International Journal for Research in Engineering Application & Management (IJREAM)
ISSN : 2454-9150 Vol-05, Issue-10, Jan 2020

Comparative Study of Machine Learning Techniques Used in Autism Spectrum Disorder

[1]Apansula Sharma, [2]Diana Nagpal, [3]Supreet Kaur

[1]Research Scholar, [2,3]Assistant Professor, GNDEC, Ludhiana, India.

[1]apansulasharma@gmail.com, [2]nagpal.d25@gmail.com, [3]ghumansupreet@gmail.com

Abstract: Autism Spectrum Disorder is a neuro developmental disorder characterized by persistent deficits in social interaction and communication and restricted, repetitive patterns of behavior, interests or activities. The paper shows the detailed comparative analysis of various machine learning techniques used in the field of autism spectrum disorder.

Keywords: Autism Spectrum Disorder (ASD), Machine Learning.

I. INTRODUCTION

Autism spectrum disorders (ASD) are a group of neurodevelopmental disorders characterized by the difficulties of social interaction and communication skills, limited repetitive interests and behaviors. The main characteristic of ASD includes inadequacy of social interaction, language and communication skills, repetitive self-stimulation, exhibiting inappropriate behaviors, excessive dependence on routines and uniformity [1]. The autism spectrum disorders are measured based on the presence of multiple symptoms that disrupt the child's ability to talk, make the relationships, explore, play, and to study.

The autism spectrum disorders belong to an "umbrella" class category of five childhood-onset Conditions called pervasive developmental disorders (PDD). They are concerning the three most common PDDs such as (1) Autism, (2) Asperger's Syndrome and (3) Pervasive Developmental Disorder. Other pervasive developmental disorders is (PDD-NOS) Childhood disintegrative disorder and Rett Syndrome. As results of each are extremely rare genetic diseases, they are sometimes thought of to be separate medical conditions that do not really belong on the autism spectrum.

II. LITERATURE SURVEY

Osman Altay et al. [2018] reviewed the classification method for ASD diagnosis was used in children aged 4-11 years. The Linear Discriminant Analysis (LDA) and The K-Nearest Neighbor (KNN) algorithms are used for classification. To test the algorithms, 30 percent of the data set was selected as test data and 70 percent as training data. As a result of the work done; In the LDA algorithm, the accuracy is 90.8%, whereas the accuracy of the KNN algorithm is 88.5%. For the LDA algorithm, sensitivity and specificity values are calculated as 0.9524 and .08667, respectively. For KNN algorithm, these values are calculated as 0.9762 and 0.80. F-measure values are calculated as 0.9091 for the LDA algorithm and 0.8913 for the KNN algorithm [2].

Elizabeth Stevens et al. [2017] used cluster analysis to a sample of 2,116 children with Autism Spectrum Disorder in order to identify patterns of challenging behaviors observed in home and centerbased clinical settings. The largest study of this type to date, and the first to employ machine learning, our results indicate that while the presence of multiple challenging behaviors is common, in most cases a dominant behavior emerges. Furthermore, the trend is also observed when we train our cluster models on the male and female samples separately. This work provides a basis for future studies to understand the relationship of challenging behavior profiles to learning outcomes, with the ultimate goal of providing personalized therapeutic interventions with maximum efficacy and minimum time and cost [3].

Nakai et al. [2017] comparing the performance of machine learning vs the clinical judgment of speech therapists in classifying children with ASD and children with TD based on single-word utterances. Participants included 30 children with ASD and 51 children with typical development. All children were between the ages of 3 and 10 years old and had no comorbid disorders. After isolating their single-word responses, an SVM classifier with cross-validation on 24 features was employed to identify ASD or TD. The SVM proved more accurate (76%) than the 10 speech therapists whose classifications were also based on the same audio recordings (69%). The SVM had a sensitivity of 81% and specificity of 73% compared to the therapists, who demonstrated a sensitivity of 54% and specificity of 80%. This study shows the potential of machine learning analysis of speech prosody as a useful screening tool [4].

Duda et al. [2016] collected and analyzed data from a web-based 15-question parent survey. Participants included 248 individuals with ASD and 174 individuals with ADHD, ages 2 to 17 years old, with no comorbidities

 DOI : 10.35291/2454-9150.2020.0024

base on parental report. A second archival dataset with SRS scoresheets was obtained from multiple repositories. The archival dataset included 2775 subjects with ASD and 150 subjects with ADHD. Subjects were diagnosed by a physician and had no comorbidities. The dataset was subsampled to maintain diagnosis proportions, and only the 15 features correlated to the survey were retained [5].

Khalid Al-jabery et al. [2016] In this paper, they present an ensemble model for analyzing ASD phenotypes using several machine learning techniques and a dimensional subspace clustering algorithm. They ensemble also incorporates statistical methods at several stages of analysis. They apply this model to a sample of 208 probands drawn from the Simon Simplex Collection Missouri Site patients. The results provide useful evidence that is helpful in elucidating the phenotype complexity within ASD. Their model can be extended to other disorders that exhibit a diverse range of heterogeneity [6].

Engchuan et al. [2015] used machine learning models to analyze genes, specifically rare copy number variation (CNV), associated with ASD. The dataset was comprised of 1892 participants with ASD and 2342 controls with at least one rare CNV. Using rare CNV data and comprehensive gene annotations, four classification methods were conducted and compared. The CF model's performance was equal or superior to the other tested classification methods. The best classifier demonstrated an AUC of 0.533, correctly categorizing 7.9% of the participants with ASD while incorrectly classifying less than 3% of the controls. Performance improved when limiting the model to participants with de novo CNVs (i.e., those occurring spontaneously as opposed to inherited from a parent) or pathogenic CNVs (i.e., those that have been previously associated with ASD). Rare genic losses were found to be more predictive than gains when analyzed alone. Finally, 20 features identified as neurally relevant were found to perform better in the model than total gene count [7].

Jiao et al. [2012] used machine learning to classify children with ASD according to symptom severity using data on genetic markers. The dataset included single nucleotide polymorphism (SNP) data for 118 children with ASD between the ages of 1.5 to 14 years old. Using the results of the Childhood Autism Rating Scale, participants were divided into two groups based on symptom severity. A total of 65 participants made up the mild/moderate group and 53 participants made up the severe group. Of the machine learning models evaluated, decision stumps and Flex Trees were found to perform best with an accuracy of 67%, sensitivity of 88%, and specificity of 42% [10].

III. MACHINE LEARNING TECHNIQUES

Machine learning techniques capture the multi-variate relationships in data and hence are well-suited to detect subtle and distributed differences in the data. So, compared to univariate techniques, machine learning techniques can perform better in capturing the brain morphology of heterogeneous conditions like ASD. Thus, they hold promise for improving our knowledge of ASD brain morphology and identifying brain biomarkers helpful for ASD diagnosis.

1. Gradient Boosting Machine (GBM): Boosting is an ensemble technique that relies on bias reduction to reduce the generalized error of an ensemble. A general boosting technique iteratively combines 30 several weak or base learners with high bias and low variance such as decision tree stumps into one strong learner. The base learners are combined so that the ensemble bias decreases while variance remains the same, thereby reducing the net ensemble error. At each iteration or boosting step, GBM constructs a new base learner to be the most parallel to the negative gradient of a loss function along the observed data so that the new base learner focuses on the weakness of the model. In other words, it performs functional approximation of a model by consecutively improving along the negative direction of a loss function.

2. Decision Tree: The decision tree is a visual representation that is used as part of a selection criteria, or even to support the selection of specific data, considering the overall structure. It represents choices and its results in the form of a tree. It can start with simple questions that will have 2 or more answers, leading to a further question, and so on. It will support to identify and classify the data. Decision trees are mostly used in Data Mining applications using machine learning.

3. Support Vector Machine: An SVM is a supervised learning algorithm that fits an optimal hyperplane in an n-dimensional space to correctly categorize the target result using the independent variables in the dataset. An SVM is a maximum margin classifier, meaning it maximizes the separation between n classes of data effectively in a high-dimensional space. SVMs are especially useful when the boundary between groups is non-linear because points can be easily transformed to a space in which the boundary is linear. Because of this feature, SVMs are generally used in classification problems in which the distinction between groups is non-linear. SVM algorithms have been used in the research included in this review to classify individuals (e.g., according to diagnosis) based on standardized assessments, genes, neuroimaging, and other measurements [8].

4. Random Forest: Random Forest is an ensemble of decision trees and its output class is the mode value of the output classes of the individual decision trees. It is an ensemble technique that relies on the reduction of the variance of the general error term. For the squared error

 DOI : 10.35291/2454-9150.2020.0024

loss, the expected generalization error of a model can be decomposed into three components. The first term noise is the irreducible error or Bayes error. It is the theoretical lower bound on the generalization error and is independent of both learning algorithm and data. The second term bias is the difference between the average prediction of the model and the prediction of the Bayes model. The third term var is the variability of the predictions at point over the models learned from all possible subsets of population. The main idea of RF is to decrease the variance term by keeping the bias constant, thereby decreasing the overall error of the ensemble. It achieves this variance reduction by averaging the high variance classifiers or decision tree classifiers. The more diverse or uncorrelated the decision trees, the more error reduction is achieved by averaging. To make the decision trees different from 28 each other, RF introduces randomness while constructing the trees, hence the name 'random forest'. The randomization is introduced at first during data sampling and then while constructing the decision trees. Each tree learns from a bootstrap replica of the data obtained by random sampling with replacement in the original data. This introduces a degree of randomness in the decision trees because they are trained with different bootstrap replicas. While growing decision trees, the quality of a node split is based only on a random subsample of the variables instead of all of them [9].

IV. COMPARATIVE ANALYSIS

This section presents the comparison of the proposed technique with the existing technique on the basis of various parameters. Following are some Parameters which evident that proposed method is the stand-alone approaches. Below Table show the performance of the model in terms of Sensitivity, Accuracy, MCC, F-measure.

Table 1. Accuracy Comparison of Proposed Technique with the Existing Technique

Algorithms	Accuracy
Naïve Bayes	67.5
Bayes Networks	59.6
Logistic	74.4
Proposed	78.31

Table 2. MCC Comparison of Proposed Technique with the Existing Technique

Algorithms	MCC
Linear SVM	0.264
Naïve Bayes	0.276
Logistic	0.197
Proposed Method	0.583

Table 3. Sensitivity of Proposed Technique with the Existing Technique

Algorithms	Sensitivity
Linear SVM	0.363
Naïve Bayes	0.683
Logistic	0.402
Proposed Method	0.902

Table 4. F-measure Comparison of Proposed method with the Existing Technique

Algorithms	F-Measure
Linear SVM	0.383
Naïve Bayes	0.419
Logistic	0.350
Proposed	0.806

Table 5. Comparison between ASD disease Classification Methods with Proposed method

	Sensitivity	Specificity	Accuracy	MCC	AuROC
Cogill and Wang's FFS method	0.784	0.737	73.90	0.385	0.805
Cogill and Wang's SFS method	0.744	0.772	76.70	0.419	0.819
Proposed Method	0.902	0.665	78.31	0.583	0.839

The table above shows the comparison of ASD Disease Classification method with proposed technique. The proposed method is better than the other classification method Sensitivity, specificity, accuracy, MCC and AuROC of the proposed method is 0.902, 0.665,78.31,0.583 and 0.839 whereas Cogill and Wangs's FFS Method Sensitivity, specificity, accuracy, MCC and AuROC is 0.784, 0.737, 73.90, 0.385 and 0.805 and Cogill and Wangs's SFS Method Sensitivity, specificity, accuracy, MCC and AuROC is 0.744,0.772,76.70,0.419 and 0.819.

V. CONCLUSION

The paper demonstrates the analytical study of different autism spectrum disorder techniques. It concludes with the comparison study of the techniques used in the field of ASD on the basis of accuracy, specificity, sensitivity, MCC and AuROC as the parameters. By comparing the

 DOI : 10.35291/2454-9150.2020.0024

different parameters, it is concluded that the paper has best results when compared to the other machine learning approaches.

REFERENCES

[1] Gok M.**, "**A novel machine learning model to predict autism spectrum disorders risk gene," *Neural Computing and Applications*, vol. 31, pp. 6711-6717, April 2018.

[2]Osman Altay, Mustafa Ulas, "Prediction of the Austim Spectrum Disorder Diagnosis with Linear Discriminant Analysis Classifier and K-Nearest Neighbour," *2018 in 6th International Symposium on Digital Forensic and Security (ISDFS)*, March 2018.

[3] Stevens, E., Atchison, A., Stevens, L., Hong, E., Granpeesheh, D., Dixon, D., & Linstead, E., "A Cluster Analysis of Challenging Behaviors in Autism Spectrum Disorder," *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, Dec. 2017.

[4] Nakai, Y., Takiguchi, T., Matsui, G., Yamaoka, N., & Takada, S., "Detecting Abnormal Word Utterances in Children With Autism Spectrum Disorders," *Perceptual and Motor Skills*, vol. 124, no. 5, pp. 961–973, Oct. 2017.

[5] Duda, M., Daniels, J., & Wall, D. P., "Clinical Evaluation of a Novel and Mobile Autism Risk Assessment," *Journal of Autism and Developmental Disorders*, vol. 46, no. 6, pp. 1953–1961, June 2016.

[6] Al-jabery Khalid, Obafemi-Ajayi, T., Olbricht, G. R., Takahashi, T. N., Kanne, S., & Wunsch, D., "Ensemble statistical and subspace clustering model for analysis of autism spectrum disorder phenotypes," *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Aug. 2016.

[7] Engchuan, W., Dhindsa, K., Lionel, A. C., Scherer, S. W., Chan, J. H., & Merico, D., "Performance of case-control rare copy number variation annotation in classification of autism," *BMC Medical Genomics*, vol. 8, pp. 1-10, Jan. 2015.

[8] Bishop, C. M, "Pattern recognition and machine learning" in New York, NY: Springer 2006.

[9] D. Cutler, T. Edwards Jr., K. Beard, A. Cutler, K. Hess, J. Gibson, and J. Lawler, "Random forests 6 for classification in ecology" in Ecology 2007.

[10] Jiao, Y., Chen, R., Ke, X., Cheng, L., Chu, K., Lu, Z., & Herskovits, E. H., "Single Nucleotide Polymorphisms Predict Symptom Severity of Autism Spectrum Disorder," *Journal of Autism and Developmental Disorders*, vol. 42, no. 6, pp. 971–983, June 2012.

[11] Hyde, K., A.-J. Griffiths, C.M. Giannantonio, A.E. Hurley-Hanson, E, "Linstead Predicting employer recruitment of individuals with autism spectrum disorders with decision trees," *17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2018.

[12] Liu, W., Li, M., & Yi, L., "Identifying children with autism spectrum disorder based on their face processing abnormality: A machine learning framework," *Autism Research*, vol. 9, no. 8, pp. 888–898, Aug. 2016.

[13] Mythili M. S., & Shanavas A.R. M., "A Study on Autism Spectrum Disorders using Classification Techniques," *International Journal of Soft Computing and Engineering*, vol. 4, no. 5, pp. 88-91, Nov. 2014.

[14] Thabtah, F., "Machine learning in autistic spectrum disorder behavioral research: A review and ways forward," *Informatics for Health and Social Care*, vol. 44, no. 3, pp. 1-20, Sept. 2019.

[15] Crippa, A., Salvatore, C., Perego, P., Forti, S., Nobile, M., Molteni, M., & Castiglioni, I., "Use of Machine Learning to Identify Children with Autism and Their Motor Abnormalities," *Journal of Autism and Developmental Disorders*, vol. 45, no. 7, pp. 2146–2156, July 2015.

 DOI : 10.35291/2454-9150.2020.0024