ISSN 2515-8260 Volume 07, Issue 09, 2020

AUTISM SPECTRUM DISORDER USING KNN ALGORITHM

Mrs. Surya .S.R [1]
Asst. Professor Dept. of Computer Applications College of Science and Humanities SRM Institute of Science and Technology,Chennai.
DR. G. Kalpana [2]
Associate Professor & Head Department of Computer Science College of Science and Humanities, SRM Institute of Science and Technology, Chennai.

Abstract:

Autism spectrum disorder (ASD) is a psychiatric disorder which leads to neurological anddevelopmental growth of a person which starts in early age and gets carried throughout their life.It is a condition associated with significant healthcare costs and early diagnosis can reduce these.Unfortunately, waiting time is lengthy for an ASD diagnosis and it is cost effective. Due to theincrease in economy for autism prediction and the increase in the number of ASD cases across theworld is in need of easily implemented and effective screening methods by GUI results. Toovercome the time complexity for identifying the disorder advanced technologies can be used suchas machine learning algorithms to improve precision, accuracy and quality of the diagnosisprocess. Machine learning helps us by providing intelligent techniques to discover the affectedpatient, which can be utilized in prediction and to improve decision making. And hence, wepropose the data set features related to autism screening of adult and child to be used for furtheranalysis and to improve the classification of ASD cases.

Keywords: ASD, GUI, Machine learning, precision, accuracy, KNN.

Introduction:

As in the recent times this autistic spectrum disorder is increasing progressively and waiting time to diagnosis these types of disorders is lengthy and procedures are not cost effective. The reported prevalence rates of autism and its related disorders have been increasing worldwide over the past decades, from approximately 4 per 10,000 to 6 per 1,000 children[1]. The reasons for this increase include wider public awareness of these disorders, broadening of the diagnostic concepts, reclassifications of disorders and improved detection. The possibility that the increase in the reported cases is a result of unidentified risk factor(s) cannot be ruled out, and therefore more research is needed to address this. Family studies have demonstrated that autism is both familial and heritable. The recurrence rate in siblings of an autistic child is 2% to 8%, which is higher than that of the general population. In order to reduce the time complexity for identifying the disorder and diagnosing it, we propose to use any of the machine learning algorithms to improve precision, accuracy and quality of the diagnosis process and which may help to reduce the time complexity[2]. In most recent times, the utility of machine picking up information of to move disciplinary points have been

exceptionally dynamic and fruitful, particularly in areas of science and neurology. A learned records representation can help visualize actualities to assist people in logical choice making and anticipate a target variable from set of enter highlights. In this classification algorithms which are available in machine learning have been used and which will help us to predict whether any person is having this type of disorder or not and improves the accuracy of prediction[3].

Autism spectrum disorder is a condition related to brain development that impacts how a person perceives and socializes with others, causing problems in social interaction and communication. The disorder also includes limited and repetitive patterns of behaviour[4]. The term "spectrum" in autism spectrum disorder refers to the wide range of symptoms and severity. Autism spectrum disorder includes conditions that were previously considered separate autism, Asperger's syndrome, childhood disintegrative disorder and an unspecified form of pervasive developmental disorder[5]. Some people still use the term "Asperger's syndrome," which is generally thought to be at the mild end of autism spectrum disorder. Autism spectrum disorder begins in early childhood and eventually causes problems functioning in society socially, in school and some children show signs of autism spectrum disorder in early infancy, such as reduced eye contact, lack of response to their name or indifference to caregivers[6]. Other children may develop normally for the first few months or years of life, but then suddenly become withdrawn or aggressive or lose language skills they've already acquired. Signs usually are seen by age 2 years. Each child with autism spectrum 8 disorder is likely to have a unique pattern of behaviour and level of severity from low functioning to high functioning[7].

A child or adult with autism spectrum disorder may have problems with social interaction and communication skills, including any of these signs[8]:

• Fails to respond to his or her name or appears not to hear you at times

• Resists cuddling and holding, and seems to prefer playing alone, retreating into his or her own world • Has poor eye contact and lacks facial expression

• Doesn't speak or has delayed speech, or loses previous ability to say words or sentences

• Can't start a conversation or keep one going, or only starts one to make requests or label items

• Speaks with an abnormal tone or rhythm and may use a singsong voice or robot-like speech

Patterns of behaviour:

A child or adult with autism spectrum disorder may have limited, repetitive patterns of behaviour, interests or activities, including any of these signs[9]:

• Performs repetitive movements, such as rocking, spinning or hand flapping

• Performs activities that could cause self-harm, such as biting or head-banging and develops specific routines or rituals and becomes disturbed at the slightest change 9

• Has problems with coordination or has odd movement patterns, such as clumsiness or walking on toes, and has odd, stiff or exaggerated body language

• Is fascinated by details of an object, such as the spinning wheels of a toy car, but doesn't understand the overall purpose or function of the object

• Is unusually sensitive to light, sound or touch, yet may be indifferent to pain or temperature

Machine Learning:

Machine learning algorithms are often categorized as supervised or unsupervised. Supervised machine learning algorithms can apply what has been learned in the past to new data using labelled examples to predict future events. Starting from the analysis of a known training dataset, the learning algorithm produces an inferred function to make predictions about the output values. The system can provide targets for any new input after enough training. The learning algorithm can also compare its output with the correct, intended output and find errors in order to modify the model accordingly. In contrast, unsupervised machine learning algorithms are used when the information used to train is neither classified nor labelled. Unsupervised learning studies how systems can infer a function to describe a hidden structure from unlabelled data. The system doesn't figure out the right output, but it explores the data and can draw inferences from datasets to describe hidden structures from unlabelled data. Semi supervised machine learning algorithms fall somewhere in between supervised and unsupervised learning, since they use both labelled and unlabelled data for training – typically a small amount of labelled data and a large amount of unlabelled data. The systems that use this method can considerably improve learning accuracy. Usually, semi-supervised learning is chosen when the acquired labelled data requires skilled and relevant resources in order to train it / learn from it. Otherwise, acquiring unlabelled data generally doesn't require additional resources. Reinforcement machine learning algorithms is a learning method that interacts with its environment by producing actions and discovers errors or rewards. Trial and error search and delayed reward are the most relevant characteristics of reinforcement learning. This method allows machines and software agents to automatically determine the ideal behaviour within a specific context in order to maximize its performance. Simple reward feedback is required for the agent to learn which action is best; this is known as the reinforcement signal[10].

The aims that to discovers and extracts hidden knowledge associated with diseases from a historical heart data set and ASD prediction system aims to exploit machine learning techniques on medical data set to assist. Then they Provides new approach to concealed patterns in the data and helps avoid human biasness and reduce the cost of medical tests[11].

Proposed System:

1. Splitting the dataset

The data use is usually split into training data and test data. The training set contains a known output and the model learns on this data in order to be generalized to other data later. It has the test dataset in order to test our model's prediction on this subset and it will do this using Sickest Learn library in Python using the train_test_split method. It will load in the data, check for cleanliness, and then trim and clean given dataset for analysis. The data set collected for predicting patient is split into Training set and Test set. Generally, 7:3 ratios are applied to split the Training set and Test set. The Data Model which was created using machine learning are applied on the Training set and based on the test result accuracy, Test set prediction is done. The data which was collected might contain missing values that may lead to inconsistency. To gain better results data need to be pre-processed so as to improve the efficiency of the algorithm. The outliers have to be removed and also variable conversion need to be done. Based on the correlation among attributes it was observed that attributes that are significant individually[12]. Advantages

➢ The goal of this problem is to predict the status of detecting the patient having ASD or not by prediction accuracy results of test dataset.

➢ To reduce doctor risk in healthcare.

➢ Therefore, a time-efficient and accessible ASD screening is imminent to help health professionals and inform individuals whether they should pursue formal clinical diagnosis.

➢ The rapid growth in the number of ASD cases worldwide necessitates datasets related to behaviour traits. However, such datasets are rare making it difficult to perform thorough analyses to improve the efficiency, sensitivity, specificity and predictive accuracy of the ASD screening process.

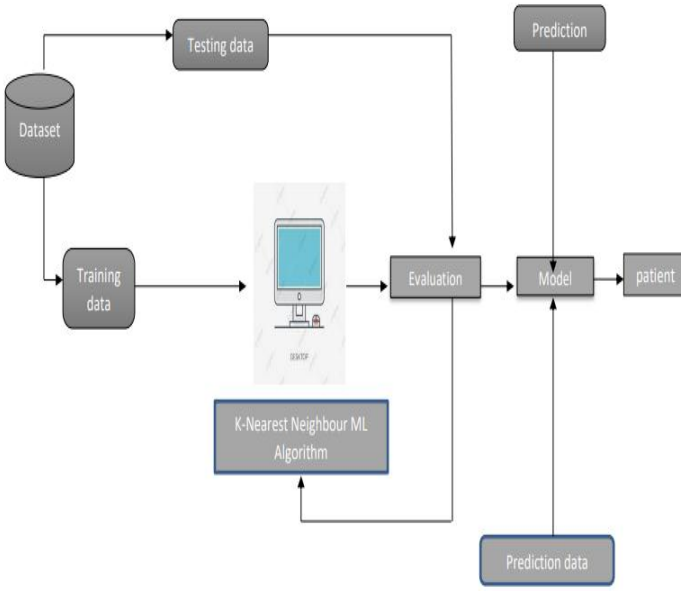

Fig: Architecture of the System

2. Modules

➢ Data validation process and pre-processing technique

➢ Comparison of machine learning for accuracy calculation

➢ GUI based prediction of ASD having or not

3. Data Validation and Pre-Processing Techniques

Validation techniques in machine learning are used to get the error rate of the Machine Learning (ML) model[13], which can be considered as close to the true error rate of the dataset. If the data volume is large enough to be representative of the population, you may not need the validation techniques. However, in real-world scenarios, to work with samples of data that may not be a true representative of the population of given dataset. To finding the missing value, duplicate value and description of data type whether it is float variable or integer. The sample of data used to provide an unbiased evaluation of a model fit on the training dataset while tuning model hyper parameters. The evaluation becomes more biased as skill on the validation dataset is incorporated into the model configuration. The validation set is used to evaluate a given model, but this is for frequent evaluation[14]. It as machine learning engineers uses this data to fine-tune the model hyper parameters. Data collection, data analysis, and the process of addressing data content, quality, and structure can add up to

a time-consuming to-do list. During the process of data identification, it helps to understand your data and its properties; this knowledge will help you choose which algorithm to use to build your model. For example, time series data can be analyzed by regression algorithms; classification algorithms can be used to analyze discrete data. (For example to show the data type format of given dataset) Data Validation/ Cleaning/Preparing Process: Importing the library packages with loading given dataset. To analyzing the variable identification by data shape, data type and evaluating the missing values, duplicate values. A validation dataset is a sample of data held back from training your model that is used to give 48 an estimate of model skill while tuning model's and procedures that you can use to make the best use of validation and test datasets when evaluating your models. Data cleaning / preparing by rename the given dataset and drop the column etc. to analyze the uni-variate, bi-vitiate and multivariate process. The steps and techniques for data cleaning will vary from dataset to dataset. The primary goal of data cleaning is to detect and remove errors and anomalies to increase the value of data in analytics and decision making

	AGE	GENDER	ETHNICITY	BORN WITH JAUNDICE	FAMILY MEMBER WITH PDD	WHO IS COMPLETING THE TEST	COUNTRY OF RESIDENCE	USED THE SCREENING APP BEFORE	SCRENING METHOD TYPE	I often notice small sounds when others dont	I find it easy to do more than one thing once	I know how to tell if someone listening to me is getting bored	I find it difficult to work out people's intensions	I find it easy to read between lines when someone is talking to me
0	31	f	INDIAN	NO	NO	MYSELF	INDIA	NO	0	disagree	agree	agree	agree	disagree
1	6	f	EUROPEAN	YES	NO	MOTHER	EUROPE	YES	1	nil	nil	nil	nil	nil
2	2	m	AMERICAN	YES	YES	FATHER	AMERICA	NO	0	nil	nil	nil	nil	nil
3	5	f	INDIAN	NO	NO	FATHER	INDIA	NO	0	nil	nil	nil	nil	nil
4	22	m	INDIAN	NO	NO	MYSELF	INDIA	NO	0	disagree	agree	agree	disagree	disagree

Fig: Data Frame for ASD

Data Pre-processing Pre-processing refers to the transformations applied to our data before feeding it to the algorithm.Data Pre-processing is a technique that is used to convert the raw data into a clean data set. In other words, whenever the data is gathered from different sources it is collected in raw format which is not feasible for the analysis. To achieving better results from the applied model in Machine Learning method of the data has to be in a proper manner. Some specified Machine Learning model needs information in a specified format; for example, Random Forest algorithm does not support null values. Therefore, to execute random forest algorithm null values have to be managed from the original raw data set[5].

	AGE	GENDER	ETHNICITY	BORN WITH JAUNDICE	FAMILY MEMBER WITH PDD	WHO IS COMPLETING THE TEST	COUNTRY OF RESIDENCE	USED THE SCREENING APP BEFORE	SCRENING METHOD TYPE	I often notice small sounds when others dont	...	I know how to tell if someone listening to me is getting bored	I find it difficult to work out people's intensions	I find it easy to read between lines when someone is talking to me	Play alone
0	26	0	5	0	0	2	5	0	0	1	...	0	0	1	2
1	4	0	4	1	0	1	4	1	1	2	...	2	3	2	0
2	0	1	1	1	1	0	1	0	0	2	...	2	3	2	0
3	3	0	5	0	0	0	5	0	0	2	...	2	3	2	1
4	18	1	5	0	0	2	5	0	0	1	...	0	2	1	2

Fig: Pre-Processed Data frame.

4. Comparison of Machine Learning

Random Forest Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of

decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of over fitting to their training set. Random forest is a type of supervised machine learning algorithm based on ensemble learning. Ensemble learning is a type of learning where you join different types of algorithms or same algorithm multiple times to form a more powerful prediction model. The random forest algorithm combines multiple algorithm of the same type i.e. multiple decision trees, resulting in a forest of trees, hence the name Random Forest. The random forest algorithm can be used for both regression and classification tasks.

The following are the basic steps involved in performing the random forest algorithm:

➢ Pick N random records from the dataset.

➢ Build a decision tree based on these N records.

➢ Choose the number of trees you want in your algorithm and repeat steps 1 and 2.

➢ In case of a regression problem, for a new record, each tree in the forest predicts a value for Y (output).

The final value can be calculated by taking the average of all the values predicted by all the trees in forest. Or, in case of a classification problem, each tree in the forest predicts the category to which the new record belongs. Finally, the new record is assigned to the category that wins the majority vote. Logistic Regression It is a statistical method for analysing a data set in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes). The goal of logistic regression is to find the best fitting model to describe the relationship between the dichotomous characteristic of interest (dependent variable = response or outcome variable) and a set of independent (predictor or explanatory) variables. Logistic regression is a Machine Learning classification algorithm that is 50 used to predict the probability of a categorical dependent variable. In logistic regression, the dependent variable is a binary variable that contains data coded as 1 (yes, success, etc.) or 0 (no, failure, etc.). In other words, the logistic regression model predicts P(Y=1) as a function of X.

Logistic regression Assumptions:

➢ Binary logistic regression requires the dependent variable to be binary.

➢ For a binary regression, the factor level 1 of the dependent variable should represent the desired outcome.

➢Only the meaningful variables should be included.

➢The independent variables should be independent of each other. That is, the model should have little.

➢The independent variables are linearly related to the log odds. Logistic regression requires quite large sample sizes.

Decision Tree

It is one of the most powerful and popular algorithm. Decision-tree algorithm falls under the category of supervised learning algorithms. It works for both continuous as well as categorical output variables.

Assumptions of Decision tree:

ISSN 2515-8260 Volume 07, Issue 09, 2020

➢At the beginning, we consider the whole training set as the root.

➢ Attributes are assumed to be categorical for information gain, attributes are assumed to be continuous.

➢ On the basis of attribute values records are distributed recursively.

➢We use statistical methods for ordering attributes as root or internal node. Decision tree builds classification or regression models in the form of a tree structure. It breaks down a data set into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. A decision node has two or more branches and a leaf node 51 represents a classification or decision. The topmost decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data. Decision tree builds classification or regression models in the form of a tree structure. It utilizes an if-then rule set which is mutually exclusive and exhaustive for classification. The rules are learned sequentially using the training data one at a time. Each time a rule is learned, the tuples covered by the rules are removed. This process is continued on the training set until meeting a termination condition. It is constructed in a top-down recursive divide-and-conquer manner. All the attributes should be categorical. Otherwise, they should be discretized in advance. Attributes in the top of the tree have more impact towards in the classification and they are identified using the information gain concept.

A decision tree can be easily over-fitted generating too many branches and may reflect anomalies due to noise or outliers.

K-Nearest Neighbour (KNN)

K-Nearest Neighbour is a supervised machine learning algorithm which stores all instances correspond to training data points in n-dimensional space. When an unknown discrete data is received, it analyzes the closest k number of instances saved (nearest neighbors) and returns the most common class as the prediction and for real-valued data it returns the mean of k nearest neighbors. In the distance-weighted nearest neighbor algorithm, it weights the contribution of each of the k neighbors according to their distance using the following query giving greater weight to the closest neighbors. Usually KNN is robust to noisy data since it is averaging the k-nearest neighbors. The k-nearest 1634eighbours algorithm is a classification algorithm, and it is supervised: it takes a bunch of 1634eighbou points and uses them to learn how to label other points. To label a new point, it looks at the 1634eighbou points closest to that new point (those are its nearest 1634eighbours), and has those 1634eighbours vote, so whichever label the most of the 1634eighbours have is the label for the new point (the "k" is the number of 1634eighbours it checks). Makes predictions about the validation set using the entire training set. KNN makes a prediction about a new instance by searching through the entire set to find the k "closest" instances. "Closeness" is determined using a proximity measurement (Euclidean) across all features.

GUI BASED PREDICTION Tkinter is a python library for developing GUI (Graphical User Interfaces). We use the tkinter library for creating an application of UI (User Interface), to create windows and all other graphical user interface and Tkinter will come with Python as a standard package, it can be used for security purpose of each users. Sensitivity False Positives (FP): A person who will pay predicted as defaulter. When actual class is no and predicted class is yes. E.g. if actual class says this passenger did not survive but predicted class tells

you that this passenger will survive. False Negatives (FN): A person who default predicted as payer. When actual class is yes but predicted class in no. E.g. if actual class value indicates that this passenger survived and predicted class tells you that passenger will die. True Positives (TP): A person who will not pay predicted as defaulter. These are the correctly predicted positive values which means that the value of actual class is yes and the value of predicted class is also yes. E.g. if actual class value indicates that this passenger survived and predicted class tells you the same thing. True Negatives (TN): A person who default predicted as payer. These are the correctly predicted negative values which means that the value of actual class is no and value of predicted class is also no. E.g. if actual class says this passenger did not survive and predicted class tells you the same thing. In the next section you will discover exactly how you can do that in Python with scikitlearn. The key to a fair comparison of machine learning algorithms is ensuring that each algorithm is evaluated in the same way on the same data and it can achieve this by forcing each algorithm to be evaluated on a consistent test harness.

• Now, the dimensions of new features in a numpy array called 'n' and it want to predict the species of this features and to do using the predict method which takes this array as input and spits out predicted target value as output. 53

• So, the predicted target value comes out to be 0. Finally to find the test score which is the ratio of no. Of predictions found correct and total predictions made and finding accuracy score method which basically compares the actual values of the test set with the predicted values.

Sensitivity is a measure of the proportion of actual positive cases that got predicted as positive (or true positive). Sensitivity is also termed as Recall. This implies that there will be another proportion of actual positive cases, which would get predicted incorrectly as negative (and, thus, could also be termed as the false negative). This can also be represented in the form of a false negative rate. The sum of sensitivity and false negative rate would be 1. Let's try and understand this with the model used for predicting whether a person is suffering from the disease. Sensitivity is a measure of the proportion of people suffering from the disease who got predicted correctly as the ones suffering from the disease. In other words, the person who is unhealthy actually got predicted as unhealthy. Mathematically, sensitivity can be calculated as the following: Sensitivity = (True Positive) / (True Positive + False Negative) The following is the details in relation to True Positive and False Negative used in the above equation. • True Positive = Persons predicted as suffering from the disease (or unhealthy) are actually suffering from the disease (unhealthy); In other words, the true positive represents the number of persons who are unhealthy and are predicted as unhealthy. • False Negative = Persons who are actually suffering from the disease (or unhealthy) are actually predicted to be not suffering from the disease (healthy). In other words, the false negative represents the number of persons who are unhealthy and got predicted as healthy. Ideally, we would seek the model to have low false negatives as it might prove to be lifethreatening or business threatening. 54 The higher value of sensitivity would mean higher value of true positive and lower value of false negative. The lower value of sensitivity would mean lower value of true positive and higher value of false negative. For healthcare and financial domain, models with high sensitivity will be desired. Specificity Specificity is defined as the proportion of actual negatives, which got predicted as the negative (or true negative). This implies that there will

be another proportion of actual negative, which got predicted as positive and could be termed as false positives. This proportion could also be called a false positive rate. The sum of specificity and false positive rate would always be 1. Let's try and understand this with the model used for predicting whether a person is suffering from the disease. Specificity is a measure of the proportion of people not suffering from the disease who got predicted correctly as the ones who are not suffering from the disease. In other words, the person who is healthy actually got predicted as healthy is specificity.

Mathematically, specificity can be calculated as the following:

Specificity = (True Negative) / (True Negative + False Positive)

The following is the details in relation to True Negative and False Positive used in the above equation.

• True Negative = Persons predicted as not suffering from the disease (or healthy) are actually found to be not suffering from the disease (healthy); In other words, the true negative represents the number of persons who are healthy and are predicted as healthy. • False Positive = Persons predicted as suffering from the disease (or unhealthy) are actually found to be not suffering from the disease (healthy). In other words, the false positive represents the number of persons who are healthy and got predicted as unhealthy. The higher value of specificity would mean higher value of true negative and lower false positive rate. The lower value of specificity would mean lower value of true negative and higher value of false positive.

Conclusion:

Computational methodologies, including machine learning are powerful tools for understanding data available in the given data set and to interpret the dataset, but there will be few miss-predictions as these are not completely accurate. Our goal with this study was to apply supervised machine learning algorithm such as KNN (90% accuracy) to a dataset derive from Autism Spectrum Disorder. The Software Development is Flexible and also User-Friendly. Further Functionalities can be added to it in the future.

References:

[1] Tabtah, F.(2017) Autism Spectrum Disorder Screening: Machine Learning Adaptation and DSM-5 Fullfillment. Proceedings of the 1 st International Conference on Medical and Health Informatics 2017, pp.1-6. Taichung City,Taiwan, ACM.

[2] Tabtah, F. (2017). ASD Tests. A mobile app for ASD screening. www.asdtests.com [accessed December 20th, 2017].

[3] Tabtah, F. (2017). Machine Learning in Autistic Spectrum Disorder Behavioral Research: A Review: To Appear in Informatics for Health and Social Care Journal. December, 2017 (in press)

[4] D. P. Wall, R. Dally, R. Luyster, J.-Y. Jung, and T. F. DeLuca, "Use of artificial intelligence to shorten the behavioral diagnosis of autism," PloS one, vol. 7, no. 8, p. e43855, 2012

[5] D. Wall, J. Kosmicki, T. Deluca, E. Harstad, and V. Fusaro, "Use of machine learning to shorten observation-based screening and diagnosis of autism," Translational psychiatry, vol. 2, no. 4, p. e100, 2012

[6] D.Godwin Immanuel, Dayana D.S, Sindarsingh Jebaseelan S.D "Hybrid Genetic Algorithm Assisted Artificial Bee Colony Approach for Voltage Stability Improvement" International Journal of Applied Engineering Research ISSN No.0973-4562 Research India Publications, Volume 10, Number 59 (2015) pp.534-541.

[7]A. Liaw, M. Wiener et al., "Classification and regression by randomforest," R news, vol. 2, no. 3, pp. 18–22, 2002.

[8]F. Hauck and N. Kliewer, "Machine learning for autism diagnostics: Applying support vector classification."

[9]W. Liu, M. Li, and L. Yi, "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, 2016.

[10] Godwin Immanuel, D., Dayana, D.S, Solar water pumping system using enhanced DC–DC converter, Journal of Advanced Research in Dynamical and Control Systems, vol.10,2018,1609-1614.

[11] T. Booth, A. L. Murray, K. McKenzie, R. Kuenssberg, M. O'Donnell, and H. Burnett, "An evaluation of the aq-10 as a brief screening instrument for asd in adults."

[12] J. Kosmicki, V. Sochat, M. Duda, and D. Wall, "Searching for a minimal set of behaviors for autism detection through feature selection-based machine learning," Translational psychiatry, vol. 5, no. 2, p. e514, 2015.

[13] F. Thabtah, "Autism spectrum disorder screening: machine learning adaptation and dsm-5 fulfillment," in Proceedings of the 1st International Conference on Medical and Health Informatics 2017. ACM, 2017.

[14] D. Godwin Immanuel and C. Chritober Asir Rajan, "An Genetic Algorithm approach for reactive power control problem," 2013 International Conference on Circuits, Power and Computing Technologies (ICCPCT), Nagercoil, 2013, pp. 74-78, doi: 10.1109/ICCPCT.2013.6528940.

[15]C. Allison, B. Auyeung, and S. Baron-Cohen, "Toward brief "red flags" for autism screening: the short autism spectrum quotient and the short quantitative checklist in 1,000 cases and 3,000 controls," Journal of the American Academy of Child & Adolescent Psychiatry, vol. 51, 2012.