


## PERFORMANCE OF MACHINE LEARNING METHODS IN DETERMINING THE AUTISM SPECTRUM DISORDER CASES

\*Ayşe DEMİRHAN

Electrical and Electronics Engineering Department, Faculty of Technology, Gazi University, Turkey,  
ayseoguz@gazi.edu.tr

 <http://orcid.org/0000-0001-9227-9210>

Received: 10.01.2018, Accepted: 03.06.2018

\*Corresponding author

Research Article

DOI:10.22531/muglajsci.422546

### Abstract

Autism spectrum disorder (ASD) is an inherited and neurological developmental disorder characterized by poor social interaction and communication weaknesses. In addition to the clinical methods, machine learning methods have been successfully applied to shorten the duration of the diagnosis and to increase the performance of the diagnosis of the ASD disease. Machine learning methods demonstrate high performance in the diagnosis of diseases with the objective algorithms they offer for the analysis of high-dimensional and multimodal biomedical data. Machine learning methods are successful in identifying the behavioral disorders such as OSB that include heterogeneous conditions because they capture the multivariate relationships in the data and therefore can detect subtle differences in data. In this study, analyzes are performed for the fast and accurate diagnosis of the ASD status using support vector machines (SVM), k-nearest neighbors (kNN) and random forest (RF) machine learning methods using ASD adolescent scan data and the performance of these methods are compared. Accuracy rates of 95%, 89%, and 100% are achieved as a result of binary classification with 10-fold cross-validation (CV) using SVM, kNN, and RF methods, respectively. Furthermore, 100% sensitivity and specificity values were obtained from the classification with RF method. With this study, it has been shown that ASD cases can be detected with complete success as a result of classification with RF method using ASD adult screening data.

**Keywords:** Autism spectrum disorder, machine learning, support vector machine, k-nearest neighbor, random forest

## MAKİNE ÖĞRENMESİ YÖNTEMLERİNİN OTİZM SPEKTRUM BOZUKLUĞU OLGULARININ BELİRLENMESİNDEKİ BAŞARIMI

### Öz

Otizm spektrum bozukluğu (OSB) sosyal etkileşim ve iletişim zayıflıkları şeklinde ortaya çıkan kalıtsal ve nörolojik bir gelişimsel bozukluktur. OSB hastalığının teşhisi için klinik yöntemlerin yanında teşhis süresini kısaltmak ve başarımı artırmak için makine öğrenmesi yöntemleri de başarıyla uygulanmaktadır. Makine öğrenmesi yöntemleri yüksek boyutlu ve çeşitli biyomedikal verilerin analizi için sundukları objektif algoritmalar ile hastalıkların teşhisi konusunda yüksek performans göstermektedir. Makine öğrenmesi yöntemleri, verilerdeki çok değişkenli ilişkileri yakaladığı ve bu nedenle verilerdeki ince farkları tespit edebildiği için OSB gibi heterojen durumlar içeren davranışsal bozuklukların tespit edilmesinde başarılı olmaktadır. Bu çalışmada OSB ergen tarama verileri kullanılarak destek vektör makineleri (DVM), k-en yakın komşu (kNN) ve rastgele orman (RO) makine öğrenmesi yöntemleriyle OSB durumunun hızlı ve doğru olarak teşhis edilmesine yönelik analizler yapılmış ve bu yöntemlerin performansları karşılaştırılmıştır. DVM, kNN ve RO yöntemleri kullanılarak 10-kat çapraz doğrulama ile yapılan ikili sınıflandırma işlemi sonucunda sırasıyla %95, %89 ve %100 doğruluk oranlarına erişilmiştir. Ayrıca, RO yöntemi ile yapılan sınıflamadan % 100 duyarlılık ve belirlilik değerleri elde edilmiştir. Bu çalışma ile OSB ergen tarama verilerini kullanarak RO yöntemi ile yapılan sınıflama sonucunda OSB olgularının tam bir başarı ile tespit edilebildiği gösterilmiştir.

**Anahtar Kelimeler:** Otizm spektrum bozukluğu, makine öğrenmesi, destek vektör makineleri, k-en yakın komşu, rastgele orman

### Cite

Demirhan, A., (2018). "Performance of machine learning methods in determining the autism spectrum disorder cases", *Mugla Journal of Science and Technology*, 4(1), 79-84.

### 1. Introduction

Autism is a developmental disorder diagnosed on the basis of social interaction and communication weaknesses and fixed and repetitive behavioral patterns.

It varies greatly depending on age and skill. An autism spectrum has been introduced to recognize this diversity. The prevalence of autism seen as a rare event in the past has risen to 1,5% for a broad spectrum. Autism spectrum

disorder (ASD) is inherited and the frequency of occurrence in males is 4 times higher than that of females. In 15-30% of children with ASD, there is a loss of ability such as development retardation and speech. ASD usually brings together sleep and eating problems as well as additional diagnoses such as attention deficit hyperactivity disorder and anxiety. IQ level of about 50% of people with ASD is in the range of mental disabilities. The information needed for the diagnosis includes a detailed story of development, a description of the child's daily behavior, an assessment of the child's social interaction and communication and intellectual functioning. The course and outcomes of ASD are mostly dependent on language and intelligence and vary considerably from person to person [1, 2].

There are numerous clinical and nonclinical methods for the diagnosis of ASD. Autism Diagnostic Interview-Revised (ADI-R) and Autism Diagnostic Observation Schedule -Revised (ADOS-R) can be given as examples of clinical diagnostic methods. These methods have close diagnostic accuracy with similar sensitivity and specificity. In addition to the clinical methods, there are self- or parent-based nonclinical methods such as Autism Spectrum Quotient (AQ) and Social Communication Questionnaire (SCQ).

The methods used for the diagnosis of ASD have several disadvantages. For example, the vast majority of these methods use the sum of the scores obtained from the tests to set the appropriate diagnosis. For this reason, these methods need to be performed with extreme caution by a pediatrician and child psychologist and a series of diagnostic assessments must be implemented by certified professionals. These meticulous diagnostic examinations usually take a few hours. The increasing demand for the appointments for ASD diagnosis exceeds the maximum capacity of the developmental pediatric clinics. This extends the waiting period from initial concerns to the diagnosis [3-5].

Machine learning has tremendous potential to enrich diagnosis and intervention studies in behavioral sciences. It can be particularly useful in research involving the rather widespread and heterogeneous syndrome of ASD conditions. The use of machine learning methods in the diagnosis of ASD is utilized to shorten the duration of diagnosis in order to provide faster access to health services, to increase diagnostic performance and to reduce the size of the input data set by determining the most successful features in ASD diagnosis [4, 6].

Kozmicki et al. [7] used eight machine learning algorithms to differentiate the children with ASD from the normal children. They performed a feature selection before classification to determine the best subset of behaviors for the diagnosis on the module 2 and module 3 of the ADOS test that are related to the vocabulary and higher levels of cognitive functioning. They achieved 98.27% and 97.66% accuracy to detect ASD risk with the 9 of the 28 behaviors of module 2 and 12 of the 28 behaviors of module 3 respectively. that 21 of the 56 features are sufficient to detect ASD risk. Abbas et al. [8]

used RF to classify the ASD cases based on the short, structured parent-report questionnaires and the short, semi-structured home videos of 162 children. They trained two independent classifiers and combined their outputs into a single screening assessment. They obtained a significant accuracy improvement over standard screening tools. Maenner et al. [9] used the words and phrases data of 1,162 children from 2008 Georgia The Autism and Developmental Disabilities Monitoring (ADDM) site to train RF for classification of the ASD. They evaluated the performance of the RF on the data of 1,450 children from the 2010 Georgia ADDM surveillance data. Their machine learning approach predicted ASD case with an 86.5% accuracy, 84.0% sensitivity and 89.4% positive predictive value.

In current clinical tests that are used in ASD diagnosis, the diagnostic period is very long and a specialist pediatrician is needed for the diagnosis. Previous studies using machine learning methods are used to analyze responses to these clinical tests. In this study, it was aimed to provide a quick preliminary diagnosis through a questionnaire which includes questions proved to be effective in the diagnosis of ASD and can be answered by the caregiver. Thus, the concerns of the families that need to wait a long time to be assessed by the developmental pediatric clinics can be addressed by this initial assessment. Machine learning methods provide an in-depth analysis of the answers given to the questionnaire.

In this study, analyzes were performed for fast and accurate diagnosis of the ASD cases by using SVM, kNN and RF machine learning methods. For this purpose, a dataset with a total of 20 properties including behavioral and characteristic data of 104 individuals was used. The performance of the machine learning methods was compared using the result obtained from 10-fold CV. As a result of the performed binary classification operation, 100% classification accuracy is achieved with RF method.

## 2. Material and Method

### 2.1. Dataset

In this study, ASD adolescent scan data from UCI Machine Learning Repository is used. This dataset contains data that are effective in determining autistic characteristics and are intended to be used for further analysis to develop the classification performance of the ASD cases. The dataset contains 20 features of adolescents' screening data for autism. In this dataset, there are 10 behavioral features that are proven to be effective in differentiating ASD cases from the controls and 10 individual features [3, 4, 10].

The features used in this study along with their types and properties are given in Table 1. Three features namely who is the person performing the test (parent, caregiver, etc.), the country of residence and whether the test has been done previously are not used in this study since they do not contribute to the classification of the ASD cases.

Table 1. Features and their descriptions.

Feature	Type	Description
Age	Number	Age in years
Gender	String	Male or Female
Ethnicity	String	List of common ethnicities in text format
Born with jaundice	Boolean (yes or no)	Whether the case was born with jaundice
Family member with Pervasive Development Disorder (PDD)	Boolean (yes or no)	Whether any immediate family member has a PDD
Screening Method Type	Integer (0,1,2,3)	The type of screening methods chosen based on age category (0=toddler, 1=child, 2=adolescent, 3=adult)
Questions (1-10) Answers	Binary (0, 1)	10 questions related to the behavioral features
Screening Score	Integer	The final score obtained based on the scoring algorithm of the screening method used. This was computed in an automated manner

## 2.2. Machine Learning Methods

In this study, SVM, kNN and RF machine learning methods have been used to detect ASD cases.

SVM is a machine learning method used for binary classification purposes. It is used widely in clinical decision support systems to diagnose diseases automatically [11, 12]. SVM uses a supervised learning algorithm that learns the difference between classes through labeled samples. The binary labeled data is mapped to a very high dimensional feature space with a nonlinear approach to perform the classification operation. The input data is separated into two classes by a separating hyperplane formed in the feature space. The decision surface is organized using the supportive and instructive examples of the training data. Training is completed by identifying two sub-spaces corresponding to the two classes to be classified [13]. Sigmoid, polynomial, and radial basis function (RBF) kernel

functions are the most widely used functions for nonlinear feature mapping of the SVM.

In this study, Gaussian RBF kernel function and sequential minimal optimization (SMO) learning algorithm are used together to train SVM. The width of the RBF function is determined by the  $\sigma$  parameter. The C editing parameter is used for the soft-margin SVM where the data cannot be separated linearly. Linearly separable conditions can be handled with high C values while a large margin is obtained with small C values. The choice of C and  $\sigma$  parameters determine the performance of the RBF function [11-13].

k-nearest neighbor (kNN) uses a nonparametric supervised classification algorithm. It compares the data sample to be classified with the existing training samples and finds the closest examples. In the next step of the algorithm, the majority class tag between the labels of the k-nearest training samples is determined and assigned to the data sample to be classified.

The key parameter of the kNN's algorithm is the k that is the number of nearest training instances in the feature domain that must be found for classification. The prediction performance of the kNN algorithm increases as the size of the training data increases. The main disadvantage of the kNN algorithm is that it has high computational costs. The kNN method is preferred in clinical decision support systems because it has a very simple algorithm and has a stable performance [12, 14].

RF is a powerful ensemble method that is the combination of many decision trees. A decision tree consists of nodes and edges. In binary decision trees, there are inner and leaf nodes where all nodes have two coming and one going edges. A test function is applied to the input of each inner node, and the outgoing edge represents the result of the test. The input data continues to progress through the inner nodes until it arrives a leaf node representing a class tag.

Optimization of the test function of the inner nodes and deciding the predictions related to the leaf nodes should be performed for training. Multiple trees are created on different sub-samples of the dataset to create a RF. Trees in the forest are trained independently. Results obtained from all the trees in the forest are put together after classification. Result of the RF is determined as the class that is the most voted one. The power of each tree and the relationship between trees determine the generalization success of the forest [15-17].

## 3. Results and Discussion

Classification of ASD cases is performed using the scan data of the 104 adolescents and SVM, kNN and RF machine learning methods. Categorical features in the dataset as detailed in Table 1 were converted into numerical values before the classification process. The {m, f} information given for gender was taken as 0 for men and 1 for women. A numeric value is assigned for each different value of the features given in text format. The answers to the questions like, whether born with jaundice and whether or not there was an autism

between and near relatives, that are answered as yes/no were expressed by Boolean 0 and 1 values. The value of each feature is a normalized between [0 1] range before training and testing.

All classification tasks have been implemented using MATLAB. Parameter selection for all methods was performed using 10-fold CV to remove parameter-dependent bias. The best parameters determined were used in the training process. The resulting model was used to test the performance of the methods. 10-fold CV was also used to evaluate the classification performance. Accuracy, sensitivity, and specificity were used as evaluation criteria. In Table 2, the success of the SVM, kNN and RF machine learning methods in classifying ASD cases is given for each CV.

RBF was selected as the feature mapping function of the SVM. SMO is used as the learning algorithm. Since the classification success of the SVM method that uses the RBF kernel function depends on the C and  $\sigma$  parameters, a grid search is performed with 10-fold CV so that the best values can be obtained for these parameters. A grid was created for the values of  $C = [2^{-9}, 2^{-8}, \dots, 2^{15}]$  and  $\sigma = [2^{-5}, 2^{-4}, \dots, 2^{15}]$ . The values that give the best results are determined on this grid [18].

The most important parameter that affects the classification success of the kNN algorithm, the k value, has been determined experimentally that gives the best result. For this purpose, different k values were used and the k value that gives the highest classification performance is used. Table 3 shows the classification accuracy, sensitivity and specificity results obtained with different k values. Euclidean distance is used as distance function of the kNN.

The parameters that should be determined when creating a RF are the number of trees in the forest, the proportion of input data sampled with replacement, and the number of randomly selected variables for each decision partition. There is no standard way to determine the number of trees in the RF. For this reason, the number of trees must be determined experimentally. In this study, the success of the forests with different numbers of trees was tested with 10-fold CV to find the best classification accuracy. The classification accuracy, sensitivity, and specificity obtained from the RFs generated with different numbers of trees are given in Table 4.

Table 3. Performance values with different k values in the kNN method

k	Accuracy	Sensitivity	Specificity
1	0,89	1,00	0,74
3	0,86	0,98	0,72
5	0,85	1,00	0,66
7	0,84	0,98	0,64
9	0,84	1,00	0,62
11	0,83	1,00	0,60
13	0,83	1,00	0,59
15	0,82	1,00	0,56

Table 2. The success of SVM, RF and kNN methods in classifying the ASD cases

	SVM			RF			kNN		
	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
CV-1	0,90	1,00	0,86	1,00	1,00	1,00	0,70	1,00	0,57
CV -2	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
CV -3	1,00	1,00	1,00	1,00	1,00	1,00	0,90	1,00	0,66
CV -4	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
CV -5	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
CV -6	1,00	1,00	1,00	1,00	1,00	1,00	0,90	1,00	0,80
CV -7	0,90	0,86	1,00	1,00	1,00	1,00	1,00	1,00	1,00
CV -8	0,90	0,86	1,00	1,00	1,00	1,00	0,80	1,00	0,33
CV -9	0,80	0,80	0,80	1,00	1,00	1,00	0,70	1,00	0,40
CV -10	1,00	1,00	1,00	1,00	1,00	1,00	0,86	1,00	0,67
<b>Mean</b>	<b>0,95</b>	<b>0,95</b>	<b>0,97</b>	<b>1,00</b>	<b>1,00</b>	<b>1,00</b>	<b>0,89</b>	<b>1,00</b>	<b>0,74</b>
<b>Standard Deviation</b>	<b>0,07</b>	<b>0,08</b>	<b>0,07</b>	<b>0,00</b>	<b>0,00</b>	<b>0,00</b>	<b>0,12</b>	<b>0,00</b>	<b>0,26</b>



Table 4. Performance of the RF with different number of trees

n	Accuracy	Sensitivity	Specificity
3	0,98	0,98	0,99
4	0,99	0,99	1,00
5	0,99	1,00	0,98
6	0,99	0,99	1,00
7	0,99	1,00	0,97
8	0,90	1,00	0,99
9	1,00	1,00	1,00
10	1,00	1,00	1,00

Table 5. Comparisons of the study with other studies that use the same dataset

Study	Method	Performance
Guttenberg and Ryota [19]	kNN	%79
Guttenberg and Ryota [19]	SVM (RBF)	%83
Guttenberg and Ryota [19]	RF	%85
Guttenberg and Ryota [19]	XGBoost	%88
Basu [20]	Naïve Bayes	%89
Basu [20]	Decision Tree	%95
Basu [20]	RF	%99
Basu [20]	SVM (Linear)	%100
This study	kNN	%89
This study	SVM (RBF)	%95
This study	RF	%100

The proportion of input data sampled with replacement is chosen 1. The number of randomly selected variables for each decision partition is determined as the square root of the total number of variables. These parameters are the values recommended by Breiman, known as the person who invented the RF method [16].

Since the dataset used in this study is released recently in December 2017 there is no published study that use this dataset. Nevertheless, the results obtained from this study were compared with those of Guttenberg and Ryota [19] and Basu [20] that are early published in arXiv and GitHub. The comparison results are given in Table 5. Both studies were performed using the Scikit-Learn package.

Guttenberg and Ryota [19] achieved a lower classification performance, even though they used the same methods in this study. The reason for this may be that using the default parameter values of the training package and not investigating the parameters sufficiently that will give the best results. It can be seen from the table that the grid search method performed for the selection of C and  $\sigma$  parameters of the SVM method

improved the performance greatly. Same applies to the kNN and RF methods, too. The classification accuracy of Basu [20] was also obtained by 10-fold CV as in this study. In the SVM method, they achieved a higher classification performance with the linear kernel function than the RBF kernel function.

#### 4. Conclusion

In this study, the performance of the machine learning methods is evaluated for the classification of ASD cases using 104 adolescent scan data. All the parameter selection processes are performed using 10-fold CV to reduce the bias that would arise from the parameters. 10-fold CV is also used for the classification tasks to eliminate the bias based on the training and test dataset selections. SVM, kNN and RF machine learning methods are used to analyze the answers given to a questionnaire which includes questions proved to be effective in differentiating the ASD cases. Accuracy, sensitivity and specificity performance metrics are used for the evaluation of the trained models. Accuracy rates of 95%, 89%, and 100% were achieved as a result of binary classification using SVM, kNN and RF methods, respectively. The lowest performance in the classification of ASD events is obtained from the kNN algorithm. The relatively low specificity rates obtained with kNN indicate that healthy people are frequently labeled with ASD by this method. The SVM method performed between RF and kNN. The results obtained with the RF method show that this method can classify ASD cases with complete success. The 100% sensitivity and specificity values obtained from RF are other indications of this success.

#### 5. References

- [1] Frith, U, Happé, F., "Autism spectrum disorder", *Current Biology*, Vol. 15, No. 19, R786-R790, 2005.
- [2] Charman, T., "Autism spectrum disorders", *Psychiatry*, Vol. 7, No. 8, 331-334, 2008.
- [3] Thabtah, F., "Machine learning in autistic spectrum disorder behavioral research: A review and ways forward", *Informatics for Health and Social Care*, 1-20, 2018.
- [4] Thabtah, F., "Autism spectrum disorder screening: machine learning adaptation and DSM-5 fulfillment", *Proceedings of the 1st International Conference on Medical and Health Informatics (ICMHI'17)*, Taichung City, Taiwan, 2017, 1-6.
- [5] Duda, M., Ma, R., Haber, N., Wall, D. P., "Use of machine learning for behavioral distinction of autism and ADHD", *Translational Psychiatry*, Vol. 6, No. 2, e732, 2016.
- [6] Bone, D., Goodwin, M. S., Black, M. P., Lee, C. C., Audhkhasi, K., Narayanan, S., "Applying machine learning to facilitate autism diagnostics: pitfalls and promises", *Journal of Autism and Developmental Disorders*, Vol. 45, No. 5, 1121-1136, 2015.
- [7] Kosmicki, J. A., Sochat, V., Duda, M., and Wall, D. P. "Searching for a minimal set of behaviors for autism detection through feature selection-based machine

- learning”, *Translational Psychiatry*, Vol. 5, No. 2, e514, 2015.
- [8] Abbas, H., Garberson, F., Glover, E., and Wall, D. P. “Machine learning approach for early detection of autism by combining questionnaire and home video screening”, *Journal of the American Medical Informatics Association*, ocy039, 2018.
- [9] Maenner, M. J., Yeargin-Allsopp, M., Braun, K. V. N., Christensen, D. L., and Schieve, L. A. “Development of a machine learning algorithm for the surveillance of autism spectrum disorder”, *PloS One*, Vol. 11, No. 12, e0168224, 2016.
- [10] Thabtah, F., ASDTests. A mobile app for ASD screening (Online). Available: [www.asdtests.com](http://www.asdtests.com) [Accessed: 09.05.2018].
- [11] Demirhan, A., “Nöro-görüntüleme tabanlı şizofreni teşhisi için desen analizi”, 25. IEEE Sinyal İşleme ve İletişim Uygulamaları (SİU 2017), Antalya, Turkey, 2017, 1-4.
- [12] Demirhan, A., “Neuroimage-based clinical prediction using machine learning tools”, *International Journal of Imaging Systems and Technology*, Vol. 27, No. 1, 89-97, 2017.
- [13] Vert, J. P., Tsuda, L., Schölkopf, B. (Editors, Schölkopf, B., Tsuda, L., Vert, J. P.), “A primer on kernel methods”, *Kernel methods in computational biology*, MIT Press, Cambridge, Massachusetts, 2004.
- [14] Hashemian, M., Pourghassem, H., “Diagnosing autism spectrum disorders based on EEG analysis: A survey”, *Neurophysiology*, Vol. 46, No. 2, 183-195, 2014.
- [15] Demirhan, A., “Random forests based recognition of the clinical labels using brain MRI scans”, 3rd International Conference on Frontiers of Signal Processing (ICFSP 2017), Paris, France, 2017, 156-159.
- [16] Breiman, L., “Random forests”, *Machine Learning*, Vol. 45, No. 1, 5-32, 2001.
- [17] Criminisi, A., Shotton, J., Konukoglu, E., “Decision forests for classification, regression, density estimation, manifold learning and semi-supervised learning”, *Technical Report, Microsoft Research Lab - Cambridge*, 2011.
- [18] Hsu, C.-W., Chang, C.-C., Lin, C.-J., “A practical guide to support vector classification”, *Technical Report, Department of Computer Science, National Taiwan University*, 2003.
- [19] Guttenberg, N., and Ryota, K., “Learning to generate classifiers”, *arXiv preprint, arXiv:1803.11373*, 2018.
- [20] Basu, K., Autism Screening Adult Data Set: A Machine Approach, (Online). Available: <https://github.com/kbasu2016/Autism-Detection-in-Adults/blob/master/report.pdf>. [Accessed: 09.05.2018].