Evaluation on Machine Learning Algorithms for Classification of Autism Spectrum Disorder (ASD)

Azian Azamimi Abdullah*, Saroja Rijal¹ and Satya Ranjan Dash²
¹School of Mechatronic Engineering, Universiti Malaysia Perlis, Pauh Putra Campus, 02600 Arau, Perlis, Malaysia
²School of Computer Applications, KIIT University, India

*azamimi@unimap.edu.my

Abstract. Autism Spectrum Disorder (ASD) was characterized by delay in social interactions development, repetitive behaviors and narrow interest, which usually diagnosed with standard diagnostic tools such as Autism Diagnostic Observation Schedule (ADOS) and Autism Diagnostic Interview-Revised (ADI-R). Previous work has implemented machine-learning methods for the classification of ASD, however they used different types of dataset such as brain images for MRI and EEG, risk genes in genetic profiles and behavior evaluation based on ADOS and ADI-R. Here a trial on using Autism Spectrum Questions (AQ) to build models that have higher potential to classify ASD was developed. In this research, Chi-square and Least Absolute Shrinkage and Selection Operator (LASSO) have been selected as feature selection methods to select the most important features for 3 supervised machine learning algorithms, which are Random Forest, Logistic Regression and K-Nearest Neighbors with K-fold cross validation. The performance was evaluated in which results Logistic Regression scored the highest accuracy with 97.541% using model with 13 selected features based on Chi-square selection method.

1. Introduction
Autism Spectrum Disorder (ASD) is a neurodevelopmental disability that refers to slow development in verbal skills, limited passion and monotonous behaviour [1]. With the increase in commonness of this problem recently, new pre-screening technology was encouraged [2]. The only reliable and non-invasive diagnosis can be done through screening tool nowadays was Autism Diagnostic Interview-Revise (ADI-R) and Autism Diagnostic Observation Schedule (ADOS). Nowadays there are many problems related to ASD detection where it was hard to find scalable biomarkers for early detection due to the wide variability of the disability and need of simpler measurements that could be applied routinely [3]. Behaviour test based on a clinical tool which was Autism Diagnostic Observation System (ADOS) was employed in this research where they used a machine learning approach to evaluate the data obtained [4]. MRI and EEG approach on ASD classification was in purpose to allow for early detection of ASD and cut down the administration time as follow in ADOS and ADI-R screening tools. The latest research done by Murat Gok by specifying on ASD risk genes of a long...
non-coding RNA (IncRNA) in order to observe whether the gene characteristic can accurately classify ASD or not [5].

2. Materials and Methods

2.1 Exploratory Data Analysis (EDA)

Data consist of 704 instances and 20 features with 1 output. Features consist of ten shortened version Autism Spectrum Quotient (AQ) self-administered questionnaire which assessed on behaviour traits which were represented in binary data (0 and 1) [6]. Python software applied to the raw dataset to visualize the dataset. The dataset observed to have 13.5% missing data. The contribution of individual diagnosed with ASD was 26.85% (189 patients) and not diagnosed with ASD was 73.15% (515 patients). The datasets undergo conversion from characters format into binary and numeric form using “Label encoder” function from Python library.

2.2 Data Pre-Processing

Data pre-processing was a crucial step in data mining techniques and it is the initial step after data exploration [7]. As for this project datasets, it contain missing values which it needs proper technique to handle it. Missing values create difficulties for researchers due it can cause a decrement in efficiency during selection or extraction process if not properly handled [8]. The amount of 13.5% missing data was removed from the datasets by dropping the empty column using Python commands. The removal of null datasets giving a total of 609 datasets left with 20 features consisting 19 features and 1 output.

2.3 Feature Selection

Feature selection was a technique used for data pre-processing and it was preferable when to perform it before fitting model in machine learning. Selection by mean here was by choosing variable and attributes or variable subset in a dataset to be fit into model and tested for performance. Chi-Square, LASSO and Random Forest method were applied in this study.

2.3.1 Chi-Square

The parameter used was $\alpha$ value, which was 0.01 and 0.05, which indicates 99% and 95% confidence interval (CI). The $\alpha$ values as a bench level for choosing the significant features. Any feature falls into the insignificant region will be expelled.

2.3.2 LASSO

The same parameter used in LASSO, which was $\alpha$ value but the usage was different from the Chi-Square method. The $\alpha$ value used in LASSO indicates the regularization value, which was the strength of penalization to choose features. The $\alpha$ parameter was selected by using Python “GridSearchCV” function to obtain the best value.

2.3.3 Random Forest

The used of “SelectFromModel” and “get.support” function was implemented to obtain the importance of each features.

2.4 Supervised Machine Learning Method

In supervised learning, it is defined as the trial to assess the relationship between input features and target features where it was provided with input and output datasets. Few machine learning algorithms used in this study, which was Random Forest, Logistic Regression and K-nearest neighbors.

2.4.1 Logistic Regression

Logistic regression has the nature of estimating probability whether it was success or failure even. This algorithm relates the categorical data dependent variables X and branched categorical output of Y [9].
Parameter C equals to 10, which regularization strength was applied in this study, which was selected using “GridSearchCV” function.

2.4.2 Random Forest
Consists of tree predictors, which each tree depends on the random values, which sampled independently. This algorithm was done by aggregating trees and the majority will be chosen [10]. Few parameters were used along this project which were the number of estimators, maximum number of features, maximum depth, minimum sample leaf, minimum sample split and boostrap which was selected by using hyper-parameter tuning “GridSearchCV”.

2.4.3 K-Nearest Neighbors
The computation of the closest distance between neighbors which represented as K value. In order to use this algorithm, we have to take count of a few elements, which are the initial subjects set, the K number (nearest neighbors) and the standard approximation between subjects. In this project the parameter K and distance of neighbors was considered.

2.5 Evaluation of Performance Metric
In this research, all performance result was obtained through 5-fold cross-validation instead of using a common split method. This method is applied to make sure the tests are less biased and will better predict the error estimation where it will split the dataset into 5 pieces and run the algorithm 5 times. Different piece of data as a test set and the other 80% of the data is the training set [11]. The classifier performance will be classified into few metric measurements which are accuracy, sensitivity, specificity and area under receiving operating characteristics (AUC) which obtained from confusion matrix. In accuracy, it can be observe and identify the accurate prediction made based on the overall number of tests [12]. The sensitivity metric define as the ratio of test diagnosed with ASD which is in true positive (TP) rate indicates how many correct predictions have been made based on the test cases. While specificity is defined as the ratio of test that is not diagnosed with ASD which indicate as true negative (TN) rate [12]. This indicates how many false predictions are made based on the test cases.

3. Results and Discussion
3.1 Feature Selection
After the pre-processing stage, we have 20 features and 1 output datasets. In the LASSO method, a trial on a few penalization values was tested in order to observe the number of features selected for each penalization coefficient. Trials have been performed for different set of alpha, \( \alpha \) which was has been chosen earlier. Table 1, Table 2, Table 3, Table 4 and Table 5 show the selected features for each selection methods.

| Table 1. LASSO | Feature with non-zero coefficient for \( \alpha=0.001 \) (17 features) |
|----------------|-----------------------------------------------|
| A9_Score       | A6_Score                                     |
| A3_Score       | A5_Score                                     |
| A7_Score       | A8_Score                                     |
| A2_Score       | A1_Score                                     |
| A4_Score       | A10_Score                                    |
| Age            | Ethnicity                                    |
| Relation       | Jundice                                      |
| Used_app_before| Contry_of_res                                 |
| Gender         | -                                            |

| Table 2. LASSO | Feature with non-zero coefficient for \( \alpha=0.01 \) (16 features) |
|----------------|-----------------------------------------------|
| A9_Score       | A6_Score                                     |
| A3_Score       | A5_Score                                     |
| A7_Score       | A8_Score                                     |
| A2_Score       | A1_Score                                     |
| A4_Score       | A10_Score                                    |
| Age            | Ethnicity                                    |
| Relation       | Jundice                                      |
| Used_app_before| Contry_of_res                                 |


3.2 Supervised Machine Learning Method

3.2.1 Logistic Regression

Based on Table 6, it can be observed that the highest performance goes to model LR4-CV with accuracy 97.541%, 96.591% specificity and 100% sensitivity. This model was able to discriminate the ASD from those who were not by 97.541%. While in terms of specificity, it was able to classify 96.591% individual, which was not possessed to ASD case. The sensitivity of the model was 100% where it was able to classify the individual with ASD class correctly. The performance of the models improved as the number of features reduced from 19 to 13 features by using Chi-Square selected features.

| Table 3. Chi-square | Table 4. Chi-square | Table 5. Feature Importance |
|---------------------|---------------------|-----------------------------|
| Features with level of significance / threshold $\alpha=0.05$ (17 features) | Features with level of significance / threshold $\alpha=0.01$ (13 features) | Feature Selected |
| A9_Score | A6_Score | A9_Score | A6_Score | A9_Score | A6_Score |
| gender | A5_Score | A10_Score | ethnicity | A4_Score | A1_Score | A10_Score |
| A7_Score | A8_Score | A7_Score | A8_Score | A7_Score | A8_Score |
| A2_Score | A3_Score | A2_Score | A3_Score | A2_Score | A3_Score |
| A4_Score | A1_Score | A4_Score | A1_Score | A4_Score | A1_Score |
| age | ethnicity | relation | jundice | contry_of_res | ethnicity |
| relation | jundice | contry_of_res | ethnicity | relation | - |

| Table 6. Logistic regression performance |
|------------------------------------------|
| Models | C value | Accuracy (%) | Specificity (%) | Sensitivity (%) | Features |
| LR-CV | 10 | 94.687 | 93.894 | 94.205 | 19 (default) |
| LR1-CV (LASSO = 0.001) | 10 | 95.082 | 95.455 | 94.118 | 17 |
| LR2-CV (LASSO = 0.01) | 10 | 97.541 | 95.455 | 100 | 16 |
| LR3-CV (Chi-square = 0.05) | 10 | 95.082 | 96.591 | 94.118 | 17 |
| LR4-CV (Chi-square = 0.01) | 10 | 97.541 | 96.591 | 100 | 13 |

3.2.2 Random Forest

In Random Forest classifier, it was found that the model with Chi-square selection method score the highest among the other models with 96.721% accuracy, 98.864% specificity and 85.294% sensitivity respectively as shown in Table 7. This performance obtained using 32 numbers of trees in the forest, 2 minimum samples leaf, and 12 maximum depths and with bootstrap. The performance trend for this classifier was increased as the features decreased for the two-selection method but not for Random Forest selection method. The performance for Random Forest selected features decreased in term of overall performance metric into 94.262% accuracy, 97.727% specificity and 79.412% sensitivity. The decrement might due to elimination of useful features of datasets.
Table 7. Random Forest performance

| Models          | Accuracy (%) | Specificity (%) | Sensitivity (%) | Features       |
|-----------------|--------------|-----------------|-----------------|----------------|
| RF-CV           | 93.442       | 97.727          | 85.353          | 19 (default)   |
| RF1-CV (LASSO = 0.001) | 90.164       | 97.727          | 79.412          | 16             |
| RF2-CV (LASSO = 0.01)  | 93.443       | 98.864          | 79.412          | 16             |
| RF3-CV (Chi-square = 0.05) | 95.082       | 98.864          | 85.294          | 17             |
| RF4-CV (Chi-square = 0.01) | 96.721       | 98.864          | 85.294          | 13             |
| RF (Feature Importance) | 94.262       | 97.727          | 79.412          | 5              |

3.2.3 K-Nearest Neighbors

In K-Nearest Neighbors classifier it was observed that the highest performance goes to model KNN4-CV with 95.902% accuracy, 97.727% specificity and 88.235% sensitivity with K=2 and hamming distance parameters. The result shows that this model has the ability to distinguish correctly patients, which does not possess ASD disorder since it has higher specificity percentage compared to sensitivity. From Table 8 it can be observed that a smaller K value and least number of features giving the best performance out of the 5 models.

Table 8. K-NN performance

| Models          | K-value | Distance | Accuracy (%) | Specificity (%) | Sensitivity (%) | Features       |
|-----------------|---------|----------|--------------|-----------------|-----------------|----------------|
| KNN-CV          | 5       | Hamming  | 95.082       | 97.727          | 88.235          | 19 (default)   |
| KNN1-CV (LASSO=0.001) | 14       | Hamming  | 95.902       | 95.455          | 97.059          | 17             |
| KNN2-CV (LASSO = 0.01) | 5        | Hamming  | 81.967       | 89.773          | 61.765          | 16             |
| KNN3-CV (Chi-square=0.05) | 16      | Hamming  | 95.902       | 96.591          | 94.118          | 17             |
| KNN4-CV (Chi-square= 0.01) | 2       | Hamming  | 95.902       | 97.727          | 88.235          | 13             |

3.2.4 Comparison on classifier

Comparison on the chosen models for each classifier will be discussed further among the three classifiers in term of accuracy, specificity, sensitivity and area under curve (AUC). Based on previous results, it can be observed that Logistic Regression score the highest with 97.541% accuracy, which it used 13 selected features based on Chi-Square selection method. RF4-CV features scored 96.721% accuracy followed by KNN2-CV selected features scored 95.902% accuracy also with Chi-Square selection method. This means the models are able to differentiate correctly patients, which does not possess ASD disorder. In term of AUC, the AUC score for the all models was summarized in Table 9. The highest AUC curve for Logistic Regression was 0.96 comes from model LR4-CV. While for Random Forest, the highest AUC was from model RF4-CV with 0.95 AUC. For K-Nearest Neighbors, the highest AUC score was 0.946 from model KNN-CV. The closer the AUC score to 1 the least misclassified case occurs, which means the higher dependability of classification.
Table 9. AUC Performance of All Classifiers

| Classifiers       | Models       | AUC   | Specificity | Sensitivity |
|-------------------|--------------|-------|-------------|-------------|
| Random Forest     | RF-CV        | 0.93  | 0.92        | 0.94        |
|                   | RF1-CV       | 0.91  | 0.92        | 0.94        |
|                   | RF2-CV       | 0.92  | 0.94        | 0.95        |
|                   | RF3-CV       | 0.94  | 0.94        | 0.95        |
|                   | RF4-CV       | 0.95  | 0.656       | 0.946       |
|                   | RFI          | 0.94  |             |             |
| Logistic Regression| LR-CV        | 0.93  | 0.94        | 0.93        |
|                   | LR1-CV       | 0.93  | 0.94        | 0.93        |
|                   | LR2-CV       | 0.94  | 0.96        |             |
|                   | LR3-CV       |       |             |             |
|                   | LR4-CV       |       |             |             |
| K-Nearest Neighbors| KNN-CV      | 0.947 | 0.779       | 0.656       |
|                   | KNN1-CV      | 0.94  |             |             |
|                   | KNN2-CV      |       |             |             |
|                   | KNN3-CV      |       |             |             |
|                   | KNN4-CV      |       |             |             |

4. Conclusion

Resulting from the selection methods, there are 6 models developed for Random Forest, 5 models for Logistic Regression and 5 models for K-Nearest Neighbors. Logistic Regression LR4-CV model gave highest performance, which 97.541% accuracy, 96.591% specificity and 100% sensitivity. While for Random Forest, the model RF4-CV scored 96.721% for accuracy, 98.864% specificity and 84.294% sensitivity. Moreover, K-Nearest Neighbors with model KNN4-CV gave highest accuracy of 95.902%, 97.727% specificity and sensitivity 88.235%. Based on the 16 models created, it was observed that the model with 13 features selected using Chi-Square method was suitable to be implemented in the classifiers especially with Logistic Regression classifier. The Logistic Regression machine learning performance promotes a high accuracy for behavior observation purpose, which can be implemented for future detection of ASD.

References

[1] C. Lord, M. Elsabbagh, G. Baird, and J. Veenstra-Vanderweele, 2018, “Autism spectrum disorder,” www.thelancet.com, vol. 392.
[2] M. Duda, R. Ma, N. Haber, and D. P. Wall, 2016, “Use of machine learning for behavioral distinction of autism and ADHD,” Transl. Psychiatry, vol. 6, p. e732.
[3] R. K. Mariette Awad, “Machine Learning and Knowledge Discovery.”
[4] T. C. McFayden, J. Albright, A. E. Muskett, and A. Scarpa, 2018, “Brief Report: Sex Differences in ASD Diagnosis—A Brief Report on Restricted Interests and Repetitive Behaviors,” J. Autism Dev. Disord., vol. 0.
[5] M. Gök, 2018, “A novel machine learning model to predict autism spectrum disorders risk gene,” Neural Computing and Applications, pp. 1–7.
[6] S. S. Abu-Naser, I. M. Nasser, I. M. Nasser, and M. O. Al-Shawwa, 2019, “Artificial Neural Network for Diagnose Autism Spectrum Disorder electronic Human Resources Management in Palestinian Universities View project Requirements of using Decision Support Systems as an Entry Point for Operations of Re-engineering in the Universities (Applied study on the Palestinian universities in Gaza Strip) View project Artificial Neural Network for Diagnose Autism Spectrum Disorder”.
[7] S. Vijayarani and M. P. Research Scholar, “Preprocessing Techniques for Text Mining-An Overview.”
[8] S. García, S. Ramírez-Gallego, J. Luengo, J. Manuel Benítez, and F. Herrera, 2016, “Big Data Analytics Big data preprocessing: methods and prospects”.
[9] K. Basu, 2018, “Machine learning approaches to the classification problem for autism spectrum disorder,” pp. 1–24.
[10] P. Kassraian-Fard, C. Matthis, J. H. Balsters, M. H. Maathuis, and N. Wenderoth, 2016, “Promises, pitfalls, and basic guidelines for applying machine learning classifiers to psychiatric imaging data, with autism as an example,” Front. Psychiatry.
[11] B. Van Den Bekerom, 2017, “Using Machine Learning for Detection of Autism Spectrum Disorder,” Twente Student Conf. IT.
[12] F. Thabtah, 2018, “Machine learning in autistic spectrum disorder behavioral research: A review and ways forward,” Informatics Heal. Soc. Care, pp. 1–20, Feb. 2018.