FORECAST THE AUTISM SPECTRUM DISORDER USING VARIOUS MACHINE LEARNING TECHNIQUES

G Karthick[1] , N Venkateswaran[2] , R Jegadeesan[3], Dava Srinivas[4], N Umapathi[5]

Article History: Received: 11.04.2023 **Revised:** 27.05.2023 **Accepted:** 13.07.2023

Abstract

Autism Spectrum Disorder (ASD) is a neurodevelopment syndrome that has a great impact on the person's behavior, mental health and also the learning skills. It limits the ability of the affected personality to use verbal, social, and cognitive skills, among other abilities, and its symptoms are differed from person to person. Diagnosing ASD is very difficult process because it does not have any medical test like blood test to predict the symptoms of the ASD. To overcome difficulties of diagnosing ASD, various machine learning methods are used to increase the performance and accuracy rate, so that diagnosing the ASD can be predicted at the early stages. As a result, it can be utilized to make decisions in a situation where there is a lot of uncertainty. In this research, machine learning techniques evaluate their performance on an ASD dataset. As the results, Random Forest Classifier and Decision Tree Classifier got the highest accuracy with 100% and Support Vector Machine (SVM) and Naive Bayes algorithms got the accuracy of 97%.

Keywords: Autism Spectrum Disorder, K-Nearest Neighbors (K-NN), Support Vector Machine (SVM), Decision Tree Algorithm (DT), Random Forest (RF) Algorithm and Naive Bayes (NB) Algorithm.

[1]Assoc.Professor, Department of ECE, Jyothishmathi Institute of Technology and Science, Karimnagar, Telangana.
[2]Assoc.Professor ,Department of CSE, Jyothishmathi Institute of Technology and Science, Karimnagar, Telangana.
[3]Professor and Head, Department of CSE, Jyothishmathi Institute of Technology and Science, Karimnagar, Telangana.
[4]Assoc. Professor, Department of CSE Jyothishmathi Institute of Technology and Science, Karimnagar, Telangana.
[5]Professor and Head, Department of ECE Jyothishmathi Institute of Technology and Science, Karimnagar, Telangana.

Email:[1]karthick.sgs@gmail.com, [2]venkywn@gmail.com,[3]ramjaganjagan@gmail.com, [4]dr.dsrinivas@jits.ac.in, [5]nrumapathi@gmail.com

DOI: 10.31838/ecb/2023.12.s3.673

1. Introduction

ASD is a neurological disorder that concerns brain development. A human with ASD is usually unable to interact socially or converge with others. When this happens, a human life is typically impacted for the remainder of his or her life. It is essential to consider that both environmental and genetic factors may likely contribute to the increase of this disorder. This illness can start as early as three years old and last a lifetime. Even if it is impossible to entirely heal a patient with this ailment, the causes can be minimized for a period if the symptoms are detected initially. The true causes of ASD have yet to be discovered. ASD is a fast-increasing problem that affects people of all ages in today's world. The subject's mental then physical health can improved if this neurological disease is detected early. It looks promising that, through machine learning-based models, we can rely on a handful of metrics to accurately identify different conditions; this is becoming increasingly popular in forecasting various human illnesses. This sparked our interest in ASD diagnosis and research to discover therapeutic techniques. ASD detection is difficult because there are a variety of mental illnesses with symptoms that are very similar to those seen in people with ASD.

ASD place a comprehensive proportion of the pediatric population. In virtual situations, it maintains be perceived in its ahead of time phases, on the other hand the greatest staggering obstruction is the distinct and time-consuming individualism of contemporary diagnosing processes. As a result, the postponement between the original suspiciousness and the thoroughgoing designation is at least 13 months. The designation buoy appropriates many hours, and the requirement for rendezvous is distance off by the competence of the country's pediatric facilities. It is crucial to detect and treat ASD early so it can be reduced or alleviated to some degree and therefore improve the quality of life of the individual. Conversely, a lot of time is lost when this medical condition stays unnoticed because of the delay between the initial worry and the diagnosis. Not only would machine learning approaches reduce the time it takes to assess ASD risk, but they would also speed up the entire diagnostic process, allowing families to access the therapy they need. Who completed the test, sex, family member, ethnicity, Class are some of the screening approaches that are used to identify ASD in children. ASD may be foreseen early on utilizing a variety of machine learning techniques.

2. Materials And Methodology

These Machine Learning Algorithms are used in medical datasets. Figure 1 displays the flow diagram of ASD dataset classification of the machine learning techniques. Figure 1 flow diagram of ASD classification of machine learning. The ASD dataset is collected from the kaggle website [9], which contains 1055 instances and 19 attributes. Next, dataset is undergoes preprocessing step. The process of preprocessing is to clean the dataset by remove the null values and redundant values which are present in the dataset and them made it to desired forma. Preprocessing the dataset is advisable to enhance the model's accuracy. Then dataset is divided into two parts i.e., Training and Testing. Training and testing measures and compares the accuracy of different classifiers. In this research paper, the dataset was divided into a training set comprising 70% of the data and a testing set comprising 30% of the data.

Next step in the block diagram is to train and test the data different machine learning algorithms are used. Python programming language has been used to implement various machine learning algorithms. Different Classification methods that are used in this paper are K-Nearest-Neighbor (KNN), Random Forest Classifier (RTC), Support Vector Machine (SVM), Naive Bayes (NB) and Decision Tree Classifier (DTC). We evaluate the performance of each algorithm by examining their confusion matrices. For each algorithm the accuracy will be measured and compared. Among this parameter precision, recall values and the F1_score will be calculated in order to help in evaluating each classifier. When there is comparison between the accuracy of training and testing, the accuracy of training should be always more for every classifier when compared with the testing accuracy, if the classifier works properly. The algorithm which got the highest accuracy is considered as the best Classifier for prediction of diagnosing ASD.

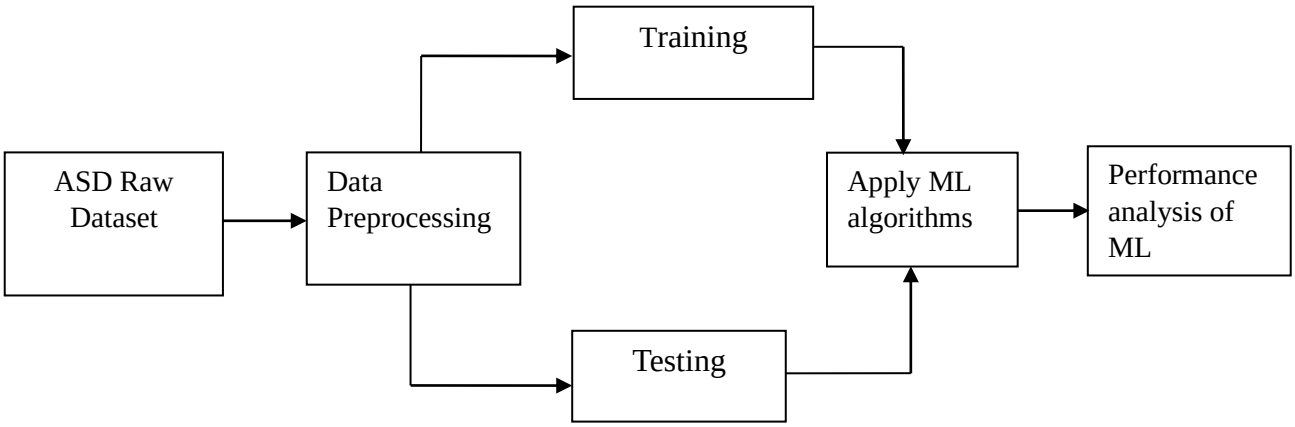

Figure 1: Flow Diagram of ASD Classification of Machine Learning

Related Work

A. Kosmicki et al. aimed at a technique for autism detection that utilizes a minimal number of features [1]. The study employed machine learning algorithms to determine the medical assessment of ASD using the Autism Diagnostic Observation Schedule (ADOS) on a subset of behaviors exhibited by adolescents. The ADOS consists of four modules. The authors employed eight various machine learning methods, including stepwise backward characteristic identification, on rating sheets from 4540 individuals. They selected 9 out of the 28 behaviors from module 2 and 12 out of the 28 behaviors from module 3 to identify ASD risk, achieving average accuracies of 98.27% and 97.6% correspondingly.

Muhammad Faiz Misman et al. give the classification of adults with ASD with the help of the Deep Neural Network (DNN) [2]. The researchers chose the DNN architecture, as it has gained popularity in recent years and has been found to improve prediction performance. Considering the different variables of adult ASD screening findings, this finds the achievement of the DNN model in the wide range of symptoms in terms of solution quality. The findings are comparable to Support Vector Machine, a prior machine learning method produced either by means of a developer. On the first dataset, the DNN mannequin successfully identified ASD diagnoses with 99.40% accuracy, while on the validation set; it did so with 96.08%. Conversely, thinking about both the first and subsequent data, the SVM model has an average state of 95.24 % and 95.08 %, respectively. The findings advice those ASD patients may additionally be labeled as such, making use of ASD grownup screening information and the DNN classification methodology. Meanwhile, lacking values precipitated 95 rows of statistics in the second dataset to be eliminated, resulting in a considerable reduction in the wide variety of cases in the dataset.

R. Mohamed Shanavas et al. discusses the relationship between autism and the various types of disorder, including Asperger's syndrome. The author compares the effectiveness of popular machine learning methods with Neural Network, Support Vector Machine (SVM), and Fuzzy methods with the Waikato Environment for Knowledge Analysis (WEKA) tool to analyze student behavior and social interactions [3]. In another study, A. Sharma et al. utilized machine learning to distinguish autism from neurotypical controls based on movements during imitation [4]. The main goal of research was to evaluate the underlying matters concerning discriminative exam circumstances and kinematic characteristics. The dataset consisted of 16 individuals with Autism Spectrum Condition (ASC) who performed a series of hand movements. Here, used 40 kinematic measures from 8 different imitation conditions to examine how well each technique performed. The study demonstrated the possibility of applying machine learning techniques to investigate high-dimensional data and diagnose autism, even with a small sample size. Sensitivity analysis was conducted using the RIPPER algorithm, which achieved the following feature selection accuracies on the AQ-Adolescent dataset: Va (87. 0%), CHI (80.5%), IG (80.5%), correlation (84.3%), CFS (84. 3%), and "no function selection" (80%).

Astha Baranwal et al. consider the ASD screening dataset for analysis [5]. The ASD screening dataset is handled in this analysis and forecast possible events in adults, babies, and adolescents. Each age group's information is analyzed, and conclusions are generated as a result. For estimation and analysis, various computer learning techniques have been implemented. The author used the autism screening datasets for adults, adolescents, and infants. Each dataset has 20 homes with ongoing segments and actual or false values. The based attributes Class ASD (0) determine whether or not a person has an ASD. The accuracy of decision trees, artificial, random forests, logistic regression, SVM and neural networks is 80%, 88%, 92%, 80%, and 76%, respectively (ANN). Apparently, the record reached is inadequate. The facts wanted to build a profitable statistical method are extraordinarily limited. Across all collections, the Decision Tree produces an over complete architecture.

Vaishali R et al. discussed a system to detect autism with best conduct sets [6]. This paper considered an ASD analysis dataset with 21 items from the UCI laptop study repository was investigated with using a swarm brain based binary firefly feature determination wrapper. The unconventional hypothesis of the investigation states that it is viable for a machine learning model to gain a higher classification accuracy with minimal characteristic subsets. By utilizing Swarm Genius and its single-objective binary firefly function choice wrapper, it is discerned that only 10 of the 21 components of the ASD dataset are required to differentiate between ASD and non-ASD patients. This approach has produced results an average accuracy from 92% to 98% with optimum feature subsets which match the average accuracy of the entire ASD diagnosis dataset.

Suman Raj et al. describe an investigation and finding of ASD using machine learning methods for predicting and examining ASD problems in children, teens, and adults using a network and a convolution neural network [7]. The proposed techniques are tested with the use of three publicly available non-clinically ASD datasets. The first dataset consists of 292 incidences and 21 features that are applicable to ASD screening in youngsters. The second dataset for ASD screening adults has 704 instances and 21 attributes. The 0.33 dataset consists of 104 cases and 21 attributes associated with teenage ASD screening. In this study, both the SVM and CNN established models show a prediction with a 98.30% accuracy score without the ASD infant dataset whilst dealing with lacking values.

Noora Saleem Jumaa et al. present the ASD Diagnosis using machine learning models, where the primary intention is to Another mission of the project is to examine any tactic that does diagnose autism in children aged four to 17 [8]. The client, who is accountable for collecting patient data and transferring it to the server, is the initial element of the ASD disorder prediction approach in this research. The device has a 0.044% average mistake rate and can become aware of ASD with a 98% accuracy rate. A mechanism has been postulated for predicting the onset of autism. As a device for testing the suggested scheme, it was used to test the use of two types of machine-learning units: collections of information collected from the (UCI) warehouse and The output of many classifying algorithms that may additionally be employed for this type of classification task can be used here. Three out of four will be utilized as the classifier in the advised technique. M. S. Murtazina and T. V. Avdeenko dedicated to using machine learning based on three different types of EEG data considered to categorize brain activity patterns. From the results SVM offers 63.33% accuracy [10].

Machine Learning Algorithms

In this paper various machine learning algorithms are used such as Support Vector Machine (SVM), Random Forest (RF), Decision Tree (DT), Naïve Bayes (NB) and K-Nearest-Neighbor (KNN).

a. Support vector machine (SVM)

SVM is the utmost common and efficient supervised classifiers, used for both classification and regression problems. All the vectors are plotted in hyperplane to define decision boundaries, separate relevant and irrelevant vectors, and categorize ASD data points. SVMs come in two varieties, each with unique applications. For classification and linear regression tasks, simple SVM is frequently utilized. In kernel SVM, you gain more elasticity for non-linear data by adding additional features to fit a hyperplane rather than just a two-dimensional space. Simple SVM is used here.

Procedure for SVM Algorithm:

Step-1: The first step is to split the dataset using the classifier to classify the classes ASD or not ASD.
Step-2: A single line is used to differentiate the two classes, although multiple lines can be employed.
Step -3: SVM algorithm determines the optimal hyperplane that partitions the two classes as closely as possible, utilizing the support vectors in defining this boundary.
Step 4 Predicting the result of the test set. Creation of the confusion matrix
Step 5 Visualization of the result of the training set.

b. Decision Tree Classifier

Decision Trees (DTs) are popular non-parametric supervised machine learning algorithms used for data categorization. The goal of DTs is to learn rules from training data to obtain a model that expects the class label of a test sample. DTs have leaf and internal nodes, with leaf nodes determining the class label based on the majority vote of training examples and internal nodes branching out based on feature questions.

Procedure for Decision Tree Algorithm:
Step-1 Determine the decision column
Step-2 Calculation of entropy for classes
Step-3 Calculation of entropy for other attributes after splitting
Step -4 Calculate the information gain for each split
Step -5 Perform the first split.

K-Nearest Neighbor (K-NN) algorithm

K-NN is a supervised learning technique which utilizes similarity between new data points and already existing information to assign the new data into an appropriate category. It stores all available data and classifies new data based on similarity, using the Euclidean distance metric to determine similarity. The kTree and k*Tree techniques were two novel K-NN classification algorithms that Shichao Zhang et.al introduced in order to choose the best k value for each test pattern for successful and efficient K-NN classification [11]. The elements that influence the algorithm can be altered to increase its effectiveness. To improve the performance of this technique, various KNN variations have been investigated. He emphasizes the use of the KNN approach and its adapted variations from earlier studies [12].

Procedure for K-NN algorithm:
Step-1: Choose the number K of neighbors
Step-2: Calculate the Euclidean distance of K neighbors
Step-3: Determine the K nearest neighbors according to the calculated Euclidean distance.
Step-4: Among these k neighbors, count the number of data points in each category.
Step-5: Assign the new data points to the category for which the number of neighbors is the highest.

c. Random Forest algorithm

The supervised learning algorithm in question is widely popular for addressing both regression and classification issues. It exploits ensemble learning, combining multiple classifiers to improve model performance. In Random Forest, a subset of the dataset's random records is selected, and unique decision trees are built for each sample. Please note that the procedures provided for each algorithm are simplified explanations of their respective processes.

Procedure for Random Forest algorithm:
Step-1 In Random Forest, n random records are selected from the dataset with k records.
Step-2 Individual decision trees are created for each sample.
Step-3 Each decision tree generates an output.
Step-4 The final output is considered based on majority decisions or averaging for classification regression.

d. Naive Bayes algorithm

The Nave Bayes algorithm is a supervised learning method used to tackle classification challenges. Its foundations are in the Bayes theorem. It displays many real-world applications such as spam filtering, sentiment analysis and article classification [13]. It is generally used in text classification tasks requiring a huge training dataset. It's a probabilistic classifier, which means it makes predictions based on an object's probability. Y.Huang et.al studies On the basis of a tiny sample of data, the Naive Bayes and SVM classification technique was used and SVM is a decent classification method and has the obvious advantages of efficiency and speed, but it takes up too much time and space [14]. one drawback is the very strong assumption of independence of class features that it makes. It is almost impossible to find such records in real life.

Procedure for Naïve Bayes Algorithm:
Step-1 Convert the data set into a frequency table by calculating the probability for each attribute.
Step-2 Create a probability table by finding probabilities.
Step-3 By applying the Naive Bayes equation, one can determine the posterior probability of each class. The predicted result will be the class with the greatest posterior probability.

$$P\left(\frac{A}{B}\right) = \frac{P(A)P\left(\frac{B}{A}\right)}{P\left(\frac{A}{B}\right)P(B)} \quad (1)$$

Step-4 The class with the best chance of being correct is deemed the result of the forecast.

3. Result And Analysis

The performance of the various machine learning techniques can be evaluated using metrics such as accuracy, precision, recall, and F1-Score.True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) have been calculated with the help of the confusion matrix.

TP (True Positive): A true positive is a prediction where system correctly classifies the event as positive.
FP (False Positive): A true positive is a prediction where system correctly classifies the event as negative.
FN (False Negative): A false negative is a prediction where system incorrectly classifies the event as negative.
TN (True Negative): A true negative is a prediction where system incorrectly classifies the event as positive.

Accuracy: It is the proportion of the total number of predictions that were correct [15].

$$\text{Accuracy} = \frac{(TP+TN)}{(TP+FP+FN+TN)} \quad (2)$$

Precision: It (also called as positive predictive value) is the probability that a positive prediction is correct [15].

$$\text{Precision} = \frac{TP}{TP+FP} \quad (3)$$

Recall: It is the ratio of correct positive results to the total number of positive results predicted by the system.

$$\text{Recall} = \frac{TP}{TP+FN} \quad (4)$$

F1_Score: It is harmonic mean of precision and recall [11] , which is given by:

$$\text{F1\_Score} = 2 * \frac{\text{Precision}*\text{recall}}{\text{Precision}+\text{recall}} \quad (5)$$

Table 1: Performance Analysis of Various Machine Learning Methods

ML Algorithm	Precision	Recall	F1-Score	Accuracy
Random Forest	100	100	100	100
Support Vector Machine	99	94	96	97.15
K-Nearest Neighbor	98	82	90	92.89
Decision Tree	100	100	100	100
Naïve Bayes	99	94	96	97.15

4. Conclusion

The basic idea behind our suggested solutions is to provide a training stage in order to lower testing stage operational costs and enhance classification performance. Comparative research employing machine learning techniques is carried out in the paper to analyze the best classifier among all classifiers using standard dataset. This implementation of multiple classifiers is attainable without putting in a lot of complex programming effort thanks to the Python programming language's powerful built-in programming frameworks. From the observations it is observed that out of 5 classifiers Random Forest and Decision Tree has the maximum accuracy rate of 100%. Support Vector Machines and Naive Bayes classifiers had a 97% accuracy rate, which was the second-

best recorded accuracy. And last but not the least KNN classifier has recorded the least accuracy rate of 93%. Finally, we conclude that even most of classifiers resulted in accuracy greater than 90%. Random Forest and Decision Tree Classifiers are considered as the best algorithms among all the classifiers. And in future we expect to have even better accuracy with least false rates and the process of evaluating can also be made fully automatic. It is clear that investigating the potential of deep learning-based models to detect ASD in humans is necessary. As mentioned previously, most of the existing works make use of conventional machine learning strategies and thus, their effectiveness is hindered.

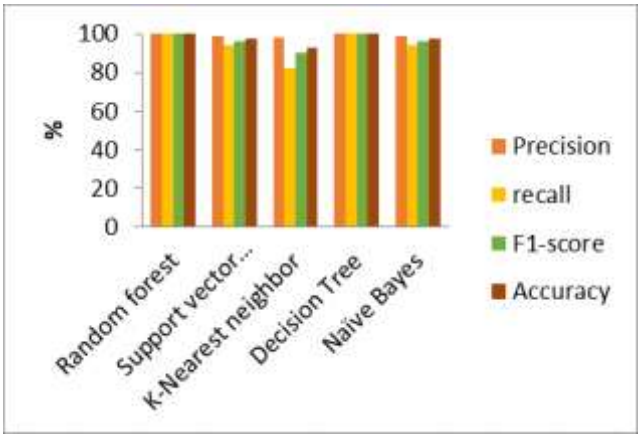

Figure 2 Graphical Representations of Various Machine Learning methods.

5. References

[1] J.A. Kosmicki, V. Sochat, M. Duda, and D. P.Wall. (2015) "Searching for a minimal set of behaviors for autism detection through feature selection-based machine learning." Translational psychiatry, 5(2): e514.

[2] M. F. Misman et al., "Classification of Adults with Autism Spectrum Disorder using Deep Neural Network," 2019 1st International Conference on Artificial Intelligence and Data Sciences (AiDAS), 2019,pp. 29-34, doi: 10.1109/AiDAS47888.2019.8970823.

[3] M.S.Mythili, and AR Mohamed Shanavas (2014) "A study on Autism spectrum disorders using classification techniques." International Journal of Soft Computing and Engineering (IJSCE), 4: 88-91.

[4] Baihua Li, Arjun Sharma, James Meng, SenthilPurushwalkam, and Emma Gowen. (2017) "Applying machine learning to identify autistic adults using imitation: An exploratory study." PloS one, 12(8): e0182652.

[5] A. Baranwal and M. Vanitha, "Autistic Spectrum Disorder Screening: Prediction with Machine Learning Models," 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), 2020, pp. 1-7, doi: 10.1109/ic-ETITE47903.2020.186.

[6] Vaishali, R., and R. Sasikala. "A machine learning based approach to classify Autism with optimum behaviour sets. (2018) " International Journal of Engineering & Technology 7(4): 18

[7] Raj, Suman; Masood, Sarfaraz (2018). Analysis and Detection of Autism Spectrum Disorder Using Machine Learning Techniques.Procedia Computer Science, 167(), 994–1004. doi:10.1016/j.procs.2020.03.399

[8] NooraSaleemJumaa, AymenDawoodSalman, RafahAlHamdani, "The Autism Spectrum Disorder Diagnosis Based on Machine Learning Techniques" Journal of Xi'an University lof Architecture & Technology, Volume XII, Issue V, 2020.

[9] Dataset of Autism spectrum disorder UCI:https://www.kaggle.com/datasets/fabdelja/autism-screening-for-toddlers.

[10] M. S. Murtazina and T. V. Avdeenko, "Classification of Brain Activity Patterns Using Machine Learning Based on EEG Data," 2020 1st International Conference Problems of Informatics, Electronics, and Radio Engineering (PIERE),

Novosibirsk, Russia, 2020, pp. 219-224, doi: 10.1109/PIERE51041.2020.9314660

[11] S. Zhang, X. Li, M. Zong, X. Zhu and R. Wang, "Efficient kNN Classification With Different Numbers of Nearest Neighbors," in IEEE Transactions on Neural Networks and Learning Systems, vol. 29, no. 5, pp. 1774-1785, May 2018, doi: 10.1109/TNNLS.2017.2673241.

[12] K. Taunk, S. De, S. Verma and A. Swetapadma, "A Brief Review of Nearest Neighbor Algorithm for Learning and Classification," 2019 International Conference on Intelligent Computing and Control Systems (ICCS), Madurai, India, 2019, pp. 1255-1260, doi: 10.1109/ICCS45141.2019.9065747.

[13] Pouria Kaviani and Sunita Dhotre, "Short Survey on Naive Bayes Algorithm", International Journal of Advance Engineering and Research Development, vol.4(11),2017.

[14] Y. Huang and L. Li, "Naive Bayes classification algorithm based on small sample set", Proc. IEEE Int. Conf. Cloud Comput. Intell. Syst., pp. 34-39, Sep. 2011.

[15] H. Rajaguru, K. Ganesan, and V. Kumar Bojan, Earlier detection of cancer regions from MR image features and SVM classifiers, Int J Imaging Syst Technol 26 (2016), 196–208.