Problem Chosen	2022	Team Control Number
C	ShuWei Cup Summary Sheet	2022111514190

Information-theoretic-based intelligent diagnostic model for AD

Alzheimer's disease, which mostly occurs in older people over the age of 64, develops with a gradual loss of independent living and death from complications. Therefore, the correct diagnosis and timely treatment of patients with mild cognitive impairment will play a crucial role in delaying the onset of AD.

For question 1, Regular expressions are used to correct the data, exclude illegal format data, and extract continuous information from some discrete features. The relationship between the null values of the data and the experimental period was observed, and the causes of the non-random missing data generation were deeply explored and processed. A prediction modeling method based on recursive idea is proposed, which can improve the data reduction by up to 24%.Discrete features are coded and the amount of change in each indicator for the same sample when tested twice replaces the original test indicator, and the impact of each indicator on the cause of Alzheimer's disease is discerned by the amount of information carried by each feature on health status, and the U-statistic between two health conditions.

For question 2, structural brain features, cognitive-behavioural features were extracted, and extremely strong missing associations between cognitive-behavioural features were observed, and missing associations between features were extracted by MissForest and predicted to be filled. Testing the characteristics of quantitative feature distribution by cumulative distribution function image and P-P plots.Outliers were removed using the 3 sigma principle. The predictions were fitted to the health types of the experimental samples, comparing the empirical and generalisation errors of various models, and XGBoost was selected for prediction modelling, And use the SHAP value to measure the marginal contribution of each feature.

For question 3, the indicators of the first participation of each experimental sample were extracted, and some of the features had strong co-collinearity, while the clustering model used spatial distance as the metric, and features with co-collinearity would have a greater impact on the clustering target, so for features with strong linear relationships, only five features with high information content for the health status of the experimental samples were retained for clustering modelling. Since feature selection mainly obeys the amount of information, the selected model should itself have the ability to extract information from the probability space of features, so the Self-organizing Map was selected for clustering.

For question 4, the time interval between the two tests for each experimental sample was calculated and the features of the amount of change in each indicator of the experimental sample constructed in problem one were extracted. The change in health status at the time of the two tests was compared and grouped, with a total of four categories for deterioration and five categories for improvement, the Tusky HSD multiple comparisons of the two conditions were carried out to evaluate the impact of various indicators on the changes of the disease.

For question 5, CN, SMC, EMCI, LMCI and AD, the main diagnostic criteria are that the patient will have memory loss, with the smallest degree of CN and the largest degree of AD and even loss of basic daily abilities, and a progressive increase in cognitive function, but that a series of interventions in the early stages, such as memory training, cognitive training, improved diet and appropriate exercise, will lead to the patient's condition improve or be brought under control.

Keywords： Information theory; Hypothesis testing;Probability space;Feature selection;Feature derivation

Content

1. Introduction

1.1 Background

Alzheimer's disease, also known as dementia, is a neurodegenerative disease characterised by memory impairment, aphasia, dyscognition, visuospatial skills, executive skills, personality and behavioural changes, and occurs in people over 64 years of age. AD is characterised by early mild cognitive impairment and late mild cognitive impairment in the early stages of the disease. Therefore, proper diagnosis and timely treatment of patients with mild cognitive impairment can play a vital role in slowing down the onset of AD.

1.2 Work

Question 1: It is required to pre-process the feature indicators given in the attached data and to determine the relevance of these data features to the diagnosis of Alzheimer's disease.

Question 2: A smart diagnosis of Alzheimer's disease was required to be devised by combining structural brain features and human action features.

Question 3: CN, MCI and AD were asked to be classified into three main categories, and the three sub-categories contained in MCI were then refined into three sub-categories based on the data characteristics by further clustering.

Question 4: The features collected were required to be analysed in relation to different time points and to reveal patterns of disease evolution over time for the different categories.

Question 5: A review of the relevant literature is required to describe the early intervention and diagnostic criteria for the five categories of CN, SMC, EMCI, LMCI and AD.

2. Problem analysis

2.1 Data analysis

This question focuses on the data indicators related to CN, MCI and AD symptoms, the development of an effective intelligent diagnostic model for Alzheimer's disease, the clustering breakdown of symptoms at different stages and the development of symptoms over time for each indicator.

2.2 Analysis of question one

For question 1, data correction was first carried out by regular expressions to remove data format anomalies caused by manual padding, followed by visualisation of the missing value distribution to remove features with more missing values, while for features containing only individual missing values, the rejection process was taken for missing samples. In order to find the association between the distribution of missing values and the test period, the absolute deviation of the distribution of missing values during the period was calculated to obtain the features for which the causes of missing values were associated with the test period. For the discrete category features for feature coding, considering the strong linear relationship between some of the features, a linear model: logistic regression and linear regression was attempted for prediction, but if the distribution of missing values of the strongly linearly related features were linked, it was obvious that the simple linear model could not handle the missing values carried in the independent variables, and in order to handle the discrete features, continuous features and missing values in the independent variables at the same time, a decision tree was used for missing value filling, and use recursive ideas to improve the model performance, for the decision tree fit poor features, temporarily not to deal with. The difference between the indicators of the comparison experiment and the difference between the indicators of the test before and after the calculation, replacing the original baseline value, so as to obtain the amount of change in indicators. The information of each characteristic indicator to DX, Gini index was calculated, and finally the significance test of CN, Dementia and MCI was performed by Mann-Whitney U test two by two.

2.3 Analysis of question two

For question 2, extract the feature indicators of brain structure and cognitive behavior based on the results of the first question, and fill in the missing values according to the distribution of missing values and correlation; observe the distribution of features, test the distribution of features for normal distribution, and if the features obey normal distribution, use the 3σ principle to eliminate outliers; according to the information content of each feature on DX, Gini index and other indicators and the results of the hypothesis test of the first question The features with low information content were eliminated according to the results of the first-question hypothesis test; the various models based on different metrics were compared, the optimal discriminant model was selected, the differences in the causes of Dementia, MCI and CN were analysed, the features with the highest causes were selected, and their interactive effects on CN, MCI and Dementia were analysed to intelligently diagnose Alzheimer's disease.

2.4 Analysis of question three

For question 3, the 25 features derived from problem 1 on the amount of variation in indicators were extracted, and in order to eliminate multicollinearity between the features, only those with a correlation coefficient greater than 0.7 were retained for those with a greater amount of information, and five features, Imageuid, Ventricles, Hippocampus, Fusiform and MidTemp, were selected for clustering in combination with statistical analysis. According to the conclusions of problems one and two, the amount of information carried by each feature varies greatly, and there is a certain probability association between each feature and the target value. The traditional linear clustering model may not be able to extract the hidden probability information carried by the data, and the SOM is considered to be used to simulate the different characteristics of the division of labour and response letter features of each neuron in different regions of the human brain, and to extract the geometric relationship between the features about probability for modelling. The data are first normalised and the first step is to cluster MCI, CN and AD, followed by subdividing MCI into three subclasses, LMCI, EMCI and SMC, observing the distance distribution of neurons for each step of classification, the two-dimensional projection of the dataset on the neural network and the sample size of each node of the network, and finally testing the sensitivity of neurons to feature taking values, and finally obtaining the sensitivity of each feature taking values to The final test is the sensitivity of the neurons to the feature values, and finally the degree of activation of each feature value on the neurons, so as to judge the impact of each indicator on the condition.

2.5 Analysis of question four

For question 4, the time difference between the two tests was first calculated by EXAMDATE and EXAMDATE_bl, while the degree of change in condition was determined based on the two characteristics DX_bl and DX, with a total of twelve types of changes in condition, a total of four categories of samples with improved condition and a total of five categories of samples with worsened condition, and an ANOVA was conducted between and within groups respectively, and finally the degree of change in condition produced by each characteristic was determined The analysis of variance (ANOVA) was conducted between and within groups to determine the effect of each characteristic on the change in condition.

2.6 Analysis of question five

For question 5, relevant medical literature was identified to summarise and describe early intervention practices and diagnostic criteria for the five categories of CN, SMC, EMCI, LMCI and AD.

3. Symbol and Assumptions

3. 1 Symbol Description

Symbol definition	Symbol Description
ρ	Pearson correlation coefficient
U	Mann-Whitney u test
r_s	spearman correlation coefficient
Ω	Regularization
TP	True Positive
FP	False Positive
TN	True Negative
FN	False Negative
Obj	Target function of XGB
α	Learning rate
x_i	Sample i
X_i	Feature i
$\hat{y}$	Predicted value of y
$Z_j^{'}$	Whether the feature j can be observed
θ_i	Attribution value of feature i
μ	Mean value
σ	Standard deviation

3.2 Fundamental assumptions

1. It is assumed that there is no misdiagnosis.

2. It is assumed that there are no accidents in the course of the experiment.

3.It is assumed that the subjects are in a stable state during the experiment and no adverse reactions occur during the data collection

4. Model

4.1 Question 1：Data pre-processing and correlation analysis models

4.1.1 Data pre-processing

Question 1 requires data pre-processing of the characteristic indicators of the attached data. The processing flow chart for pre-processing is shown in Figure 1.

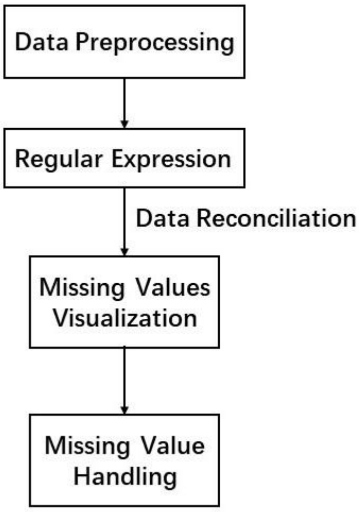

Figure1 Data pre-processing process

There are various causes of missing values, which can be divided into mechanical and human causes according to the means of collection. In addition to the null values in the dataset itself, there are also artificially populated values such as <NAN> as missing values, and the first step is to deal with the artificially generated null values.

As the features provided in the appendix data have a specific format, in order to ensure that each piece of data is valid, the first step in data pre-processing is to correct the data by means of regular expressions, so as to find outliers with specific format features and set them to null. For partially continuous data containing null values and artificially set null values, the values are extracted by means of regular expressions and set for some of the continuous data containing null values and artificially set null values.

The specific characters used are shown under Table 1.

Table 1 Basic elements used in regular expressions

Field	regular expression
PTID	\d{3}_[A-Za-z]{1}_\d{4}
Time Series Feature	\d{4}.\d{1,2}.\d{1,2}
Continuous feature	[-]?\d+[\.]?[\d+]?$

Constructing regular expressions is done in the same way as creating mathematical expressions. That is, multiple metacharacters and operators are used to combine small expressions together to create larger expressions[1].

After converting some of the artificially filled missing to null values, the overall distribution of missing values and the distribution of duplicate samples in the dataset were observed and the results of the statistics were as follows Table 2.

Table 2 Statistics of the data set

Number of Variables	116
Number of Rows	16222
Missing Cells	587341
Missing Cells (%)	31.2%
Duplicate Rows	0
Duplicate Rows (%)	0%
Total Size in Memory	32.5 MB
Average Row Size in Memory	2.1 KB
Variable Types	Numerical: 92 Categorical: 24

The overall distribution of missing values for the features in the dataset is shown in Figure 2 below：

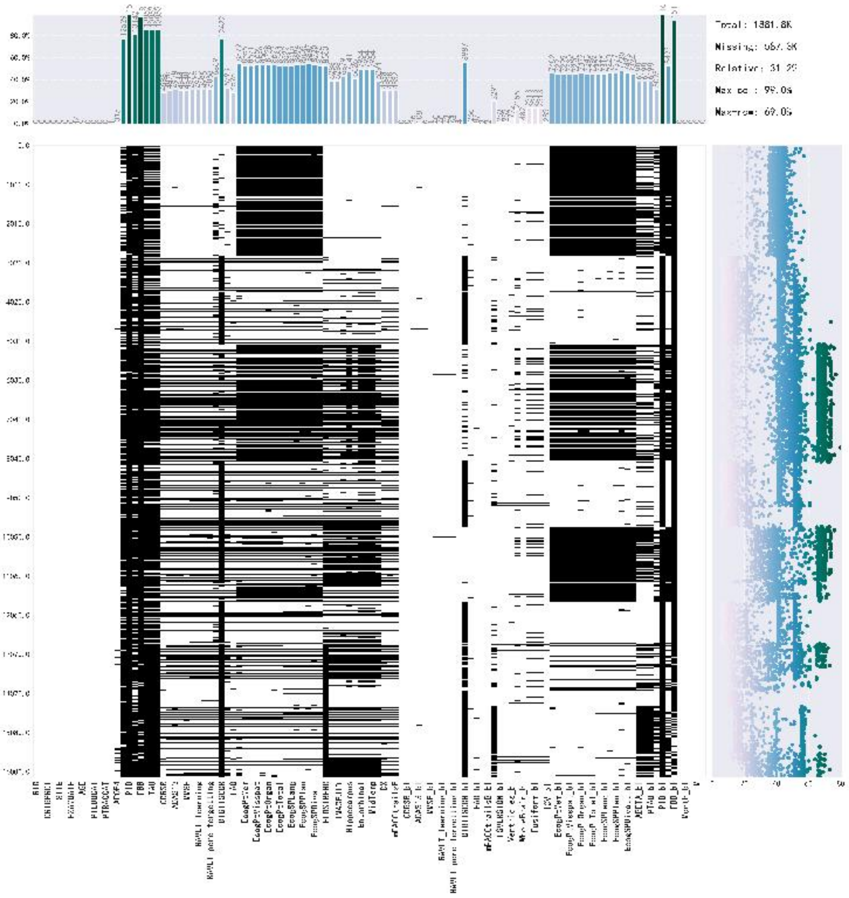

Figure 2 Distribution of missing feature values

Features with more than 70% of missing values were excluded, while samples containing only individual missing values (less than 1% of missing values) were excluded.

The missing values can be divided into completely random missing, random missing and completely non-random missing in terms of distribution, and the missing values between some features may have a connection, in order to explore the non-random missing relationship among them, the missing value connection of each feature indicator is shown in the form of a thermal Figure, as shown in Figure3 below, it can be seen that the bluer the colour represents the higher degree of positive correlation, and the missing values between features have a non-random missing relationship, and the redder the colour represents the higher degree of negative correlation, and the missing values between features present a random missing relationship.

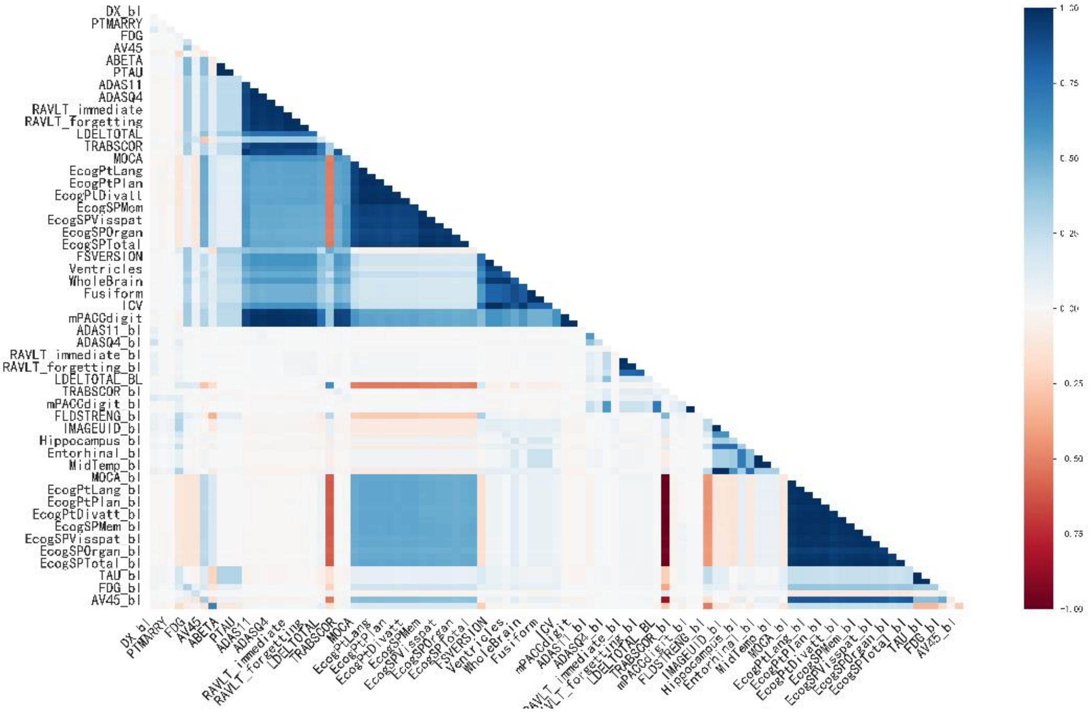

Figure 3 Characteristic indicators: Thermal Figure

This shows that there is an extremely strong non-random missing relationship between some of the features, showing a more obvious joint missing relationship. The cause of this may be related to the experimental period, artificial sampling or the condition.

The experiment was divided into four phases: ADNI1, ADNI2, ADNI3 and ADNIGO. In order to verify whether the missing data were related to the experimental period, the overall distribution of missing values in the four phases was first observed, and the pie charts of the percentage of the four phases are shown in Figure 4 below. ADNI1, ADNI2, ADNI3 and ADNI1GO correspond to orange, blue, green and red respectively, with a percentage of 30.90%, 42.81%, 21.33%, and 4.96%.。

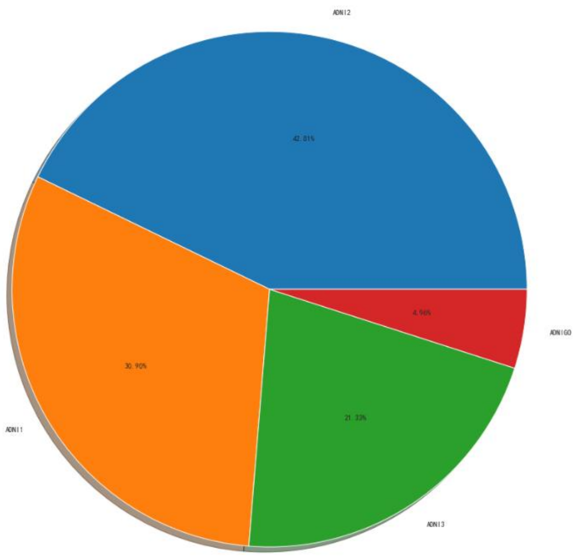

Figure 4 Pie chart of the number of experimental periods as a percentage

Assuming that a feature indicator was added or removed because of a different experimental period, the feature should be completely empty for a given period, with a total of 32 missing values of 1 for the period. The absolute deviation of the percentage of missing values between weeks was calculated, and those with little difference in the percentage of missing values were excluded, using 0.2 as the threshold.

This resulted in 31 feature indicators being added in the Phase 2 experiment (ADNIGO) and one feature being added in the Phase 4 experiment (ADNI3).

Table 3 Missing Values Phase IV Experiment Data Statistics Table (partial)

	ADNI1	ADNI2	ADNIGO	ADNI3
MOCA	1.000000	0.350468	0.419154	0.316474
EcogPtMem	1.000000	0.339381	0.421642	0.244509
⋮	⋮	⋮	⋮	⋮
EcogSPTotal_bl	1.000000	0.223470	0.527363	0.096821
AV45_bl	1.000000	0.226782	0.531095	0.427168

4.1.2 Relevance analysis

In order to explore the characteristics of the relationship between each characteristic indicator and thus derive the correlation between the characteristic indicators and the diagnosis of Alzheimer's disease, this paper uses Pearson correlation coefficients to describe the correlation between the characteristic indicators.

With N data pairs (x_i , y_j), (i=1, 2, ..., n), the Pearson correlation coefficient ρ is expressed as:

$$\rho = \frac{\sum_{i=1}^{N}(x_i - \frac{1}{N}\sum_{j=1}^{N}x_j)(y_i - \frac{1}{N}\sum_{j=1}^{N}x_j)}{\sqrt{\sum_{i=1}^{N}(x_i - \frac{1}{N}\sum_{j=1}^{N}x_j)^2}\sqrt{\sum_{i=1}^{N}(y_i - \frac{1}{N}\sum_{j=1}^{N}y_j)^2}} \tag{1}$$

The range of ρ is -1≤ρ≤1. The positive and negative values of ρ represent whether the two variables change in the same direction, i.e. ρ>0 means the two variables are positively correlated, i.e. when one group of variables increases or decreases, the other group of variables also increases or decreases; ρ<0 means the two variables are negatively correlated, i.e. the two variables change in opposite directions. The closeness of the relationship between the two is reflected in the value of ρ. A larger ｜ρ｜ means a stronger correlation: ｜ρ｜ =1 means the two variables are perfectly correlated; 0.95<｜ρ｜<1 means the variables are significantly correlated; ｜ρ｜=0 means the two groups of variables are independent of each other and not correlated[2]。

The linear relationship between the features is shown in Figure 5 below.

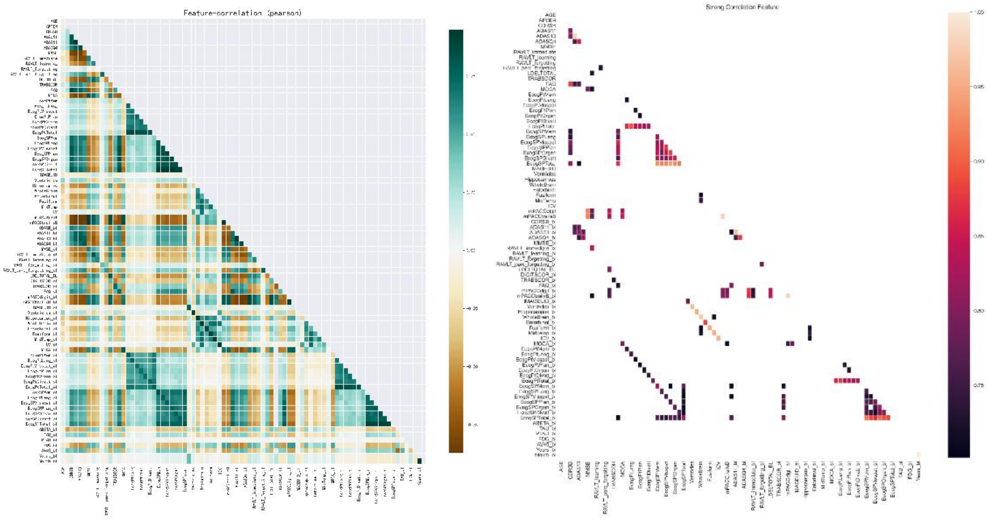

Figure 5 Linear correlation coefficient

The percentage of target features (DX) missing in each phase of the experiment is shown in Figure 9 below, with 23.66%, 48.16%, 21.38% and 6.80% for ADNI1, ADNI2, ADNI3 and ADNIGO respectively.

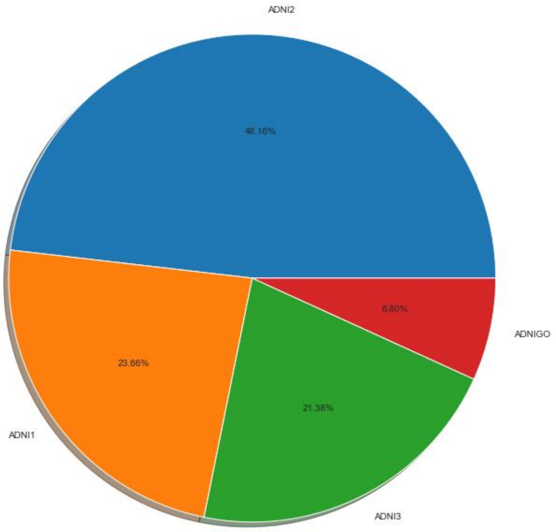

Figure 6 Pie chart of DX share of experiments in each period

Next, the features were encoded, i.e. the discrete features were made continuous, the values of binary attributes such as PTGENDER and PTETHCAT were replaced by 0 and 1, and the values of the representational values contained in the features were still extracted with regular expressions and replaced with the original sample values (e.g. if the value in the feature FSVERSION is n Tesla MRI, then the original sample values can be replaced with n values). The format of update_stamp is guessed as time difference: since the data format is number1:number2, and number2 < 60, max(number2) > 24, number2 should be seconds or minutes data, number1 is minutes data or hours data, which is 60 decimal, and its continuousization is number1 * 60 + number2, for other high base class unordered features, for the time being, the sequence code is used instead.

For data showing a strong linear relationship, a linear model was attempted to fill in the predictions; data categories less than 100 were judged to be discrete features and were filled in with predictions using logistic regression; those greater than 100 were judged to be continuous features and were predicted using linear regression.

It can be seen that the prediction accuracy of most of the features is above 0.7, and there are 11 prediction accuracies above 0.99 in the test machine. In the process of trying to predict the filling, since such linear models cannot ignore the missing values of other features, for example, it is known from Figure 7 that features a and b show a strong correlation, and when trying to predict the null of b with a, a also contains null values, at this time, it is impossible to complete the The prediction is not completed and therefore only part of the data is filled in.

Therefore, if the missing values are filled by this method, the model should be insensitive to the missing values, while the decision tree has this feature. In order to achieve better and more accurate filling, and in the filling of the linear model, it was found that some strongly correlated features have similar distribution of missing values, and if the method of using non-missing features to predict features containing missing values is used directly, it is very easy to produce a part of important information lost and the model will be lost due to The accuracy of the model declined significantly due to the neglect of strongly correlated features, so a prediction method based on the recursive idea was adopted: the

1. first ranking the model performance capabilities of each missing feature in descending order.
2. populating round by round in that order.
3. starting with the second round, adding to the sample in each round the features for which the predictive fill was completed in the previous round.

If there are initially m features without missing values and n features with missing values, the original method uses m features to fit the target features in each round, while the method contains $m+i-1$ features in the ith round of training data, expanding the sample size, using the method, and the first to predict the features with high accuracy, which can better improve the prediction model for features with low accuracy, and in general, the original model score The lower the ranking, the more obvious the improvement, in this dataset, the lowest score in the original score > 0.7 is mPACCdigit, the accuracy of 0.716182, and the most obvious improvement, after using the method, the goodness of fit improved to 0.956825, at this time there are 29 features due to poor fit, do not deal with for the time being.

The information content of each feature for DX, the Gini index, was calculated and the Mann-Whitney u test was performed two-by-two for the three lesions (CN,Dem,Mci) with the following formula.

$$U = n_1 n_2 + \frac{n_1(n_1+1)}{2} - W_1 \tag{2}$$

The predicted data were tested in terms of the degree of uncertainty in the data, the probability of a randomly selected sample being misclassified, and the two independent samples rank sum test, some of which are shown in Table 4 below, and the full results of the test are shown in the attached stats.csv.

Table4 Statistical table of test prediction results (partial)

	iv	gini	⋯	Mann_CN_Mci_Stats	chi2
mPACCtrailsB	9.339392	0.511739	⋯	34014378	7.72E-193
mPACCdigit	9.010023	0.511545	⋯	34136462	3.42E-117
⋮	⋮	⋮	⋱	⋮	⋮
DX_bl			⋯	7919298	0
EXAMDATE_bl			⋯	20102765	0.001279

4.2 Question 2 Intelligent diagnostic model for Alzheimer's disease

4.2.1 Data pre-processing

Combined with the pre-processing results of question 1, the distribution of missing values was observed by extracting the structural brain features and cognitive-behavioural features, see Figure 10 and Figure 11, it can be found that there are still more features containing missing values, and the distribution of missing values of cognitive-behavioural features has a very strong correlation, with some of the missing values correlating above 0.7. In response to the strong correlation of this type of data, MissForest was used to fill in the predictions for the missing values.

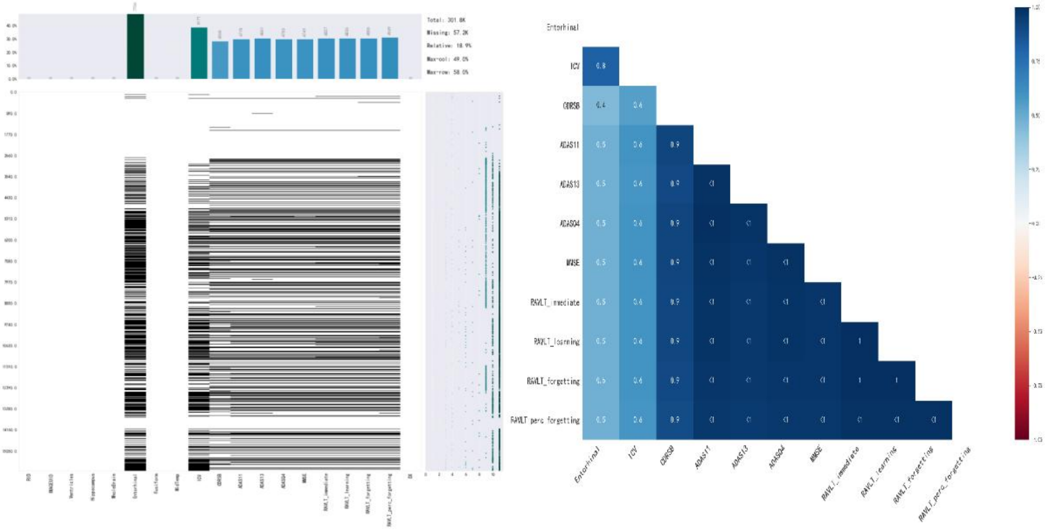

Figure 7 Missing value distribution and links

Next, Spearman correlation coefficient is used to explore the correlation between brain structural characteristics and cognitive behavioral characteristics. The formula is as follows:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2 - 1)} \tag{3}$$

By observing the joint distribution of Spearman coefficient, brain structure and cognitive behavior characteristics among the characteristic indicators, see Figure 8., It is found that some features have correlation, and the distribution of features with strong correlation of Spearman coefficient is observed. It is found that ADAS11 and ADAS13 almost completely show a one-dimensional linear relationship.

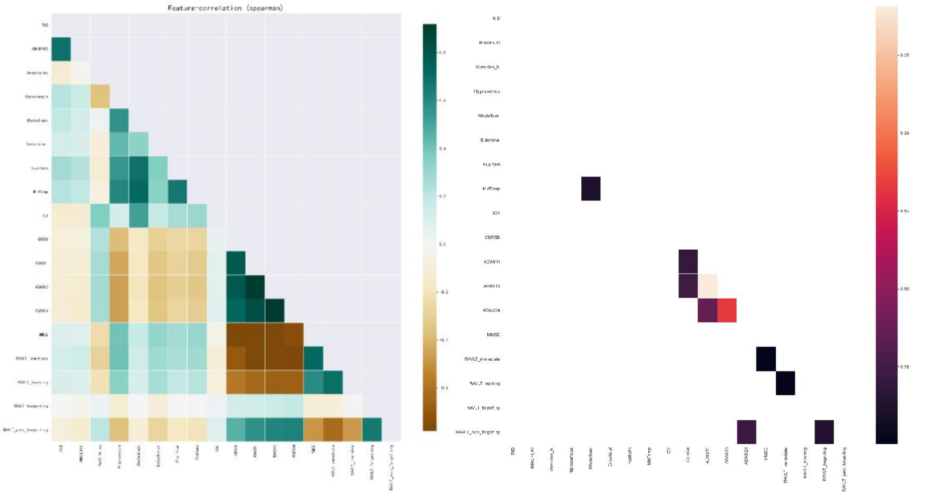

Figure 8 Spearman coefficient heat map

From the above figure, it can be seen that there is no excessive linear association between structural brain features and cognitive-behavioural features, while there may be mutual influence and interrelationship among the features within structural brain features and cognitive-behavioural features, and the joint distribution of features within each group of structural brain features and cognitive-behavioural features is observed separately.

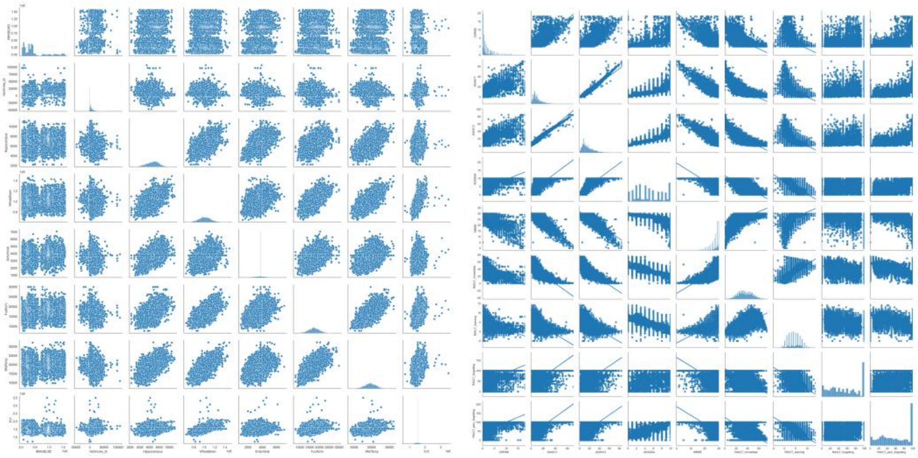

Figure 9 Joint distribution Structural brain features (left) Cognitive behavioural features (right)

During the exploratory data analysis phase in question one, it was found that the distribution of the above indicators might show a normal distribution, the characteristic cumulative distribution was observed and the conjecture was verified by means of a P-P diagram.

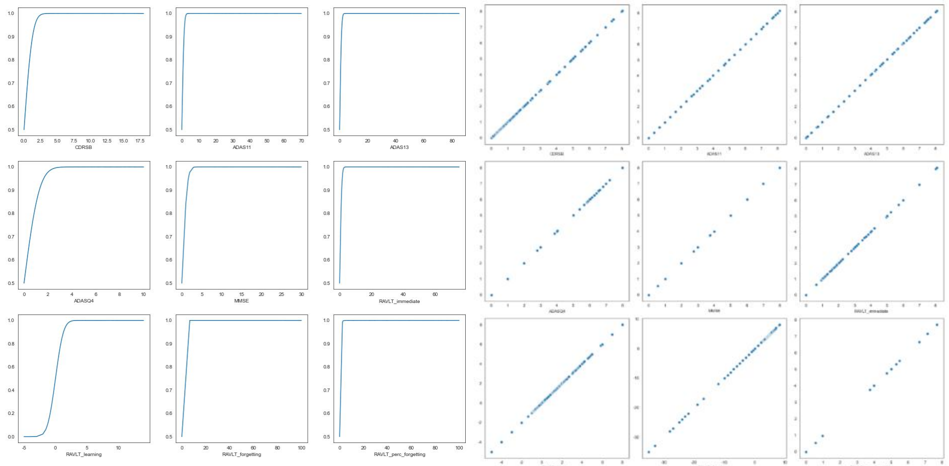

Figure 10 Cumulative distribution chart (left) P-P chart (right)

The joint distribution of strongly correlated features is shown in Figure 11.

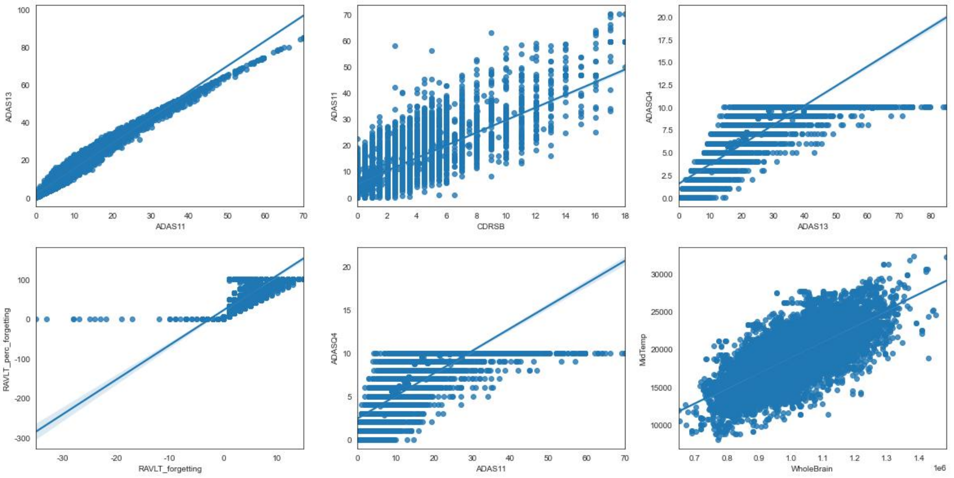

Figure 11 Map of the distribution of indicators of strong correlation characteristics

The RAVLT_perc_forgetting should be a percentage number with a distribution interval of [0, 100], which appears to be partially negative, and the negative numbers are replaced with 0. Also, the characteristic indicators that obey the normal distribution are excluded from the sample using the 3sigma principle.

Based on the results of the hypothesis test in question 1, the ICV, WholeBrain and other indicators were excluded because they accepted the original hypothesis (i.e., they were not significant) and had a low information content.

4.2.1 Intelligent diagnostic models

The accuracy, precision, recall and F1 score of probability-based models (Bayse), distance-based models (SVM), information-theoretic-based models (decision trees) and integrated models (RF, XGB, GBDT) are calculated as follows

$$
\begin{aligned}
acc &= \frac{(TP+TN)}{(TP+TN+FP+FN)} \\
pre &= \frac{TP}{(TP+FP)} \\
recall &= \frac{TP}{(TP+FN)} \\
F1 &= 2 * \frac{pre * recall}{pre + recall}
\end{aligned}
\tag{4}
$$

Comparative data is shown in Table 5 below.

Table 5 Comparison table of discriminant model data

	Bayes	SVM	DT	RF	GBDT	XGBoost
accuracy	0.641898	0.791406	0.773724	0.853850	0.833483	0.878245
F1	0.626951	0.791147	0.774055	0.853916	0.833631	0.878242
Precision	0.663245	0.791437	0.774585	0.856693	0.836584	0.879591
Recall	0.641898	0.791406	0.773724	0.853850	0.833483	0.878245

As can be seen from the table above, the XGBoost model has an accuracy of 87.8245% and an F1 value of 0.878242 when comparing accuracy, F1, precision and recall, thus showing that the intelligent diagnostic model based on XGBoost is superior to other models and more suitable for intelligent diagnosis of Alzheimer's disease. XGBoost [3] is a specific implementation of the Fradient Boosting method that uses a more accurate approximation to find the best tree model. the core of XGB is an integrated algorithm based on the Gradient Boosting Tree (GBDT) implementation[4] . the objective function of XGB is：

$$Obj=\sum_{i=1}^{n} l(y_i,\hat{y}_j)+\sum_{k=1}^{K}\Omega(f_k) \quad (5)$$

The first part of the right-hand side of the equation is the training error of the model, and the second part of the equation is the regularization term. the XGBoost counterpart contains multiple cart trees, defining the complexity of each tree.

$$\Omega（f）=\gamma\mathrm{T}+\frac{1}{2}\lambda\|\varpi\|^2 \quad (6)$$

SHAP is one of the best methods used to measure the degree of marginal contribution of features, and is used in this paper to evaluate the final selected XGBoost model, observing the difference between Dementia and MCI causes, and the percentage of each feature indicator is shown in Figure 12 below, which shows that CDRSB has the highest marginal contribution to the model prediction.

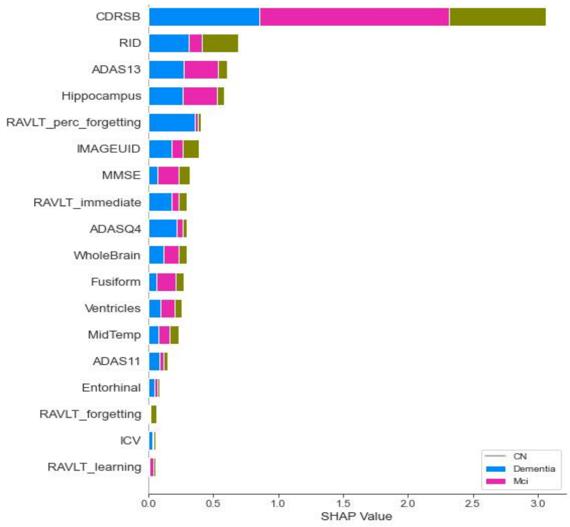

Figure 12 Share of characteristic indicators

CDRSB was selected to explore the interaction effects on CN, MCI, and Dementia.

The distribution of the correlation effects of CDESB on CN, Dem, and MCI respectively are shown in Figure 13, Figure 14, and Figure 15 below. SHAP interprets the predicted value of the model as the sum of the attributed values of each input feature, i.e. the personalised feature attribution result, which is formulated as follows.

$$g(Z^{'}) = \varphi_0 + \sum_{j=1}^{M} \varphi_j Z_j^{'} \tag{7}$$

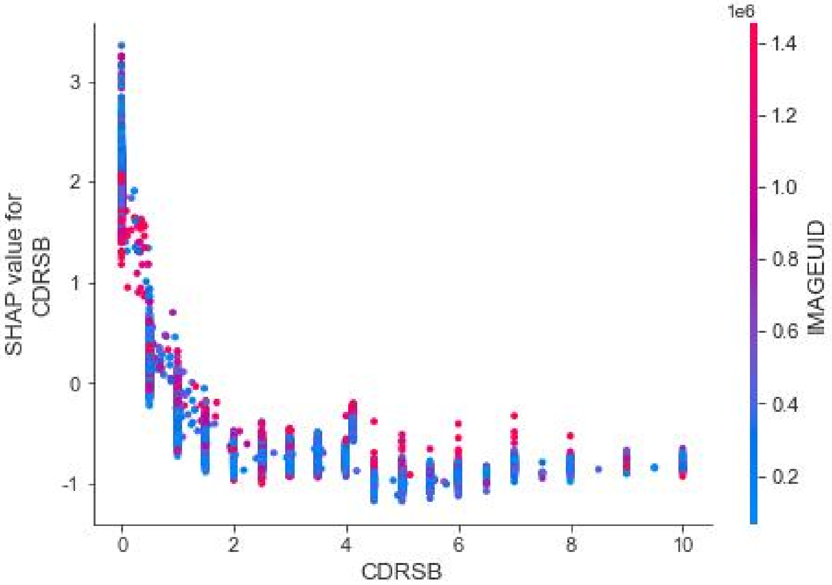

Figure 13 CDRSB_CN_Dependence

From Figure 13, it can be found that when CDRSB is 0, the greater the effect on CN, and as CDRSB grows, the interactive effect of CDRSB on CN then decreases, indicating a decreasing relationship of CDRSB for the normal elderly (CN).

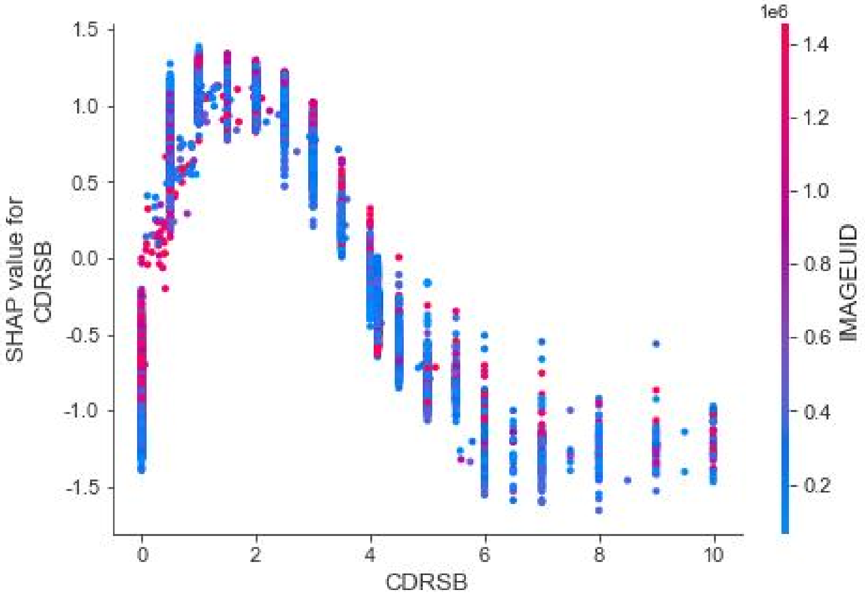

Figure 14 CDRSB_AD_Dependence

As can be seen in Figure 14, the greater the effect on AD when the CDRSB is 2, after which the interactive effect of CDRSB on AD decreases as the CDRSB increases, indicating that the probability of being diagnosed with Alzheimer's disease (AD) is greatest at a CDRSB of 2.

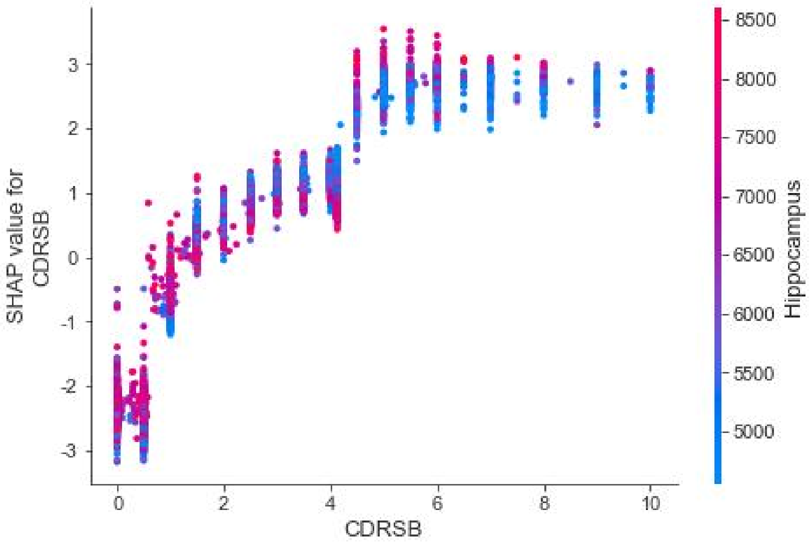

Figure 15 CDRSB_Mci_Dependence

As can be seen in Figure 15, the greater the effect on MCI when the CDRSB is 5, the greater the effect on MCI as the CDRSB grows before this point, and the decrease in the effect on MCI as the CDRSB grows after this point, indicating that the probability of being diagnosed with mild cognitive impairment (MCI) is greatest when the CDRSB is 5. .

Meanwhile, IMAGEUID, Hippocampus was mainly distributed with CDRSB in the [0,2] interval and, in general, SHAP values decreased with increasing IMAGEUID.

4.3 Question 3: Clustering models for different categories of disease conditions

4.3.1 Data pre-processing

In problem 1 in the construction of features phase, new RAVLT_perc_forgetting, RAVLT_perc_forgetting_bl for percentage data, the interval is [0,100], replace some of the data where underflow occurs with 0 for prediction filling, and also replace the original baseline features with the difference value of the comparison test, i.e. the bl value of the feature indicator is the amount of change of each indicator of the comparison test.

4.3.2 Feature extraction

Firstly, for the symptoms of baseline, brain structure features with suffix bl, cognitive-behavioural features and some of the features of the comparison test were extracted, a total of 25. From the conclusions of questions 1 and 2, it can be seen that some of the features have extremely strong linear relationships, and in order to remove the multiple covariance among the features, the amount of information carried by each feature indicator for the brain state was calculated, and for the strongly correlated features, i.e. the correlation coefficient is greater than 0.7, only the features carrying a larger amount of information are retained, and after eliminating features with low correlation, the heat map of the correlation coefficient matrix of each feature is shown in Figure 16 below.

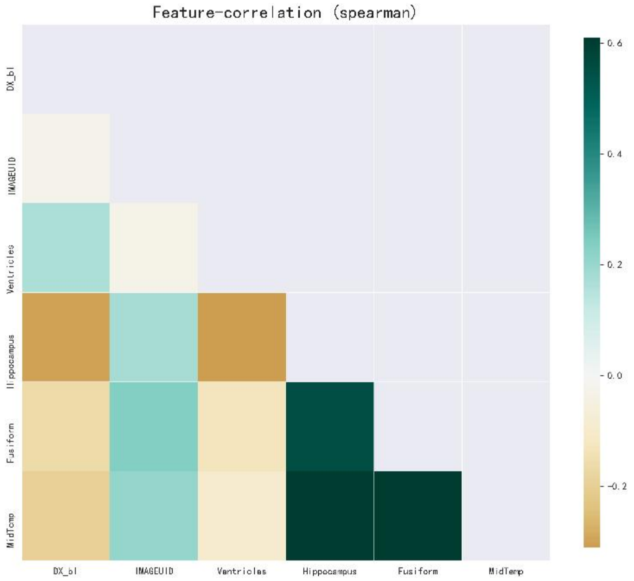

Figure 16 Correlation coefficient heat map

Also based on the statistics of each feature, the final five features Imageuid, Ventricles, Hippocampus, Fusiform and MidTemp were selected for clustering modelling. To ensure the accuracy of the results, the correlation between the features was observed and the multicollinearity was basically eliminated.

4.3.3 SOM clustering prediction

It was observed that most of the feature distributions were skewed towards a normal distribution, and in order to eliminate the influence between different magnitudes, the data were normalised and the LMCI, SMC and EMCI in DX_bl were tentatively grouped into MCI classes.

$$x^{'} = \frac{x-u}{\sigma} \tag{8}$$

The overall modelling in this paper is based on information theory. As can be seen from problem one and two, the amount of information carried by each feature varies greatly, so the influence of information probability on the clustering categories should be taken into account when modelling clustering.

SOM requires no supervision, has a high learning capacity and can automatically classify input patterns. It mainly uses the idea of dimensionality reduction to map high-dimensional input data into a one- or two-dimensional low-dimensional space for decision making [5]. Figure 17 shows a schematic diagram of the SOM network structure.

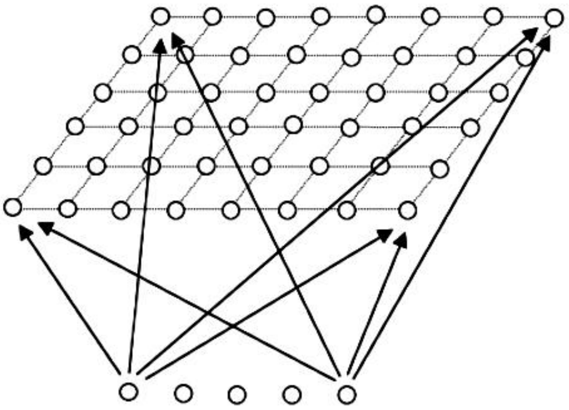

Figure 17 Schematic diagram of the SOM network structure

The SOM was selected for clustering prediction, comparing three different types of clustering algorithms, K-Means, K-Modes and FCM, and the data for the classification performance evaluation metrics of the three models are shown in Table 6, Table 7 and Table 8 below.

Table6 K-means classification performance evaluation metrics

	precision	recall	f1-score	support
0	0.32	0.22	0.26	4806
1	0.53	0.40	0.46	9427
2	0.27	0.86	0.41	1650
accuracy			0.40	15883
macro avg	0.37	0.50	0.38	15883
weighted avg	0.44	0.40	0.39	15883

Table7 K-modes classification performance evaluation metrics

	precision	recall	f1-score	support
0	0.31	1.00	0.47	4806
1	0.84	0.01	0.02	9427
2	0.01	0.00	0.00	1650
accuracy			0.31	15883
macro avg	0.39	0.34	0.16	15883
weighted avg	0.59	0.31	0.15	15883

Table8 SOM classification performance evaluation indicators

	precision	recall	f1-score	support
0	0.98	0.84	0.90	4806
1	0.91	0.99	0.95	9427
2	1.00	0.88	0.93	1650
accuracy			0.93	15883
macro avg	0.96	0.90	0.93	15883
weighted avg	0.94	0.93	0.93	15883

The correlation accuracy improved by 138.462%, 520.0% and 144.737% year-on-year, with the final SOM having an average F1 score of 0.93, an average recall of 0.93, an average precision of 0.94 and an accuracy of 0.93, giving the best model results.

After normalizing the current input pattern vector X in the self-organizing network, the corresponding inner star power vector of each neuron in the competitive layer is given by the following equation.

$$\hat{X} = \frac{X}{\|X\|}, \hat{W}_j = \frac{W_j}{\|W_j\|} \tag{9}$$

Only the winning neuron is entitled to adjust its weight vector, and the weight vector learning adjustment is as follows.

$$y_i(t+1) = \left\{ \begin{matrix} 1, j = j^* \\ 0, j \neq j^* \end{matrix} \right\} \tag{10}$$

Only the winning neuron is entitled to adjust its weight vector, and the weight vector learning adjustment is as follows.

$$\left\{ \begin{matrix} W_{j^*}(t+1) = \hat{W}_{j^*}(t) + \Delta W_{j^*} = \hat{W}_{j^*}(t) + \alpha(\hat{X} - W_{j^*}), j \neq j^* \\ W_j(t+1) = W_j(t) \end{matrix} \right\} \tag{11}$$

0<α≤1 is the learning efficiency, and α generally decreases as learning progresses in multiple dimensions, i.e. the degree of adjustment becomes smaller and smaller, tending to the centre of the clusters.

The visualisation of the weight matrix after SOM clustering is shown in the U-Matrix matrix, as shown in Figure 18 below, and its distance is calculated using the Euclidean distance with the following formula.

$$\|X - X_i\| = \sqrt{(X - X_i)^T (X - X_i)} \tag{12}$$

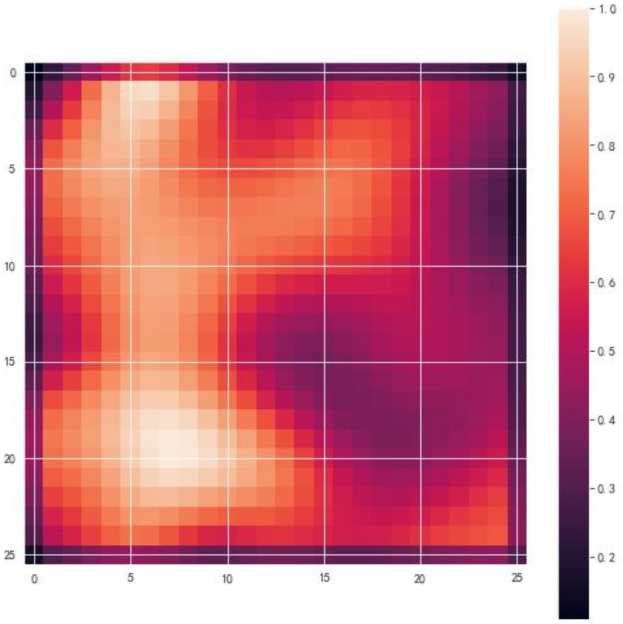

Figure 18 Visualisation of the weighting matrix

As can be seen from the figure, the SOM is more densely distributed on the left side of the neurons and the data is probably mostly distributed on the right side.

The MCI was subdivided into three subclasses, and the data analysis resulted in the three subclasses of MCI as shown in Figure 19 below.

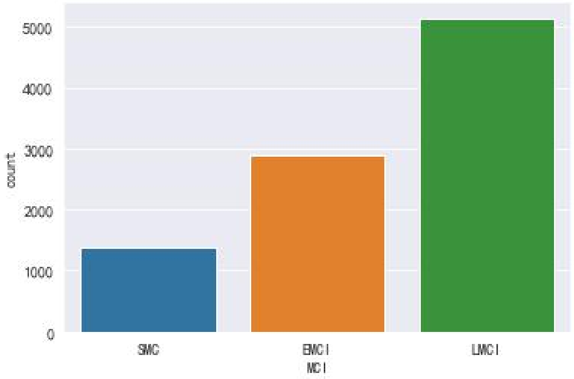

Figure 19 Histogram of the distribution of the three MCI sub-categories

After clustering again using the SOM and observing the individual neuron distance distribution, the U-Matrix matrix is shown in Figure 20 below, from which it can be seen that the neurons are mainly distributed in the upper left and lower right nodes, so the data may be projected mainly around the right diagonal.

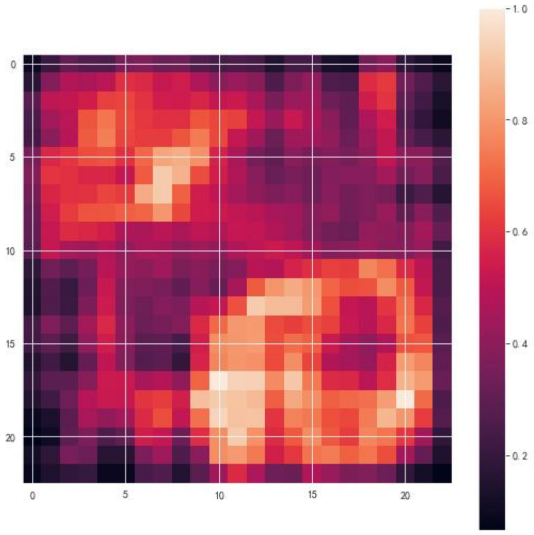

Figure 20 Visualization of SOM secondary clustering weight matrix

A two-dimensional projection of the dataset onto the neural network is shown in Figure 21below, and the exact number of samples for each node in the network is shown in the Appendix.

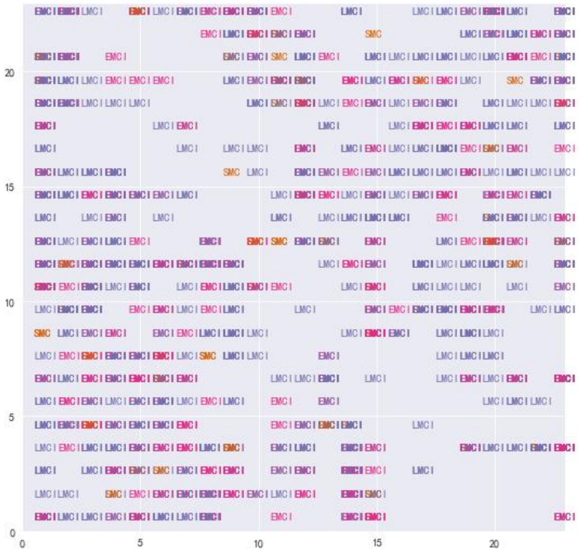

Figure 21 SOM to MCI secondary segmentation distribution map

To test the sensitivity of the neurons to the individual feature taking values, the right diagonal has a significantly deeper colour depth than the left diagonal, with a sparser distribution of data in the top left and bottom right corners.

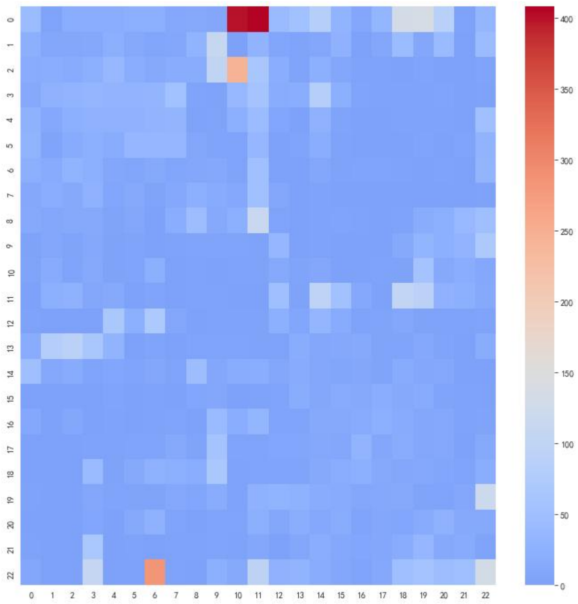

Figure 22 Data node distribution

The neuronal sensitivities of the five features Imageuid, Ventricles, Hippocampus, Fusiform, and MidTemp are shown below.

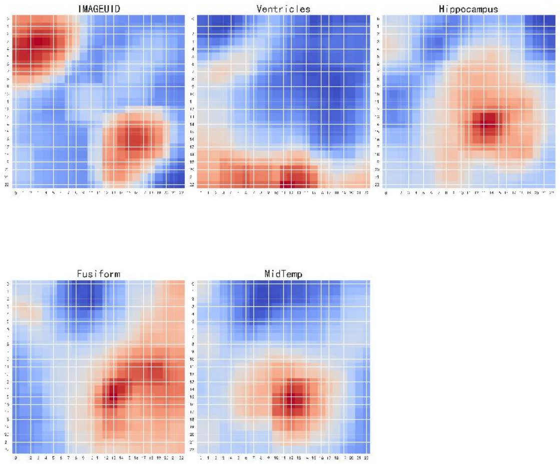

Figure 23 Neuronal sensitivity

As can be seen from the figure, Ventricles works best for clustering, with the three categories being more evenly distributed in the lower left, diagonal and upper right, followed by Hippocampus and MidTemp with one category more evenly distributed in the central region, two other categories spread out from the periphery of the central region, and IMAGEUID with a more scattered distribution.

4.4 Question 4: Modelling changes over time for different categories of diseases

4.4.1 Disease Change Grouping

Based on the original test result (DX_bl), the retest result (DX), the original test date (EXAMDATE_bl) and the retest date (EXAMDATE), the time difference between the two test times, EXAMDATE and EXAMDATE_bl, allows DX_bl and DX to be stitched together into 13 condition evolution results.

These 13 disease evolutions were further divided into three groups, i.e. four conditions, SMC_CN, AD_MCI, EMCI_CN and LMCI_CN, for the improving group; five conditions, LMC_AD, EMCI_AD, SMC_AD, CN_AD and CN_MCI, for the deteriorating group; and four conditions, CN_CN, SMCI_MCI, LMCI_MCI and AD_MCI, for the stable group. _MCI and AD_AD.

4.4.2 Multi-factor analysis of variance

Based on the discussion of the effect of each characteristic on the change in condition at different time points, the five conditions that improved and the nine conditions that worsened were selected for intra-group and inter-group ANOVAs, so that the effect of each characteristic on the degree of improvement and the degree of worsening could be determined within the group and the significant effect of the characteristic on whether the condition worsened or improved could be determined between groups.

Multi-factor ANOVA is used to determine whether a dependent variable is influenced by more than one independent variable (also known as a factor), and it tests whether there is a significant difference between the means of the dependent variable between different combinations of the levels at which multiple factors are taken.

Multi-factor ANOVA decomposes the total variance of the observed variables into: (using two control variables as an example).

$$SST = SSA + SSB + SSAB + SSE \tag{13}$$

SST represents the total variation of the observed variable and its expression is:

$$\mathrm{SST}=\sum_{i=1}^{\mathrm{k}}\sum_{j=1}^{r}\sum_{k=1}^{n_{ij}}(x_{ijk}-\overline{x})^2 \tag{14}$$

SSA and SSB represent the variation caused by the independent action of the control variables A and B, respectively, and their expressions are:

$$\text{SSA}=\sum_{i=1}^{\text{k}}\sum_{j=1}^{r}n_{ij}(\overline{x}_i^A-\overline{x})^2$$
$$\text{SSB}=\sum_{i=1}^{\text{k}}\sum_{j=1}^{r}n_{ij}(\overline{x}_i^B-\overline{x})^2 \tag{15}$$

SSE is the variance due to random factors and is expressed as：

$$\text{SSE}=\sum_{i=1}^{\text{k}}\sum_{j=1}^{r}\sum_{k=1}^{n_{ij}}(x_{ijk}-\overline{x}_{ij}^{AB})^2 \tag{16}$$

SSAB is an explainable variation of the interaction.

In multivariate ANOVA, an F-number is used to test for significance.

$$F=\frac{SSE}{SSA} \tag{17}$$

4.4.3 Tusky HSD multiple comparison ANOVA

Turkey's HSD is based on pairwise comparisons of studentised extreme differences. By calculating the HSD statistic, if the difference between the two group means is greater than this extreme difference, the difference is considered to be significant and therefore the null hypothesis is rejected and the two group means are considered to be different. The formula for calculating the critical HSD was：

$$HSD=q_\alpha(k,v)\sqrt{\frac{S_c^2}{n}} \tag{18}$$

4.4.3.1 Perform within-group ANOVA

In the better group, features CDRSB_bl,AGE,ADAS11_bl,ADAS13_bl, ADASQ4_bl,MMSE_bl,RAVLT_immediate_bl,RAVLT_forgetting_bl,RAVLT_perc_forgetting_bl,TRABSCOR_bl,FAQ_bl,mPACCdigit_bl,mPACCtrailsB_bl 、 IMAGEUI-D_bl、Ventricles_bl、Hippocampus_bl、WholeBrain_bl、Entorhinal_bl、Fusiform_bl、MidTemp_bl、MOCA_bl、EcogPtTotal_bl、EcogSPTotal_bl、timestamp showed significant differences between the groups of the 24 sets of data; There were no significant differences between the three groups of data RAVLT_learning_bl, ICV_bl, and Month_bl.

Table 9 Intra-group variability by characteristics of disease improvement

Feature	Difference
CDRSB_bl	LMCI_CN < EMCI_CN <= AD_MCI < SMC_CN
AGE	EMCI_CN < SMC_CN <= LMCI_CN < AD_MCI
ADAS11_bl	LMCI_CN < EMCI_CN < SMC_CN <= AD_MCI
ADAS13_bl	LMCI_CN < EMCI_CN < SMC_CN < AD_MCI
ADASQ4_bl	LMCI_CN <= EMCI_CN < SMC_CN <= AD_MCI

MMSE_bl	EMCI_CN <= SMC_CN <= AD_MCI <= LMCI_CN
RAVLT_immediate_bl	AD_MCI <= SMC_CN < LMCI_CN <= EMCI_CN
RAVLT_forgetting_bl	EMCI_CN <= SMC_CN < LMCI_CN < AD_MCI
RAVLT_perc_forgetting_bl	LMCI_CN < EMCI_CN <= SMC_CN <= AD_MCI
TRABSCOR_bl	EMCI_CN <= LMCI_CN <= SMC_CN < AD_MCI
FAQ_bl	LMCI_CN <= EMCI_CN <= SMC_CN < AD_MCI
mPACCdigit_bl	AD_MCI <= SMC_CN < EMCI_CN < LMCI_CN
mPACCtrailsB_bl	AD_MCI <= SMC_CN < EMCI_CN < LMCI_CN
……	……

In the analysis of variance between groups in the worse group, it was found that the features were not significant between the columns of the TRABSCOR_bl and Month_bl groups, and the other groups showed significance. Table 10 below, see annexes solve5_info_bader.txt and solve_5_bader_ANOVA.xlsx for details.

Table 10 Intra-group variability by characteristics of disease deterioration

Feature	Difference
CDRSB_bl	CN_MCI < EMCI_AD <= LMCI_AD <= SMC_AD <= CN_AD
AGE	EMCI_AD <= LMCI_AD <= SMC_AD <= CN_MCI <= CN_AD
ADAS11_bl	CN_MCI < LMCI_AD <= SMC_AD <= EMCI_AD < CN_AD
ADAS13_bl	CN_MCI <= SMC_AD <= LMCI_AD <= EMCI_AD < CN_AD
ADASQ4_bl	LMCI_AD <= CN_MCI <= SMC_AD <= EMCI_AD < CN_AD
MMSE_bl	CN_AD < LMCI_AD <= SMC_AD <= EMCI_AD < CN_MCI
RAVLT_immediate_bl	CN_AD <= SMC_AD <= EMCI_AD <= LMCI_AD <= CN_MCI
RAVLT_forgetting_bl	SMC_AD <= CN_AD <= CN_MCI < EMCI_AD < LMCI_AD
RAVLT_perc_forgetting_bl	LMCI_AD <= EMCI_AD < CN_MCI <= SMC_AD <= CN_AD
TRABSCOR_bl	SMC_AD <= LMCI_AD <= CN_MCI <= EMCI_AD <= CN_AD
FAQ_bl	CN_MCI <= SMC_AD <= EMCI_AD <= LMCI_AD < CN_AD
mPACCdigit_bl	CN_AD <= SMC_AD <= EMCI_AD <= LMCI_AD < CN_MCI
mPACCtrailsB_bl	LMCI_AD < EMCI_AD <= CN_MCI < CN_AD <= SMC_AD
……	……

4.4.3.2 Perform analysis of variance between groups

Analysis of variance between the better group and the worse group of the disease showed that CDRSB_bl, AGE, ADAS11_bl, ADAS13_bl, ADASQ4_bl, MMSE_bl, RAVLT_immediate_bl,RAVLT_forgetting_bl,RAVLT_perc_forgetting_bl, TRABSC-OR_bl, FAQ_bl, mPACCdigit_bl， mPACCtrailsB_bl， IMAGEUID_bl，Ventricles_bl ， Hippocampus_bl,WholeBrain_bl,Entorhinal_bl,Fusiform_bl,MidTemp_bl,MOCA_bl, EcogSPTotal_bl,timestamp showed significant differences between the groups of the 24 sets of data; There were no significant differences between the three sets of data: IMAGEUID, EcogPtTotal_bl, and Month_bl.

According to the analysis of variance between and within groups, we select three characteristics: EcogSPTotal_bl, Month_bl and IMAGEUID_bl for analysis.

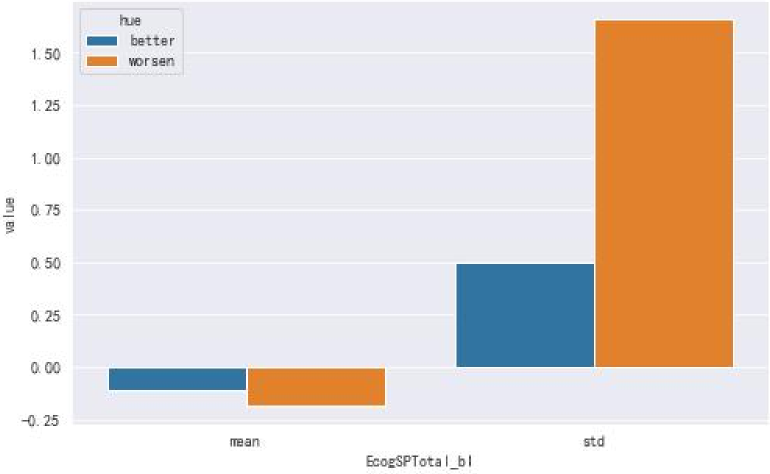

Figure 24 Histogram of mean and variance of EcogSPTotal_bl in the Better and worsen groups

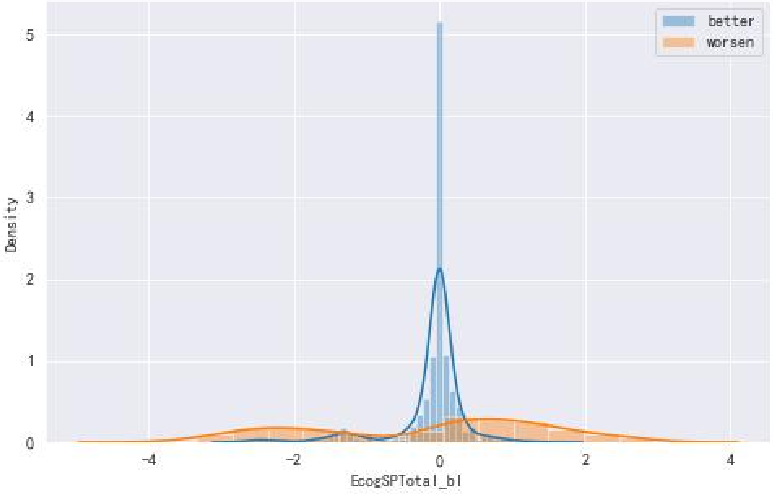

Figure 25 Histogram and kernel density curve of the effect of EcogSPTotal_bl on condition

From Figure 24 and Figure 25 above, it can be found that the variance and mean values of Month_bl are not significantly different between the improving and deteriorating groups, but the kernel density plots show a normal distribution for both groups, with the data in the deteriorating group being more concentrated compared to the data in the improving group.

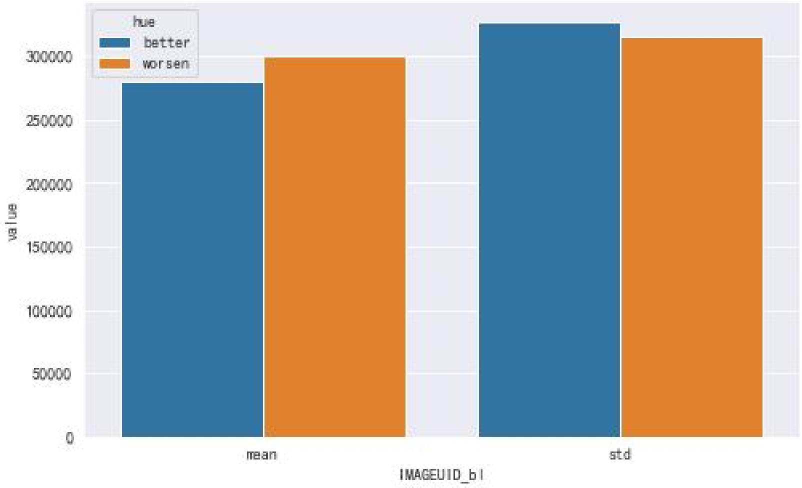

Figure 26 Histogram of mean and variance of IMADEUID_bl in the Better and worsen groups

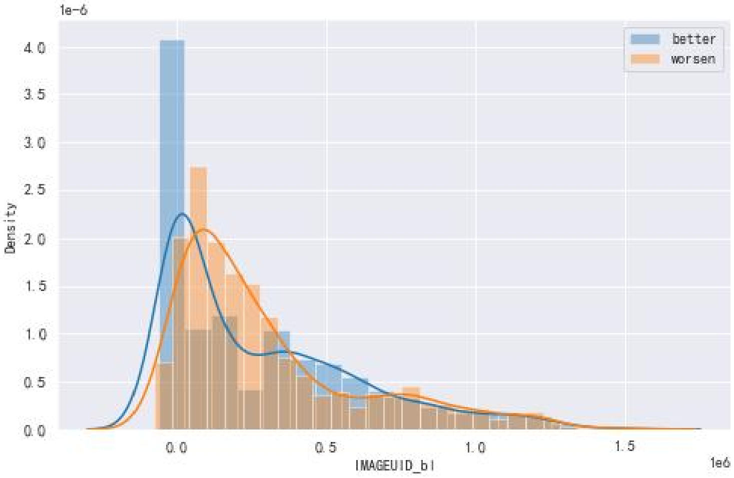

Figure 27 Histogram and kernel density curve of the effect of IMAGEUID_bl on the condition

As can be observed in Figures 26 and 27 above, the data in IMAGEUID_bl are large but the values are scattered, and in the improving and deteriorating groups, the kernel density curve its waveform shifts to the right, with a left skewed distribution, increasing in horizontal width and decreasing in the number of peaks. This indicates a gradual increase in their disparity.

4.5 Question 5: Early intervention and diagnostic criteria for five types of disease

4.5.1 Early intervention and diagnostic criteria for cognitively normal older people (CN)

Normal older people (CN), identified as older people over 60 years of age.

Early intervention.

1. active prevention of lifestyle diseases such as obesity, fatty cancer, hypertension, diabetes and other diseases.
2. attention to dietary habits and modification of the structure of the diet
3. change of lifestyle habits and control of smoking and alcohol in moderation
4. using the brain more diligently to keep the brain fully active.
5. exercising scientifically to keep the body healthy
6. improve the quality of sleep to provide adequate mental reserves
7. actively participate in social activities and increase outside communication.

Diagnostic criteria.

1. having sufficient energy to carry the heavy workload of daily life with ease.
2. optimism in the world, positive attitude and willingness to take responsibility.
3. good rest and good sleep.
4. resilient and able to adapt to a variety of changes in the environment
5. able to resist the common cold and infectious diseases.
6. of moderate weight, well proportioned and standing with head, shoulders and arms in a coordinated position.
7. bright eyes, responsive, eyes and eyelids not glowing
8. clean teeth, no caries, no pain, no bleeding of the gums
9. shiny hair, free of dandruff
10. fullness of muscle and elasticity of skin.

4.5.2 Early Intervention and Diagnostic Criteria for Memory Mastery Complaint (SMC)

Memory complaints (SMC)[6]， Primarily refers to a way in which an individual complains or complains about their memory loss, and is one of the early manifestations of AD. In early Alzheimer's disease, when patients suffer from SMC, they will repeatedly complain about language, lack of language expression due to memory loss, memory reproduction ability decline, communication with others is a cover-up behavior, there will also be language conflicts, and even refusal to communicate with others avoidance behavior.

Early intervention:

1. Strengthen the education of memory knowledge of AD patients with SMC;
2. The memory content described by the patient should be recognized, increase his self-confidence, and avoid negative emotions and decline in life ability;
3. It can also appropriately exercise the patient's memory ability, and carry out intervention behaviors such as recall awakening, experience presentation, and experience sharing for the patient.

Diagnostic criteria:

1. Personal complaints or memory function decline;
2. The age of onset is over 60 years old;
3. Feel that their memory is significantly lower than that of their peers;
4. Later stage may suffer from Alzheimer's disease;

4.5.3 Early intervention and diagnostic criteria for early mild cognitive impairment (EMCI).

Early mild cognitive impairment (EMCI) refers to progressive loss of memory or other cognitive functions that does not affect the ability to perform daily living and does not meet the diagnostic criteria for dementia.

Early intervention:

1. Appropriate physical activity;
2. Cognitive training for the elderly;
3. Increase your intake of vegetables, fruits, grains and fish, and reduce your intake of meat

Diagnostic criteria:

1. Patients or insiders report, or experienced clinicians find cognitive impairment: the patient complains of memory loss, or the family finds that the patient forgets a lot, repeatedly looks for things, and gets lost;
2. Objective evidence of impairment of one or more cognitive domains;
3. Complex instrumental daily abilities can have slight impairment, large ability to maintain independent daily living;
4. The diagnosis of dementia has not yet been reached[7]。

4.5.4 Early intervention and diagnostic criteria for late mild cognitive impairment (LMCI)

Late mild cognitive impairment (LMCI), defined as progressive decline in memory or other cognitive functions that do not interfere with the ability to perform daily activities and that do not meet the diagnostic criteria for dementia but whose criteria are close to those for dementia.

Early intervention.

1. implementation of effective implementation of MCI health education activities.

2. organising planned participation in cultural and physical activities in the community to distract the patient.

3. Reasonable adjustment of dietary structure, with attention to a balanced diet of meat and vegetables, quality, "three lows and one precaution", and appropriate supplementation of foods with high protein and vitamin content.

4. insist on appropriate exercise, participate in group activities during free time, and pay attention to the work and rest schedule.

5. Help the patient to carry out activities that stimulate the brain to function, or to strengthen memory by increasing the amount of information, repeated prompts or using some recording devices or tools.

6. giving the patient confidence, repetition and patience in cognitive training [8].

Diagnostic criteria.

1. cognitive decline, with cognitive impairment identified by the patient himself or other knowledgeable persons, and abnormal cognitive function identified after a definitive systematic examination.

2. normal basic everyday abilities, where the patient can perform basic everyday abilities and complex instrumental everyday abilities independently, although there may be minor impairment in the latter.

3. close to a diagnosis of dementia but not reaching it[9]。

4.5.5 Early intervention and diagnostic criteria for Alzheimer's disease (AD).

Alzheimer's disease (AD) is a neurodegenerative disease with an insidious onset that is often accompanied by symptoms such as memory loss, inmobility, language and memory [10], often referred to as Alzheimer's disease, and patients are generally older than 65 years old.

Early intervention:

1. Nostalgia therapy, through oral or non-verbal means, let patients recall past events, stimulate patients' memories of previous things to improve patients' cognition is an obstacle;

2. Reasonable lifestyle changes, strengthen physical exercise, increase mental activity, improve diet, you can try Mediterranean diet;

3. Maintain social contact;

4. Avoid head trauma;

5. Avoid suffering from high blood pressure, diabetes and hypercholesterolemia.

Diagnostic criteria:

1. Meet the diagnostic criteria for organic mental disorders;

2. Comprehensive intelligent damage;

3. No sudden stroke-like seizures, no signs of focal neuroolefin damage in the early stage of the disease;

4. No clinical or special examination suggests that intellectual impairment is caused by other physical or brain diseases;

5. The function of the higher cortex is impaired, and there may be aphasia, agnosia and uselessness;

6. Apathy, lack of active activities, or irritability and loss of social behavior;

5. Model evaluation and optimisation

5.1 Model Benefits

1. the paper as a whole is based on information theory, which can handle not only continuous features but also discrete features compared to general linear models.

2. by calculating indicators such as information content and Gini index, the dimensional disaster caused by high-dimensional data can be effectively avoided and quality features can be better selected.

3. XGBoost adds regularization to the objective function, while introducing column sampling of random forests, which can better filter irrelevant variables.

4. using the idea of recursion to predict the filling of random missing values, which effectively improves the data completeness.

5. deep mining of non-random missing features, analysis of the causes of missing values for various types of features, and handling of joint missing problems based on missing feature associations.

6. corrected invalid and illegal data through regular expressions, while extracting continuous information from some discrete features.

7. the probability space was extracted through Self-organizing Map with high interpretability.

8. the introduction of Shaply Additive explanations value based on game theory to judge the marginal contribution of each feature from both local and global perspectives.

5.2 Model disadvantages

1. In general, the model based on information theory has more advantages in processing discrete features, while its interpretation and accuracy for continuous features may not be as good as that of the general linear model;

2. It is difficult to adjust the parameters of XGBoost, which is prone to overfitting and other problems;

3. In the Self-organizing Map, some neurons fail to win at all times, resulting in neuron death and additional algorithm overhead;

4. When the features had moderate permutation importance, SHAP value could not determine whether it was due to the large impact on only a small number of predictions.

References

[1] 陈红英,张昌明,何晶等.基于正则表达式的遥测数据预处理研究[J].舰船电子工程,2015,第35卷(12):130-133.

[2] 徐岩,程姝,薛艳静.基于故障暂态电流Pearson相关系数的直流配电网保护[J].华北电力大学学报(自然科学版),2021,第48卷(4):11-19.

[3] CHEN T Q,GUESTRIN C.Xgboost:a scalable tree boosting system[C].Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining.ACM,2016:785-794.

[4] 郑文珍,赖俊龙,翁雯璇等.基于医保欺诈检测的RF与XGB模型比较[J].金融科技时代,2021,(6):68-73.

[5] 舒服华.基于SOM的商业银行盈利能力聚类研究[J].金融理论与教学,2022,(1):9-14.

[6] 符娇英,毛洁,颜平康,陈晓俊.阿尔茨海默病患者记忆抱怨主诉的记忆护理综述[J].中国乡村医药,2022,29(15):78-80.DOI:10.19542/j.cnki.1006-5180.006494.

[7] 中国痴呆与认知障碍诊治指南写作组,中国医师协会神经内科医师分会认知障碍疾病专业委员会.2018中国痴呆与认知障碍诊治指南(五):轻度认知障碍的诊断与治疗[J].中华医学杂志,2018,98(17):1294-1301

[8] 殷淑琴，宋喻等.湖州市老年人轻度认知功能障碍早期干预研究[J].临床研究，2013,51(3)：67-68

[9] 莫颖敏，韩敏.轻度认知功能障碍早期诊断与治疗进展[J].中国临床新医学，2011,4(9)：895-898

[10]Andrieu S,Coley N,Lovestone S,et al.Prevention of Sporadic Alzheimer’s Disease:Lessons Learned from Clinical Trials and Future Directions[J].Lancet Neurology,2015,14(9): 926-944.

Appendix

Table1 Information value of feature

	iv	gini	Entropy	unique
mPACCtrailsB	9.339392123	0.511739191	1.029327104	9567
mPACCdigit	9.010022553	0.511544509	1.029628909	4699
CDRSB	8.641348057	0.471509219	1.026282426	32
ADAS13	8.475049519	0.51970186	1.028968829	209
LDELTOTAL_BL	8.057378673	0.480130157	1.029617192	24
FAQ	6.561653451	0.557455754	1.025300782	32
ADAS11	6.175200608	0.572346866	1.029084285	166
ADASQ4	5.581550767	0.571017072	0.990560359	12
LDELTOTAL	5.530530694	0.564048221	1.02509318	27
MMSE	5.445234698	0.540464499	1.00096367	32
EcogSPMem_bl	5.001322154	0.532490428	1.008519072	50
EcogSPLang_bl	4.791462499	0.545786084	1.019169816	61
EcogSPTotal_bl	4.692974585	0.582081498	1.029747622	3797
RAVLT_immediate	4.614250081	0.56262908	1.023427382	77
RAVLT_forgetting_bl	4.384045795	0.565610119	1.020514016	117
EcogSPVisspat_bl	4.047297529	0.558913739	1.004485985	51
FAQ_bl	3.876726973	0.570918138	1.029680184	61
CDRSB_bl	3.816494711	0.57065619	1.029811657	54
EcogSPOrgan_bl	3.783586903	0.556487537	1.007133344	37
MMSE_bl	3.702827569	0.610821306	1.029723813	39
EcogSPPlan_bl	3.677588669	0.552590235	1.023469415	30
EcogSPTotal	3.579526727	0.585549039	1.028593045	835
MOCA_bl	3.433446307	0.601109865	1.029496596	41
EcogSPDivatt_bl	3.407249712	0.561383384	1.014088732	19
EcogSPMem	3.356241727	0.586659508	1.0222682	61
DIGITSCOR_bl	3.244613442	0.571531666	1.02315563	68
RAVLT_perc_forgetting	3.10608671	0.580874731	1.017781912	73
RAVLT_forgetting	2.911634088	0.568199454	1.017368777	66
FDG_bl	2.911070294	0.578992336	1.029486503	1497
ADAS11_bl	2.863757835	0.590152742	1.029211951	436
MOCA	2.655261901	0.59291396	1.024385022	32
EcogSPVisspat	2.618130928	0.598011966	1.027246372	56
EcogSPPlan	2.600050126	0.595979872	1.025358255	32
EcogSPLang	2.584437167	0.593454568	1.028397864	80
EcogSPOrgan	2.561365061	0.596464231	1.019879543	38
Hippocampus	2.3679928	0.578962402	1.029415671	5082
RAVLT_learning	2.266168341	0.582548106	1.023433076	21
AV45_bl	2.2272368	0.590300511	1.029801863	1076
EcogSPDivatt	2.164971577	0.595402382	1.020048372	20

TRABSCOR	2.115727999	0.599727873	1.029627213	293
mPACCtrailsB_bl	1.934558047	0.58175994	1.029804562	10133
mPACCdigit_bl	1.770203052	0.587660595	1.029816792	9216
EcogPtMem_bl	1.721471346	0.567378778	1.016526702	45
ABETA_bl	1.636692008	0.591618231	1.029753825	1002
ADAS13_bl	1.618945845	0.59144353	1.029737009	462
RAVLT_immediate_bl	1.592538038	0.612577442	1.029751332	104
TRABSCOR_bl	1.585405891	0.609128089	1.029669911	506
ADASQ4_bl	1.47383601	0.610600708	1.027499764	20
MidTemp	1.447847392	0.594782639	1.029820058	5750
EcogPtPlan_bl	1.424491381	0.590407871	1.029162382	22
RAVLT_perc_forgetting_bl	1.318168471	0.614296567	1.029810955	1577
Entorhinal	1.26559423	0.602425378	1.029687272	3118
EcogPtTotal_bl	1.236728274	0.601271908	1.02976369	3213
TAU_bl	1.196726327	0.601430778	1.029398213	1143
EcogPtVisspat_bl	1.087863223	0.597609968	1.029339796	45
Fusiform	1.040282862	0.603237061	1.029406523	5524
EcogPtOrgan_bl	0.966799445	0.605090719	1.029783758	35
EcogPtLang_bl	0.903408604	0.591422457	1.0290487	51
EcogPtDivatt_bl	0.864658803	0.598484285	1.026854779	16
APOE4	0.799691377	0.601615151	0.984364366	3
Entorhinal_bl	0.767514905	0.618231177	1.029832698	3284
Ventricles	0.692432133	0.609390171	1.029806938	8731
update_stamp	0.608478988	0.610990839	1.029832477	107
EcogPtMem	0.573312773	0.604404302	1.025768811	61
EcogPtTotal	0.532488709	0.603969141	1.026736429	623
EcogPtPlan	0.502449051	0.611587698	1.026173484	31
MidTemp_bl	0.499468868	0.614228577	1.029803701	3831
EcogPtVisspat	0.487273067	0.612307505	1.02573326	55
RAVLT_learning_bl	0.402407784	0.617560285	1.029767504	27
EcogPtOrgan	0.380336649	0.613968898	1.024356374	38
WholeBrain	0.360965491	0.614926324	1.029818369	8778
EcogPtLang	0.313915888	0.611275822	1.026514517	73
Fusiform_bl	0.313418448	0.617029338	1.029828573	3708
PTAU_bl	0.275493179	0.621829649	1.029828308	191
IMAGEUID	0.272188801	0.617972037	1.029823079	9718
Ventricles_bl	0.243416387	0.619845378	1.029785139	8158
AGE	0.230822692	0.61625539	1.029730836	346
EcogPtDivatt	0.224150184	0.615529514	1.025160233	20
Hippocampus_bl	0.213470628	0.619748272	1.029832056	3851
M	0.209815776	0.619905243	1.029719308	35
VISCODE	0.209815776	0.619905243	1.029719308	35
Years_bl	0.20277575	0.61978218	1.029750845	2781
Month	0.201046365	0.619841548	1.029702284	35

IMAGEUID_bl	0.188836844	0.620304899	1.029815182	12010
FSVERSION	0.178186723	0.612875445	1.01214763	3
PTEDUCAT	0.161476065	0.619047069	1.029693761	16
WholeBrain_bl	0.125359697	0.621979268	1.029824756	8340
ICV_bl	0.117479966	0.622012154	1.029831916	6351
PTMARRY	0.112445067	0.6187796	1.020606458	5
ICV	0.086779518	0.621682277	1.029827744	8959
SITE	0.086030918	0.621388	1.029786487	66
FSVERSION_bl	0.079071714	0.616022588	1.017757272	3
FLDSTRENG	0.071533772	0.619894965	1.022036947	2
PTGENDER	0.040363777	0.619611582	1.024284413	2
FLDSTRENG_bl	0.037682202	0.61954276	1.022431303	2
Month_bl	0.035662674	0.623300428	1.02981537	868
PTRACCAT	0.030895283	0.620751074	1.025329801	7
PTETHCAT	0.004315944	0.623710382	1.029550649	2

Table2 Tusky HSD

group 1	group 2	meandiff	p-adj	lower	upper	reject	target
bader	better	-3.7597	0.0	-3.9484	-3.5711	True	CDRSB_bl
bader	better	-3.6017	0.0	-4.101	-3.1024	True	AGE
bader	better	-7.4709	0.0	-8.0317	-6.91	True	ADAS11_bl
bader	better	-10.9304	0.0	-11.5844	-10.2763	True	ADAS13_bl
bader	better	-1.8671	0.0	-2.0268	-1.7075	True	ADASQ4_bl
bader	better	4.6434	0.0	4.3126	4.9743	True	MMSE_bl
bader	better	9.198	0.0	8.4625	9.9336	True	RAVLT_immediate_bl
bader	better	1.1658	0.0	0.9657	1.3659	True	RAVLT_learning_bl
bader	better	-30.7869	0.0	-32.9128	-28.661	True	RAVLT_forgetting_bl
bader	better	-12.8377	0.0	-15.4192	-10.2562	True	RAVLT_perc_forgetting_bl
bader	better	-26.0081	0.0	-32.7076	-19.3086	True	TRABSCOR_bl
bader	better	-9.3598	0.0	-9.8232	-8.8964	True	FAQ_bl
bader	better	8.5433	0.0	8.0859	9.0008	True	mPACCdigit_bl
bader	better	7.8689	0.0	7.4343	8.3035	True	mPACCtrailsB_bl
bader	better	-20394.5733	0.1033	-44931.8251	4142.6785	False	IMAGEUID_bl
bader	better	-8576.4591	0.0	-9332.6	-7820.3182	True	Ventricles_bl
bader	better	589.5576	0.0	546.1698	632.9454	True	Hippocampus_bl
bader	better	24703.0139	0.0	21688.3119	27717.7159	True	WholeBrain_bl
bader	better	604.0073	0.0	464.9156	743.0989	True	Entorhinal_bl
bader	better	961.3604	0.0	845.2709	1077.4498	True	Fusiform_bl
bader	better	1368.1782	0.0	1249.9706	1486.3859	True	MidTemp_bl
bader	better	125982.2641	0.0	71626.6265	180337.9017	True	ICV_bl
bader	better	9.2897	0.0	8.67	9.9094	True	MOCA_bl
bader	better	0.5957	0.0	0.5255	0.6659	True	EcogPtTotal_bl

bader	better	0.0737	0.1373	-0.0235	0.1709	False	EcogSPTotal_bl
bader	better	0.0656	0.0908	-0.0104	0.1417	False	Month_bl
bader	better	-688.0439	0.0	-771.5677	-604.5201	True	timestamp

Program Code 1	Recursive prediction-filling

```
cols = []


 for col in acc_pre.index:
    missing = (counterpart_2.shape[0] - counterpart_2.count()) / counterpart_2.shape[0]
    missing = missing[missing > 0.]
    dt = DecisionTreeClassifier(random_state=42)

    if counterpart_2[col].nunique() > 100:
        dt = DecisionTreeRegressor(random_state=42)

    temp = counterpart_2[counterpart_2[col].notna()]
    nan = counterpart_2[counterpart_2[col].isna()]

    X = temp[[i for i in temp.columns if i not in missing.index and i not in [col, 'PTID', 'EXAMDATE',
                                                                                  'EXAMDATE_bl']]]
    y = temp[col]
    X_ = nan[[i for i in temp.columns if i not in missing.index and i not in [col, 'PTID', 'EXAMDATE',
                                                                                  'EXAMDATE_bl']]]
    cols.append(X.columns)
    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=.3, random_state=42)

    try:
        dt.fit(X_train, y_train)
    except ValueError:
        y_train = y_train.astype(np.str)
        y_test   = y_test.astype(np.str)
        dt.fit(X_train, y_train)

    print(col, end='\t')
    print(dt.score(X_test, y_test))
    try:
        y_pre = dt.predict(X_)
        print(counterpart[col].isnull().sum(), end='\t')
        counterpart.loc[X_.index, col] = y_pre
        counterpart_2.loc[X_.index, col] = y_pre
        print(counterpart[col].isnull().sum())
    except ValueError as e:
        print(e)
```

Program Code 2	Feature encoding

```
science_dict = {
    'ADNI1': 1,
    'ADNI2': 2,
    'ADNI3': 4,
    'ADNIGO': 3}

counterpart['PTGENDER'] = counterpart['PTGENDER'].map(
    {
        'Male': 0,
        'Female': 1
    })
counterpart['PTETHCAT'] = counterpart['PTETHCAT'].map(
    {
        'Not Hisp/Latino': 0,
        'Hisp/Latino': 1
    })
def get_number(value):
    value = str(value)
    re_str = '[0-9]+[\.]?[\d+]*'
    res = re.findall(re_str, value)
    if len(res) > 0:
        return float(res[0])
    return np.nan

counterpart['FLDSTRENG'] = counterpart['FLDSTRENG'].apply(get_number).astype('float')
counterpart['FLDSTRENG_bl'] = counterpart['FLDSTRENG_bl'].apply(get_number).astype('float')
counterpart['FSVERSION'] = counterpart['FSVERSION'].apply(get_number).astype('float')
counterpart['FSVERSION_bl'] = counterpart['FSVERSION_bl'].apply(get_number).astype('float')
counterpart['COLPROT'] = counterpart['COLPROT'].map(science_dict).astype('float')
counterpart['ORIGPROT'] = counterpart['ORIGPROT'].map(science_dict).astype('float')
```

Program Code 3	Data check

```
import re
# 数据格式校验
def pid_ack(pt_id):
    re_str = r'\d{3}_[A-Za-z]{1}_\d{4}'
    res = re.findall(re_str, pt_id)
    if len(res) > 0:
        return res[0]
    return np.nan
def time_ack(time):
    re_str = r'\d{4}.\d{1,2}.\d{1,2}'
    time = str(time)
    res = re.findall(re_str, time)
    if len(res) > 0:
        return res[0]
    return np.nan
def num_ack(num):
    num = str(num)
    re_str = '[-]?\d+[\.]?[\d+]?$'
    res = re.findall(re_str, num)
    if len(res) > 0:
        return float(res[0])
    return np.nan
def object_act(value):
    if value == '<NA>' or value == 'NA':
        return np.nan
    return value

data['PTID'] = data['PTID'].apply(pid_ack)
data['EXAMDATE'] = pd.to_datetime(data['EXAMDATE'].apply(time_ack), infer_datetime_format=True)
data[['ABETA', 'TAU', 'PTAU', 'ABETA_bl', 'TAU_bl', 'PTAU_bl']] = data[[
                                            'ABETA', 'TAU', 'PTAU', 'ABETA_bl', 'TAU_bl', 'PTAU_bl'
                                          ]].applymap(num_ack).astype(np.float64)


data = data.applymap(object_act)
data
```

Program Code 4	Self-organizing Map

```
N = x.shape[0]
M = x.shape[1]

size = math.ceil(np.sqrt(5 * np.sqrt(N)))

max_iter = 300

som = MiniSom(size, size, M, sigma=10, learning_rate=0.1,
              neighborhood_function='gaussian')

som.pca_weights_init(x)
som.train_batch(x, max_iter, verbose=False)
winmap = som.labels_map(x, true)
 def classify(som,data,winmap):
    from numpy import sum as npsum
    default_class = npsum(list(winmap.values())).most_common()[0][0]
    result = []
    for d in data:
        win_position = som.winner(d)
        if win_position in winmap:
            result.append(winmap[win_position].most_common()[0][0])
        else:
            result.append(default_class)
    return result

y_pred = classify(som, x, winmap)print(classification_report(true, np.array(y_pred)))
```

Program Code 5	Analysis of variance

```
def oneWayAnova(df,cata_name,num_name,alpha_anova=0.05,alpha_tukey=0.05):
    info = ''
    df[cata_name]=df[cata_name].astype('str')

    s1=df[cata_name]
    s2=df[num_name]

    fml=num_name+'~C('+cata_name+')'

    model = ols(fml,data=df).fit()
    anova_table_1 = anova_lm(model, typ = 2).reset_index()
    p1=anova_table_1.loc[0,'PR(>F)']

    if p1>alpha_anova:
        print(num_name + '组间【无】显著差异')
        info = info + '\n' + str(num_name) + '组间【无】显著差异\n'
    else:
        print(num_name + '组间【有】显著差异')
        info = info + '\n' + str(num_name) + '组间【有】显著差异\n'

    df_p1=df.groupby([cata_name])[num_name].describe()

    mc = MultiComparison(df[num_name],df[cata_name])
    df_smry = mc.tukeyhsd(alpha=alpha_tukey).summary()
    m = np.array(df_smry.data)
    df_p2 =pd.DataFrame(m[1:],columns=m[0])

    df_p1_sub=df_p1[['mean']].copy()
    df_p1_sub.sort_values(by='mean',inplace=True)

    output_list=[]

    for x in range(1,len(df_p1_sub.index)):
        if (df_p2.loc[((df_p2.group1==df_p1_sub.index[x-1])&(df_p2.group2==df_p1_sub.index[x]))|
                      ((df_p2.group1==df_p1_sub.index[x])&(df_p2.group2==df_p1_sub.index[x-1])),
                      'reject'].iloc[0])=="True":
            smb='<'
        else:
            smb='<='
        if x==1:
            output_list.append(df_p1_sub.index[x-1])
            output_list.append(smb)
            output_list.append(df_p1_sub.index[x])
        else:
            output_list.append(smb)
            output_list.append(df_p1_sub.index[x])
    out_sentence=' '.join(output_list)
    print(out_sentence)
    info += out_sentence
    info += '\n\n'

    return df_p1,df_p2, info
```