Research on identifying the optimum behavior set of Autism

# Liping Ma1, Haitao Yang1, Jixin Wen1, Aijun He (✉)1

1 School of Electronic Science and Engineering, Nanjing University, Nanjing 210023, China;

**Abstract Autism Spectrum Disorder (ASD) is one of the fastest-growing neuro-disorder diseases in the last several years with a morbidity of 1 in 69. The disease has a lifelong effect on people’s ability to communicate and interact with others. Therefore, providing a quick, simple, and effective disease predicting assessment to these families is vital, which may increase the likelihood they will seek professional assessment. Using six different feature selection methods, we trained and tested two machine learning models on clinically ASD datasets to diagnose ASD. The dataset of children ASD screening has 561 instances and 57 attributes. After applying various machine learning techniques and handling missing values, the results strongly suggest that Sequential Backward Selection (SBS) based logistic regression (LR) model has the best performance on all these datasets. We found that 16 out of the 57 behavior items from the dataset are sufficient to detect ASD risk with 0.9048 AUC and 88.49% accuracy. A greater than 71% reduction in the number of behaviors yet maintained competitive AUC, accuracy, sensitivity, and specificity rates suggests a role for computational and statistical methods to streamline ASD risk detection and screening. The result pinpointed several most relevant behaviors that could help guide future efforts focused on expeditious observation-based screening to speed up children’s ASD diagnosis.**

# **Keywords** autism; neuro-disorder; feature selection; machine learning

# **Introduction**

The World Health Organization (WHO) estimated that 0.76% of the world’s children have ASD [1]. Although the average age of diagnosis is around 4–5 years old [2,3], studies have suggested that early behavioral signs of ASD may be detected before clinical diagnosis as early as before the first year of life [4]. Over the past decade, the health and education communities have witnessed a serious concern for children with ASD across the country [5].

At present, the diagnosis of ASD is not difficult in 3a grade hospitals in China. The forms of autism detection in children's hospitals mainly include Sensory Integration Assessment, Childhood Autism Rating Scale (CARS), ICD-10 ADHD, ICD-10 Autism, IQ Test. The number of questions corresponding to the five scales is 58, 15, 18, 16, 72, which can comprehensively evaluate autism and ensure the distinction between autism and other mental disorders. These forms require responses to a large number of questions which makes many of them lengthy and inefficient [6]. Besides, the diagnosis of a low age and slightly autism is quite difficult, and it requires professionals. Some surveys have found that at the age of 1-2 years, nearly half of them are missed diagnoses. Thus, even for experienced professionals, the diagnosis of autism is sometimes to be a challenge. Propelled with the rise in the use of machine learning techniques in the research dimensions of medical, applying machine learning can accelerate families to make their preliminary judgment of autism before deciding whether to go to the hospital for further diagnosis.

Limited examinations of machine learning and computational intelligence perspectives on ASD have been previously conducted, in particular regarding the reduction of feature sets in screening tools. Allison et al. [7], proposed to adapt the AQ (child, adolescent, and adult versions) and the Q-CHAT into short versions for use in primary or social care settings by busy frontline health care professionals. This study selected the top 10 features from adult AQ, adolescent AQ, child AQ, and Q-CHAT by calculating a discrimination index (DI) for each item. Nevertheless, the way items have been chosen were based on a DI, which is a simple rather than intelligent method. Other studies by Wall et al. [6,8] used artificial intelligence (decision tree classifiers) to reduce the number of items for ADOS-R in a clinical environment. The previous study claimed 8 of the 29 items contained in ADOS-R (Module 1) to classify autism with 100% accuracy. Another one showed that 7 of the 93 items contained in the ADI-R to classify autism with 99.9% accuracy. However, a later study by Bone et al. [9], revealed serious pitfalls in Wall et al. [6,8] related to conceptual and methodological issues that call into question the reliability and validity of their claims. Kosmicki et al. [10], applied stepwise backward feature selection and tested with eight machine learning algorithms to minimize behavior numbers. In this paper, 9 behaviors were effectively extracted from 28 behaviors in the ADOS module (2) and 12 from 28 behaviors in the module (3) to predict an ASD possibility with 98.27% and 97.66% accuracy respectively. However, the data used in the study were from AC, AGRE, NDAR, and SSC databases, which were highly unbalanced in the proportion of ASD and non-ASD patients. Before 2017, there were few public autism datasets. Fadi thabtah collected ASD screening datasets for children, adults, and adolescents through a mobile application, and published them on the UCI machine learning repository for public access and research. Vaishali R, Sasikala R. et al. [11], used binary firefly feature selection on ASD datasets, which has 21 features obtained from the UCI machine learning repository. It is reported that 10 features among 21 features of the ASD dataset are sufficient to distinguish between ASD and non-ASD patients and the average accuracy is approximately equal to that produced by the entire ASD diagnosis dataset. However, for swarm intelligence wrapper, the risk, time complexity, and search complexity of overfitting are not discussed. Fadi Thabtah et al. [12], proposed a new computational intelligence method called Variable Analysis (VA), which can reduce the correlation between features based on normalized scores of Chi-Square (CHI) and the Information Gain (IG) methods. VA minimized the feature number of the AQ-10 Adult, AQ-10 Adolescent, and AQ-10 Child to 6, 8, and 8 items while maintaining acceptable levels of specificity, sensitivity, and predictive accuracy by two machine learning algorithms (the RIPPER and C4.5). However, IG belongs to information theory. This method focuses not only on the class correlation but also on the internal correlation of features. Therefore, the effect of eliminating redundancy is the characteristic of IG, not VA.

Limited research has been conducted on identifying and evaluating ASD traits in the clinical environment. We show recent results and challenges when machine learning is adopted for ASD classification which future studies can consider to improve the quality of the outcome. ? This paper attempts to construct an autism prediction model with different feature selection methods, to select optimum behavior sets for autism self-screening in ordinary families. In other studies, accuracy was used as the main evaluation index, while AUC was replaced in this study. AUC is the area under the ROC curve and as a preferred evaluation measure over accuracy when evaluating and comparing classifiers [13]. It should also be noted that a lot of existing literature data are from public databases. These public databases have the problem of data imbalance or questionable data sources. The data used in this paper are all from real cases of children's Hospital, which has certain guiding significance for the preliminary screening of autism.

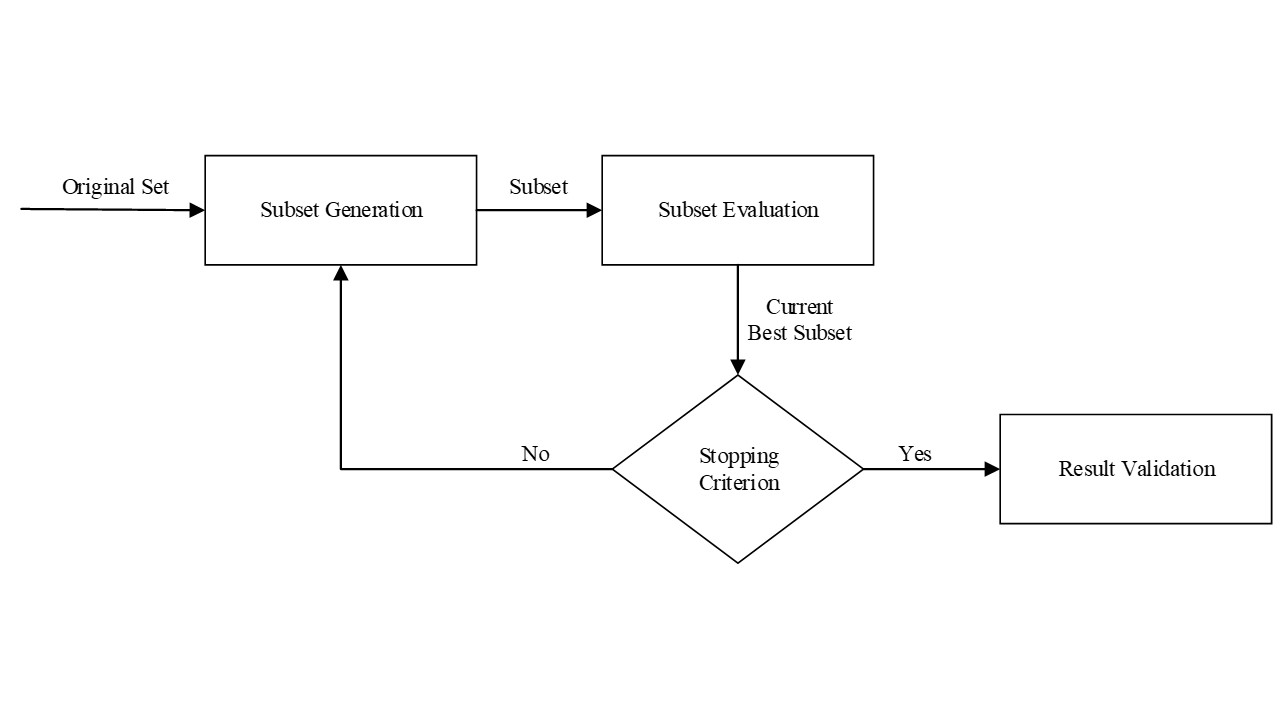
The paper is structured as follows: firstly, the feature selection method is described. Then, the datasets and results analysis is conducted and explained in detail. Conclusions are given in the last.

# **Feature selection methods**

Feature selection has been an active and fruitful field of research and development for decades in machine learning[14], statistical pattern recognition[15], and statistics [16]. It reduces the number of features, removes irrelevant, redundant, or noisy data, and brings immediate effects for applications. A typical feature selection process consists of four basic steps (shown in Fig. 1), namely, subset generation (search strategy), subset evaluation, stopping criterion and result validation [17].

Subset generation is a process of generating candidate feature subsets for evaluation according to a certain search strategy [18]. Each candidate subset is evaluated according to a certain evaluation criterion and compared with the previous best subset. If the new subset proves to be better, it replaces the previous best subset. The process of subset generation and evaluation is repeated until a given stopping criterion is satisfied. Once the stopping criterion is satisfied, an optimal subset will be output for subsequent validation. At present, the research on feature selection methods mainly focuses on the search strategy and evaluation criterion. There are three types of search strategies: complete, sequential, and random search[19]. Evaluation criteria broadly fall into three categories: the filter, the wrapper, and the embedded[20].

A complete search is to enumerate all the feature combinations in the feature set to find the optimal result. The complexity is as high as , so it is not used in practical application. Sequence search omits a large number of unnecessary search paths, reduces the amount of computation, and improves efficiency, thus risks losing



**Fig. 1**  A basic framework of feature selection.

optimal subsets. Algorithms with the sequential search are simple to implement and fast in producing results as the order of the search space is usually or less. Random starts with a randomly selected subset and proceeds in two different ways. One is to follow the sequential search, which injects randomness into the above classical sequential approaches. The other is to generate the next subset in a completely random manner. The use of randomness helps to escape local optima in the search space, and the optimality of the selected subset depends on the resources available. But in some specific problems, it will take more time than a sequence search because it does not make use of prior knowledge.

The filter applies the classifier-irrelevant metrics like -test, information gain (IG) to estimate the discriminative power of features [21]. Thus the computational cost is often lower than that of wrappers and embedded methods. In addition to their generality, filters are appropriate choices when dealing with high dimensional data. Wrappers utilize the results of a specific learning algorithm to select features. Correlation and dependencies between the features are considered while selecting the features. Considering the bias of the prediction algorithm helps in optimizing the performance of the algorithm. The main drawback of the wrapper method is computational expensiveness due to searching for the optimal set from a large space of dimensionality. Wrappers have a high risk of overfitting. Embedded search the optimal feature subset while building a classifier. Advantages of the embedded method are the same as the wrapper method but it is better in terms of computational complexity than it.

## **Sequential Backward Selection**

Sequential backward selection (SBS), in which features are sequentially removed from a full candidate set until the removal of further features does not improve the model. This sequential feature selector removes features to form a feature subset greedily.  The main limitation of SBS is its inability to reevaluate the usefulness of a feature after it has been discarded.

## **Chi-square test**

A chi-square () statistic is a test that measures how a model compares to actual observed data.  can be used to test whether two variables are related or independent from one another or to test the goodness-of-fit between an observed distribution and a theoretical distribution of frequencies. computes the correlation between variable-class using their expected and observed probabilities based on Eq. (1).

(1)

indicates the deviation between the observed value and the theoretical value, is the observed frequency, is the expected frequency if no relationship existed between the variables.

## **Information gain**

Information gain is calculated for a split by subtracting the weighted entropies of each branch from the original entropy. Entropy is defined as Eq. (2).

(2)

Where i is the number of different values that x can take. Information gain is the amount of entropy ( disorder) we removed by knowing an input feature beforehand. Mathematically, Information gain is defined as Eq. (3).

(3)

The more information gain, the more entropy is removed, and the more information does the variable X carries about Y.

## **Recursive Feature Elimination**

Recursive Feature Elimination (RFE) is a typical example of the wrapper method. The goal of RFE is to select features by recursively considering smaller and smaller sets of features. First, the estimator is trained on the initial set of features and the importance of each feature is obtained either through any specific attribute or callable. Then, the least important features are pruned from the current set of features. That procedure is recursively repeated on the pruned set until the desired number of features to select is eventually reached.

## **Random forest**

Embedded, performs feature selection during model training, and uses the resulting performance as a guide for selection. For example, feature\_importances\_ attributes in the random forest can list the contribution of each feature to tree establishment. We can base on the evaluation of this contribution, the most useful features for model building are found.

## **Variable Analysis**

Fadi Thabtah et al. [12], proposed a new computational intelligence method called Variable Analysis (VA). Combining filtering methods CHI and IG, which can further reduce the deviation of the results. Since each method employs a specific metric to evaluate the worthiness of variables. To make the results of the two methods comparable, VA normalizing their scores to one score per variable and utilizing the new score as a new metric for ranking the variables. Through the above treatment, the VA can effectively reduce variable-to-variable correlations and maintains good variable-to-class correlation.

(4)

(5)

(6)

(7)

Features with a greater value of will be ranked higher.

# **Datasets and results**

## **Data sets**

The data of 1420 children from May to July 2020 were collected from a children's Hospital in Nanjing. As shown in Table 1, the forms of autism detection in children's hospitals mainly include Sensory Integration Assessment, Childhood Autism Rating Scale (CARS), ICD-10 ADHD, ICD-10 Autism, IQ Test. We deal with nearly 7000 datasets in Python using Pandas and another handy open-source Python library, Numpy. There are some cases of missing values. Unless the nature of missing data is ‘missing completely at random’, the best avoidable method in many cases is deletion. To make sure that data is internally consistent, that is, each data type has the same content and format, data standardization was used. After data pre-processing, a case sample of 347 individuals with ASC and a control sample of 214 controls with no ASC diagnosis were selected. These children are aged between 2 and 12 years. Of the 347 diagnosed autistic children, 264 were boys and 83 were girls. The prevalence was significantly higher in boys than in girls. Given this, the attribute of gender is added. A total of 57 attributes in Table 2 are used for prediction.

**Table 1** List of ASD datasets

|  |  |  |  |
| --- | --- | --- | --- |
| No. | Dataset Name | Number of Attributes | Number of Instances |
| 1 | Sensory Integration Assessment | 8 | 561 |
| 2 | CARS | 14 | 561 |
| 3 | ICD-10 Autism | 15 | 561 |
| 4 | ICD-10 ADHD | 18 | 561 |
| 5 | IQ test | 1 | 561 |

The study used six different feature selection methods on datasets and built a model with 80% of the data for training and the remaining 20% for testing. Table 3 shows 20 sample data instances that have been collected from the hospital.

## **Result**

*Evaluation indices*

Common metrics derived from the confusion matrix (Table 4), such as AUC, accuracy, specificity, and sensitivity, are used to evaluate the performance of the features selected by the six methods [22]. Two machine learning algorithms have been employed, named logistic regression (LR) and support vector machines (SVM), to produce ASD classification systems from the different subsets of features chosen by SBS, CHI, IG, RFE, RANDOM FOREST, and VA. The reason for employing two different predictive algorithms is to generalize the results obtained, especially the goodness of the distinctive features.

**Table 2** List of attributes in the dataset

|  |  |
| --- | --- |
| No. | Attributes Description |
| 1 | Extent of vestibular and binauralization |
| 2 | Cerebral nerve physiological inhibition |
| 3 | Tactile defense and temper sensitivity |
| 4 | Developmental movement and daily operation |
| 5 | Spatial form and visual awareness |
| 6 | Proprioception (gravity insecurity) |
| 7 | Learning, emotion, and self-image |
| 8 | Psychological stress and behavioral performance |
| 9 | Interpersonal relation |
| 10 | Imitation (words and actions) |
| 11 | Emotional response |
| 12 | Physical capacity |
| 13 | Relationship with non-living objects |
| 14 | An appropriate response to environmental change |
| 15 | Visual response |
| 16 | Auditory response |
| 17 | Anxiety reaction |
| 18 | Language-communication |
| 19 | Nonverbal communication |
| ... | ... |
| 56 | IQ |
| 57 | sex |

True Positive (TP) positive samples predicted as positive by the model, True Negative (TN) negative samples predicted as negative by the model, False Positive (FP) negative samples predicted by the model as positive, False Negative (FN) positive samples predicted as negative by the model.

(1)

(2)

(3)

(4)

(5)

FPR represents the proportion of all negative cases predicted to be positive, TPR represents the proportion of positive cases predicted to be positive.

The receiver operating characteristic curve (ROC curve) has a good characteristic: when the distribution of positive and negative samples in the test set changes, the ROC curve can remain unchanged. ROC curve is a comprehensive index reflecting the sensitivity and specificity of continuous variables. It reveals the relationship between sensitivity and specificity by composition method. It calculates a series of sensitivities and specificities by setting several different critical values of continuous variables and then draws a curve with the false positive rate (FPR) as the horizontal axis and the correct positive rate (TPR) as the longitudinal axis [23].

The area under curve (AUC) is the area under the ROC curve. The higher the AUC value, the better the classifier effect.

Accuracy is calculated based on the better threshold, not the better threshold in the overall distribution, but an attribute index of a random sample. The ROC curve compares the classifiers’ performance across the entire range of class distributions and error costs. AUC is a better measure than accuracy based on formal definitions of discriminant and consistency. Therefore, using AUC as a preferred evaluation measure over the accuracy, specificity, and sensitivity when evaluating and comparing classifiers [24].

*Results and analysis*

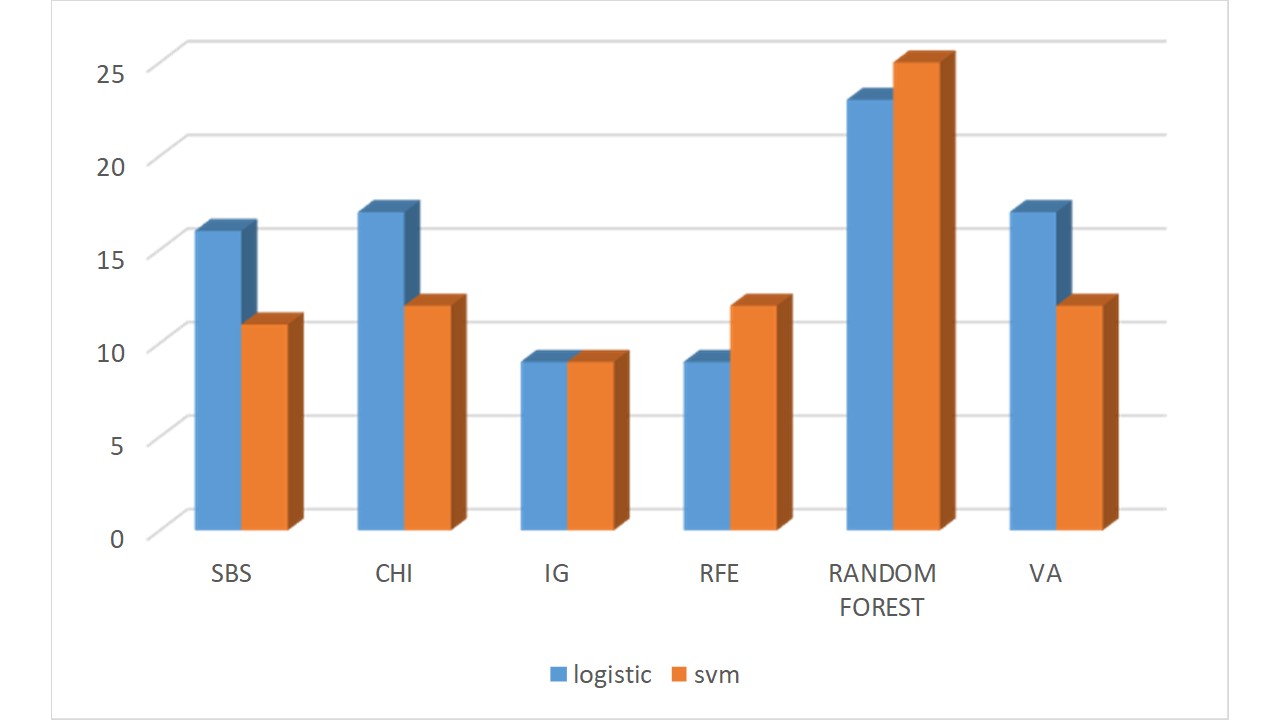
Fig. 3 shows the number of features selected by different methods under the two machine learning algorithms. In the LR model, SBS selected 16 features while CHI, IG, RFE, RANDOM FOREST, and VA are associated with 17, 9, 9, 23, and 17 features respectively. In the SVM model, the best number of features chosen by SBS, CHI, IG, RFE, RANDOM FOREST, and VA are 11,12,9,12,25 and 12 respectively. It is clear from the derived figures that IG consistently selects a lesser number of features considered in comparison with the other methods. This indicates that IG not only focuses on class-relevance but also on feature inner correlations to explicitly eliminate redundancy[24].

Fig. 4 and Fig. 5 display the AUC and accuracy rates derived by the LR and SVM classifiers from the different distinctive feature sets’ data. These results show that the highest moments of AUC and accuracy occur in the LR model, which are as high as 0.9048 and 88.49% respectively. The AUC and accuracy rate of the classifier derived from the SBS was 4.5% and 9.9% higher than the original set of features when the LR

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Table 3**  Sample 10 data instances collected from Children's Hospital | asd | n | n | y | y | n | y | y | y | n | y | The corresponding table of 1-8 belongs to Sensory Integration, which is divided into five grades. 0 for normal, 1 for marginal, 2 for mild, 3 for moderate, and 4 for severe; Item 23-37 belongs to ICD-10 Autism, Item 38-55 belongs to ICD-10 ADHD, 0 and 1 represent yes and no respectively. |
| 57 | f | m | m | m | m | f | m | m | m | m |
| 56 | 80 | 80 | 55 | 77 | 79 | 55 | 47 | 81 | 72 | 53 |
| 55 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| 54 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| 53 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 |
| 52 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 51 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 50 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 |
| 49 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 |
| 48 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 47 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 46 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 1 |
| 45 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 1 |
| 44 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 43 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 1 |
| 42 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| 41 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 0 |
| 40 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 39 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 38 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 |
| 37 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 |
| 36 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 |
| 35 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 1 |
| 34 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 |
| 33 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 |
| 32 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 |
| 31 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 |
| 30 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 |
| 29 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 1 |
| 28 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| 27 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 |
| 26 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 1 |
| 25 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 1 |
| 24 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 0 |
| 23 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 22 | 2 | 2 | 2 | 2 | 2 | 2 | 3 | 2 | 2 | 3 |
| 21 | 2 | 3 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| 20 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 1 | 2 | 2 |
| 19 | 3 | 3 | 3 | 3 | 3 | 4 | 3 | 2 | 3 | 3 |
| 18 | 1 | 1 | 2 | 2 | 1 | 2 | 2 | 2 | 1 | 3 |
| 17 | 2 | 2 | 2 | 1 | 1 | 3 | 2 | 2 | 1 | 2 |
| 16 | 2 | 1 | 3 | 2 | 1 | 3 | 2 | 2 | 2 | 2 |
| 15 | 2 | 2 | 2 | 2 | 1 | 2 | 2 | 2 | 1 | 2 |
| 14 | 2 | 2 | 3 | 2 | 2 | 3 | 2 | 2 | 1 | 2 |
| 13 | 2 | 2 | 3 | 3 | 3 | 2 | 2 | 3 | 1 | 2 |
| 12 | 3 | 3 | 2 | 2 | 2 | 2 | 2 | 2 | 1 | 3 |
| 11 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 1 | 2 |
| 10 | 1 | 3 | 3 | 3 | 2 | 3 | 3 | 2 | 3 | 3 |
| 9 | 2 | 3 | 3 | 2 | 2 | 2 | 3 | 2 | 2 | 3 |
| 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 7 | 0 | 2 | 2 | 2 | 2 | 1 | 0 | 0 | 0 | 1 |
| 6 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 0 | 0 | 0 | 0 | 0 | 2 | 2 | 0 | 0 | 0 |
| 4 | 0 | 1 | 1 | 0 | 0 | 3 | 0 | 0 | 0 | 1 |
| 3 | 2 | 0 | 2 | 0 | 2 | 1 | 2 | 1 | 0 | 2 |
| 2 | 0 | 1 | 1 | 1 | 2 | 1 | 1 | 2 | 1 | 1 |
| 1 | 1 | 0 | 4 | 2 | 1 | 2 | 0 | 0 | 2 | 0 |

**Table 4** Confusion matrix

|  |  |  |
| --- | --- | --- |
|  | Predicted ASD Value | |
| Actual ASD Value | True Positive (TP) | False Positive (FP) |
| False Negative (FN) | True Negative (TN) |

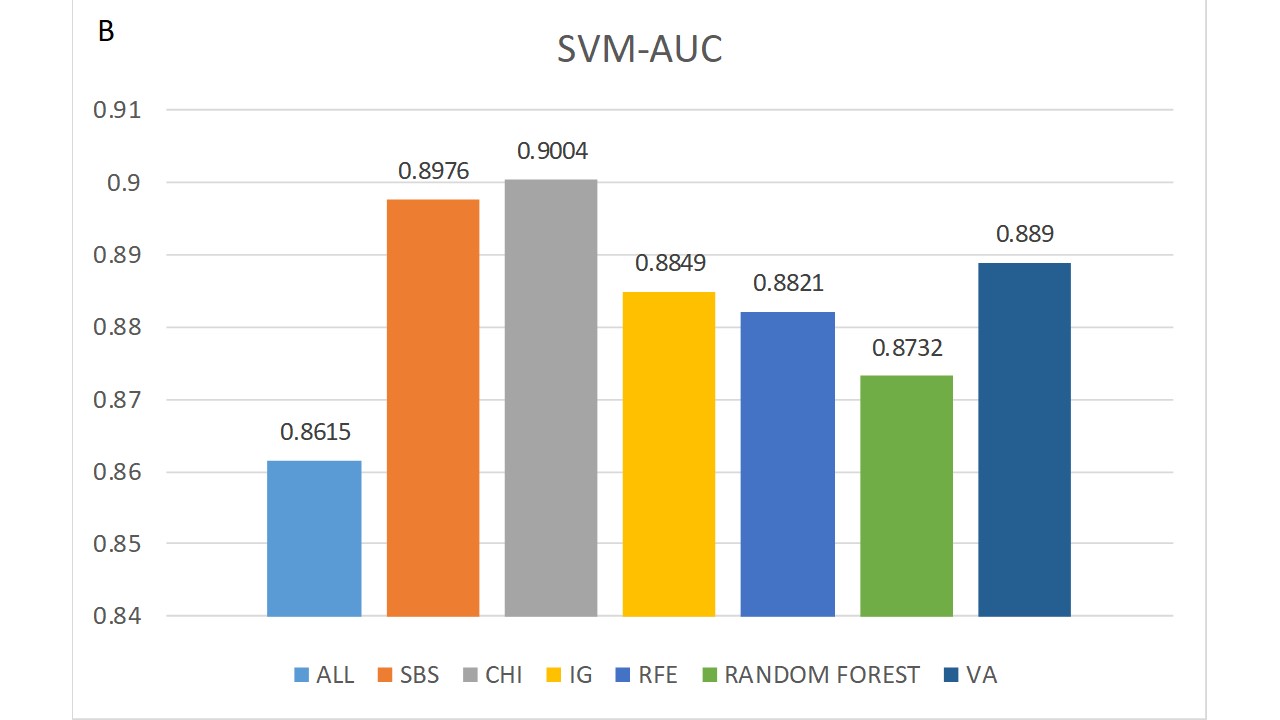
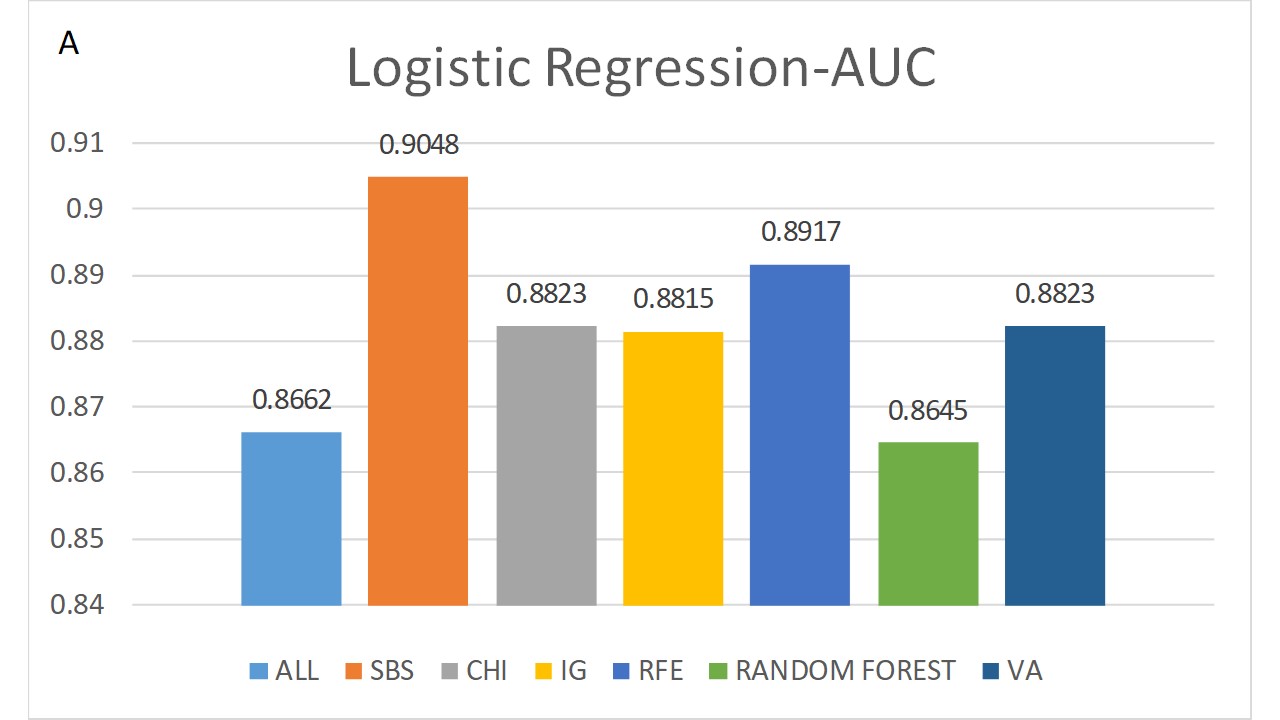


**Fig. 2** Number of variables selected from the ASD dataset based on the best AUC

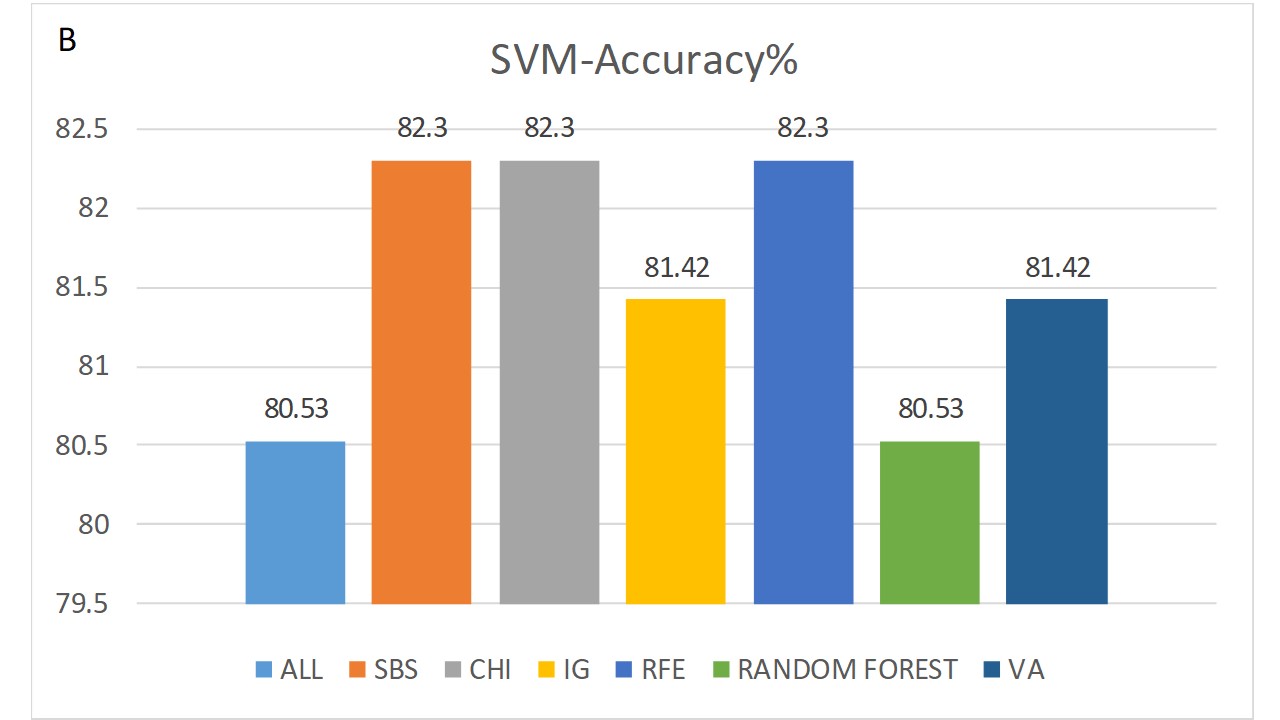
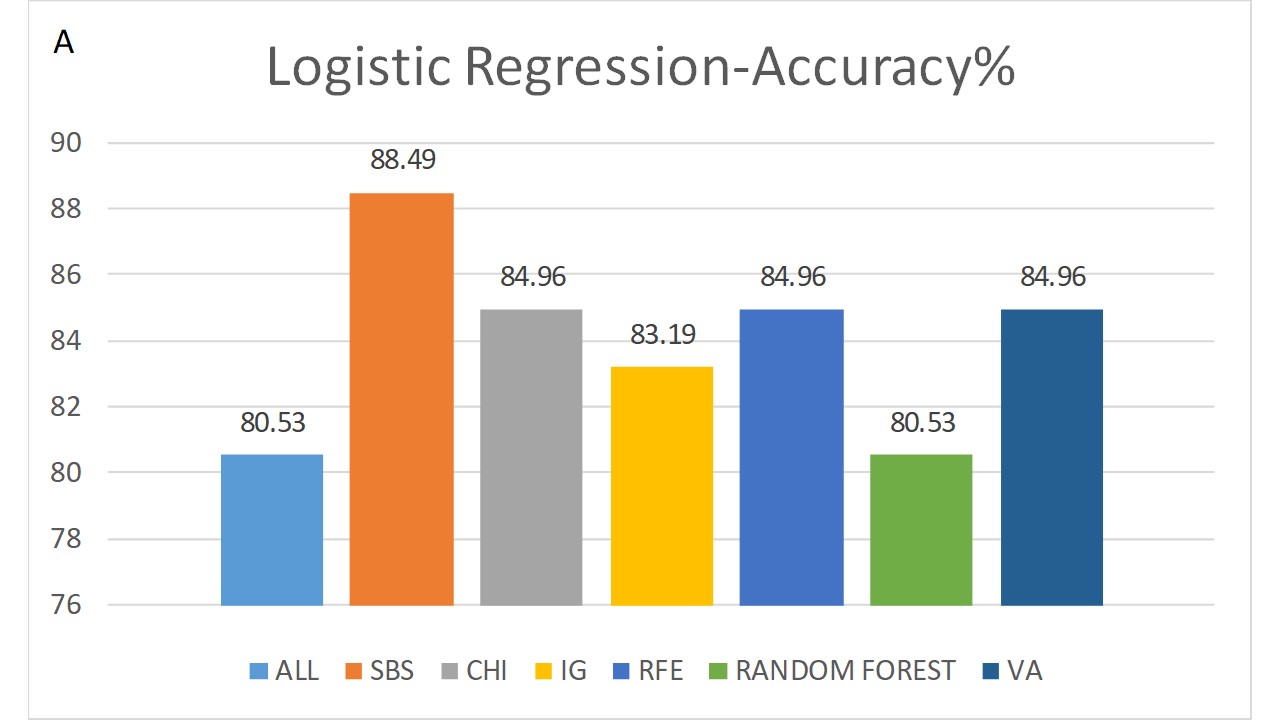
algorithm was utilized. SBS maintained an acceptable number of features and, more importantly, was able to significantly improve the AUC and accuracy. Seven more than the minimum number of features can be tolerated in light of the less than double in AUC and accuracy by the other methods considered. On the other hand, if the number of features is too small, the fault tolerance of the results will be greatly reduced. The selected features are mainly used for family preliminary screening, while ordinary people are not as professional as doctors, they may make mistakes in answering some questions. Once they make an incorrect answer to a question, it may deeply affect the final judgment. Therefore, the number of 16 features is appropriate, which not only can achieve rapid detection but also has a higher fault tolerance for non-professionals.

Fig. 6 and Fig. 7 demonstrate the sensitivity and specificity rates derived by the LR and SVM algorithms. Often in medical research, including autism, acceptable levels of sensitivity and specificity should be at least 80% [25]. The sensitivity rates derived by the LR and SVM algorithms from SBS’s features were 90.48% and 89.76% respectively and were higher than the results obtained by the remaining methods. Specifically, in the LR model, SBS was superior to “no feature selection,” CHI, IG, RFE, RANDOM FOREST, and VA, having achieved higher sensitivity by 13.94%, 10.27%, 6.63%, 2.95%, 13.94%, and 10.27% respectively, thereby proving that SBS can handle noisy data better than the other methods. Similarly, the specificity selected by SBS is the highest of the two models, which are 88.49% and 82.3%. These rates clearly show a good level of specificity and the superiority of SBS.

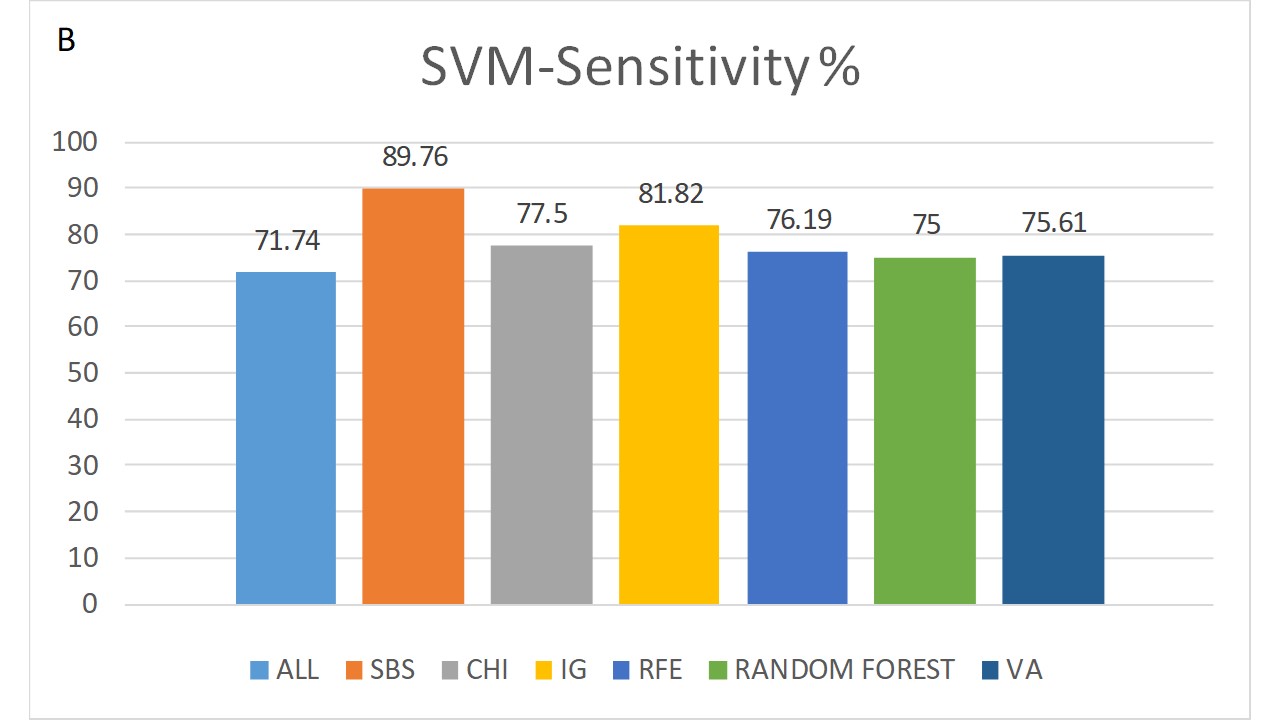
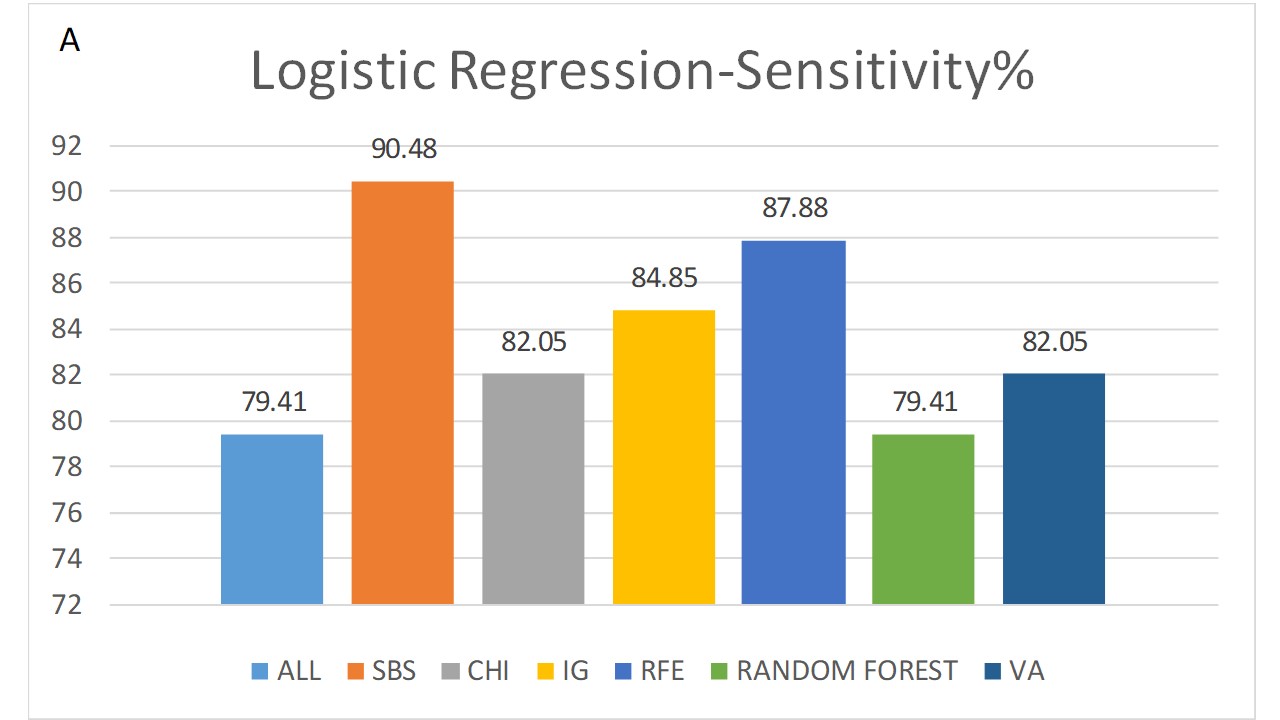
Overall, SBS is the best in all aspects, except the number of features is not the least, but a certain number of features can ensure the detection accuracy of ordinary families. Table 5 lists the 16 features selected from the SBS.



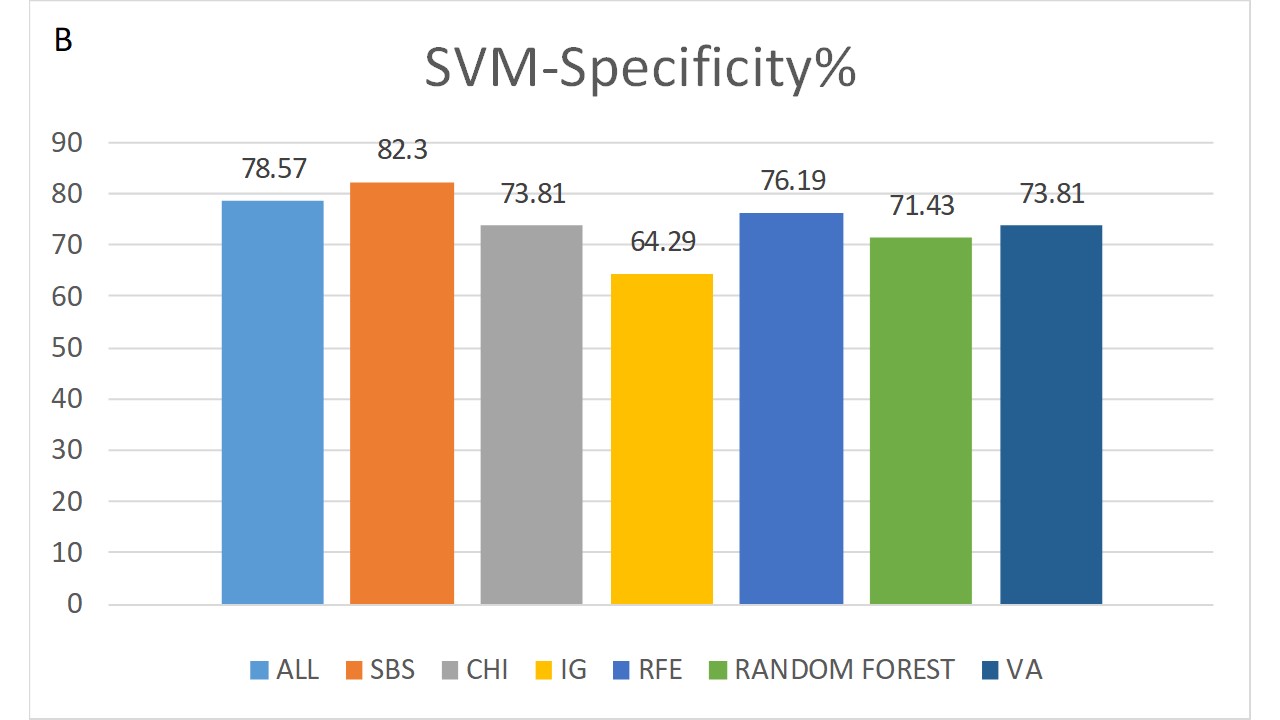
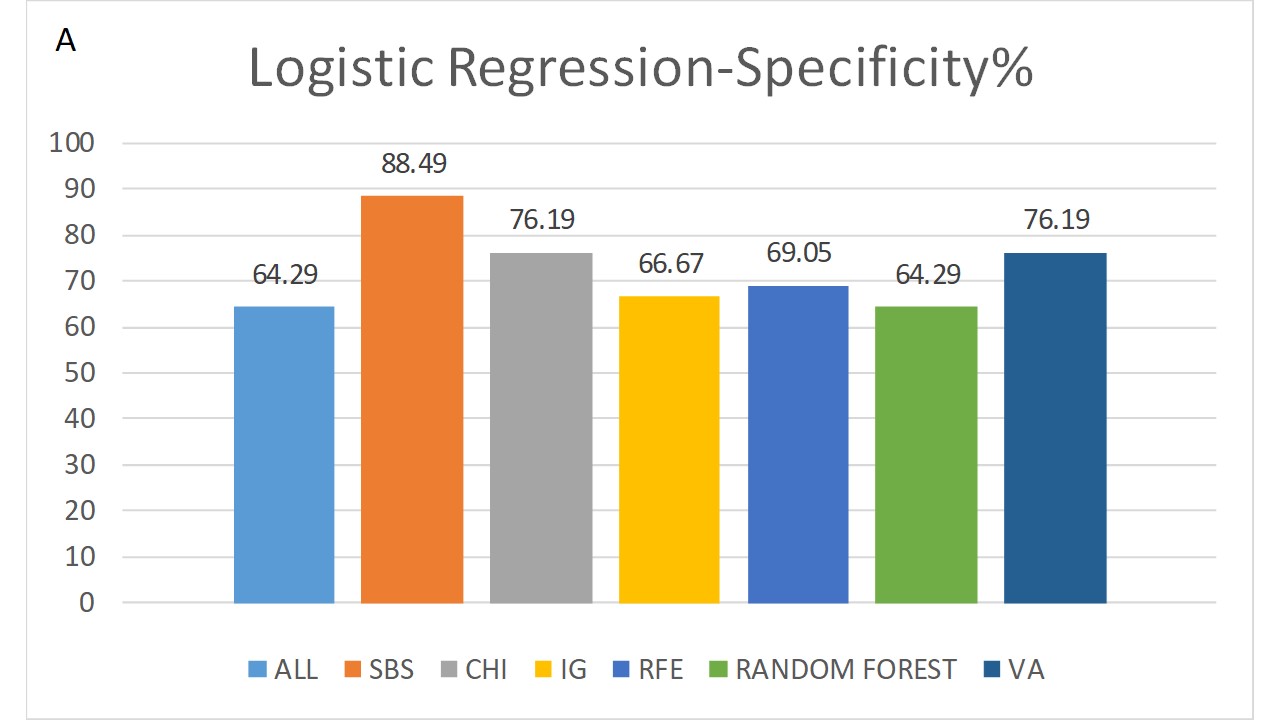
**Fig. 3**  AUC of LR and SVM algorithms against the selected subsets of data of the considered methods.



**Fig. 4** Classification accuracies of LR and SVM algorithms against the selected subsets of data of the considered methods.



**Fig. 5**  Sensitivity rates of LR and SVM algorithms against the selected subsets of data of the considered methods.



**Fig. 6**  Specificity rates of LR and SVM algorithms against the selected subsets of data of the considered methods.

**Table 5**  Selected features

|  |  |  |
| --- | --- | --- |
| No. | Attributes Description | Attribute belonged dataset |
| 1 | Emotional response | CARS |
| 2 | social communication | ICD-10 Autism |
| 3 | Interpersonal relation | CARS |
| 4 | Indulging in one or more rigid and limited interests with extraordinary attention | ICD-10 Autism |
| 5 | Symbolic or imaginative games | ICD-10 Autism |
| 6 | Lack of peer relationship with children of similar age | ICD-10 Autism |
| 7 | Adaptation to environmental change | CARS |
| 8 | Behavioral defects in verbal communication, such as eye contact, facial expression | ICD-10 Autism |
| 9 | Lack of social or emotional interrelationship | ICD-10 Autism |
| 10 | Obstinate to certain special routine or ritual actions that have no practical value | ICD-10 Autism |
| 11 | Visual response | CARS |
| 12 | Imitation (words and actions) | CARS |
| 13 | Auditory response | CARS |
| 14 | lack of spontaneous children's pretend play or social play | ICD-10 Autism |
| 15 | Lack of opportunities to spontaneously seek and share fun or achievements | ICD-10 Autism |
| 16 | Lying in bed, often twist and turn over | ICD-10 ADHD |

As can be seen from Table 5, 9 of the 16 attributes are from ICD-10 Autism, 6 from CARS, and 1 from ICD-10 ADHD. Both ICD-10 Autism and CARS are used to detect autism, which also shows the effectiveness of the two scales. On the contrary, none of the features in the Sensory Integration Assessment and IQ test are left. The hospital can safely remove the two tests when conducting autism screening in the future, which can save a certain amount of time and energy.

Among the 16 features, attributes 3, 4, 6, 9, and 16 are also included in other feature selection methods. This shows that no matter which method is used, the universality of these five features is stronger. Perhaps in the future autism detection, we can consider the importance of strengthening these items. As three of the five features come from ICD-10 Autism, which can also reflect that the ICD-10 Autism is more important than other scales. The remaining features 3 come from Autism CARS and 16 from the ICD-10 ADHD. It is worth noting that the total number of features in the ICD-10 ADHD was 18. Under the selection of these methods, only one feature of attribute 16 was selected. Therefore, this finding may also help us not to check ICD-10 ADHD in the future when detecting autism in China, and only need to evaluate attribute 16 additionally.

Moreover, it is worth mentioning that the 16 selected features can be preliminarily screened by parents under the observation of themselves, without the company of professionals. For example, the first feature: emotional response. Parents can observe the children's emotions through the game and score them as normal, mild, moderate, and severe. It's an easy job to do. On the other hand, such an evaluation method may have a certain degree of subjectivity. When designing the self-assessment form later, it may be possible to remove this concern through the reverse design form scoring mechanism.

# **Conclusion**

Data and characteristics determine the upper limit of machine learning, algorithms and models just keep approaching this upper limit. Therefore, this paper focuses on the performance of different feature selection methods in ASD screening. In general, an AUC of 0.5 suggests no discrimination, 0.7 to 0.8 is considered acceptable, 0.8 to 0.9 is considered excellent, and more than 0.9 is considered outstanding[26]. The results show that SBS has the best effect, the AUC is 0.9048 and the accuracy rate is 88.49%. Moreover, this method can effectively reduce the dimension of features by 71.93%. The selected 16 features can help early detection of ASD features, to promote the acquisition of necessary support. It can also be used in the preliminary screening of autism combined with the WeChat applet or app. If diagnosed with autism, it needs to go to professional institutions for further diagnosis and treatment. This study also can provide some guidance for future autism testing form updating, and can effectively promote the physical, social, and educational well-being of patients and families, and increase the possibility of patients' rehabilitation.

# **Compliance with ethics guidelines**

Liping Ma, Haitao Yang, Jixin Wen, Aijun He declares that they have no conflicts of interest. All procedures followed were in accordance with the ethical standards of the responsible committee on human experimentation (institutional and national) and with the Helsinki Declaration of 1975, as revised in 2000 (5). Informed consent was obtained from all patients for being included in the study.

# **References**

1. Baxter AJ, Brugha TS, Erskine HE, Scheurer RW, Vos T, Scott JG. The epidemiology and global burden of autism spectrum disorders. Psychol Med. 2015 Feb;45(3):601-13
2. Daniels AM, Mandell DS. Explaining differences in age at autism spectrum disorder diagnosis: a critical review. Autism. 2014 Jul;18(5):583-97
3. Zwaigenbaum L, Penner M. Autism spectrum disorder: advances in diagnosis and evaluation. BMJ. 2018;361, k1674
4. Maestro S, Muratori F, Cesari A, et al. Course of autism signs in the first year of life.Psychopathology. 2005;38(1):26–31
5. Huang AX, Jia M, Wheeler JJ. Children with autism in the People’s Republic of China: diagnosis, legal issues, and educational services. J Autism Dev Disord 2013;(43): 1991–2001
6. Wall DP, Dally R, Luyster R, Jung JY, Deluca TF. Use of artificial intelligence to shorten the behavioral diagnosis of autism. PLoS One. 2012;7(8):e43855. doi: 10.1371/journal.pone.0043855
7. Allison C, Auyeung B, Baron-Cohen S. Toward brief “Red Flags” for autism screening: The Short Autism Spectrum Quotient and the Short Quantitative Checklist for Autism in toddlers in 1,000 cases and 3,000 controls [corrected]. J Am Acad Child Adolesc Psychiatry. 2012 Feb;51(2):202-212.e7
8. Wall DP, Kosmicki J, Deluca TF, Harstad E, Fusaro VA. Use of machine learning to shorten observation-based screening and diagnosis of autism. Transl Psychiatry. 2012 Apr 10;2(4): e100
9. Bone D, Goodwin MS, Black MP, Lee CC, Audhkhasi K, Narayanan S. Applying machine learning to facilitate autism diagnostics: pitfalls and promises. J Autism Dev Disord. 2015 May;45(5):1121-36
10. Kosmicki JA, Sochat V, Duda M, Wall DP. Searching for a minimal set of behaviors for autism detection through feature selection-based machine learning. Transl Psychiatry. 2015 Feb 24;5(2): e514
11. Ravindranath V, Ra S. A machine learning based approach to classify Autism with optimum behaviour sets. International Journal of Engineering and Technology. 2018;7(3.18)
12. Thabtah F, Kamalov F, Rajab K. A new computational intelligence approach to detect autistic features for autism screening. Int J Med Inform. 2018 Sep; 117:112-124
13. Ling CX, Huang J, Zhang H. AUC: A Better Measure than Accuracy in Comparing Learning Algorithms. 2003; Berlin, Heidelberg. Springer Berlin Heidelberg; 329-341
14. Robnik-Sikonja and I. Kononenko. Theoretical and empirical analysis of Relief and ReliefF.Machine Learning. 2003;53:23–69
15. Mitra, C. A. Murthy, and S. K. Pal. Unsupervised feature selection using feature similarity. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2002;24(3):301–312
16. Hastie, R. Tibshirani, and J. Friedman. The Elements of Statistical Learning. Springer, 2001
17. Molina LC, Belanche L, Nebot N. Feature Selection Algorithms: A Survey and Experimental Evaluation. Proceedings of the 2002 IEEE International Conference on Data Mining (ICDM 2002), 9-12 December 2002, Maebashi City, Japan; 2002
18. Blum AL, Langley P. Selection of relevant features and examples in machine learning. Artificial Intelligence. 1997;97(1):245-271
19. Sun Z, Bebis G, Miller R. Object detection using feature subset selection. Pattern Recognition. 2004;37(11):2165-2176
20. Liu H, Yu L. Toward integrating feature selection algorithms for classification and clustering. IEEE Transactions on Knowledge & Data Engineering. 2005;17(4):491-502
21. A YZ, B RZ, C ZC, D JG, A DX. Evaluating and Selecting Features via Information Theoretic Lower Bounds of Feature Inner Correlations for High-Dimensional Data. European Journal of Operational Research. 2020
22. I. Witten, E. Frank, Data Mining: Practical Machine Learning Tools and Techniques,(2005)
23. Fawcett T. An introduction to ROC analysis. Pattern Recognition Letters. 2005;27(8):861-874
24. Bennasar M, Hicks Y, Setchi R. Feature selection using Joint Mutual Information Maximisation. Expert Systems with Applications. 2015;42(22):8520-8532
25. Towle PO, Patrick PA. Autism Spectrum Disorder Screening Instruments for Very Young Children: A Systematic Review. Autism Res Treat. 2016
26. Mandrekar, Jayawant N . Receiver Operating Characteristic Curve in Diagnostic Test Assessment[J]. Journal of Thoracic Oncology, 2010, 5( 9):1315-1316