ISSN 2347 - 3983
Volume 7, No. 11 November 2019
International Journal of Emerging Trends in Engineering Research
Available Online at http://www.warse.org/IJETER/static/pdf/file/ijeter407112019.pdf
https://doi.org/10.30534/ijeter/2019/407112019

Autism Detection and Subgrouping using Machine Learning Algorithms

Swarna Kuchibhotla[1], Neha Reddy Bonthu[2], Bhardwaj Sri Kaushal Kavuri [3] , Pavan Kumar Datla[4]
[1]Associate Professor, Department of Computer Science and Engineering,Koneru Lakshmaiah Education Foundation,Vaddeswaram,drkswarna@kluniversity.in, India.
[2]Student ,Department of Computer Science and Engineering,Koneru Lakshmaiah Education Foundation,Vaddeswaram,nehareddybonthu725@gmail.com, India.
[3]Student ,Department of Computer Science and Engineering,Koneru Lakshmaiah Education Foundation,Vaddeswaram,kaushal9989@gmail.com, India.
[4]Student ,Department of Computer Science and Engineering,Koneru Lakshmaiah Education Foundation,Vaddeswaram,pavan.datla.1099@gmail.com, India

ABSTRACT

This paper is about the detection of the disorder and to determine whether they belong to the Autism Spectrum Disorder or not. The detection process was performed using the datasets collected from resources and for the sub-grouping of the disorders, distinguishable symptoms of both behavioral and physical are collected. The results obtained determine disorder and conclude whether it comes belongs to the spectrum of ASD or not and the available treatments.

Key words : Autism Spectrum Disorder (ASD), DSM-5 (The Diagnostic and Statistical Manual of Mental Disorders), NSCH (National survey of child health).

1. INTRODUCTION

Autism is a neurological and developmental disability that affects social interactions, communication, and behavior. It generally begins at the early stages of childhood and can last for a lifetime. Various screening processes are conducted by medical experts based on observable factors. Yet, existing methods involve certain factors like a large number of questions to get an accurate classification, a lot of manual effort and time. Through referring to the existing database, we can provide an alternative diagnosis with the aid of machine learning techniques.

Machine learning is a research-based area Concerned with automatic (Spatio-temporal) identification of data patterns. In the study of "big data," it has been primarily used to analyze data syntactically and semantically.

Essentially, this means an automated search for the best template, provided a specific task and information. A large number of algorithms are developed for various tasks in which the methodology was borrowed from bio-inspired artificial intelligence investigations.

This study further involves the process of sub grouping as,

Earlier the symptoms of autism always led the psychiatrists to a diagnosis of schizophrenia. Donald Triplett was the first person to be diagnosed with autism in 1933 under Dr. Kannner . Until 1943, even though the diagnosis was happening, autism was not a diagnostic term. The importance of sub grouping of disorders has started since then. Yet currently, there a multitude of mental health disorders were classified based on several characteristics such as emotional, social and physical behaviors, that are placed under a single term known as Autism Spectrum disorder.

1.1 Sub-grouping

DSM is the diagnostic manual used by clinicians which involves the classification of several disorders and to diagnose them. The current version available is DSM-5 that was used along with the web resources to perform the process of sub grouping. The specifiers used for the process of classification of disorders are developmental disorders, gender, cognitive profile, history of health, Genetic correlation, Potential environment contributors.

The proposed study aims at building a mobile screening tool that may be used by the stakeholders regarding the type of

disorder, Whether the disorder is in the spectrum of autism and treatments that are currently preferred by the medical experts for the disorders.

2. LITERATURE SURVEY

Several studies were done earlier for the screening process on data using machine learning classifiers. Yet, most of them used different classifiers. For the Machine learning techniques, [8,10,11] are referred to obtain the suitable method of classification technique based on the nature of datasets obtained. In [1] and [4] same datasets were used generated from a mobile screening app where the former used ten-fold cross-validation using if-then rules and obtained an accuracy around 90% and the latter performed KNN (K-Nearest Neighbor) and LDA(Linear Discriminant analysis) and best of them turned out to be LDA with an accuracy of 90%.

ADI-R (Autism Diagnostic Interview-Revised) dataset is used in both [6] and [5], wherein [5] additionally SRS (Social Responsiveness Scale) dataset is used. In [6] fifteen different classifiers are used out of which ADT classifier bestowed with 99% accuracy. In [2], the ADDM network survey dataset is used and displayed an accuracy of 86.5% using a random forest classifier. Though [1,5,8] are based on the screening process using classifiers, the significance of DSM-5 was mentioned in these studies.

The role of DSM in the decision making of clinicians and medical experts is cited in [3,7]. In [7], DSM-IV and DSM-V are compared and the stringent nature, advancements in DSM-V that meet the criteria of ASD diagnosis are incorporated. In [3], the ideas of sub grouping of ASD and the significance of it during diagnosis are acknowledged. The basis of sub grouping in [3] is from clinical samples of [7,9] and some existing datasets. The specifiers in DSM-5 are mentioned which aid the phenotypic characterization.

3. METHODOLOGY

In this study , data is collected from three sources and it is in raw format. Preprocessing is performed and the missing values in data were and replaced with their arithmetic mean values to convert the data into an efficient format. The replacement of the values is to ensure a smooth process of classification of data.

Each dataset is converted into two formats, one including just questionnaire and other the whole dataset. Classifiers like Random Forest, Naive Bayes, Neural networks and Support Vector Machine are applied to the given data. There was a slight variation inaccuracy in both formats of datasets but the best among the classifiers was Neural Networks.

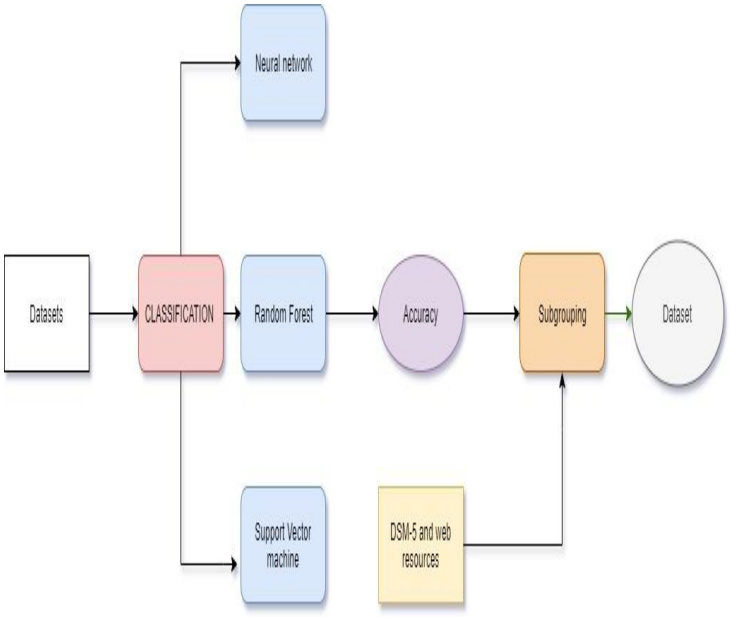

Figure 1: Illustration of classifying the dataset using effective algorithms and sub grouping

The next phase of this study is to perform sub grouping of disorders in ASD, using datasets collected. These datasets were under sampled and the information of patients with ASD is collected. Based on this information collected and with the help of DSM-5 sub grouping is implemented. DSM (Diagnostic and statistical manual of mental disorders) Is a mental disorder category consistent with certain parameters. It is recognized in the area of mental health as a generic guide for clinical practice.

It is intended to serve as a practical, functional and flexible guide to organize information that can help to treat disorders accurately.

Certain specifiers are required to subgroup these disorders. Specifiers like

1. **Developmental pattern :**

To determine the development of an individual as it can mask some symptoms of the disorder.

2. Gender:

As many researchers claim that boys or men are most affected by mental disorders compared to women or girls, but some disorders like Rett's syndrome, Tardive dyskinesia, etc girls are the most affected gender.

3. Cognitive profile:

Cognitive profile is defined as the individual skills required for communication, self-care, etc. The uneven cognitive skills often noticed in children are known as, splinter skills which are used to categorize the disorders.

4. Clinical Phenotypes:

Physical disorders or diseases that are observable without any implementation of a mechanism. Jaundice, seizures, etc are the most common clinical phenotypes considered during the screening process.

5. Genetic Correlates:

The cause of most of the disorders is due to the Genetic mutation, a lasting change in the gene-forming sequence of DNA. It can be either hereditary mutation that is inherited from bloodline And are there throughout the life of a person or through acute mutations that occur at some point in person life not in every cell in the body. Thus, the family medical history is often asked by the clinician or medical experts during diagnosis.

6. Environment Contributors :

Environmental factors like physical or mental abuse, trauma, poor relationships, dysfunctional home life, etc are often considered to be the reason as he long term stress can lead to mental or physical illness. So, the participants of most of the surveys conducted by child health care are parents, guardians, caretakers or teachers with whom children spend most of their time.

4. FEATURE EXTRACTION

Feature extraction in machine learning is used for the dimensionality reduction of the data which is suitable for building the architecture of the system. The transformed attributes are a linear combination of original attributes. The attributes in the datasets are questionnaires that are reduced into these features.

4.1. Social Interaction

Social interaction of children is to be known as it is the first indication of autism. It is to analyze how comfortable are children communicating with an individual or with a block of society.

4.2. Focus :

Children may deviate their attention for a bit of time which is normal. The focus attribute is to determine the amount of information obtained from both vocal and non-vocal formats.

4.3. Gender :

Gender attribute is mentioned in the dataset as researchers have consistently found that more boys/ men than women/girls are prone to have autism.

4.4. Multi-Tasking :

Performing multiple tasks at a single -go requires certain cognitive ability and both physical and mental coordination. It is to determine whether comfort levels of children while multitasking.

4.5. Age :

Age attribute is used just to classify the data in which age group is most likely to have autism as previous researchers mostly concluded based on this feature.

4.6. Medical History :

Medical history is essential to analyze both physical and mental disorders. As, diseases such as Jaundice , seizures, etc are the most prevalent ones in autistic children.

5. CLASSIFICATION ALGORITHMS

Classification is supervised learning which is a method of determining the data point category. There are currently seven types of classification algorithms in machine learning. Based on the application and nature of dataset classification algorithms are applied. The algorithms used in this study are,

5.1 Random Forest:

In the random forest approach, many decision trees are created. Every insight is fed into each decision tree. The most popular outcome is used as the final product for each test. For each classification method, a new observation is fed into all the trees and a majority vote is taken. Random Forest implements the random algorithm for forest classification and

regression from Breiman. It can be used to determine proximities between data points in an unsupervised mode.

5.2 Naive Bayes:

Naive Bayes is a probabilistic classifier inspired by the Bayes theorem, which classifies each pair of features as independent and assumes that all values are equal. Bayes theorem compares two conditional probabilities in the simplest form for events A and B as follows:

$$P(A \cap B) = P(A, B) = P(A)P(B|A) = P(B)P(A|B)$$

$$\Longrightarrow P(B|A) = P(B)P(A|B)P(A)$$

In R programming, it Calculates the conditional a-posterior probabilities of a variable of classification class given independent predictor variables using the Bayes rule.

5.3 Neural Networks:

Neural Network (or Artificial Neural Network) can learn by examples. ANN is a model for processing information inspired by the biological neuron network. This consists of a large number of highly interconnected processing components called the neuron to solve problems. This takes the non-linear path and processes information along the nodes in parallel. A neural network is a complex system of adaptation. Adaptive means that by changing input weights it can alter its internal structure.

5.4 Support Vector Machine:

Support Vector machine classifies using a separating hyper plane for the given set of data points and distinctly classifies them. The points are mapped in space such that there is a clear gap between sets and it is as wide as possible. In SVM, the hyper plane which maximizes the boundaries or margin is considered to be the best.

6. EXPERIMENTAL RESULTS

Four different classifiers runs have been performed on three datasets, in which two datasets from UCI and Kaggle were split and classification is performed on both the partitioned dataset and the complete dataset. The partitioned dataset contains only the attributes of the questionnaire, only a slight variation was noticed in the results obtained from these datasets.

For the NSCH dataset, it consists of around 22000 samples in which the people with and without ASD are in the ratio of 1:900. The dataset was to be under sampled and preprocessed. The results obtained were almost similar compared with the others.

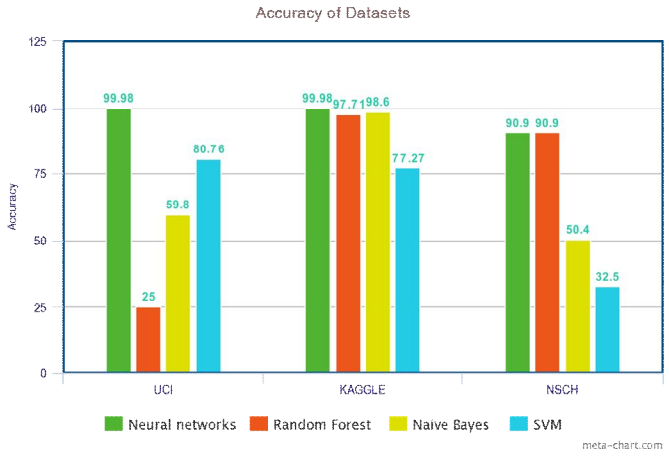

Figure 2: Illustration of classifying the dataset using effective algorithms and sub grouping.

In Figure 2, the accuracy values of datasets using four different classifiers are represented in which neural networks outperformed the remaining classifiers.

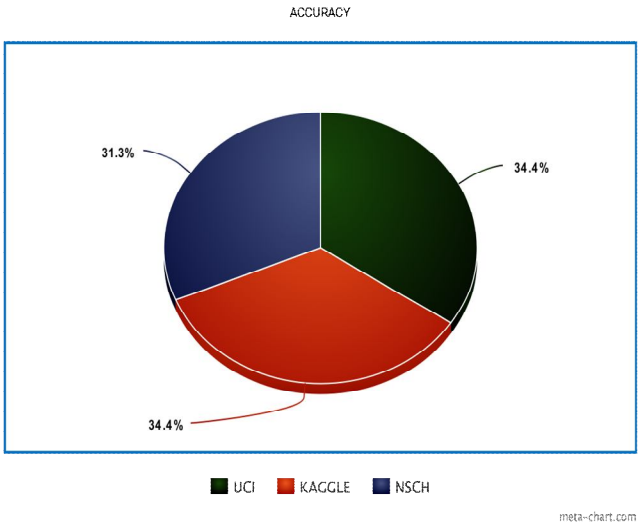

Figure 3: Highest accuracy obtained in three datasets .

From the subgrouping dataset generated, 90% of the disorders didn't make it into the spectrum. In most of the disorders, boys or men are the most affected gender to these disorders.

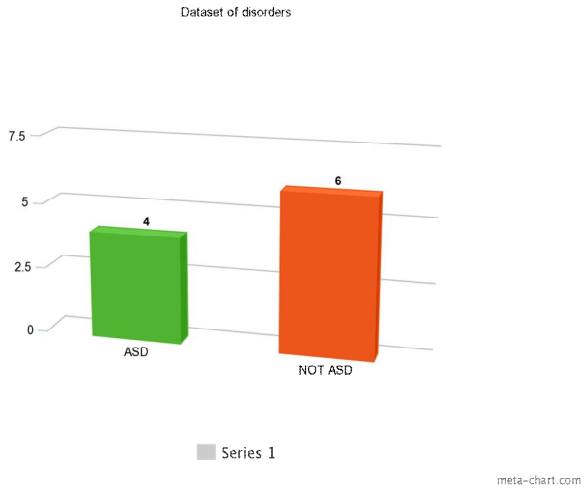

Figure 4: Results of Dataset created.

7. CONCLUSION AND FUTURE SCOPE

Cleaning the data set (which recorded the characteristics associated with ASD) was difficult in that we had mainly categorical variables and only two numerical variables, but eventually, we were able to build these models and found that all machine learning algorithms are performed by the algorithm which performs best in all aspects and neural networks.

Although the information association made the prediction very easy, we believe that this research will serve as useful aid for doctors to identify new autistic cases.

In our consideration, we need to have larger datasets to build a precise and robust model. Here, after cleaning the data, the number of instances was not sufficient to say this model is optimal. With this current data collection, nothing can be changed by looking at the quality of our learning systems, as models are already at their peak. After discussing this issue with a researcher working directly on child autism, we realized that collecting a lot of well-documented ASD-related data is extremely difficult. Recently, this ASD dataset has been made public (available since December 2017), and so little work has been done. With this in mind, our research has led to well-developed models that can be accurately detect ASD in people with unique behavioral and health knowledge attributes. Such models may serve as benchmarks for any machine learning researcher / practitioner interested in further exploring this dataset or other Autism screening disorder related data sets.

REFERENCES

1. F.Abdeljaber , **Detecting Autistic Traits using Computational Intelligence & Machine Learning Techniques** , 2018 .

2. Matthew J.Maenner, Kim Van, Kim Van *.et.al* , **Development of a Machine Learning algorithm for the Surveliance of Autism Spectrum Disorder** 2016. https://doi.org/10.1371/journal.pone.0168224

3. Meng-Chuan Lai, Bhismadev Chakrabati *.et.al* , **Subgrouping the Autism "Spectrum" : Reflections on DSM-5** 2013

4. Osman Altay, Mustafa Ulas , **Prediction of the Autism Spectrum Disorder Diagnosis with Linear Discriminant analysis and K-Nearest Neighbor in Children** 2018. https://doi.org/10.1109/ISDFS.2018.8355354

5. Matthew P.Black, Catherine Lord *.et.al* , **Use of machine learning to improve autism screening and diagnostic instruments : effectiveness, efficiency and multi instrument fusion** 2016.

6. Rhiannon Luyster , Jae-Yoon Lung *.et.al* ,**Use of artificial intelligence to shorten the behavioral diagnosis of autism** 2012.

7. Vicki Gibbs, Felicity Chandler *.et.al* , **An exploratory study comparing diagnostic outcomes for autism spectrum disorders under DSM-IV-TR with the proposed DSM-5 revision** 2012. https://doi.org/10.1007/s10803-012-1560-6

8. Narasinga Rao, M.R., Sajana, T., Bhavana, N., Sai Ram, M. & Nikhil Krishna, C. 2018**, "Prediction of chronic kidney disease using machine learning technique",** Journal of Advanced Research in Dynamical and Control Systems, vol. 10, pp. 328-332.

9 .Swarna Kuchibhotla, HD Vankayalapati and KR Anne, **"An optimal two stage feature selection for speech emotion recognition using acoustic features,"** International Journal of Speech Technology, pages 1-11, 2016. https://doi.org/10.1007/s10772-016-9358-0

10. R.B.Saroo Raj, Ankush Rai *.et.al* **Weather forecasting System using Machine Learning ,** International Journal of Emerging Technologies in Engineering Research , pages 1-4,2018.

11. Aruna A, Mayukh Dasgupta *.et.al* **Efficient Distribution of Water using Machine Learning Techniques** , International Journal of Emerging Technologies in Engineering Research , pages 1-3, 2018.