A Study on Various Techniques For Predicting Autism Using Weka Tool

R. Ramya[1], B. S. E. Zoraida[2]
[1] Dept of Computer Science
[2]Assistant Professor, Dept of Computer Science
[1, 2] Bharathidasan University, Tiruchirappalli – 620023.

Abstract- *Autism is spectrum disorder and developmental disability that affects children prior to the age of three in three different areas Verbal and nonverbal communication, Social interaction and Academic performance. There is extensive degree of variation in affects human beings. Every child on the Autism Spectrum Disorder (ASD) has unique abilities, symptoms and challenges. The child health organization is identified and proves that males are higher risks for neurodevelopment disorder such ASD than girls. The occurrence of ASD based on the gender. The autism can be classified as three different levels like, low-functioning, moderate and high-functioning. ASD diagnosing is based on patient symptoms dataset. This paper focuses on predicting the level of ASD using various data mining techniques in WEKA tool.*

Keywords- Autism Spectrum Disorder (ASD), Naïve Bayes, Logistic Regression, IBk (K-Nearest neighbor), SMO (Sequential Minimal Optimization) and REPTree (Decision Tree), Waikato Environment for Knowledge Analysis (WEKA) tool.

I. INTRODUCTION

Autism is neurodevelopment spectrum disorder and then characterized by lack of social interactions, verbal and non-verbal communication and limited behavior. The spectrum reflects the wide variation in challenges and strengths possessed by each person with ASD. ASD have one of the five subgroups are Aspenger's Syndrome, Autistic Disorder, PDD-NOS (Pervasive Developmental Disorder Not Otherwise Specified), Rett's Syndrome, Childhood Disintegrative Disorder. Above these five subgroups based on three different levels as Low-functioning autism, moderate autism and High-functioning autism. The Autism with Males are higher risk for neurodevelopment disorder than ASD with females. The female functions are lot better than males with similar mutation affecting brain development. The autism early stages symptoms are very common because difficult to identify these stage. The symptoms of autism typically appear during the first three years on the life and some children show signs from birth. The autism diagnosing is based on patient symptoms details in the dataset

The ASD patient dataset is downloaded from the VAERS website and its form collects information about the vaccine, the person vaccinated, and the adverse event. The VAERS is yearly 30,000 reports are filled and each report is assigned a VAERS_ID identification number. In this paper focuses on predicting the level of autism using data mining techniques like Naïve Bayes, Logistic Regression, IBk (K-Nearest neighbor), SMO (Sequential Minimal Optimization) and REPTree (Decision Tree) techniques to find the Accuracy, TP Rate (True Positive Rate), FP Rate (False Positive Rate) and F- Measure.

II. LITERATURE REVIEW

The research challenges encompass various work is done in the area of autism for prediction, classification, expressive facial dynamics in children with autism and behaviors of parents of children with autism, to provide awareness to the level of autism, etc. in this present works the data mining techniques being studied under the autism. Data mining is the extraction of useful information from the large volume of data. Data mining has been applied in various fields like medicine, marketing, banking, etc. In medicine, predictive data mining is used to diagnose the disease at the earlier stages itself and helps the physicians in treatment planning procedure. Michael Siller *et al.* proposed behaviors of children with autism show during play interactions with their children. We were particularly interested in the extent to which the caregiver's verbal and nonverbal behaviors were synchronized with the child's focus of attention as well as his or her ongoing activity [1]. S. Wheelwright *et al.* investigate the association between scores on the Empathy Quotient (EQ), Systemizing Quotient-Revised (SQ-R) and Autism Spectrum Quotient (AQ) in together a large sample of classic participants, and a sample of adults with Autism Spectrum Conditions (ASC) [2]. Laurence Chaby *et al.* providing an automatic, detailed and objective measure of multimodal socio-emotional behaviors, we thought that our method would become a valuable tool for examining language, emotional and social interactions in

clinical populations like autism spectrum disorders [3]. E. M. Albornoz *et al.* describe a soft computing in evolutionary method for automatic selection of features of speech in a classification task and based on a genetic algorithm that selects the best combination of phonic features using SVM as classifier [4]. Pawan Sinha *et al.* describe how theoretical considerations and a review of empirical data lead to the hypothesis that some salient aspects of the autism phenotype may be manifestations of an underlying impairment in predictive abilities [5]. Priyanka Juneja *et al.* proposed autistic patients are here analyzed by using interpretation value analysis and it is taken a parameter based fuzzification that will perform the analysis based on some parameters [6]. Ahmad Al-Khoder *et al.* did a comparison on various data mining tool and conclude that WEKA was the best tool to run the selected classifiers followed by R, Rapid Miner, and finally KNIME respectively [7]. Ionut Taranu describes Data mining has great importance for area of medicine, and it needs of the healthcare organizations. Data mining can be used to make successful decisions that will improve success of healthcare organization and health of the patients [8]. Arpit Arya that uses Naïve Bayes Classifier and Logistic Regression model. WEKA (Waikato Environment for Knowledge Analysis) tool is used to predicting autism for large scale dataset. WEKA is a very easy yet powerful tool for knowledge analysis, most researchers use WEKA for analysis related to data mining [9]. Tanaya Guha *et al.* describes decreasing complexity in facial expression dynamics of disscussion with High Functioning Autism relative to their typically developing peers. Major difference is analyzed for expressions related to joy, disgust and sadness [10]. Sumi Simon *et al.* did a comparison on data mining classification methods for ASD prediction. The work aimed at comparing the performance through accuracy of different data mining methods. SVM, J48, BVM and decision tree algorithms are the most commonly used classifiers are best suited to classify the autistic data and it provides high accuracy and low error rate [11]. M. S. Mythili *et al.* proposed advanced ASD predictive methods among children using fuzzy cognitive map and feature extraction techniques [12].

III. PROPOSED WORK

The implementation work is predicting three different levels of autism (Low, Middle and High) using data mining techniques in WEKA environment. The prediction of ASD has been divided into three different levels as Low-functioning autism, moderate autism and High-functioning autism. Figure 1 shown in the process flow architecture of proposed system.

A. Implementation using WEKA Tool:

WEKA is open source software. It is collection of machine learning algorithms used for data mining. The machine learning algorithm can be applied on the dataset directly. The WEKA contains tool and methods for Data Preprocessing, Predictive Modeling, Clustering, Attribute Selection, Association and Visualization.

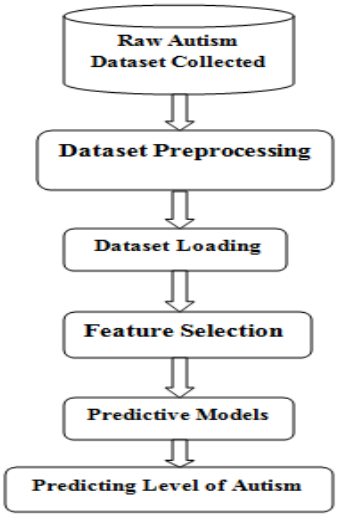

Figure 1: Process Flow Architecture

B. Autism Dataset Collection:

The ASD dataset is downloaded from *"https://vaers.hhs.gov/data/data"* in CSV file format. In this ASD dataset contains several fields. In table 1 is description for each attribute in ASD dataset. Every record had to be provided with class Label (Low, Middle, and High).

Table 1: ASD Dataset Explanation

No.	Attributes	Explanation
1	VAERS_ID	It is each data assigned a unique identification number in VAERS website.
2	Gender	It is defined autism patient gender.
3	Symptom 1	It is defined symptom stage 1
4	Symptom 2	It is defined symptom stage 2
5	Symptom 3	It is defined symptom stage 3
6	Symptom 4	It is defined symptom stage 4
7	Symptom 5	It is defined symptom stage 5
8	Level of Autism	It is label attribute and it is predicting three different levels as low, medium, high.

C. Data Preprocessing:

The Data Preprocessing is one of the most important steps of prediction because raw dataset is not ready to use directly. Data preprocessing technique is used to remove the noises of data in the ASD dataset. The Data Preprocessing, data will be clean, clear, noiseless data and redundant values

D. Dataset Loading:

The Autism dataset will be loaded on the WEKA environment. WEKA prefers to load dataset in the ARFF (Attribute Relation File Format) format. In this format supports numeric and categorical values but also supports dates and string values. Dataset loading important step is converted CSV (Comma Separated) format into ARFF format and then use WEKA environment.

E. Feature Selection:

The Feature Selection is a choosing only the relevant attribute in the dataset. Many feature selection techniques are supported in WEKA environment. Feature selection is divided into two parts are Attribute Evaluator and Search Method. The attribute evaluator is the technique by which each attribute in ASD dataset is evaluated in the context of the output variable. The search method is the technique by choosing only the relevant attribute.

F. Predictive Modeling:

The Prediction is a supervised learning task where the data is used to directly predicting the class Label in the ASD dataset. In this paper have been chooses the top of five Predictive Models in WEKA.

i. ***Naïve Bayes:*** In machine learning, naïve bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes theorem with strong independence assumptions between features. Naïve bayes are highly scalable, requiring a number of parameters linear in the number of variables in a learning problem.

ii. ***Logistic Regression:*** Logistic Regression is one of the most popular machine learning algorithms for binary classification. This is because it is simple algorithm that performs very well on wide range of problems. The logistic regression model takes real-valued inputs and makes a prediction as to the probability of the input belonging to the default class.

iii. ***IBk (K-Nearest Neighbor):*** The IBk algorithm does not build a model, instead it generates a prediction for a test instance just-in-time. The IBk algorithm uses a distance measure to locate k close instance in the training data for each test instance and uses those selected instances to make a prediction.

iv. ***SMO (Sequential Minimal Optimization):*** SMO is an algorithm for solving the quadratic programming problem that arises during the training of support vector machines. SMO is an iterative algorithm for solving the optimization problem.

v. ***REPTree:*** REPTree can support classification and regression problems. They work by creating a tree to evaluate an instance of data, start at the root of the tree and moving down to the leaves until a prediction can be made. The process of creating a decision tree works by greedily selecting the best split point in order to make predictions and repeating the process until the tree is a fixed depth.

IV. RESULT AND ANALYSIS

The predicting level of ASD using to different five data mining techniques like, Naïve Bayes, Logistic Regression, IBk, SMO, and REPTree. All of these classifiers are executed in WEKA environment and find the accuracy, Precision, TP Rate, FP Rate and F-Measure.

Table 2: Accuracies of Classifiers

No.	Classifiers	Accuracy (Correctly classified instances)
1	Naïve Bayesian	91.01%
2.	SMO	98.85%
3.	Logistic Regression	99.6%
4.	IBk	97.3%
5.	REPTree	82.3%

In Table 2 is illustrating Accuracy of correctly classified instances in five different data mining classifiers.

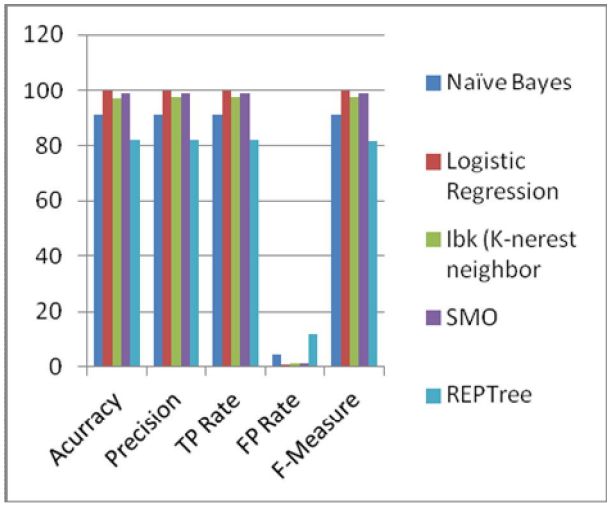

Figure 2: Result and Analysis for Different Data Mining Techniques

In Figure 2 shown in results of different five classifiers. Here X-Axis represents classifiers components and Y-Axis represents the classifiers. The SMO model is best suited to predict the ASD dataset and it is provides highest Accuracy, TP Rate, Precision, F-Measure and Lowest FP Rate.

V. CONCLUSION

In this paper analyzing the patient symptoms based on ASD dataset for diagnosing different level of autism is done. Five different data mining techniques like, Naïve Bayes, Logistic Regression, IBk, SMO, and REPTree techniques are used to find the Accuracy, TP Rate, Precision, FP Rate and F-Measure. The result shows data mining techniques in SMO techniques is providing highest Accuracy and lowest Error-Rate in ASD Dataset for predicting autism.

REFERENCES

[1] Michael Siller and Marian Sigman, “The Behaviors of Parents of Children with Autism Predict the Subsequent Development of Their Children’s Communication”, Journal of Autism and Developmental Disorders, Vol. 32, No. 2, April 2002.

[2] S. Wheelwright, S. Baron-Cohen, N. Goldenfeld, J. Delaney, D. Fine, R. Smith, L. Weil, A. Wakabayashi, “Predicting Autism Spectrum Quotient (AQ) from the Systemizing Quotient-Revised (SQ-R) and Empathy Quotient (EQ)”, 2006.

[3] Laurence Chaby, Mohamed Chetouani, Monique Plaza, David Cohen, “Exploring Multimodal Social-Emotional Behaviors in Autism Spectrum Disorders”, 2012 ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust.

[4] E. M. Albornoz, L. D. Vignolo, C. E. Martínez & D. H. Milone, "Genetic Wrapper Approach for Automatic Diagnosis of Speech Disorders related to Autism", 14th IEEE International Symposium on Computational Intelligence and Informatics (CINTI), nov, 2013.

[5] Pawan Sinha, Margaret M. Kjelgaard, Tapan K. Gandhi, Kleovoulos Tsourides, Annie L. Cardinaux, Dimitrios Pantazis, Sidney P. Diamond, and Richard M. Held, “Autism as a disorder of prediction”, 2014.

[6] Priyanka Juneja and Anshul Anand, " International Journal of Computer Science and Mobile Computing", Vol.3 Issue.7, July- 2014, pg. 585-593

[7] Ahmad Al-Khoder, Hazar Harmouch, “Evaluating four of the most popular Open Source and Free Data Mining Tools”, IJASR International Journal of Academic Scientific Research ISSN: 2272-6446 Volume 3, Issue 1 (February - March), PP 13-23.

[8] Lonut Taranu, “Data mining in healthcare: decision making and precision”, Database Systems Journal vol. VI, no. 4/2015.

[9] Arpit Arya, “Predicting Autism over Large-Scale Child Dataset”, 2015.

[10] Tanaya Guha, Member, IEEE, Zhaojun Yang, Student Member, IEEE, Ruth B. Grossman, Shrikanth S. Narayanan, Fellow, IEEE, “A Computational Study of Expressive Facial Dynamics in Children with Autism”, IEEE TRANSACTIONS ON AFFECTIVE COMPUTING, VOL. XX, NO. X, MARCH 2016.

[11] Sumi Simon, Chandra J and Saravanan N, "Empirical Evaluation of Data Mining Classification Methods for Autistic Children", International Journal of Trend in Research and Development (IJTRD) in 2016.

[12] M. S. Mythili and A. R. Mohamed Shanavas,” An Improved Autism Predictive Mechanism Among Children Using Fuzzy Cognitive Map And Feature Extraction Methods (Feast), VOL. 11, NO. 3, FEBRUARY 2016.

[13] “Vaccine Adverse Event Reporting System? (VAERS), https://vaers.hhs.gov/data/data.

[14] “WEKA tool”, http://cs/waikato.ac.nz/