Intelligent Frame Work to Predict Autism in Infants using Machine Learning

S.V. Evangelin Sonia, R. Prasanth, B. Mythreyi, A. Harshavardhan, A. Manoharan

Abstract: Autistic Spectrum Disorder (ASD) is a behavioral impairment that interferes with the use of auditory, communicative, cognitive, abilities and social skills. ADS was introduce researched using advanced approaches based on machine learning to speed up the diagnostic period or increase the responsiveness, reliability or precision of the diagnosis process. Normal medical checkup data (training data) of the baby, test data is used to classify the child with autism. The assessment results showed that, for both types of datasets, the proposed prediction model produces better results in terms of precision, responsiveness, precision and false positive rate (FPR). **In the modern day, Autism Spectrum Disorder (ASD) is gaining some momentum quicker than ever. Detecting signs of autism by screening tests is very costly and time consuming. Autism can be identified faster by combining artificial intelligence and Machine Learning (ML).**

Keywords*: *Autistic Spectrum Disorder, Machine Learning, Classifiers, Predictions.

I. INTRODUCTION

Autistic Spectrum Disorder (ASD) is a brain-development condition that prevents normal growth from some speech and social behaviors. However, its causes have been attributed to genetic and neurological factors. Despite its genetic basis, ASD is primarily diagnosed using behavioral factors such as social interaction, imaginative ability, repetitive behaviors and people-to-people communication.

Children with ASD face more severe early developmental problems compared with other groups of children. These behavioral issues vary and include difficulties in responding to sensory stimuli (hearing, thinking, feeding, etc.), lagging difficulties in language comprehension and communication, disrupting early learning and finding it difficult to communicate with others. Autism is a disorder that can be characterized by actually being different from their neurotypical peers. Parents can determine the prevalence of autism in children using Machine Learning algorithms through classic signs such as lack of eye contact, repeated gestures that sensory problems.

Revised Manuscript Received on April 21, 2020.

Correspondence Author:

S.V.Evangelin Sonia*, Assistant Professor, Department of Computer Science and Engineering, Sri Shakthi Institute of Engineering and Technology, Coimbatore, India.
Email: evangelinsonia.vs@gmail.com

R.Prasanth, Department of Computer Science and Engineering, Sri Shakthi Institute of Engineering and Technology, Coimbatore, India.
Email: sundarprasanth1999@gmail.com

B. Mythreyei, Department of Computer Science and Engineering, Sri Shakthi Institute of Engineering and Technology, Coimbatore, India.
Email:mythreyeib@gmail.com

A.Harshavardhan, Department of Computer Science and Engineering, Sri Shakthi Institute of Engineering and Technology, Coimbatore, India.
Email: vardhan0211@gmail.com

A.Manoharan, Department of Computer Science and Engineering, Sri Shakthi Institute of Engineering and Technology, Coimbatore, India.

ASD is clinically diagnosed by examining three aspects of activity (American Association of Psychologists, 2000): communication and speech, social interaction and restricted behavior. A variety of methods of clinical and non-clinical treatment for ASD have been employed. Examples ofclinical screening approaches are the Schedule-Revised Autism Experimental Assessment (ADOS-R) and an interview on Autism Screening (ADI). Some of the latest clinical ASD diagnostic methods have demonstrated competitive performance, such as ADOS-R and ADI-R, which have achieved reasonable sensitivity, precision, and validity outcomes in many experimental research trials. They therefore need extremely careful use in addition to accuracy, and need the availability of expert clinicians.

To enhance ASD's diagnostic mechanism, researchers have recently begun to implement intelligent methods of machine learning. The primary aim of these ASD machine learning studies is to improve a case's diagnostic time enhancing diagnostic precision and dimensionality of the input data set to identify the top-rated ASD features and provide easier access to health care facilities. More specifically, In these studies, we look critically at innovations related to the development of new machine learning methods for classifying ASD, decreasing diagnostic time and minimizing features, improving diagnosis, accuracy, specificity and responsiveness.

Recent ASD studies on machine learning to critically assess developments in this research, particularly the development of new machine learning methods for automatic ASD classification. When implementing machine learning for ASD classification, we show recent results and
challenges which future studies will consider for improving the quality of the study. We believe that machine learning will be the next generation of screening tools, where automated mathematical models will replace hand crafted methods of classification. Such models can guide clinical experts to make fast yet accurate diagnostic decisions.

II. OBJECTIVE

Autism Spectrum Disorder (ASD) is a behavioral condition that inhibits the use of verbal, communicative, cognitive, and social skills. It was tested using the Intelligent Methods of Machine Learning to improve the process of identifying Autism in Children. The current way of predicting Autism in child is done using only the screening process which is a time consuming and a tedious process. This ML based detection will also improve the specificity, sensitivity, accuracy of the detection even better than the current screening process if the Data provided is appropriate. The Data considered here for training the model is being obtained from the infant's regular medical checkup data.

So, the ultimate objective of this study is to apply artificial intelligence and Machine learning techniques and predict whether the child has Autism Spectrum Disorder.

III. LITERATURE REVIEW

J Am Acad Child Adolesc Psychiatry(2012) has stated that the Frontline health professionals need a "red flag" tool to help them making a decision on whether to make a referral for a full diagnostic assessment for an Autism Spectrum Disorder (ASD) in children as well as Adults. Their aim was to identify the 10 items based on the Autism Spectrum Quotient (AQ) with good accuracy.

According to Neda Abdelhamid(2014) the process of doing Associative classification (AC) is one of the promising data mining approaches that is used to integrate classification and association rules discovery to build effective classification models (classifiers). In the last decade, several Associative classification algorithms such as Association Based Classification (CBA), Predicted Association Rule (CPAR) classification, Multi-class Classification using Association Rule (MCAR), Live and Let Live (L3), and others have been proposed.For the various test cases, such algorithms use different procedures for rule learning, rule sorting, rule pruning, classifier construction, and class allocation. This paper portrays the light and critically compares common Associated Classification algorithms with reference to the aforementioned process.

Duda M (2016) reported that the Autism Spectrum Disorder (ASD) and attention deficit hyperactivity disorder (ADHD) continue to rise in prevalence, affecting over 10 per cent of today's pediatric population together. The testing approaches used remain arbitrary, inefficient and time-intensive. For the delays over a year between initial detection and diagnosis, there is often a lack of precious time in which medications and therapeutic measures may be implemented, since such conditions remain undetected. Techniques to determine risk for these and other developmental disabilities rapidly and reliably are required to streamline the diagnosis and therapy process faster.

IV.METHODOLOGY USED

The data sets are taken from autistic dataset for toddlers. Preprocessing the data is an important phase in the process of data mining. We have used two types of pre-processing methods Standard Scalar and Min Max Scalar. The concept behind Standard Scalar is to transform the data so its distribution has a mean value of 0 and a standard deviation of 1. The Min Max scalar[10] standardizes the functionality by scaling each function to a given range. This estimator scales and individually translates each feature, so that it is on the training set in the given range, i.e. between zero and one. Feature Selection is the mechanism where certain features are picked that most relate to the predictive variable or performance. The Recursive Feature Elimination (RFE) approach is used for selecting the features most relate to the predictive variable or performance. Machine learning algorithms construct a mathematical model based on sample data, referred to as "training data," such that forecasts or decisions can be made without complex task programming. The ML algorithms used in this paper are the algorithms Logistic Regression, Support Vector Machine, Random Forest, K-Nearest Neighbor. The 10-fold cross validation with train split approach is used to verify the attrubutes. As training data, 80 percent data is used and 20 percent data are used as test data. The effectiveness of the models is assessed in a clinical environment where limitations on operational quantities, such as precision and reliability of the model, have to be met and sepsis detected correctly before impacting.

A. Model Diagram

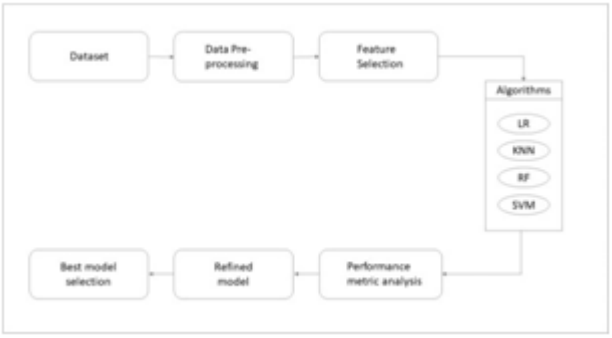


Fig 1: Model Diagram

B. Data Set

Data used in the study is autistic dataset for toddlers. The description of the dataset attributes are given below:

Attribute Name	Datatype
A1	int64
A2	int64
A3	int64
A4	int64
A5	int64
A6	int64
A7	int64
A8	int64
A9	int64
A10	int64
Age_Mons	int64
Qchat-10-Score	int64
Sex	Object
Ethnicity	Object
Jaundice	Object
Family_mem_with_ASD	int64
Class/ASD Traits	Object
Dtype	Object

C. Data Preprocessing

Data processing is a crucial step in making it suitable for the most common forms of ML.A large amount of data is typically needed. To get better results from the model applied in Machine Learning projects, the data format has to be right.

Machine learning primarily relies on the results of the tests. Pre-processing the data is an important step in the process of data mining.

The expression "garbage in, garbage out" refers in particular to projects involving data mining and machine learning. Techniques for collecting data are often poorly supervised, resulting in in-range values (e.g., income:
−100), odd data combinations (e.g., sex: male, pregnant: yes), missing values etc. Analysis of data which has not been thoroughly analyzed for these issues can produce misleading results. The study and precision of the data is therefore primarily prior to an experiment. Information processing is often the most critical step of a machine learning project, particularly in computational biologyIf there is much unnecessary and redundant information, or noisy and inaccurate data, then it becomes more difficult to discover knowledge during the training process. It will take considerable time to handle the preparation and processing of data steps. The final training set is a feature of the preprocessed data.Pre-processing data requires cleaning, gathering events, normalizing, converting, extracting and gathering features etc. The preprocessing methods employed here are Normal Scalar and Min Max Scalar. The definition of Standard Scalar is to transform the data, so that its distribution has a mean value of 0 and a standard deviation of
1. Scaling every feature to a certain amount, The Min Max scalar renders functionality standardized. This estimator scales and independently translates each function, so that it is within the defined range, i.e. between zero and one, on the training set.

D. Feature Selection

Feature Selection is the mechanism where certain features are selected which mostly relate, automatically or manually, to the predictive variable or output. With irrelevant characteristics in the data, the accuracy of the model can be decreased and the model learned based on irrelevant characteristics. The Recursive Elimination Function (RFE) algorithm is used for selecting features in the implementation. The Recursive Elimination Function (RFE) algorithm is used to pick Implementation features. The RFE approach reflects a function selection strategy. Works by deleting the attributes recursively, and building a pattern on the remaining attributes. The precision of the model is used to distinguish what attributes (and attribute combinations) contribute most to the prediction of the target attribute.

E. Machine Learning Algorithms

Machine learning (ML) is the theoretical analysis of computer systems using algorithms and mathematical models without the use of explicit instructions to perform a particular task. This is seen as a branch of artificial intelligence. Machine learning algorithms create a mathematical model that is based on sample data, known as "training data," so predictions or decisions can be made without complex task programming. The logistic regression(LR), support vector machine(SVM), random forest(RF), k-nearest neighbors(K-NN), the ML algorithms used in this paper are.

E.1 Logistic regression (LR)

Logistic regression is named at the core of the method for the function employed, the logistical equation. Statisticians developed the logistic equation, also known as the sigmoid equation, to describe the characteristics of ecological population growth, gradually expanding and optimizing the potential of the ecosystem.
It is an S-shaped curve that can take any real-evaluated number and map it to a value between 0 and 1, but never precisely at those limits 1/(1 + e^-value)
Where e is the basis for the natural logarithm (Euler number or EXP) (function in your spreadsheet) and where e is the actual numerical value you wish to transform.

E.2 Support Vector Machine

XGBoost is a Machine Learning algorithm based on an ensemble of decision-trees using a gradient boosting method. Artificial neural networks tend to outperform all the other algorithms or systems when solving problems involving unstructured data (images, text, etc.). Nonetheless, decision-based algorithms based on a tree are considered best-in-class when it comes to small to medium structured / tabular data. A wide variety of applications: Can be used to solve problems with regression, classification, ranking, and user-defined prediction. Taking into account the portability, it can run smoothly on Windows, Linux and OS X. All major programming languages which include C++, Python, R, Java, Scala and Julia are supported languages.It supports the AWS, Azure, and Yarn clusters and functions well with Flink, Spark, and other ecosystems [4].

E.3 K-Nearest Neighbor

K closest neighbor(KNN) is a simple algorithm which stores all cases and classifies new cases based on a measure of similarity. KNN algorithm also referred to as 1) case-based reasoning 2) k closest neighbor 3) example-based reasoning 4) memory-based learning 5). 6) Lazy learning[4].Since 1970 KNN algorithms have been used in many applications such as statistical estimation and pattern recognition, and so on. KNN is a non-parametric classification method that is typically divided into two forms 1) less structure of NN techniques 2) structure based NN techniques.For less NN techniques in structure, entire data is classified into sample data for training and analysis. Evaluate the distance from training point to the sample point, and the lowest distance point, is called the nearest neighbor. NN methods based on structure are based on data structures such as orthogonal tree structure (OST), ball tree, k-d tree, axis tree, nearest future line, and central line. Nearest neighbor classification is mostly used when all attributes are continuous.

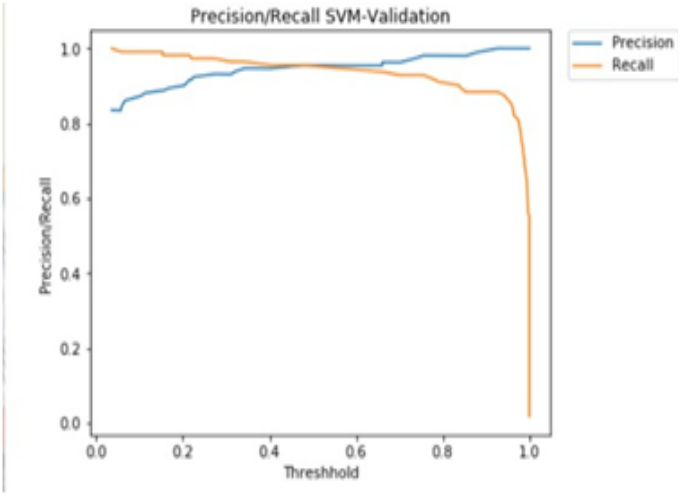
E.3 Random Forest

Random forests are a mixture of tree predictors, such that each tree depends on the values of an independently sampled random vector and on all forest trees with the same distribution. Error of generalization of the woods converges a.s. When the number of trees in the forest increases, to a limit. The generalization error of a tree classifier forest depends on the power and relation of the individual trees within the forest. Using a random collection of features to isolate each node yields error rates that are more stable than Adaboost in terms of noise, but higher.

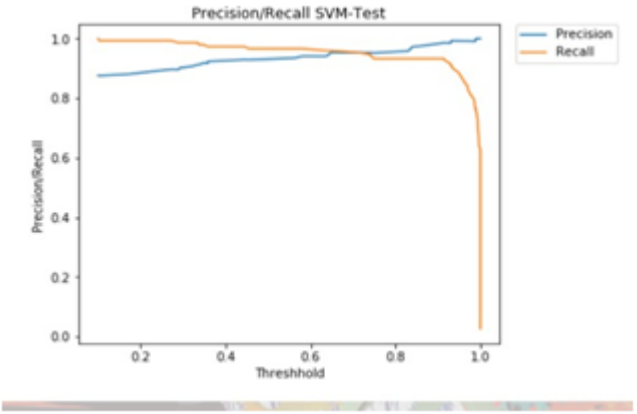
V. EXPERIMENTAL RESULTS

The experiment started with the Collection of Data from toddler Autism Dataset through Kaggle. Once, it is retrieved, the Data is sent for Pre-processing in order to clean up the data.

The Data pre-processing step includes removal of missing values or identification of an apt value for filling the missing values using any of the statistical measures.
Sometimes, the data might have irrelevant data that may end up in inappropriate prediction or hindering the performance of the model. So, all those measure are being taken to avoid such issues.
Then, the target value is identified of its type and the applicable algorithms are found out. Since, our project deals with the prediction of whether the infant will be affected by the Autism, we have used Supervised learning algorithm. The classification algorithms are identified and the Data is passed to them.
When irrelevant features are passed to the model, then the performance will be affected. Hence, the feature selection and feature engineering has to be done before training the model for better results.
Once this process is done, then the proper dataset is passed to all the applicable algorithms and the model is well trained with the training Data.
Now, the same data is supplied to all the algorithms and the accuracy of each model is checked. As the last step, one best model is identified among all the algorithms and its accuracy is tuned for best result.
Here, we have taken 4 algorithms for the analysis such as Logistic regression, vector machine support, K-Nearest Neighbours, Random Wood. All these 4 algorithms are trained and defined as best results are provided by the Logistic regression and Support vector machine. Logistic regression gave the most optimal result and the second came SVM. But SVM is considered as the best algorithm. The reason for this is that, in Logistic regression, a threshold value will be considered and if the output value is very small, the output will be considered as “No”. But that is a wrong prediction. This case will not happen in Support vector Machine. Considering all these, Support Vector Machine is taken for final tuning. The model is well trained and the output was able to achieve 100% accuracy in its prediction. The below picture shows the curve when training data is passed to the model :



The below figure depicts the curve when Test data is being passed:



VI.CONCLUSION

Some of the big challenges of Autism research today is strengthening diagnosis of current testing methods so that as a quicker service, the individuals can get more reliable and better outcomes. This can be done in many ways. By reducing the diagnosis time or increasing predictive accuracy of the diagnosis without issues in the quality or sensitivity of the test. Through this treatment, the implementation of machine learning has proven its best outcome in the industry.
Nevertheless, the conceptual, implementation, evaluation and data related problems were not thoroughly considered in this report. When these are integrated into the existing tool, especially SVM, then it might greatly serve experts and researchers as effective decision making tools.

REFERENCE

1. Gorodetski, I. Dinstein and Y. Zigel, "Speaker diarization during noisy clinical diagnoses of autism," 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Berlin, Germany, 2019, pp. 2593-2596.
2. K. A. A. Mamun et al., "Smart autism — A mobile, interactive and integrated framework for screening and confirmation of autism," 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Orlando, FL, 2016, pp. 5989-5992.
3. M. Elbattah, R. Carette, G. Dequen, J. Guérin and F. Cilia, "Learning Clusters in Autism Spectrum Disorder: Image-Based Clustering of Eye-Tracking Scanpaths with Deep Autoencoder," 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Berlin, Germany, 2019, pp. 1417-1420.
4. N. Büyükoflaz and A. Öztürk, "Early autism diagnosis of children with machine learning algorithms," 2018
5. 26th Signal Processing and Communications Applications Conference (SIU), Izmir, 2018, pp. 1-4.
6. Abbas, F. Garberson, E. Glover and D. P. Wall, "Machine learning for early detection of autism (and other conditions) using a parental questionnaire and home video screening," 2017 IEEE International Conference on Big Data (Big Data), Boston, MA, 2017, pp. 3558-3561.
7. Tyagi, R. Mishra and N. Bajpai, "Machine Learning Techniques to Predict Autism Spectrum Disorder," 2018 IEEE Punecon, Pune, India, 2018, pp. 1-5.
8. E. Linstead, R. Burns, D. Nguyen and D. Tyler, "AMP: A platform for managing and mining data in the treatment of Autism Spectrum Disorder," 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Orlando, FL, 2016, pp. 2545-2549.
9. Tyagi, R. Mishra and N. Bajpai, "Machine Learning Techniques to Predict Autism Spectrum Disorder," 2018 IEEE Punecon, Pune, India, 2018, pp. 1-5.

10. M. Che, L. Wang, L. Huang and Z. Jiang, "An Approach for Severity Prediction of Autism Using Machine Learning," 2019 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), Macao, Macao, 2019, pp. 701-705.
11. S. Sartipi, M. G. Shayesteh and H. Kalbkhani, "Diagnosing of Autism Spectrum Disorder based on GARCH Variance Series for rs-fMRI data," 2018 9th International Symposium on Telecommunications (IST), Tehran, Iran, 2018, pp. 86-90.
12. S. R. Dutta, S. Giri, S. Datta and M. Roy, "A Machine Learning-Based Method for Autism Diagnosis Assistance in Children," 2017 International Conference on Information Technology (ICIT), Bhubaneswar, 2017, pp. 36-41.
13. J. Kim et al., "Optimal Feature Selection for Pedestrian Detection Based on Logistic Regression Analysis," 2013 IEEE International Conference on Systems, Man, and Cybernetics, Manchester, 2013, pp. 239-242.
14. Huang, Y. Chen, W. Chen, H. Cheng and B. Sheu, "Gastroesophageal Reflux Disease Diagnosis Using Hierarchical Heterogeneous Descriptor Fusion Support Vector Machine," in IEEE Transactions on Biomedical Engineering, vol. 63, no. 3, pp. 588-599, March 2016.
15. Yi, Q. Xiong, Q. Zou, R. Xu, K. Wang and M. Gao, "A Novel Random Forest and its Application on Classification of Air Quality," 2019 8th International Congress on Advanced Applied Informatics (IIAI-AAI), Toyama, Japan, 2019, pp. 35-38.
16. Ping-Hung Tang and M. Tseng, "Medical data mining using BGA and RGA for weighting of features in fuzzy
17. k-NN classification," 2009 International Conference on Machine Learning and Cybernetics, Hebei, 2009, pp. 3070-3075.
18. S. S. Alwakeel, B. Alhalabi, H. Aggoune and M. Alwakeel, "A Machine Learning Based WSN System for Autism Activity Recognition," 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA), Miami, FL, 2015, pp. 771-776.
19. Linstead, R. German, D. Dixon, D. Granpeesheh, M. Novack and A. Powell, "An Application of Neural Networks to Predicting Mastery of Learning Outcomes in the Treatment of Autism Spectrum Disorder," 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA), Miami, FL, 2015, pp. 414-418.
20. Short, D. Feil-Seifer and M. Matarić, "A comparison of machine learning techniques for modeling human-robot interaction with children with autism," 2011 6th ACM/IEEE International Conference on Human-Robot Interaction (HRI), Lausanne, 2011, pp. 251-252.

AUTHORS PROFILE

S.V.Evangelin Sonia was born in Nagercoil, India, in 1987.She received the B.E degree in Computer Science and Engineering from the Anna University, India, in 2009, and the M.E. degree in Computer Science and Engineering from the Anna University, India, in 2011 and pursuing Ph.D degree in Computer Science and Engineering from the Anna University, India.In 2011,She joined as an Assistant Professor in Vins Christian College of Engineering ,Nagercoil,India. Since June 2016, she has been with the
Department of Computer Science and Engineering, Sri Shakthi Institute of Engineering and Technology,Coimbatore, where she is an Assistant Professor. Her current research interests include Data Mining,Data Analytics and Machine Learning.She is a Life Member of Institution of Engineers (India) [IEI].

R.Prasanth was born in Coimbatore ,India ,in 1999. He is currently pursuing his Bachelors degree in Computer Science and Engineering in Sri Shakthi Institute of Engineering and Technology.
Internship in Ducen Pvt.Ltd since July 2019. Her current role in the Company is Product Owner.

A.Harshavardhan was born in Coimbatore, India, in 1998. He is currently pursuing his Bachelors degree in Computer Science and Engineering in Sri Shakthi Institute of Engineering and Technology.He is also doing his Internship In Closerlook Digital Software Services Pvt Ltd. His current role in the company is IoT and Web Developer.

J.Manoharan was born in Tirupur, India, in 1998. He is studying his Bachelors degree in Computer Science and Engineering in Sri Shakthi Institute of Engineering and Technology.

B.Mythreyei was born in Coimbatore, India, in 1998. She is currently pursuing her Bachelors degree in Computer Science and Engineering in Sri Shakthi Institute of Engineering and Technology. She is also doing her