**The Construction of Predictive Model of Childhood Autism and its Value Research**

**Prediction Modeling of Children Autism and Application in Diagnosis**

# Abstract

Autism is a widespread disorder in neurological development, often onset in childhood, mainly manifested as disorders in social interactions, narrowness of interests, repeated and mechanical actions, and body movements. Existing diagnoses of autism largely rely on scales, but the indicators used in most scales are simply numerical summations of values or intervals. Autism Spectrum Disorders (ASD) are complex neurological diseases, where early intervention and treatment proved to be conducive to the recovery of patients, which is not possible without early and accurate diagnosis. This paper tries to predict using machine learning techniques.

Through chi-square test, this research employs machine learning to quantify the major characteristic values of the diagnostic scales. Based on logistic regression analysis and the idea of SMO (Sequential Minimal Optimization) + SVM (Support Vector Machine), our research designed two prediction models that exhibit statistically significant accuracy. The prediction models and some conclusions obtained in the process have practical implications regarding the prediction of autism, the popularization of diagnosis, and early intervention and treatment. This research proves innovative in that it incorporates advanced machine learning methods into disease prediction, which contributes to the development of intelligent healthcare and facilitates the design and implementation of precise, personalized health schemes.

***Key Words****：*Autism prediction; Neurodevelopmental disorders; Diagnostic scale; Machine learning; Support Vector Machine (SVM)

Contents

[Abstract 1](#_Toc56714303)

[1. Introduction 3](#_Toc56714304)

[2. Research goals 5](#_Toc56714305)

[3. Methods and results 5](#_Toc56714306)

[3.1 Data collation 5](#_Toc56714307)

[3.2 Machine Learning 9](#_Toc56714308)

[3.2.1 Mean normalization 9](#_Toc56714309)

[3.2.2 Feature selection 10](#_Toc56714310)

[3.2.3 Logistic regression analysis 16](#_Toc56714311)

[3.2.4 Relationship between Logistic regression and Support Vector Machine 20](#_Toc56714312)

[3.2.5 Support Vector Machine 20](#_Toc56714313)

[3.2.6 Sequential Minimal Optimization 23](#_Toc56714314)

[3.3 Model validity test 24](#_Toc56714315)

[4. Discussion 26](#_Toc56714316)

[5. Conclusion 28](#_Toc56714317)

[6. Future Outlook 28](#_Toc56714318)

[References 29](#_Toc56714319)

[Acknowlegements 30](#_Toc56714320)

# Introduction

Autism is the abbreviation of Autism Spectrum Disorder, a kind of neurological developmental disorder, often onset in childhood, mainly manifested as impaired social interactions, speech inability, narrowness of interests, and repetitive behaviors[1][2]. According to WHO, 1 of every 160 children in the world suffer from autism, and the prevalence rate has been increasing over the past 50 years[3]. As shown in the data published by CDC in 2016, 1.85% of 8-year-old children in the US were diagnosed with autism[4]. Globally, autism has brought tremendous burden to patients and there families, both economically and mentally[5-6]. Currently, there is no therapy or drug that completely heals autism, but appropriate behavioral therapy can effectively help patients recover from some of the symptoms. Since children brains in early stages tend to have higher plasticity and with the development of healthcare, studies have shown that earlier diagnosis and intervention of autism lead to better result of rehabilitation[7]. Hence, a more efficient and accurate procedure for autism diagnosis is pivotal for relieving symptoms effectively and designing personalized healthcare plan[8].

On account of the absence of organic lesion and direct biochemical indicators, the diagnoses of autism largely depend on the behavior assessment of rating scales. Due to the complicated nature of autism, myriad scales with complex variants are involved. This thesis selects a few of the most widely used and highly recognized rating scales, namely, the autism diagnosis scale from International Classification of Diseases, Children Autism Rating Scale, Sensory Integration Ability Scale, ADHD Rating Scale, SCT Attention Test, and PPVT IQ Test. The outcomes of rating scales are usually a numerical value or interval, which are usually results after simple accumulation, since such practice controls the cost and efficiency in hospitals and organizations where large flow of people come by. The ultimate diagnoses are made by professional medical personnels, combining rating scale results, patient or parent description, and direct observation of patient behaviors.

Nowadays, for neurological disorders featured by sophisticated symptoms and lengthy traditional diagnosis procedure, artificial intelligence has come to be more widely utilized in different branches of medical science[9]. For instance, with the help of statistical models and machine learning algorithms, some researchers has already constructed relatively accurate suicidal models of depression patients[10], which is objective and efficient. The concept of Autism Spectrum Disorder stems from clinical observation, yet different patients perform differently about their defects in social interactions, speech inability, and repetitive behaviors; on the other hand, besides the core symptoms regarding the nervous system, some of the autism patients also have peripheral symptoms in immune system and digestive system. The heterogeneity and manifold nature of autism result in the sophistication of related variables, highlighting the necessity of multivariate analysis for diagnosis or even exploring the inherent laws[11]. Big data analysis as well as machine learning can integrate data and offer an insight, predicting paroxysm and behaviors more efficiently, and synthesizing complicated indicators. Hence, the application of machine learning in the construction of autism prediction model avails early diagnosis and intervention.

# 2. Research goals

By organizing and integrating variants in autism related rating scales, we hope to construct a prediction model of certain sensitivity with the help of machine learning, thus providing reliable help for efficient and accurate diagnosis. We also hope to study the results we get while practicing machine learning and look for inner laws to have a better understanding of autism symptoms, and to offer a reference for the design of personalized healthcare plan for diagnosed patients.

# 3. Methods and results

## 3.1 Data collation

We collected original data from patients’ database in Nanjing Tianyou Children's Hospital, which ranged from May to July 2020. There were 542 data that meet the conditions. The original forms were as followed:

1. Sensory Integration Evaluation

Table 1 Evaluation Form of Sensory Integration ability

|  |  |
| --- | --- |
| Test results | |
| Items | Evaluation |
| Degree of vestibular and bicerebral differentiation | Normal |
| Neurophysiological depression of brain | Mild |
| Tactile defense and temper sensitivity | Mild |
| Exercise and daily operation during development | Edge |
| Spatial form and visual perception | Edge |
| Proprioception (gravity insecurity) | Normal |
| Learning, emotion and self- image feeling | Normal |
| Psychological stress and behavior | Normal |
| Total score and total evaluation | Mild |

2. Autism CARS Assessment

Table 2 Evaluation Form of Childhood Autism Rating Scale (CARS)

|  |  |  |
| --- | --- | --- |
| No. | Items | Score |
| 1 | Interpersonal relationship | 3 |
| 2 | Imitation ability (words and actions) | 3 |
| 3 | Emotional response | 2 |
| 4 | Physical ability | 2 |
| 5 | The relationship with inanimate objects | 3 |
| 6 | Appropriate response to environmental change | 2 |
| 7 | Visual response | 2 |
| 8 | Auditory response | 2 |
| 9 | Near sensory response | 2 |
| 10 | Anxiety response | 2 |
| 11 | Language communication | 3 |
| 12 | Nonverbal communication | 2 |
| 13 | Activity level | 2 |
| 14 | Intellectual function | 2 |
| Preliminary results | | 32 |

3. ICD-10 Autism

Table 3 Autism Assessment Form According to ICD-10

(ICD-10 means International Classification of Diseases, 10th Edition)

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Test contents | | | | | | | | Results |
| **1** | Obvious defects in social communication, language communication and hobbies | | | | | | | | ≥6 |
| Items | No. | Results | Items | No. | Results | Items | No. | Results |
| Behavior defects of verbal communication | (1) | √ | Qualita-  tive defects of verbal communi-cation | (1) | √ | Stereotyp-ed, repetitive and limited behaviors | (1) | √ |
| (2) | √ | (2) |  | (2) |  |
| (3) | √ | (3) | √ | (3) |  |
| (4) |  | (4) | √ | (4) | √ |
| **2** | Abnormal or delayed function occurred before 3 years old | | | | | | | (1) | √ |
| (2) | √ |
| (3) | √ |
| **3** | It's not Ritter's syndrome or child disintegration disorder. | | | | | | | |  |

The original data of the hospital is nested in four levels of folders. The last level of folder’s content were all the examination reports performed by a certain patient, 5-8 examinations for each. Assuming that 500 patients came to check and had, supposedly, 5 examinations, we estimate that there would be as many as 2500 documents that need to be extracted. Therefore, if the data are manually extracted, it would certainly result in huge work amount and sometimes, errors. Therefore, in order to provide reliable basis for the follow-up work, we used python programs to do the data preprocessing, which would be much faster and more accurate. The implementation steps are as follows:

Since python cannot process doc documents, we first converted all doc documents to the docx format (note: the original doc documents will be replaced). The code is shown in Appendix A. Here we use the docx package and the client package in win32com. After opening the .doc file, add ‘x’ after the file name and save it as a .docx file.

After the document conversion is completed, we are ready to get the data from tables, taking reading all "Autism CARS" tables as an example.

First, we design the header of the table according to the key information to be extracted from the "Autism CARS.docx" document. For example, in "Autism CARS.docx", we want to get the following information: name, number, interpersonal relationship, imitation (Words and actions), emotional response, physical ability, relationship with non-living objects, appropriateness to environmental changes, visual response, auditory response, close sensory response, anxiety response, verbal communication, nonverbal communication, activity level, For the form of intelligence functions, these are used as table headers. To manipulate tables in python, we import the Workbook in the openpyxl package and then traverse the "Autism CARS.docx" document in all folders. To read the file, we import the Document in the docx package in advance, and then extract the content of the corresponding header in the document in each row and column, saving it in the corresponding position of the excel table, in the end using the instruction workbook.save('$$path\\Autism CARS.xlsx') to save the table.

The current analysis is mainly based on the three tables of "Autism CARS", "ICD-10 Autism", and "Sensitive Integration Evaluation Report Form", and then we combine the three tables together to obtain a summary table. The data of the summary table is shown in the figure:





Figure 1 Partial data of the summary table

Before doing machine learning, we should first manually convert the required data into txt form. The txt file contains all the data in the three tables of "Sensory Integration Evaluation Report", "Autism CARS", and "ICD-10 Autism" as features, plus the autism status as the last column (0 means no autism, 1 means autism). txt part of the data is shown in the figure:

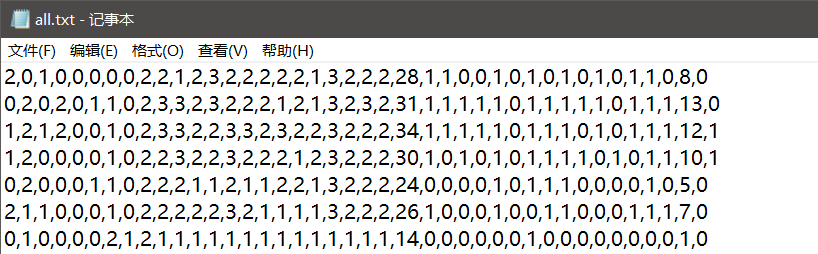


Figure 2 Partial data of feature quantity

There are a total of 542 rows and 40 columns of data. Each row represents a set of data. The first 9 columns in each row represent various data of sensory integration assessment, the middle 15 columns represent various data of autism CARS, and the last 15 columns represent ICD—— 10 various data of autism, the last column represents the diagnosis result (1 means diagnosis of autism, 0 means diagnosis of non-autism).

## 3.2 Machine Learning

### 3.2.1 Mean normalization

Different indicators often have different dimensions and units, which will affect the results of data analysis. In order to eliminate the dimensional impact between different indicators, it is necessary to carry out data standardization processing to make the data comparable. Standardized data is the preparatory work for data mining. After data standardization, the original data is suitable for comprehensive comparison Evaluation.

There is a total item of evaluation in the characteristic quantity. The difference between this column of data and other data is a little big, so normalization is done.

We use the method of mean normalization, first of all, select a column of data, get the maximum value xmax and minimum value xmin of the column data,then each data in the column is normalized () to make each xi in the range of (0,1), so as to realize the feature scaling and make the gradient descent method converge faster.

### 3.2.2 Feature selection

When the data processing is completed, we need to select meaningful feature and model for training. At present, there are 40 feature quantities. If the correlation between features and target is too high, the features will be selected first, while the features with low correlation can be discarded. For regression and classification problems, chi-square test and other methods can be used to test the characteristics to measure the correlation between independent variables and dependent variables.

The idea of chi-square test is to determine whether the original hypothesis is true by observing the deviation between the actual value and the theoretical value. First, assume that the two variables are independent (this is the original hypothesis), and then observe the degree of deviation between the actual value and the theoretical value. If the deviation is small enough, the deviation is considered to be a natural sample error, and the original hypothesis is accepted. If the deviation is large enough, the original hypothesis will be denied. For example: to test whether smoking or not is independent of health, the original hypothesis is that smoking and health are independent. If the p value measured by chi-square test is 0, it means that smoking and health are independent, indicating that smoking and health are related. The smaller the p value, the stronger the correlation.

The P values obtained by chi-square test of the three tables are as follows:

Table 4 Correlation of characteristic quantity of sensory integration assessment

|  |  |
| --- | --- |
| Characteristics of sensory integration assessment | p value |
| Vestibular and bicerebral degree | 0.25008036 |
| Neurophysiological depression of brain | 0.39117843 |
| Tactile defense and temper sensitivity | 0.04221487 |
| Exercise and daily operation during development | 0.06650252 |
| Spatial form and visual perception | 0.13543413 |
| Proprioception (gravity insecurity) | 0.15007178 |
| Learning, emotion and self- image feeling | 0.07120045 |
| Psychological stress and behavior | 0.10306098 |
| Total score | 0.31163739 |

Table 5 Correlation of cars characteristics in autism

|  |  |
| --- | --- |
| characteristics of autism | p value |
| Interpersonal relationship | 0.00172878 |
| Imitation ability (words and actions) | 0.02058875 |
| Emotional response | 0.00224271 |
| Physical ability | 0.01818358 |
| The relationship with inanimate objects | 0.03735618 |
| Appropriate response to environmental change | 0.00399149 |
| Visual response | 0.06769808 |
| Auditory response | 0.01841675 |
| Near sensory response | 0.00089975 |
| Anxiety response | 0.03174428 |
| Language communication | 0.30700938 |
| Nonverbal communication | 0.68931038 |
| Activity level | 0.20431492 |
| Intellectual function | 0.15601021 |
| Total score | 0.02099851 |

T Table 6 correlation of ICD-10 autism characteristics

|  |  |  |
| --- | --- | --- |
| ICD-10 | number | p value |
| Behavior defects of verbal communication | 1 | 1.20750784e-02 |
| 2 | 1.29875093e-05 |
| 3 | 1.44507467e-03 |
| 4 | 5.08721794e-06 |
| Qualitative defects of verbal communication | 1 | 8.50329841e-01 |
| 2 | 5.47869906e-01 |
| 3 | 3.95254880e-01 |
| 4 | 5.31564751e-03 |
| Stereotyped, repetitive and limited behaviors | 1 | 4.17188566e-05 |
| 2 | 7.63923534e-06 |
| 3 | 3.69259673e-04 |
| 4 | 4.38980731e-03 |
| Abnormal or delayed function occurred before 3 years old | 1 | 2.53877948e-06 |
| 2 | 4.40109005e-01 |
| 3 | 1.23744700e-04 |

After testing, the following features are selected as input parameters：

1,2,3,4,5,6,7,8,9,10,11,12,15,18,24,26,27,28,33,34,35,37,39

Among these characteristics, 1-9 were the characteristic values of P < 0.4 in sensory integration assessment, which were vestibular and biconcentration degree, neurophysiological inhibition state, tactile defense and temper sensitivity, exercise and daily operation, spatial morphology and visual perception, proprioception (gravity insecurity), learning, emotion and self-image, psychological stress and behavior performance.

10,11,12,15,18,24 indicated the characteristic values of P < 0.01 in cars table, which were interpersonal relationship, imitation (words and actions), emotional response, appropriateness to environmental change, anxiety response and intellectual function.

26,27,28,33,34,35,37,39 represent the characteristic values of P < 0.002 in ICD-10 autism table. Among them, 26, 27, 28 represent the last three indicators of "speech communication behavior defect", 33, 34, 35 represent the first three indicators of "stereotyped, repetitive, limited behavior, interest and activity", and 37 and 39 represent the first and third indicators of "abnormal or delayed function and appearing before 3 years old".

Selecting part of the feature quantity can effectively improve the training speed. When selecting the feature quantity, the initial idea is to combine all the feature quantities to sort, and select the smaller p value, but after this selection, the training accuracy rate is low, and the training model has no practical value. When the lower p value is used to train the characteristic quantity in the three tables respectively, and all the characteristic quantities are selected, the correct rate of training on small samples is consistent, and the training speed is fast. According to the above principle, the feature quantity selected according to the above principle is almost the same as the correct rate of the whole characteristic quantity participating in the training, which can achieve a good balance between speed and performance.

The above-mentioned method of feature selection using chi square test is modified in actual use to achieve better results. Therefore, we have also experimented with random forest method, random forest Forest (RF) is a statistical learning theory, which is a classifier for training and forecasting samples. It extracts multiple samples from the original samples, builds decision tree models for each sample, and then combines the predictions of multiple decision trees to obtain the final prediction results by voting, as shown in Figure 3.

The main advantage of random forest algorithm is fast operation speed, the effect of processing big data is very good, and it does not need to be independent of multiple variables, and it can not only verify the importance of features, but also can be used as a classifier.

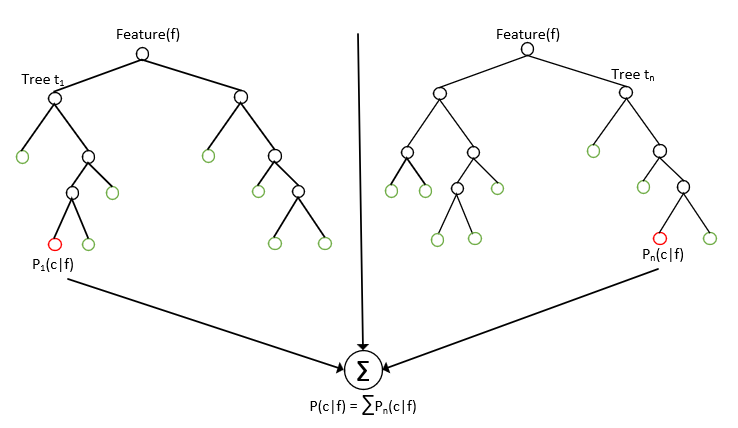


FIG. 3 Schematic diagram of random forest principle

The random forest can be interpreted as a number of independent variables (X1, X2,... , Xk) on the dependent variable Y. If the dependent variable Y has n observed values, there are k independent variables associated with it. When constructing the tree, the random forest will randomly select n observed values from the original data, among which some observed values are selected several times and some are not selected, which is the method of re-sampling. At the same time, the random forest randomly selects some variables from k independent variables to determine the classification tree nodes. In this way, the tree constructed each time may be different. In general, random forests randomly generate hundreds to thousands of trees, and then select the tree with the highest degree of repetition as the final result.

The principle for calculating the importance Of parameters in random forest is as follows: when the sample set stone is constructed, about 1/3 Of the samples will not appear in it every time, and then these samples will not participate in the establishment Of decision tree. We call this 1/3 Of the data OOB (Out Of Bag), which can be used to replace the error estimation method Of test set.

1. For each decision tree in the random forest, the OOB(out-of-pocket data) data is used to calculate its out-of-pocket data error, which is denoted as errOOB1.

2. Randomly add noise interference to feature X of all samples of OOB data outside the bag (the value of sample at feature X can be changed randomly), and then calculate its out-of-bag data error again, denoted as errOOB2.

3. Assuming random there are N trees in the forest, so characteristics for the importance of X in, can use this expression as the importance of the characteristics of corresponding measurements because: if give some characteristics of random noise interference, the accuracy greatly drop outside the bag, and shows the characteristics for the classification of the sample has great influence on the results, in that its importance is quite high.

Python was used to complete the random forest algorithm and calculate the importance of the homogenized data. Variables and importance were obtained as shown in the table:

Table 7 Importance ranking of Random Forest variables

|  |  |
| --- | --- |
| Variable name | Importance |
| IQ | 0.065876 |
| Interpersonal relationship | 0.064973 |
| Emotional response | 0.057371 |
| Appropriate response to environmental change | 0.038441 |
| ICD-10 autism 4.1 | 0.032204 |
| Imitation ability (words and actions) | 0.026246 |
| ICD-10 autism 1.2 | 0.025348 |
| ICD-10 autism 3.1 | 0.024262 |
| Near sensory response | 0.023473 |
| ICD-10 autism 4.3 | 0.023141 |
| Degree of vestibular and bicerebral differentiation | 0.022092 |
| Neurophysiological depression of brain | 0.021629 |
| Anxiety response | 0.021268 |
| Exercise and daily operation during development | 0.020971 |
| Physical ability | 0.020739 |
| The relationship with inanimate objects | 0.019898 |
| ICD-10 autism 1.4 | 0.019835 |
| Tactile defense and temper sensitivity | 0.018999 |
| Auditory response | 0.018976 |
| Spatial form and visual perception | 0.018464 |
| Total score and total evaluation | 0.017716 |
| Language communication | 0.017411 |
| Visual response | 0.016681 |
| ICD-10 autism 1.3 | 0.01663 |
| ...... | ...... |

Analysis data in the table, important degree represents the classification to predict the effects of each variable size, in the end we take important degree greater than 0.0166 24 variables as a feature of selected and obtained good effect in the back of the prediction, or even better than the effect of the chi-square above, chose before 24 variables, because for the row in the back of the characteristics of random forest algorithm to calculate the importance of indicators and there is a difference of 0.0166.

As mentioned above, in addition to evaluating the importance of features, the RANDOM forest algorithm can also be used as a classifier. Samples are divided into training sets and test sets in a ratio of 4:1.

Train set -- a sample of data used for model fitting that cannot be used to reflect the true capabilities of the model.

Test Set -- Used to evaluate the generalization capability of the final model. But it can not be used as the basis for the selection of parameters and features.

Finally, when the number of word trees is 1500 and the random number seeds are 4, the best accuracy rate is 81.5%.

### 3.2.3 Logistic regression analysis

Linear regression is the most basic regression model, which USES linear functions to describe the relationship between variables and maps the continuous or discrete independent variables to the continuous real field.

Prediction function:

To obtain the prediction function, the parameter  must be calculated and the error square cost function is introduced:



Our goal is to minimize by learning the algorithm to find the ideal model parameter. The flow chart is as follows:

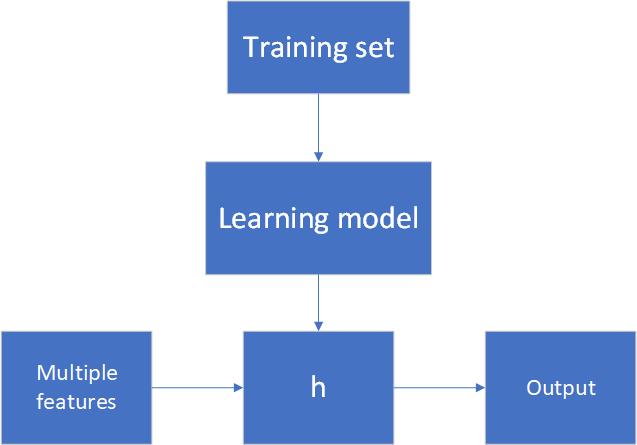


FIG. 4 Linear regression flow chart

The process of logistic regression is similar to linear regression, but the difference is the output value of linear regression may be greater than 1 or less than 0, while for a dichotomy problem, the output value will always be between 0 and 1. Therefore, if the output value is mapped to 0-1, the prediction function will become:, and the cost function will change.

The purpose of Logistic regression is to learn a 0/1 classification model from features, and this model takes the linear combination of features as independent variables, since the value range of independent variables is from minus infinity to infinity. Therefore, the estimated value of the logistic function (or sigmoid function) represents the probability that y=1 is obtained under the input value x, that is, the probability that the sample is a positive class,. Thus, when we want to determine which class a new feature belongs to, we only need to require that if  is greater than 0.5, it belongs to the class y=1, and if not, it belongs to the class y=0.

Prediction function:，where。

The cost function：



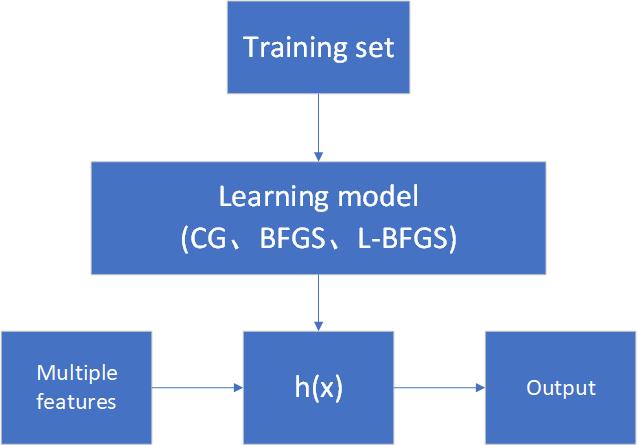


Figure 5 Logistic regression flow chart

The general steps of logistic regression analysis are as follows:

First, the data will be read by pandas, and after the data is processed, values will be converted to the matrix for operation. Then, sigmoID function, cost function and gradient will be defined, and the solution parameters will be automatically optimized in a more advanced algorithm.

The code will now be implemented using three high-level algorithms: Conjugate gradient method, quasi-Newton method (BFGS, L-BFGS).

Conjugate gradient method is an iterative method, which is often used to numerically solve partial differential equations, so it is suitable for sparse matrix systems or unconstrained optimization problems. The advantage is that it is relatively easy to implement and the optimization speed is very fast when the training sample is large enough. The disadvantage is that some parameters need to be adjusted artificially, such as learning rate and convergence criterion.

Quasi - Newton method is the most effective optimization algorithm. The advantage is that only the first derivative is needed in the iteration, and the Hessen matrix is not needed. For the general case, it has the superlinear convergence rate, and it also has the second-order convergence rate of N steps. But its disadvantage is that the required storage capacity is large, for large problems will face storage difficulties.

Table 8 Accuracy rates of the three algorithms

|  |  |
| --- | --- |
| Logistic regression | Accuracy |
| BFGS | 0.7786 |
| NCG | 0.7786 |
| L-BFGS-B | 0.7786 |

Since logistic regression adopts linear regression, the regression function has only 26 effective parameters when the sample size is 542, and the exact sample number obtained by the three regression algorithms is exactly 422, so the accuracy rate is exactly the same under the three logistic regression algorithms.

Fitting of the data from each data point to create a better way more features, we will be mapped to all these features on the polynomial of x, until the sixth power through polynomial kernel function is low dimensional vector mapped to high-dimensional space, therefore may be seen fitting, then join L2 regularization conditions in the cost function can make up for a fitting on the impact of time shown in the table below:

Table 9 Accuracy of the three algorithms after adding polynomial kernel function

|  |  |
| --- | --- |
| Logistic regression | Accuracy |
| BFGS | 0.8063 |
| NCG | 0.8007 |
| L-BFGS-B | 0.7952 |

It can be seen from the table that, after the introduction of polynomial kernel function, the number of parameters of regression function is greatly improved, and the accuracy is also improved to some extent. Because the regularization condition is added, the over-fitting will not occur due to the excessive number of parameters. In this case, the correct number of SAMPLES of BFGS, Newtonian conjugate gradient and L-BFGS was 437, 434 and 431 respectively, and the quasi-Newton method (BFGS) performed the best with an accuracy of 80.63%.

### 3.2.4 Relationship between Logistic regression and Support Vector Machine

Try to transform the logistic regression. First, replace the result tags y = 0 and y = 1 with y = -1,y = 1, then replace the b in () with , and finally replace  with . So . In other words, except that y goes from y=0 to y=-1, the linear classification function is no different from  of logistic regression.

### 3.2.5 Support Vector Machine

Support Vector Machine (SVM) is a linear supervised Machine learning method for classification and regression. It is mainly used for classification and will not cause overfitting problems [12]. Due to the existence of kernel functions, SVM algorithm can solve the nonlinear separable problem [13] by converting data to a higher space, and achieve good classification effect [14] by selecting appropriate kernel functions.

Instead of one hyperplane, SVM should fit two hyperplanes, and the distance between the two hyperplanes should be as large as possible. It mainly focuses on the edge data points (support vectors) close to the classification boundary. Convex optimization method is adopted to maximize the interval between the positive examples and the negative examples, so as to minimize the number near the hyperplane of the boundary, avoid the misjudgment of sample points near the boundary, and improve the accuracy.

Used  to represent the functional interval between points and the partition plane, and use  to represent the geometric interval between points and the partition plane. Then, the current research goal is to find w and X in the classifier definition, that is, to find their support vectors. Once the number with the minimum interval is found, the interval should be maximized. The function can be written as:

 (1)

It is a problem to find the optimal solution after given constraints. By introducing Lagrange multiplier, the hyperplane can be written into the form of data points based on constraints, and the optimization objective function can also be written as:

 (2)

The constraint condition is：

 (3)

Among them, the constant C is used to control the weights of the two goals of "maximizing the interval" and "ensuring that the function interval of most points is less than 1.0", which is called the relaxation variable. In the high-dimensional feature space, it is not necessary to carry out the feature transformation in accordance with the above transformation, but to carry out the inner product operation, and the inner product operation can use the function in the original space, without knowing the form of the transformation. According to the theory of functional analysis, if a selected kernel function satisfies the relevant conditions stipulated by Mercer, it will take the inner product of a certain high-dimensional feature transformation space. Thus, as long as the appropriate inner product function can be found in the optimal classification surface, the linear classification after nonlinear transformation can be realized without increasing the computational complexity [15].

The introduction of kernel function provides a method to solve the linear inseparability problem effectively. The complexity of the evaluation object is predicted, and then the kernel function and the optimal parameter value are determined according to the complexity of the evaluation object. This is the core content of establishing support vector machine model, which is related to the validity of the model. In practice, kernel and parameter values are usually selected based on sample data. The common kernel function types are shown in table [16].

Table 10 Common kernel function types

|  |  |  |
| --- | --- | --- |
| Kernel function type | expression | Parameter value |
| Linear kernel |  |  |
| Kernel polynomial |  |  |
| Laplacian nucleus |  |  |
| Gaussian kernel |  |  |
| Sigmoid core |  |  |

When the number of features is large and the number of training sets is small, logistic regression or SVM (linear kernel function) without kernel is generally selected.

In the end, two hyperplanes B and C with the furthest distance are obtained by SVM, as shown in FIG. 5. Then, hyperplane A is selected as the decision hyperplane because it has the best "tolerance" for local disturbance of the training sample and the strongest classification robustness. , for example, due to the limitation of the training set or noise interference, training sets of samples may be worse than the figure 5 training sample is closer to two classes in the current separation boundary, at the time of classification decision errors will occur and the hyperplane affected A minimum, that is to say, the results generated classification hyperplane A is one of the most robust, is the most reliable, the generalization ability of the strongest to did not see sample. Hyperplane A is the final classification hyperplane obtained by SVM.

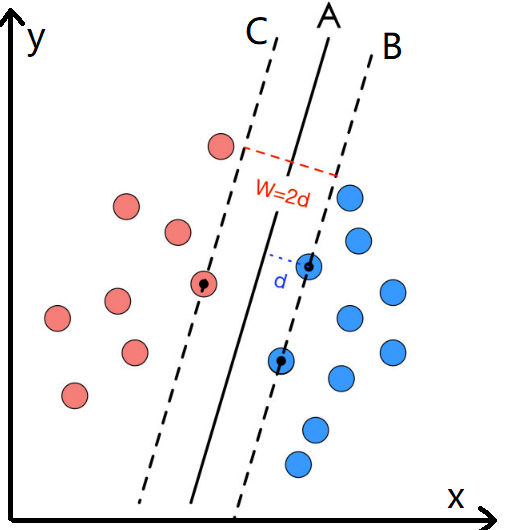


Fig.6 HYPERplane of SVM

### 3.2.6 Sequential Minimal Optimization

Developed by Platt, SMO (Sequential Minimal Optimization) is one of the most effective solutions to the SVM training phase, efficiently solving the dual problem in SVM. It is derived by solving a series of quadratic programming (QP) problems. In each iterator, two variables alpha(1, 2) are selected in the working set for optimization. If alpha 1 is determined, alpha 2 is then obtained [16]. These QPS decomposed one by one tend to be easier to calculate, so smOs generally have lower time and space complexity. Convergence, in order to determine whether can check the KKT conditions, due to the KKT conditions itself is relatively rigid, so need to set a tolerance value tol, in general the value of this parameter between 0.001 to 0.01, that is, all samples within the scope of the tolerance values meet the KKT conditions that training can end, implementation steps that is repeated until the following process of convergence:

(1) Choose two Lagrangian multipliers and ;

(2) Fix other Lagrangian multipliers （k is not equal to I and J), and only optimizeand ;

(3) Update the value of intercept B according to the optimized  and  [15];



Figure 7 SMO algorithm flow chart

## 3.3 Model validity test

In the construction of SMO+SVM prediction model, the key is to determine penalty parameter C, error accuracy toler of SVM stop training, iteration times and kernel parameters. If the setting of these major parameters is unreasonable or the parameter values are too large or too small, it will directly lead to the problem of overfitting or insufficient fitting in the training process, and then affect the accuracy of the evaluation model.

The total number of existing samples is 542, 500 samples are randomly selected as the training set, and the other 42 samples are selected as the test set. By importing the Sklearn algorithm package in Python and using SVM for training data, the highest accuracy rate obtained by the test is 87.805%, as shown in the figure below:

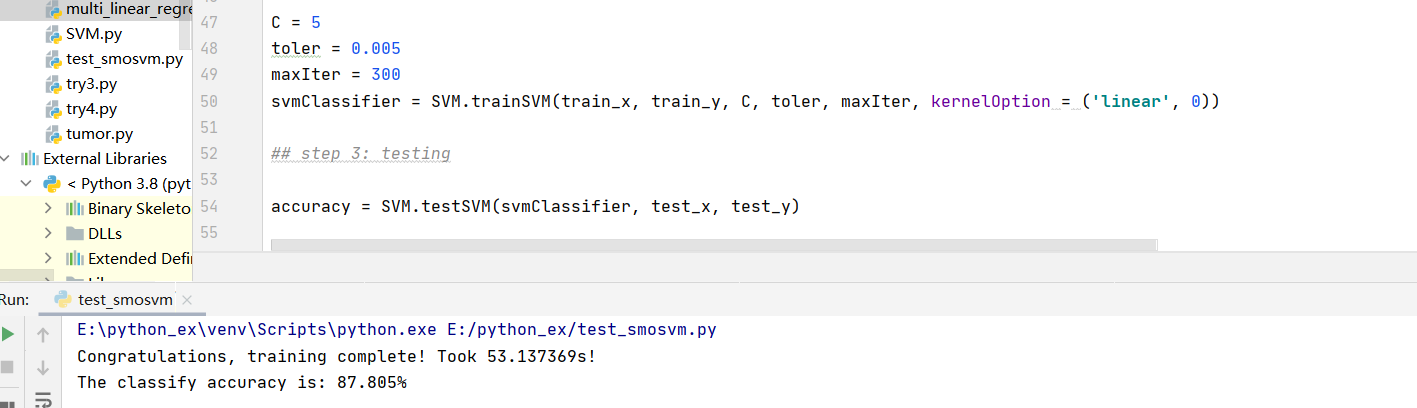
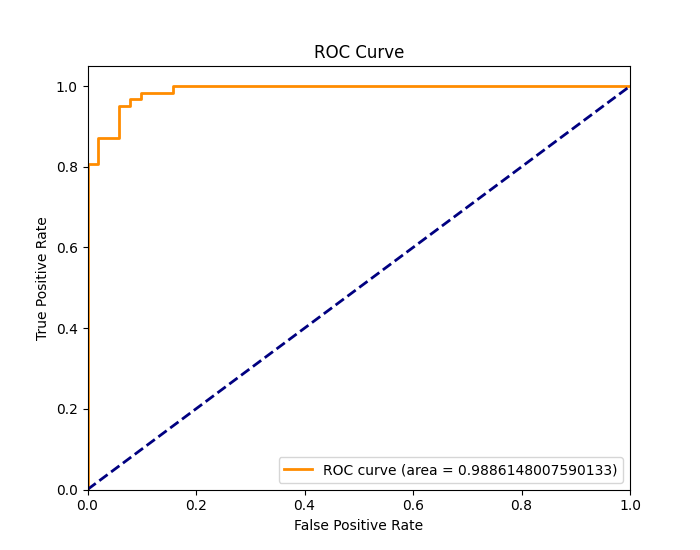


Figure 8 Partial codes and accuracy after training

As shown in the figure below is the confusion matrix and the corresponding ROC curve of the random forest classification algorithm



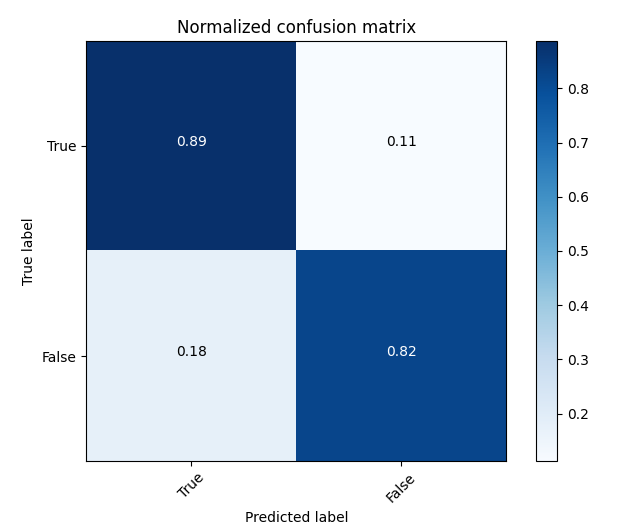


Fig. 9 confusion matrix and ROC curve of random forest

Table 11 summarizes the accuracy of three methods for autism prediction, including Logistic regression analysis, random forest classification and SVM+SMO. It can be seen that the SVM+SMO method has the highest accuracy.

|  |  |
| --- | --- |
| algorithm | accuracy |
| Logistic regression analysis | 80.6% |
| Random forest classification | 81.5% |
| SVM+SMO | 87.8% |

Table 11 Comparison of accuracy of the three algorithms

# 4. Discussion

The prevalence rate of autism has almost reached 1% worldwide, making the early precaution and diagnosis of autism extraordinarily meaningful for innumerable families or even the whole community. The diagnosis of autism highly depends on several rating scales, from which results are calculated by simple accumulations to control the cost and ensure efficiency, yet the fixed standard cannot properly adapt to new changes as well as the undiscovered inner links. Considering the extent and complexity of related indicators, and a need for adaptation, the cutting-edge artificial intelligence tool is worth a try. The relatively accurate results that come out quickly is helpful for publicizing self-test of potential patients and earlier intervention. This research, based on a certain extent of patient data, using logistic regression, the idea of SMO+SVM, and random forest, constructed several prediction models and acquired accuracy that is statistically significant.

We designed updatable prediction models that can adapt to change of the group’s overall characteristics once latest data set is added, almost eliminating deviation between individuals. The result will be produced only a few seconds after a click, effectively reducing the calculation time, compared with the traditional diagnosis procedure. The more data is added to the income of the model, it’s likely that the more accurate the result will become.

Our study included 5 models based on 3 different methods. Previous researches has shown that a model has excellent application value if its error rate is under 10% or 20%[8]. The accuracy of model under the SVM+SMO idea has reached an accuracy of 87.8%, much higher than the international standard of usable scales, which is 81%. Such model is helpful for monitoring the condition of already diagnosed patients and pursue precise medical care. On the other hand, logistic regression models, though having accuracy not that remarkable, perform better after some adjusting, and is exceptionally timesaving. These ones are quite suitable for risk screening at the outpatient, saving labor and time. Indeed, the increase of speed sacrifices attention devoted in time, which may lead to situations that the subject doesn’t completely understand the variables and fill out results arbitrarily. This requires further effort for improvements.

In the feature screening part, the most important variables like IQ and interpersonal relationship are consistent with existing neurological research results, confirming the significance of our work and potential research value of discovering unknown determinants in the future. By analyzing features in different scales, eliminating the overlapped ones, and compositing a more concise and effective new rating scale, basic prediction will become easier and more acceptable. With the convenience, it’s much more practical to disseminate self-test to a wider population.

Also, considering the insufficiency of data, there is still room for improvement. We believe that if there is more available data, the quality of our model will scale new heights.

# 5. Conclusion

The machine learning method has achieved higher accuracy than the traditional scale in the establishment of the scale-based autism prediction model for children, which has certain research significance and contributes to the improvement of autism awareness and screening efficiency, new scale compilation and later follow-up treatment. It is hoped that this study will play its role in the future. It is also hoped that in the future more predictive models will add more objective factors such as eye tracking and functional magnetic resonance imaging to the scale to explore its relationship with autism. Accurate prediction of autism is to intervene in the development of the disease earlier, bring help to the patient and the family within the capacity, save time and financial resources, and more importantly, strive for recovery and healthy body

# 6. Future Outlook

The results of autism prediction models can create AI aided diagnosis opportunity for hospitals, providing reliable reference for doctors, saving time and money. Although machines cannot replace people anyway, especially in the information acquired during doctors’ inquiry, the prediction models still offer convenient and relatively reliable reference.

Furthermore, in our society, huge autism population is underrepresented, with much of the parents unaware of the prevalence rate of autism. We would like to encourage more people to do self tests, and if they are autism patients, early intervention will speed up recovery. To achieve this goal, we can design a program package or even a smartphone app for the public. This will be instrumental to the dissemination of such a tool.

And with an updatable prediction system, medical personnels can track the condition of already diagnosed patients. With fresh information, therapists and experts will be able to design personalized treatment plans for each single case from time to time. As the cases accumulate, this system will become more mature, benefiting hundreds of thousands of autism patients and bringing hope and love to innumerable families.

# References

1. American Psychiatric Association: Diagnostic and Statistical Manual of Mental Disorders: DSM-V. (5th edition). Washington, D.C.: American Psychiatric Publishing, 2013
2. Society of Neuroscience. 2018. *Brain Facts*. 136pp.
3. Mayada et al. Global prevalence of autism and other pervasive developmental disorders. [Autism Res](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3763210/). 2012 Jun; 5(3): 160–179.
4. Maenner MJ, Shaw KA, Baio J, et al. Prevalence of Autism Spectrum Disorder Among Children Aged 8 Years — Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2016. MMWR Surveill Summ 2020, 69 (No. SS-4): 1-12.
5. Vargason, T., Grivas, G., Hollowood-Jones, K. L., & Hahn, J. (2020). Towards a Multivariate Biomarker-based Diagnosis of Autism Spectrum Disorder: Review and Discussion of Recent Advancements. Seminars in Pediatric Neurology, 100803.
6. 柳毅恒. 基于循环神经网络的自闭症谱系障碍预测框架[D]. 2018.
7. Aron L, Loprest P: Disability and the education system. Future Child 22:97-122, 2012
8. 孙宾宾. 2-6岁孤独症谱系障碍儿童眼动注视特征研究及诊断预测模型的建立[D]. 2018.
9. LU H L,TANG J,HUANG Z W, et al．Development and value evaluation of a simple prediction model of suicidal behavior in depressive disorder［J］．Chinese General Practice,2020,23(26)：3247-3252．
10. Walsh, C. G., Ribeiro, J. D., & Franklin, J. C. (2017). Predicting Risk of Suicide Attempts Over Time Through Machine Learning. Clinical Psychological Science, 5(3), 457– 469.
11. McPartland JC: Considerations in biomarker development for neuro- developmental disorders. Curr Opin Neurol 29:118-122, 2016
12. Nell Critianini，John Shawe-Tayer. 支持向量机导论[M].北京:电子工业出版社.2004:82-98.
13. 王华忠，俞金寿.核函数方法及其模型选择[J].江南大学学报(自然科学版)，2006，5(4):126 -130.
14. 周晓剑，马义中，朱嘉钢． SMO 算法的简化及其在非正定核条件下的应用［J］． 计算机研究与发展2010，47( 11) : 1962－1969．
15. 赵长春，姜晓爱，金英汉．非线性回归支持向量机的 SMO 算法改进［J］． 北京航空航天大学学报，2014，40( 1) : 125－130．
16. 李航．统计学习方法［M］．北京: 清华大学出版社，2012．

# Acknowlegements

Contestant Tianyi Chen is a 12th grade student from Nanjing Foreign Language School. She has an enthusiasm for life sciences, awarded national second prize in the China Student Biology Olympiad and silver prize in British Biology Olympiad. She found her interest in neuroscience after attending a summer program in Columbia University, studying the Fundamentals of Neuroscience. She attended China Brain Bee in 2020 and won national second prize, during which she got to know the team’s tutor BAI Tao.

Contestant Naicheng He is an 11th grade student from Nanjing Foreign Language School. He loves maths and computer science, especially interested in data analysis and python programming. He won first prize of China division in Duke Mini Mathematics competition 2020…

Tutor Tao Bai is a PhD candidate in Laboratory of Neurobiology of Context and Behavior, Institute of Neuroscience, Chinese Academy of Science. His research interests are the neural mechanism of PTSD and the produce of discontinuous event-related learning and memory.

The contestants interned at Nanjing Tianyou Children Hospital in the summer of 2020 and realized the difficulty that patients with neurological diseases and their parents have to handle. If diagnosed early and immediately receive rehabilitation, the patients would recover better. However, due to the weak consciousness of self test in the society, many families miss the best time for treatment. Equipped with neurological, data analyzing, and machine learning knowledge, the team wonder if they could help by utilizing the data, and started to do the research under the guidance of Mr. Tao Bai.

Chen and He together finished the work of data gathering and processing as well as thesis writing. The team acquired the admission of ethics committee of Nanjing Tianyou Hospital to use the data for non-profit scientific research, and personal information of the patients is hidden. Chen was in charge of researching related neurological knowledge document, while He analyzed the data and finished work regarding coding. Mr. Bai had brainstorms with contestants, guiding them to find the research direction and provided suggestions, but all of the important choices and decisions are made by the contestants. The team hereby thanks Mr. Bai for his dedicated guidance.