart of AD Classification Algorithm Based on Deep Feature and Nonlinear Dimension Reduction

7.2.2 SVM based classification algorithm

Support vector machine is a machine learning method based on statistical learning theory proposed by vipnik. Its basic model is defined as the linear classifier with the largest interval in the feature space. Its learning strategy is to maximize the interval, that is, the support vector, and finally transform it into the problem of solving convex quadratic planning.

The problem of slope deformation prediction can be seen as a nonlinear functional relationship regression problem between slope deformation and time. For nonlinear regression, the basic idea of support vector machine is to raise the dimension of feature space in high-dimensional space, that is $\varphi : R^n \to \Gamma$, to find the best function $f(x)$ with the maximum deviation from the measured values of samples in high-dimensional space, which is usually expressed as:

$$f(x) = w\varphi(x) + b \tag{7.1}$$

In formula (7.1): w is the adjustable weight vector and b is the deviation vector. Define the insensitive loss function:

$$L_c = \begin{cases} 0, \mathrm{if}\ \ y - (w\varphi(x) + b) \le \varepsilon \\ |y - (w\varphi(x) + b)| - \varepsilon, \mathrm{else} \end{cases} \tag{7.2}$$

In formula (7.2): ε is the insensitive coefficient, which controls the precision of fitting function and y is the true value vector.

By introducing nonzero relaxation variables, which are ξ_i and ξ_i^*, the regression problem is transformed into a convex quadratic optimization problem with linear constraints.

$$\begin{cases} \min \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \\ s.t. y_i - w \cdot x_i - b \le \varepsilon + \xi_i \\ y_i - w \cdot x_i - b \ge -\varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \ge 0, i = 1, 2, \cdots, n \end{cases} \tag{7.3}$$

In formula (7.3): ξ_i and ξ_i^* indicate the punishment degree of sample judgment error. C is the penalty coefficient, which is used to control the degree of punishment for sample errors. The greater the value of C, the greater the punishment for judgment errors.

Lagrangian multiplier a_i and a_i^* are introduced, which can be transformed into a dual problem. For the partial derivative w, make it equal to 0 to obtain its dual form as follows:

$$\begin{cases} \min \frac{1}{2}\sum_{i,j=1}^{n}(a_i^* - a_i)(a_j^* - a_j) \\ +\varepsilon\sum_{i=1}^{n}(a_i^* + a_i) - \sum_{j=1}^{n}(a_j^* - a_j) \\ s.t.\sum_{i=1}^{n}(a_i^* - a_i) = 0 \\ 0 < a_i^*, a_i < C, i = 1, 2, \cdots, n \end{cases} \tag{7.4}$$

The linear fitting function is derived from formula (7.4):

$$f(x) = \sum_{i=1}^{n}(a_i^* - a_i)(x_i, x) + b \tag{7.5}$$

For nonlinear problems, a kernel function $K(x_i, x)$ can be introduced to express the nonlinear situation as:

$$f(x) = \sum_{i=1}^{n}(a_i^* - a_i)K(x_i, x) + b \tag{7.6}$$

7.2.3 Classification method based on BP neural network

As early as the 1960s, scientists expected to use computers to simulate the process of human brain thinking, so the concept and model of "neural network" began to emerge and gradually developed. At the beginning of the 21st century, the development of neural networks reached a climax. The principle of neural network is shown in Figure 6.2, where the middle circle represents a neuron, $x_1 \sim x_n$ represent the input signal (either the initial input vector or the input vector from other neurons), $w_{i1} \sim x_{in}$ represents the connection weight vector from the previous neuron (or input) to the next neuron (or output), θ represent the threshold value, $y_i = f(net_i)$ represents the activation function in the neuron, and the general form of the activation function net_i is $net_i = \sum_{j=1}^{n} w_{ij}x_j - \theta = \sum_{j=0}^{n} w_{ij}x_j$.

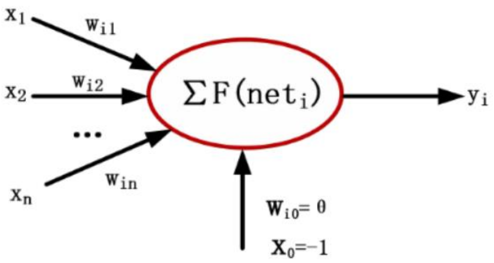

Figure. 7.3 Schematic Diagram of Neural Network

Among many neural networks, BP neural network (BPNN) is the most widely used. BP neural network is a typical supervised learning machine learning algorithm, which can be used not only for regression prediction analysis of data, but also for data classification and recognition. The core algorithm idea is as follows:

Step 1: forward propagation, that is, first transmit the training data from the input layer to the output layer through the hidden layer, and then compare the current output of the output layer with the expected output. If the error does not meet the expected requirements, go to step 2 -- back propagation.

Step 2: back propagation: the output error is propagated to the input layer one by one, and the connection weight between each two neurons is corrected by using the error, and then the training data is propagated back to the output layer.

7.3 Solution of model

In this chapter, SVM was used to classify 2400 data, and the classification results were shown in FIG. 7.4 and FIG. 7.5. The accuracy of SVM clustering algorithm for CN, MCI and AD was 0.905. The accuracy of SMC, EMCI and LMCI clustering was 0.745.

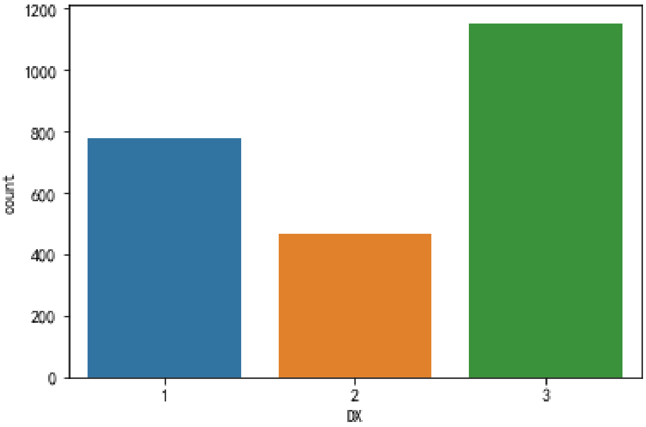

Figure 7.4 Classification of CN (1 represents CN),AD (2 represents AD) and MCI (3 represents MCI)

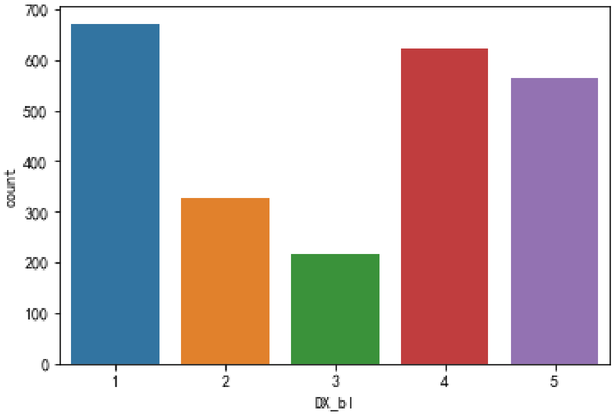

Figure 7.5 Classification of CN (1 represent CN), AD (2 represents AD), SMC (3 represents SMC), EMCI (4 represent EMCI) and LMCI (5 represents LMCI)

8. Model establishment and solution of question 4

8.1 Selection of variables and data

The development of disease progression models helps to understand the underlying pathological mechanisms of AD and can explain to some extent the differences in the rate of disease progression between patients by assessing their associated influencing factors. Simulation of AD clinical trials through disease progression model can significantly improve efficiency and reduce unnecessary development costs. For example, in the early stages of drug development, the screening of promising drug candidates is completed with the guidance of a model. Table 8.1 below shows the classes of variables that remain after data pre-processing.

Table 8.1 Variable classification

type	
Cognitive assessment	CDRSB, LDELTOTAL, MMSE, FAQ, MOCA, ADAS13, ADASQ4, mPACCtrailsB, mPACCdigit, EcogPtLang, EcogPtOrgan, EcogPtTotal, EcogPtDivatt, EcogSPMem RAVLT_immediate, RAVLT_forgetting, RAVLT_perc_forgetting
Anatomical quantification values	IMAGEUID
Other	VISCODE, AGE, PTETHCAT

From the table we can see that AD disease and VISCODE are associated. For the remaining 21 variables after screening, most of which are related to cognitive assessment, six factors such as CDRSB, ADAS13, EcogPtTotal, RAVLT_immediate, IMAGEUID, and PTETHCAT were selected for analysis. CDRSB (Clinical dementia rating sum of boxes) is the largest impact factor; The last factor, PTETHCAT, stands for Ethnicity. The first 5 variables can be analyzed in relation to time, and PTETHCAT should be analyzed in relation to the disease situation.

The data in sub-section 8.3 were obtained by categorising the 5 disease conditions at the first visit, which ensures that each volunteer starts with a similar window to observe the disease conditions under time variation; the data in sub-section 8.4 categorised the pre-processed data by variable factor condition, counting the level of each factor under different disease conditions, as reflected by box plots.

In contrast, this subsection screens the information from the processed data for the first visit of these 646 volunteers to obtain the racial prevalence, as shown in Table 8.2 below.

Table 8.2 Diseases classification by Ethnicity

Type	CN	SMC	EMCI	LMCI	AD
Unknow	-	2	-	-	-
Hisp/Latino	10	2	7	1	6
Not Hisp/Latino	140	102	130	129	122

As can be seen from Table 8.2, Ethnicity has about the same probability of having MCI and AD, with a higher percentage of people who are not Hispanic/Latino, and race is an influential factor.

8.2 Model establishment

8.2.1 Empirical linear models

The simplest model of AD disease progression is the linear model, which usually assumes that the disease state is linear over time and that the rate of disease progression can be defined by the slope in a linear equation. Using the Alzheimer's Disease Scale Cognitive Subscale (ADAS13) as an example, the relevant equation is as follows.

$$\frac{d\text{ADAS13}}{dt} = k \qquad (8.1)$$

i.e. $\Delta\text{ADAS13} = kt$.

It denotes changes in ADAS13 scale scores over time and k denotes the linear rate of disease progression. When using linear models to describe disease change in AD, studies of models based on individual data as well as those based on literature data have mostly shown a high correlation between the rate of disease progression and the severity of the patient's dementia.

8.2.2 logistic growth model

To overcome the limitations of the linear model assumptions, logistic growth models can also be used to describe changes in disease status due to the changing rate of disease progression over time and the upper limit of the scale scores. The advantage of using a logistic growth model is that it takes into account the non-linear nature of AD disease progression and thus provides a more accurate estimate of the inflection point at which the rate of disease progression suddenly changes, while also limiting the predictive value of the scale scores to a theoretical range. Using the Alzheimer's Disease Scale Cognitive Subscale (ADAS11) as an example, the relevant equation is as follows[13].

$$\frac{d\text{ADAS13}}{dt} = k \times \text{ADAS13}^{\alpha} \times [1 - \left(\frac{\text{ADAS13}}{70}\right)^{\beta}]^{\gamma}, \text{ADAS13}(0) = \text{ADAS13}_0 \qquad (8.2)$$

Where r is the rate constant for disease progression, the value of α, β, γ determines the shape of the curve and the inflection point for the rate of disease progression, and the total score for the ADAS13 scale is 70.

8.2.3 Disease onset trajectory model

Due to the very long time of AD onset, only a specific stage of the patient's disease development may be observed in different cohorts; Moreover, the onset of AD is insidious and is determined to be AD when seeking medical treatment, and it has often been ill for a long time. Therefore, the disease onset trajectory model normalizes the onset time of patients in different dementia stages and combines them into new development traces, so as to depict the complete trajectory of AD patients from normal cognitive degradation to continuous change of dementia state. The model assumes that the patient's cognitive deterioration occurs prior to enrolment and therefore requires extrapolation and adjustment for the true onset time, using the equations shown below:

$$\text{A} = \text{A}_0 \times e^{\phi \times \left(t' + \gamma\right)} \qquad (8.3)$$

where A represents the scale score observed at the moment after the patient was enrolled in the group, and A_0 is the patient's score on the scale at the time of enrollment (i.e. baseline period, $t' = 0$). ϕ is the rate of disease progression, which is a first order rate change constant. γ is the model-implied time elapsed between the onset of disease at the beginning of the patient's illness and entry into the group.

Empirical models are easy to construct and have good interpretability of the model parameters, making them extremely useful in clinical trials. A simple linear model can broadly describe disease progression in AD patients, but needs to take full account of the dynamics of the slope parameter with disease severity. logistic growth models take into account the non-linear component of AD disease progression and the upper limit of the scale. Combining individual data from multiple sources for meta-analysis helps us to further explore the key information (i.e. covariates) affecting disease progression and to improve the accuracy and extrapolation of model predictions.

8.3 Trends in disease change

In Figures 8.1 to 8.5 below 24, 36 and 48 represent 24, 36 and 48 months have passed since the first examination. One year is the interval and 12 was not chosen because the sample size was so small that there was essentially little change.

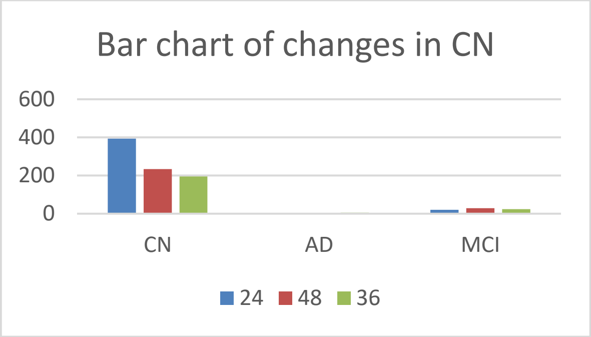

Figure 8.1 Change of prevalence of CN at initial diagnosis

As can be seen in Figure 8.8, as time passes, patients who were CN, become MCI patients and slowly get worse and begin to transform into AD patients, with a smaller proportion of patients suffering from AD past 24 months than past 36 and 48 months.

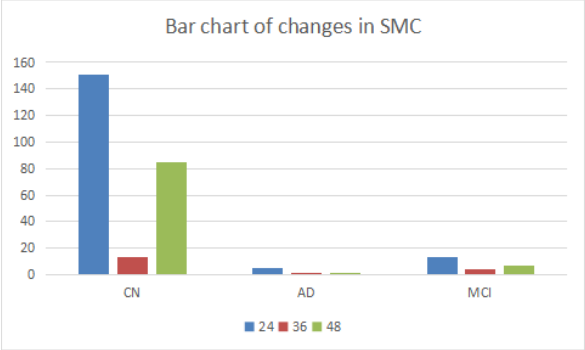

Figure 8.2 Change of prevalence of SMC at initial diagnosis

As can be seen in Figure 8.2, patients initially diagnosed with SMC are at high risk of converting to CN. SMC is a mild cognitive impairment that can be managed and converted to CN with intervention.

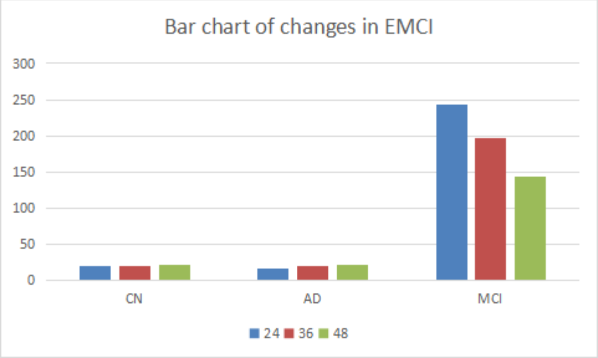

Figure 8.3 Change of prevalence of EMCI at initial diagnosis

As can be seen in Figure 8.3, EMCI is among the longer stages of pre-AD. Patients in EMCI have a small probability of converting to CN and are generally in the middle of MCI, slowly converting to AD patients over time. Patients in EMCI should have intensive interventions to slow disease progression and delay AD.

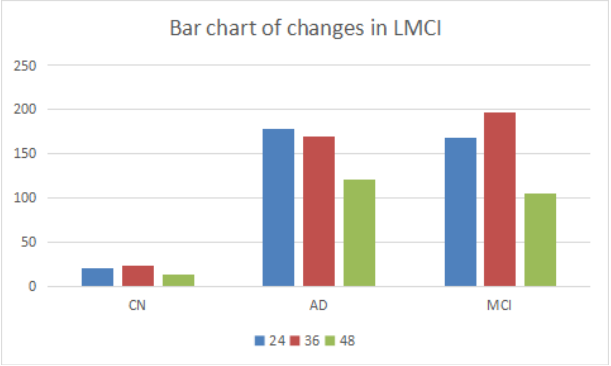

Figure 8.4 Change of prevalence of LMCI at initial diagnosis

Figure 8.4 shows that LMCI has about a half probability of converting to AD, and the probability of converting to AD is greater over time. It is also clear that the condition varies from person to person, with some progressing rapidly and others slowly, and that LMCI is a precursor to AD and should be taken very seriously, as once it has progressed it is AD.

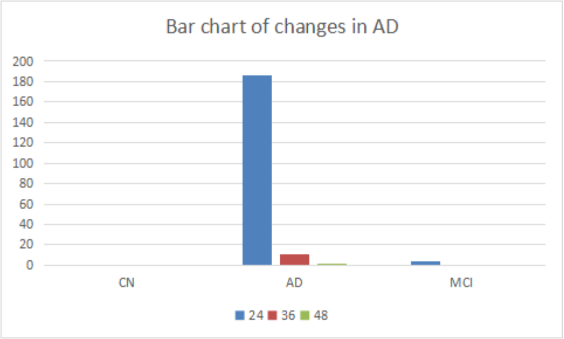

Figure 8.5 Change of prevalence of AD at initial diagnosis

As can be seen in Figure 8.5, a diagnosis of AD, unless misdiagnosed, is basically AD later on, which also indicates the irreversible nature of AD. Even with AD, there are levels of severity of the condition, and patients with AD should have strong interventions to prevent further deterioration, with family members prepared to care for the patient.

8.4 Variable factors over time

8.4.1 CDRSB

The Clinical Dementia Rating - Box Sum (CDRSB) score is widely used as a global measure of disease progression in AD trials. Ratings are made after semi-structured interviews with patients and informants and do not rely directly on psychometric tests, thus avoiding learning effects. A CDRSB box plot of the disease classification is shown below.

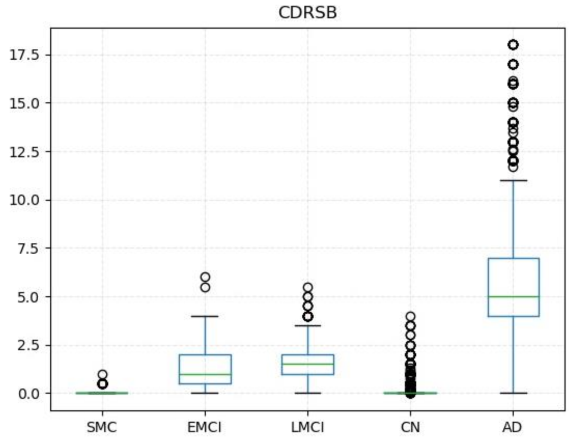

Figure 8.6 CDRSB box plot for different diseases

As can be seen from Figure 8.6 above, the CDRSB levels for CN and SMC remain constant, basically at 0. The CDRSB for EMCI and LMCI is slightly higher, both within 5, while the CDRSB for AD patients is higher, remaining at 4 and above. This means that CN, like SMC, basically does not show dementia behaviour, MCI has some degree of dementia in the middle and late stages, and when CDRSB exceeds 5, dementia can basically be judged. With the passage of time, patients with AD will become increasingly ill.

8.4.2 ADAS13

The ADAS13 was included as a global measure of cognitive function. ADAS13 is a test battery developed to assess severity of cognitive impairment associated with AD and includes subtests and clinical evaluations assessing memory function, reasoning, language function, orientation and praxis. The ADAS13 is a modified version of the original ADAS11, adding a cancellation task and a delayed free recall task21. The higher the scores, the more severe impairment of cognitive function

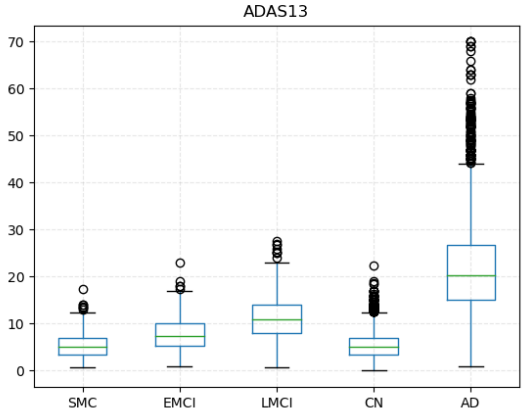

Figure 8.7 ADAS13box plot for different diseases

As can be seen in Figure 8.7, the AMC and CN scores are not very different, with slightly higher scores in the mid to late MCI and the highest scores in AD patients, i.e. the higher the ADAS score, the greater the chance of getting AD and the more severely impaired cognitive function.

8.4.3 EcogPtTotal

The ECog is a short questionnaire that assesses participants' ability to perform normal everyday tasks on a scale of 5 compared to their activity level 10 years ago. Three domains were assessed: memory, language and executive function. the EcogPtTotal is an overall rating of the questionnaire.

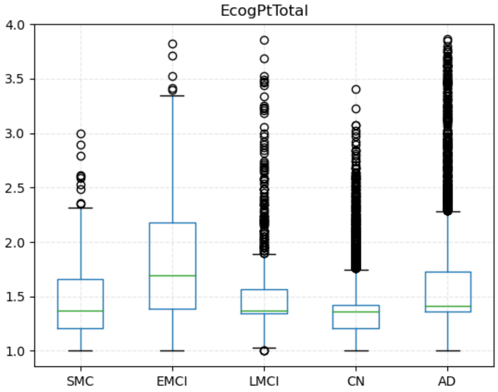

Figure 8.8 EcogPtTotal box plot for different diseases

As can be seen from Figure 8.8, the CN scores are low, but the scores are also not very high for patients with other diseases and slightly higher for those with AD. As a side note, the questionnaire is not very accurate!

8.4.4 RAVLT_immediate

The RAVLT was included as a measure of memory function. In this test, the participants are asked to recall words from a list of 15 nouns immediately after each of five learning trials and after a short and a long delay. Immediate recall (RAVLT-Im): the number of correct responses across the immediate recall of the five learning trials. The lower the scores, the more severe impairment of cognitive function..

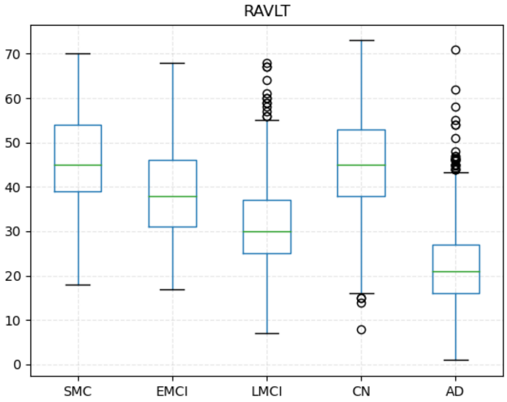

Figure 8.9 RAVLT_immediate box plot for different diseases

As can be seen in Figure 8.9, the scores for CN and SMC patients are very similar, i.e. this questionnaire area may not differentiate between SMC and CN patients. patients with late MCI score lower and lower as their condition worsens, and AD patients score the lowest. Overall, it appears that for AD patients, the scores get lower and lower as time passes and the condition becomes more severe.

8.4.5 IMAGEUID

IMAGEUID refers to the code of the machine imaging for a particular test patient, and the code varies from patient to patient.

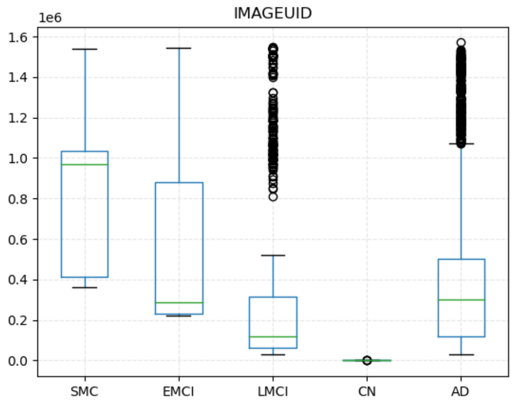

Figure 8.10 IMAGEUID box plot for different diseases

As can be seen from Figure 8.10, only CN is 0. Patients with other diseases can be clearly distinguished from those with CN, but outside of CN, it is not possible to distinguish between the specific diseases, SMC and EMCI are similar, LMCI and AD are similar, but the more severe the disease, the lower the IMAGEUID, and there is some discrimination.

9. Model establishment and solution of question 5

In recent years, meta-evidence from populations carrying genetic risk and from clinically cognitively normal elderly people suggests that the occurrence of Alzheimer's disease (AD) is a continuous pathological process that began many years before the clinical diagnosis of dementia. As new drug clinical trials for the dementia stage of AD have repeatedly failed, researchers have gradually realized that the focus of AD research should shift to the pre-AD dementia stage, namely SMC, LMCI, and EMCI. The introduction of the concept of AD dementia stage is important for early warning of AD. On the one hand, at this stage, the neurons are not yet largely apoptotic, and the disease process is somewhat reversible. On the other hand, the early screening of other chronic diseases, such as cardiovascular and cerebrovascular diseases and diabetes mellitus, can identify patients before the onset of clinical symptoms, which is expected to greatly delay the progression of the disease. Studies have shown that delaying the onset of AD by five years could reduce the number of people with AD by 57 percent and reduce medical costs from $627 billion to $344 billion.

So how to and normal people (CN) to prevent the disease and SMC, EMCI, LMCI, AD to delay the disease?

9.1 Prevention of disease onset in normal people(CN)

Doing more brain-healthy activities. Education level may have an impact on Alzheimer's disease. People with less than 10 years of education have a much greater risk of developing the disease.

Social can't stop. Previous studies have found that single people are more likely to have such symptoms than married people, with a 40 percent higher risk. Elderly people must not simplify, and communicating with their wives, relatives and friends every day has been proved to greatly reduce the risk of Alzheimer's disease, after all, loneliness is also a sad thing.[13]

Keeping weight and heart healthy. A study from Vanderbilt University shows that the heart health directly affects the brain memory area, heart health affects the brain aging degree, the heart pumping less blood in the elderly of the brain temporal lobe blood flow is poor, the brain premature failure for up to 20 years, and the temporal lobe is the earliest occurrence of Alzheimer's pathology.

taking more exercise. Exercise is always the number one way to stay healthy, whether you are an adult or an older adult. Exercise also plays a role in alleviating the cognitive deterioration that older adults are prone to.A survey of 33,000 people showed that people who exercise regularly in their daily life reduced the risk of cognitive deterioration by 38 percent[13].

No Smoking and Drinking. Smoking is able to increase the risk of AD, especially in those carrying APOE. Heavy alcohol drinking causes alcohol dementia, while mid-year heavy drinking increases the risk of AD 3-fold.

9.2 Diagnosis and treatment and delay of S M C, EMCI, and LMCI

Subjective memory complaints (SMC), in which individuals complain of memory decline while objective psycho-psychological tests are normal , was first proposed by Brazilian scholar Reisgberg et al. in 1982 and is important in slowing down the dementia process. The evaluation tool mainly relies on several questionnaires: SMCQ (subjective Memory Complains Questionnaire), MMQ (Multifactorial Memory Questionnaire), MAC-Q (Memory Assessment Complaint Questionnaire), SMQ (Short Memory Questionnaire), and BECSI (Brief) Elderly Cognitive Screening Inventory)[14]. SMC in the elderly include demographic sociology, health and nutritional status, psychological factors, and social support. Currently, non-pharmacological interventions are mainly used for older adults with SMCs, which are low cost to implement and easy to implement intervention strategies, mainly including 3 types of cognitive interventions, exercise interventions, and lifestyle interventions. The elderly should be encouraged to effectively implement memory strategies in their daily lives to keep their brains and minds active; exercise is relatively simple and effective because it is not limited to venues, and the intervention methods and strategies can be adjusted according to the interests and physical status of the elderly and are highly implementable; lifestyle interventions can significantly improve the language and memory functions of the elderly with SMC and improve subjective physical and mental health.

LMCI is a post-state of MCI and a pre-state of Alzheimer's disease (AD). There is no significant difference in brain structure between the EMCI stage and healthy individuals. LMCI and EMCI belong to MCI, and mild cognitive impairment is between normal aging and dementia, manifesting as reduced cognitive function incompatible with age and education, but not yet affecting the ability to perform daily life and not yet meeting the clinical diagnostic criteria for dementia. Specific criteria for recognition include: subjective reports of memory decline, preferably supported by information from a knowledgeable person; impairment of memory function (or other cognitive functions) on objective tests that are incompatible with age and education level; relatively intact overall cognitive function; no impairment of daily living functions; and no dementia[15]. MCI Patients with outwardly apparent behavioral problems include not only cognitive impairment, but also often accompanying psycho-behavioral The main problems are depression, irritability, apathy, sleep abnormalities, and agitation/aggression. Currently, there are two main types of MCI intervention strategies, namely pharmacological interventions and non-pharmacological interventions (somatic motor training, cognitive interventions and psychological interventions, etc.).

9.3 Diagnosis and treatment and delay of A D disease

Patients with dementia in Alzheimer's disease and related disorders may exhibit a variety of psychiatric behavioral abnormalities. These symptoms are related both to the disease itself and to changes in the external environment. Patients often exhibit: aggressive behavior, which includes verbal and physical aggression; anxiety and agitation, in which patients may become fidgety and walk back and forth or wander constantly; confusion, in which patients may appear to not recognize familiar people, places or things[18]. Forgetting the relationship between things, confusing where he lives, forgetting the purpose of commonly used objects; repetitive behavior, the patient may repeat an event over and over again, ask the same question repeatedly, repeat an action; suspicious, becoming suspicious of people around him, even suspecting that people around him are thieves, being distrustful, or exhibiting other inappropriate behaviors. Sometimes also misinterpreting what he sees or hears; hallucinations, where they see, hear, smell, taste, or feel things that are not actually there; disinhibition, where some patients with dementia act impulsively without thinking about the consequences; indifference, showing no interest in things they normally enjoy, showing no interest in friends and family, being reluctant to talk, and lacking emotional responses.

For the diagnosis of Alzheimer's disease, clinical AD diagnosis can be based on the AD diagnostic criteria proposed by NINCDS-ADRDA, 1984 edition or NIA-AA, 2011 edition. When molecular imaging tests and cerebrospinal fluid tests for AD are available, the diagnosis can be made according to the 2011 edition of NIA-AA or IWG-2 diagnostic criteria.

For the treatment of AD, cholinesterase inhibitors (ChEIs), excitatory amino acid receptor antagonists (memantine hydrochloride), herbal medicines and other therapeutic drugs are available[16]. ChEIs are the first choice for patients with a clear diagnosis of AD, and high doses of ChEIs can be used for patients with moderate to severe AD; patients with a clear diagnosis of moderate to severe AD can be treated with memantine or memantine in combination with donepezil[17]. The combination of ChEls and memantine is especially recommended for patients with severe AD with obvious psycho-behavioral symptoms; after explaining the benefits and possible risks of treatment to patients, the Ginkgo biloba brain protein hydrolysate olanzide or piracetam can be used as synergistic adjuvant therapy for AD patients.

References

[1] Atri A, Molinuevo JL, Lemming 0, et al. Memantine in patients with Alzheimer&apos; s disease receiving donepezil; new analyses of efficacy and safety for combination therapy[J]. Alzheimers Res Ther, 2013,5(1):6.

[2] Anki T, Wake R, Miyaoka T, et al. The effect of combine treatment of memantine and donepeail on Alzheimer&apos; s disease patients and its relationship with cerebral blood flow in the prefrontal area[J]. Int J Geriatr Psychiatry, 2014,29 (9) :881-889.

[3] Schmidt R, Hofer E, Bouwman FH, et al. EFNS-ENS/EAN .Guideline on concomitant use of cholinesterase inhibitors and memantine in moderate to severe Alzheimer &apos;s disease[J]. Eur J Neurol, 2015 ,22(6) :889-898.

[4] Morris JC. Early-stage and preclinical Alzheimer disease [J].Alzheimer Dis Assoe Disord, 2005 , 19(3) :163-165.

[5] Hsieh S w, Hsiao SF, Liaw LJ, et al. Effects of multiple training modalities in the elderly with subjective memory complaints: a pilot study [J].Medicine (Balti-more),2019 , 98(29):e16506.

[6] Chan A S, Cheung W K,Yeung M K,et al. A Chinese chair based mind-body intervention improves memory of older adults[J] Front Aging Neurosci, 2017,9:190.

[7] Yu Woo J. Cognitive assessment of older people: do sensory function and frailty matter? [J]. Int J Environ Res Public Health,2019, 16(4) : 662.

[8] Merchant C, Tang MX, Albert S, et al. The influence of smoking on the risk of Alzheimer&apos;s disease[J]. Neurology, 1999 ,52(7):1408-1412.

[9] Anttila T, Helkala EL, Viitanen M, et al. Alcohol drinking in middle age and subsequent risk of mild cognitive impairment and dementia in old age: a prospective population based study[J]. BMJ, 2004, 329(7465): 539.

[10] Anstey KJ, Mack HA, Cherbuin N. Alcohol consumption as a risk factor for dementia and cognitive decline: meta- analysis of prospective studies[J]. Am J Geriatr Psychiatry, 2009,17(7): 542-555.

[11] CUI R X, LIU M H. Hippocampus analysis by combination of 3D DenseNet and shapes for Alzheimer&apos;s disease diagnosis[J]. IEEE J Biomed Health Inform, 2018, 23(5): 2099-2107

[12] JAIN R, JAIN N, AGGARWAL A, et al. Convolutional neural network based Alzheimer’s disease classification from magnetic resonance brain images[J]. Cogn Syst Res, 2019, 57: 147-159.

[13] ALZHEIMER&apos;S ASSOCIATION. 2018 Alzheimer&apos;s disease facts and figures[J]. Alzheimers Dement, 2018, 14(3): 367-429.

[14] BRON E E, SMITS M, VAN DER FLIER W M, et al. Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural MRI: the CADDementia challenge[J]. Neuroimage, 2015, 111: 562-579.

[15] SORENSEN L, SHAKER S B, DE BRUIJNE M. Quantitative analysis of pulmonary emphysema using local binary patterns[J]. IEEE Trans Med Imaging, 2010, 29(2): 559-569.

[16] UCHIYAMA Y, KATSURAGAWA S, ABE H, et al. Quantitative computerized analysis of diffuse lung disease in high-resolution computed tomography[J]. Med Phys, 2003, 30(9): 2440-2454.

[17] PLIS S M, HJELM D R, SALAKHUTDINOV R, et al. Deep learning for neuroimaging: a validation study[J]. Front Neurosci-Switz, 2014, 8: 229.

[18] KHVOSTIKOV A, ADERGHAL K, BENOS-PINEAU J, et al. 3D CNN-based classification using sMRI and MD-DTI images for Alzheimer disease studies[J]. 2018. arXiv: 1801.05968.

Appendix

Source code	Environment: python

```
import pandas as pd
import numpy as np
from pandas.core.frame import DataFrame
import seaborn as sns
from matplotlib.pyplot import figure
import matplotlib.pyplot as plt
from colorama import Fore
from sklearn.metrics import mean_absolute_error, mean_squared_error
import math
from datetime import datetime, date
import chinese_calendar
from sklearn.impute import SimpleImputer
from sklearn.ensemble import RandomForestRegressor
from sklearn.model_selection import train_test_split
import warnings  # 忽略警告
warnings.filterwarnings('ignore')
plt.rcParams['font.sans-serif'] = ['SimHei']  # 用来正常显示中文标签
plt.rcParams['axes.unicode_minus'] = False  # 用来正常显示负号
np.random.seed(4) # 设置随机数种子

#%% 数据读取
data= pd.read_csv('ADNI2_3.0T.csv')

#%% 缺失值初步处理
'''下面函数的作用： 删除数据一样的列和全为空值的列'''
def dropNullStd(data):
    beforelen = data.shape[1] #原始表的列数
    colisNull = data.describe().loc['count'] == 0 #转换为 bool 型
    for i in range(len(colisNull)):
        if colisNull[i]:
            data.drop(colisNull.index[i], axis = 1, inplace = True)

    stdisZero = data.describe().loc['std'] == 0
    for i in range(len(stdisZero)):
        if stdisZero[i]:
            data.drop(stdisZero.index[i], axis = 1, inplace = True)

    afterlen = data.shape[1] #删除后表的列数
    print("剔除列的数目：", beforelen - afterlen)
    print("剔除后数据的形状为：", data.shape)
    return data


""" 1 删除一样的列和全为空值的列 """
data1 = dropNullStd(data)
data1.dtypes
# 剔除列的数目： 3， 剔除后数据的形状为： (2400, 53)

'''2 删除相同的行'''
```

```
data2 = data1.drop_duplicates(subset = data1.columns.tolist()[1:-1], keep = 'first')
data2.dtypes
print("删除的行数为：{}".format( data1.shape[0] - data2.shape[0]))
# 删除的行数为：0，剔除后数据的形状为：（2400, 53)

'''3 删除缺失值超过 50%的列'''
data3=data2.dropna(thresh=0.5*data2.shape[0], axis=1)
data3.dtypes
print("删除的列数为：{}".format( data2.shape[1] - data3.shape[1]))
# 删除的列数为：4，剔除后数据的形状为：（2400, 49)

#%% 特征工程
data3.dtypes.value_counts()
object_data=data3.select_dtypes(include=['object']) # 提取需要量化的列
object_data.columns
#[ 'VISCODE', 'PTGENDER', 'PTETHCAT', 'PTRACCAT','PTMARRY','DX', 'DX_bl']

'''量化处理'''
data3['VISCODE'].unique()
dic2 = {'bl':0,
         'm06':6,
         'm12':12,
         'm24':24,
         'm36':36,
         'm48':48,
         'm60':60}
data3['VISCODE'] =data3['VISCODE'].map(dic2)

data3['PTGENDER'].unique()
dic3 = {'Male':1,
         'Female':2}
data3['PTGENDER'] =data3['PTGENDER'].map(dic3)

data3['PTETHCAT'].unique()
dic4 = {'Unknown':0,
         'Hisp/Latino':1,
         'Not Hisp/Latino':2}
data3['PTETHCAT'] =data3['PTETHCAT'].map(dic4)

data3['PTRACCAT'].unique()
dic5 = {'Unknown':0,
         'White':1,
         'Black':2,
         'Asian':3,
         'Am Indian/Alaskan':4,
         'Hawaiian/Other PI':5,
         'More than one':6}
data3['PTRACCAT'] =data3['PTRACCAT'].map(dic5)

data3['PTMARRY'].unique()
dic6 = {'Unknown':0,
         'Married':1,
         'Divorced':2,
         'Widowed':3,
         'Never married':4}
data3['PTMARRY'] =data3['PTMARRY'].map(dic6)
```

```
data3['DX'].unique()
dic7 = {'CN':1,
        'Dementia':2,
        'MCI':3}
data3['DX'] =data3['DX'].map(dic7)

data3['DX_bl'].unique()
dic8 = {'CN':1,
        'AD':2,
        'SMC':3,
        'EMCI':4,
        'LMCI':5}
data3['DX_bl'] =data3['DX_bl'].map(dic8)

data3.to_csv('Quantified data.csv',index = False, header=True,encoding
="utf_8_sig",errors='strict')

#%% 缺失值填充
# 随机森林缺失值填充
data_rf= pd.read_csv('Quantified data.csv')
data_rf.isna().sum()  # 统计缺失值个数
sindex = np.argsort(data_rf.isna().sum().values.tolist()) # 将有缺失值的列按缺失值的多少由小
到大排序
# 进入 for 循环进行空值填补
for i in sindex:                                         # 按空值数量,从小到大进行排序来遍
历
    if data_rf.iloc[:,i].isna().sum() == 0:              # 将没有空值的行过滤掉
        continue                                          # 直接跳过当前的 for 循环
    df = data_rf                                          # 复制 df 数据
    fillc = df.iloc[:,i]                                 # 将第 i 列的取出，之后作为 y 变量
    df = df.iloc[:,df.columns != df.columns[i]]           # 除了有这列以外的数据，之后作为 X
    df_0 = SimpleImputer(missing_values=np.nan,                # 将 df 的数据全部用 0 填充
                         strategy="constant",
                         fill_value=0).fit_transform(df)
    ytrain = fillc[fillc.notnull()]                      # 在 fillc 列中,不为 NAN 的作为 Y_train
    ytest = fillc[fillc.isnull()]                       # 在 fillc 列中,为 NAN 的作为 Y_test
    xtrain=df_0[ytrain.index,:]                           # 在 df_0 中(已经填充了 0),中那些 fillc
列不为 NAN 的行作为 Xtrain
    xtest=df_0[ytest.index,:]                             # 在 df_0 中(已经填充了 0),中那些 fillc
等于 NAN 的行作为 X_test


    rfc = RandomForestRegressor()
    rfc.fit(xtrain,ytrain)
    ypredict = rfc.predict(xtest)                        #Ytest 为了定 Xtest,以最后预测出
Ypredict

    data_rf.loc[data_rf.iloc[:,i].isnull(),data_rf.columns[i]] = ypredict
    # 将 data_copy 中 data_copy 在第 i 列为空值的行,第 i 列,改成 Ypredict
data_rf.isna().sum()  # 统计缺失值个数

data_rf.to_csv('Missing_value processed data.csv',index = False, header=True,encoding
="utf_8_sig",errors='strict')
```

```
#%% 异常值处理
data_abnor= pd.read_csv('Missing_value processed data.csv')
ID=data_abnor.loc[:,'RID']
data_y1=data_abnor.loc[:,'DX']
data_y2=data_abnor.loc[:,'DX_bl']
data_x = data_abnor.drop(['RID','DX', 'DX_bl'], axis=1)

df1 = np.abs((data_x - data_x.mean())) > (3*data_x.std())    # 灵活运用广播机制，得到判断数据框
numabnormal = []                        # 用来存每一列中的异常值个数
pff = pd.DataFrame()                    # 用来存储能处理的、异常值比较少的变量
dff = pd.DataFrame()                    # 用来存储能不处理的、异常值比较多的变量
for m in range(df1.shape[1]):         # 循环用来计算每一列的异常值个数
    number = np.sum(df1.iloc[:,m])
    numabnormal.append(number)

t = pd.DataFrame(numabnormal) < 0.01*data_x.shape[0]          # 判断哪些变量中的异常值可以进行处理
t1 = t.iloc[:,0]
for i in range(df1.shape[1]):
    if t1[i] == True:
        pff = pd.concat([pff,data_x.iloc[:,i]], axis = 1)    # 可以处理的就放在一起
    if t1[i] == False:
        dff = pd.concat([dff,data_x.iloc[:,i]], axis = 1)    # 不可以处理的就放在一起

def deal_abnormal(data):
    data1 = np.abs((data - data.mean())) > (3*data.std())
    for i in range(len(data)):
        for j in range(len(data.columns)):
            if data1.iloc[i, j] == True:
                data.iloc[i, [j]] = data.iloc[:, j].median()      # 用中位数进行填充
    return data

pff1 = deal_abnormal(pff)                    # 处理异常值
data_x1 = pd.concat([dff,pff1],axis = 1)       # 再将数据进行合并

data4=pd.concat([ID,data_y1,data_y2,data_x1],axis = 1)
data4.to_csv('Outlier processed data.csv',index = False, header=True,encoding
="utf_8_sig",errors='strict')


#%% 数据归一化
def Norm(data):
    return (data - data.min()) / ( data.max() - data.min())

data_norm = Norm(data4.drop(['RID'], axis=1))
data_norm1=pd.concat([ID,data_norm],axis = 1)

data_norm1.to_csv('Normalized data.csv',index = False, header=True,encoding
="utf_8_sig",errors='strict')


#%% 特征向量筛选
```

```
putin= pd.read_csv('Normalized data.csv')
putin1=putin.drop(['RID','DX_bl'], axis=1)

# 热力图
putin1.columns
f, ax = plt.subplots(nrows=1, ncols=1, figsize=(20, 20))

shifted_cols = putin1.columns[:]
corrmat = putin1[shifted_cols].corr('spearman')
sns.heatmap(corrmat, annot=True, vmin=-1, vmax=1, cmap='coolwarm_r')
ax.set_title('Correlation Matrix of Lagged Features', fontsize=16)
plt.tight_layout()
plt.show()

putin_y1=putin.loc[:,'DX']
putin_y2=putin.loc[:,'DX_bl']
putin_x = putin.drop(['RID','DX', 'DX_bl'], axis=1)
x_train = putin_x
y_train = putin_y1


''' 计算 spearman 相关系数 '''
spearman=putin.drop(['RID','DX_bl'], axis=1).corr('spearman')
spear_DX=abs(spearman.loc[:,'DX'])
spear_DX1 = spear_DX.sort_values(ascending=False)
spear_DX1.to_csv('spearman correlation coefficient.csv',encoding = 'gbk',index =True)


''' 随机森林筛选变量 '''
forest = RandomForestRegressor(random_state = 123).fit(x_train, y_train)
print("train accurary:{:.4f}".format(forest.score(x_train,y_train)))
#train accurary:0.9206 训练的精度已经很高了，当然，有可能过拟合

features = x_train.columns
importances = forest.feature_importances_
features_random = pd.DataFrame()
features_random["特征"] = features
features_random['特征重要性'] = importances

features_random1 = features_random.sort_values(by="特征重要性", ascending=False)
features_random1.to_csv('Random Forest Results.csv',encoding = 'gbk',index = False)


#%% 方差分析
from statsmodels.formula.api import ols
from statsmodels.stats.anova import anova_lm
from statsmodels.stats.multicomp import pairwise_tukeyhsd

data_var = pd.read_csv('Outlier processed data.csv')
data_var.columns

formula = '''DX~
PTETHCAT+PTRACCAT+PTMARRY+CDRSB+ADAS11+ADAS13+MMSE+FAQ+MOCA+
                    EcogPtLang+EcogPtVisspat+EcogPtPlan+EcogPtOrgan+EcogPtTotal+
                    EcogSPVisspat+IMAGEUID+Ventricles+mPACCtrailsB+VISCODE+
                AGE+PTGENDER+PTEDUCAT+APOE4+AV45+ADASQ4+RAVLT_immediate+
                    RAVLT_learning+RAVLT_forgetting+RAVLT_perc_forgetting+LDELTOTAL
+TRABSCOR+EcogPtMem+EcogPtDivatt+EcogSPMem+EcogSPLang+EcogSPPlan+
                    EcogSPOrgan+EcogSPDivatt+EcogSPTotal+Hippocampus+WholeBrain+
                    Entorhinal+Fusiform+MidTemp+ICV+mPACCdigit'''
anova_results = anova_lm(ols(formula,data_var).fit())
anova_results.to_csv('results of ANOVA.csv',encoding = 'gbk',index = True)

#%% AD 诊断模型
data_m= pd.read_csv('model_input.csv')
Y = data_m.iloc[:,1].values
X = data_m.iloc[:,3:].values

# 划分训练集测试集
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size = 0.2, random_state = 0)

''' LSTM 模型 '''
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Dense, LSTM

# 模型搭建，参数配置
lstm=Sequential()
lstm.add(LSTM(units=160, return_sequences=True, input_shape=(X_train.shape[1],1)))
lstm.add(LSTM(50, return_sequences=False))
lstm.add(Dense(10))
lstm.add(Dense(1))

lstm.compile(loss='mean_squared_error', optimizer='adam')   # 配置训练方法

lstm.fit(X_train, Y_train, batch_size=2400, epochs=50, validation_data=(X_test, Y_test)) # 模型
训练
lstm.summary()

y_pred = lstm.predict(X_test)
Y_pred = y_pred.astype(int)

# 模型评价
from sklearn import metrics
from sklearn.metrics import precision_score, recall_score, f1_score
print('测试集查准率：{:.4f}'.format(metrics.precision_score(Y_test,Y_pred,average='micro')))
print('测试集查全率：{:.4f}'.format(recall_score(Y_test, Y_pred, average='macro')))
print('测试集 F1 值：{:.4f}'.format(f1_score(Y_test, Y_pred, average='macro')))
'''
测试集查准率：0.7958
测试集查全率：0.8628
测试集 F1 值：0.7432'''

test_pred=pd.concat([pd.DataFrame(Y_test), pd.DataFrame(Y_pred)],axis = 1)
test_pred.to_csv('AD diagnostic results.csv',encoding = 'gbk',index = True)

#SVM
from sklearn.tree import DecisionTreeClassifier
tree = DecisionTreeClassifier(random_state=0)
tree.fit(X_train, y_train)
print("Accuracy on training set: {:.3f}".format(tree.score(X_train, y_train)))
print("Accuracy on test set: {:.3f}".format(tree.score(X_test, y_test))) #精度 0.905
```

```
from sklearn.model_selection import train_test_split
X_train2, X_test2, y_train2, y_test2 = train_test_split(data.loc[:, data.columns != 'DX_bl'],
data['DX_bl'], stratify=data['DX_bl'], random_state=66)
tree.fit(X_train2, y_train2)
print("Accuracy on training set: {:.3f}".format(tree.score(X_train2, y_train2)))
print("Accuracy on test set: {:.3f}".format(tree.score(X_test2, y_test2))) #精度 0.745
```