International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 06 Issue: 06 | June - 2022 Impact Factor: 7.185 ISSN: 2582-3930

Autistic Spectrum Disorder Screening: Prediction with Machine Learning Models

Laxmi Shabadi[1], Shanti Rajashekhar Jangam[2], Shwetha V Shetty[3], Soujanya Ganesh Naik[4], Tanzeem Taj[5]

[1] Assistant Professor, Department of Computer Science Engineering, SJBIT, Bengaluru, 560060
[2,3,4,5]Student, Department of Computer Science and Engineering, SJBIT, Bengaluru, 560060

***Abstract* — In present scenario autism spectrum disorder (ASD) is having faster momentum than ever before. Detecting autistic behavior and traits through screening test is time consuming, labor demanding and expensive. With growth of AI and ML, autism can be predicted in an early stage. Various studies have been carried out using different techniques, but these studies didn't provide any proper conclusion about predicting autism traits in terms of different age groups. Therefore, this paper aims to propose an effective prediction model based on ML technique and to develop a user interface forpredicting ASD for people of any age. As outcomes of this research, an autism prediction model was developed by merging Random Forest-CART (Classification and Regression Trees) and Random Forest-ID3 (Iterative Dichotomizer 3) and also a user interface was developed based on the proposed prediction model. The proposed model was evaluated with AQ10 dataset and 250 real datasets collected from people with and without autistic traits. The evaluation results showed that the proposed prediction model provides better results in terms of accuracy, specificity, sensitivity, precision and false positive rate(FPR) for the data.**

***Keywords* — Autism Spectrum Disorder, Machine Learning, Diagnosis**

I. INTRODUCTION

Autism spectrum disorder (ASD) is a cognitive disability that will cause serious communication, social and behavioral challenges and its trait is seen in the first two years of life and it gradually develops through time. People with autism face different kinds of struggles with concentration, learning disabilities, mental health issues like depression, anxiety and so on, motor difficulties, sensory issues and others. Experimentation tells that both environment and genes play significant roles. Present day explosion rate of autism around world is more and it is growing rapidly.

As stated by WHO disorder can live without depending on anyone while others require lifelong support and

care. Identification of autism requires a significant amount of cost and time.

The objective of this work is to propose an autism prediction model using ML techniques and to develop a mobile application that could effectively predict autism traits of an individual of any age. In other words, this work focuses on developing an autism screening application for predicting the ASD traits among people of age groups 4-11 years, 12-17 years and for people of age 18 and more.

In machine learning and statistics, classification is a supervised learning approach in which the computer program learns from the data input given to it and then uses this learning to classify new observation. Here, the multiclass problem is implemented using the machine learning approaches like Naive Bayesian probabilistic approach, Decision Tree Classifier, K Nearest Neighbor Classifier and Support Vector Machine Classifier which are well suitable for classification problems. The reasons for applying these methods are their ability in simulating expertise as well as classifying an input even if the number of samples is small with appreciated accuracy.

The System Architecture of the proposed approach is carried out 5 phases including data collection, data synthetization, developing the prediction model and evaluating the prediction model. The data collection helps to develop an effective predictive model AQ-10 datasetwas used which consists of three different datasets based on AQ-10 screening tool questions. The data synthetization that uses statistical techniques to combine results from different studies and obtain a quantitative estimate of the overall effect of a particular intervention or variable on a defined outcome. Developing a prediction model to generate prediction of autism traits, algorithms had been developed and their accuracy were tested. After attaining results from various types of supervised learning like Linear Regression, SVM, Naive Bayes, AdaBoost, KNN; Random Forest was found to be highly feasible with higher accuracy than the other algorithms. The proposed predictive model was tested with the AQ10 dataset and data collected from real- world in terms of the accuracy, specificity, precision, sensitivity and false positive rate. For the AQ-10 dataset, leave-one-out technique was also applied to check effectiveness of the proposed model.

International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 06 Issue: 06 | June - 2022 Impact Factor: 7.185 ISSN: 2582-3930

Thus, a time efficient, accurate and easy screening test tool is very much required which would predict autism traits in an individual and identify whether or not they require comprehensive autism assessment. The objective of this work is to propose an autism prediction model using ML techniques and to develop a mobile application that could effectively predict autism traits of an individual of any age.

In other words, this work focuses on developing an autism screening application for predicting the ASD traits among people of age groups 4-11 years, 12-17 years and for people of age 18 and more. The rest of the paper is organized as follows. Section 2 discusses the related research previously done in this area. Section 3 presents the proposed methodology. Detailed implementation of the proposed system is given, system architecture, implementation of the proposed system. Section 4 represents the results of the proposed system. Finally Section 5 concludes the report by mentioning the conclusion and future work.

II. RELATED WORK

A. Use of Artificial Intelligence to shorten the Behavioral Diagnosis of Autism:

They used machine learning techniques to study the complete sets of answers to the ADI-R available at the Autism Genetic Research Exchange (AGRE) for 891 individuals diagnosed with autism and 75 individuals who did not meet the criteria for an autism diagnosis. Their analysis showed that 7 of the 93 items contained in the ADI-R were sufficient to classify autism with 99.9% statistical accuracy.

They further tested the accuracy of this 7 question classifier against complete sets of answers from two independent sources, a collection of 1654 individuals with autism from the Simons Foundation and a collection of 322 individuals with autism from the Boston Autism Consortium. In both cases, their classifier performed with nearly 100% statistical accuracy, properly categorizing all but oneof the individuals from these two resources who previously had been diagnosed with autism through the standard ADI-R.
Ability to measure specificity was limited by the small numbers of non-spectrum cases in the research data used, however, both real and simulated data demonstrated a range in specificity from 99% to 93.8%. Their study was limited by the content of existing repositories and as a consequence they had a small number of matched controls for construction and validation of the classifier.

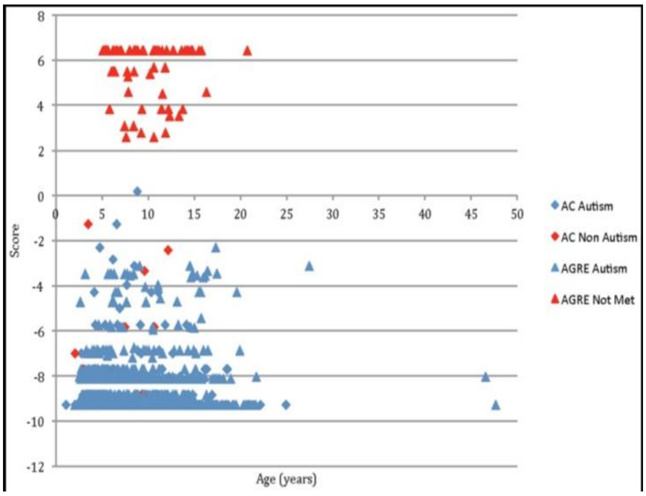


Figure 2.1. Decision tree score and classification

B. Machine Learning for Autism Diagnostic applying support vector classification:

In this paper, they present an approach for developing a support vector machine (SVM) that can facilitate autism diagnostics. Models are trained and tested on a dataset of about 2,500 records with autism diagnostic observation schedule (ADOS) and autism diagnostic interview revised (ADI-R) items. . The results show that a small combination of selected ADOS and ADI-R items is sufficient to achieve a good performance in diagnosing ASD. Their models reach between 85.6% and 94.3% sensitivity and between 80.9% and 89.3% specificity with 10 features at the most. Feature selection is performed by using a greedy backward-elimination process and SVM is trained using a linear kernel as well as a radial basis function (RBF) kernel. Cross validation (CV) is applied to ensure a high generalization performance.

C. Autism Screening using Deep Embedding Representation:

In this paper proposed by Haishuai Wangl, Li Li, Lianhua Chi and Ziping Zhao where they apply novel feature engineering and feature encoding techniques, along with a deep learning classifier for ASD screening. Algorithms were created via a robust deep learning classifier and deep embedding representation for categorical variables to diagnose ASD based on behavioral features and individual characteristics. The proposed algorithm is effective compared with baselines, Achieving 99% sensitivity and 99% specificity. The results suggest that deep embedding representation learning is reliable methods for ASD screening.

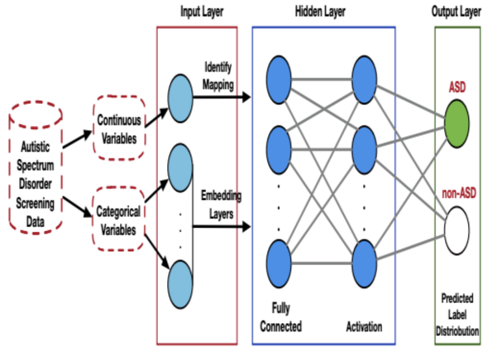


Figure 2.2. Overall architecture of DENN

D. Engagement Detection with Autism Spectrum Disorder using Machine Learning:

This paper aims to propose an autism prediction model using ML techniques and to develop a web application that could effectively predict autism traits of an individual. In other words, this work focuses on developing an autism screening application for predicting the ASD traits among people of age group 3 years and below. The proposed model was evaluated with AQ-10 datasets (1054 datasets) and 50 real datasets collected from people with and without autistic traits. The evaluation results showed that the proposed prediction model provides better results in terms of accuracy, specificity, sensitivity, precision and f1-score for both kinds of datasets. The primary limitation of the study is lack of sufficiently large data to train the prediction model. Another limitation is the screening application is not designed for age group above 3 years.

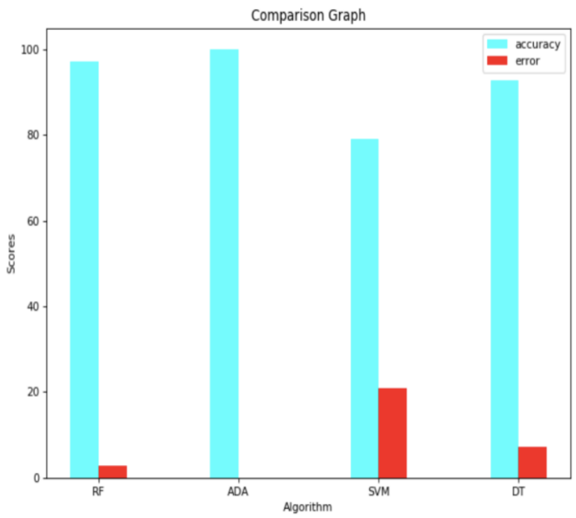


Figure 2.3. Comparison of Algorithm

III. METHODOLOGY

1. System Architecture.

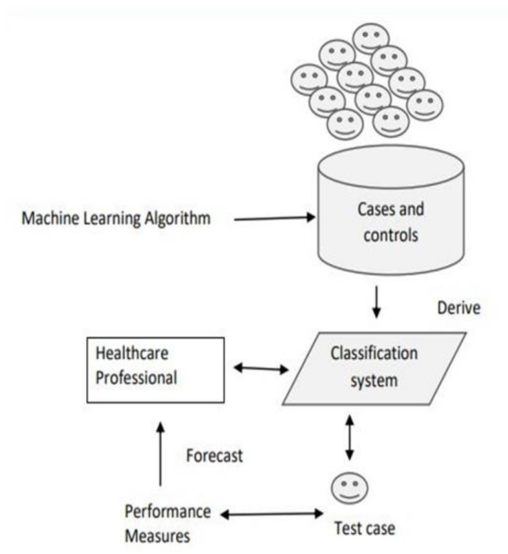


Figure 3.1.1 System Architecture

The project is distributed in four different steps, The Data Collection, Data Preprocessing, Developing the prediction model and evaluating the prediction model. The algorithm with best accuracy is used for the classification.

A. Data Collection:

To develop an efficient prophetic model, AQ-10 dataset was used that consists of dataset supported AQ-10 screening tool queries .These 3 information sets contain data of age teams of 4-11 years (child), 12-16 years (adolescent) and last ages of eighteen or additional (adult). AQ-10 or syndrome verbal description Quotient tool is employed to spot whether or not a personal should be referred for a comprehensive syndrome assessment. AQ-10 screening queries specialise in totally different domains such as- attention to detail, attention switch, communication, imagination and social interaction. the evaluation technique of the questions is that only one purpose will be scored for every of the ten questions. Users might score zero or one purpose on every question

International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 06 Issue: 06 | June - 2022 Impact Factor: 7.185 ISSN: 2582-3930

based mostly on their answer. Datasets of kid, adolescent and adult contain 292, 104 and 704 instances severally. The datasets contain twenty-one attributes that area unit a mixture of numerical and categorical information, that includes: Age, Gender, Ethnicity, If born with Jaundice, family members with autism, Whois finishing the take a look at, Country of Residence, Used the screening app before, Screening technique kind, Question 1-10, Result and sophistication.

B. Data Preprocessing

The collected information was synthesized to get rid of extraneous features. as an example, the ID column was irreverent to develop a prediction model, so it absolutely was removed. To handle null values, list wise deletion technique was applied wherever a specific observation was deleted if it had one or additional missing values. Then to extract spare options from the dataset, decision tree algorithmic program was used. Results showed dropping 'age desc', 'relation', 'age' and 'app used before ' columns would result in more correct classification and then those columns were dropped.

C. Developing the Prediction Model

To generate prediction of syndrome traits, algorithms had been developed and their accuracy were tested. Various algorithms like naïve bayes, KNN algorithm, random forest, decision tree and ensemble techniques like XG boost and AdaBoost is used.

D. Evaluating the Prediction Model

The planned prophetic model was tested with the AQ- 10 information set and data collected from the real-world in terms of the accuracy, specificity, precision, sensitivity and false positive rate. For the AQ-10 dataset, leave-one-out technique was additionally applied to ascertain effectiveness of the planned model. Again, to validate the planned model, virtually one hundred information of ASD cases and a hundred and fifty information of Non- ASD cases were collected from the associate degree institute of education for the people.

2. IMPLEMENTATION

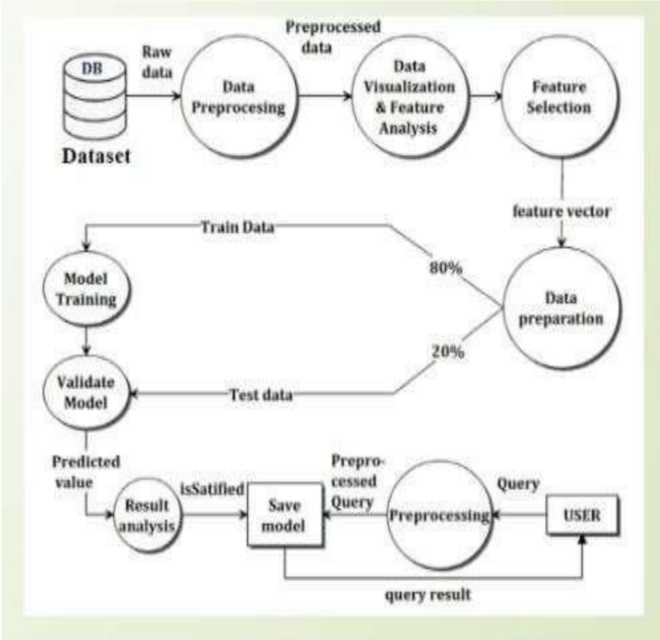


Figure 3.2.1. Data Flow Diagram

Raw Dataset: Raw data, also known as primary data, are data (e.g., numbers, instrument readings, figures, etc.) collected from a source. In the context of examinations, the raw data might be described as a raw score.

Labelling: In machine learning, data labeling is the process of identifying raw data (images, text files, videos, etc.) and adding one or more meaningful and informative label**s** to provide context so that a machine learning model can learn from it.

Data Preprocessing: Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviours or trends, and is likely to contain many errors.

Training: The model is initially fit on a training dataset, which is a set of examples used tofit the parameters of the model. The model is trained on the training dataset usinga supervised learning

International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 06 Issue: 06 | June - 2022 Impact Factor: 7.185 ISSN: 2582-3930

method, for example using optimization methods such as gradient descent or stochastic gradient descent. In practice, the training dataset often consists of pairsof an input vector (or scalar) and the corresponding output vector (or scalar), where the answer key is commonly denoted as the target (or label). The current model is run with the training dataset and produces a result, which is then compared with the target, for each input vector in the training dataset. Based on the result of the comparison and the specific learning algorithm being used, the parameters of the model are adjusted. The model fitting can include both variable selection and parameter estimation.

Testing: A test dataset is a dataset that is independent of the training dataset, but that follows the same probability distribution as the training dataset. If a model fit to the training dataset also fits the test dataset well, minimal overfitting has taken place . A better fitting of the training dataset as opposed to the test dataset usually points to overfitting. A test set is therefore a set of examples used only to assess the performance (i.e. generalization) of a fully specified classifier. To do this, the final model is used to predict classifications of examplesin the test set. Those predictions are compared to the examples' true classifications to assess the model's accuracy.

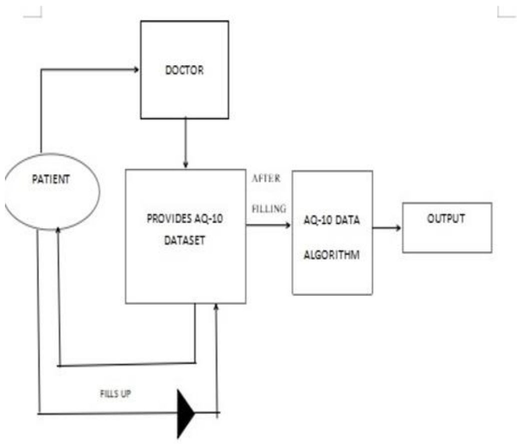


Figure 3.2.2 Implementation

In the process of prediction initially, the patient/user will input the data for AQ10 dataset which has 20 attributes in total where 10 general questions and other 10 attributes are personal questions like age, country etc. This is provided by the doctor or any medical assistant. After filling up the data given by the user, several required machine learning algorithms were used to predict whether the patient/user had ASD traits or not.

Classification Procedure for Diagnosis of ASD

The primary form of the classification problem is a binary classification, for example, ASD or non-ASD, as illustrated in Figure 5.2.1 For clarity, binary classification involves grouping the entire data elements or features into two distinct classes according to already defined classification rules. This figure shows possible responses designed for a test case prediction. Previous research used machine learning algorithms for prediction purposes. A confusion matrix was utilized as a binary classification.

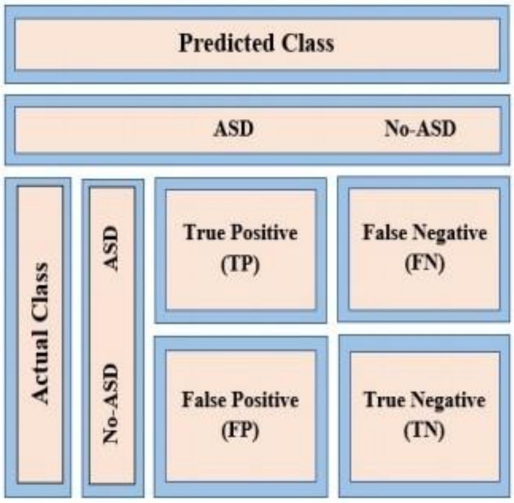


Figure 3.2.3 Confusion Matrix for ASD Classification

ASD diagnosis is a classification problem that consists of a few steps. The process is described in Figure 3.2.4. Here, the training dataset is input as cases and controls that have been diagnosed earlier.

International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 06 Issue: 06 | June - 2022 Impact Factor: 7.185 ISSN: 2582-3930

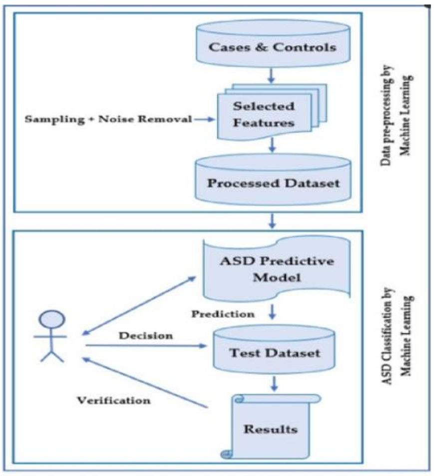


Figure 3.2.4 ASD classification with machine learning

Generally, these cases and controls are created by a diagnostic tool such as ADOS-Ror ADI-R in a clinic. Clinical specialists or psychologists control these tools. After identifying the training dataset, there is an optional step where feature selection is made by selecting a reduced set of features to shrink data dimensionality. Besides narrowing down the problem and recognizing the essential ASD features, the process of managing the computing resources is employed and incorporated during data management.

A subsequent step is noise removal, which is optional. Noise can be in the form of replicated records, missing values, or imbalanced data. Recently, many studies used sampling techniques to enhance data concerns. After completing the above preprocessing steps, machine learning algorithms areemployed in the subsequent steps to classify ASD cases. Currently, software packages such as Weka and R are utilized by researchers. In order to use these software packages, related preprocessed data should be loaded. There are different options to select various filters and machine learning algorithms for data classification. Machine learning tools embedded inthese types of software have more benefits in terms of having multiple evaluation metrics. Here, users can reach their expected goals, such as accuracy, sensitivity, specificity, false positive rate, false negative rate, processing rate, receiver operating characteristic (ROC), and F-measure.

On comparing the accuracy of various algorithms, the algorithm with the best accuracy is used for classifying the autistic patients. The chatbot is implemented for efficient screening.

3. Chatbot Implementation

A bot is a program that automatically completes an action based on specific triggers and algorithms. A chatbot is a computer program that's designed to simulate human conversation. Users communicate with these tools using a chat interface or via voice, just like they would converse with another person. Chatbots interpret the words given to them by a person and provide a pre-set answer.

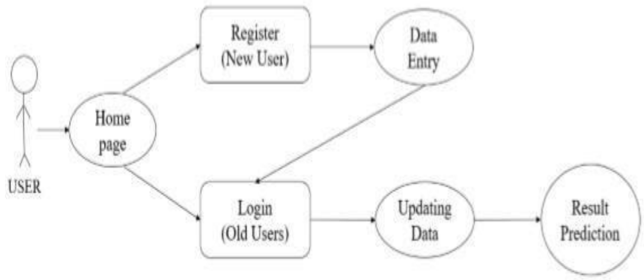


Figure 3.3.1 Use-case Diagram for Autism Prediction Model

1. Prepare the Dependencies: The first step in creating a chatbot in Python with the ChatterBot library is to install the library in your system. It is best if you create and use a new Python virtual environment for the installation.

2. Import Classes: Importing classes is the second step in the Python chatbot creation process. All you need to do is import two classes – ChatBot from chatterbot and ListTrainer from chatterbot.trainers.

3. Create and Train the Chatbot:This is the third step on creating a chatbot in python. The chatbot you are creating willbe an instance of the class "ChatBot." After creating a new Chatterbot instance, you can train the bot to improve its performance. Training ensures that the bot has enough knowledge toget started with specific responses to specific inputs.

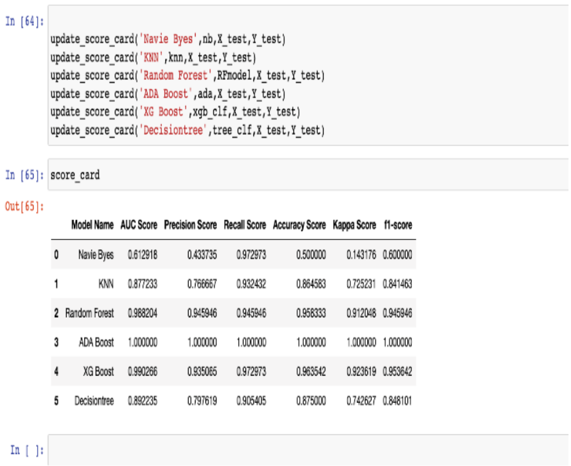
4. Communicate with the Python Chatbot: To interact with your Python chatbot, you can use the .get_response() function.

International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 06 Issue: 06 | June - 2022 Impact Factor: 7.185 ISSN: 2582-3930

IV RESULTS AND ANALYSIS

1. Accuracy Analysis

On comparing various machine learning models it was found that the ensemble technique yielded the best accuracy in comparison to other algorithms used. The AdaBoost algorithm in ensemble technique yielded the highest accuracy among all. Fig 4.1 shows the scorecard of various algorithm that was implemented on the datasets.



```
In [64]:
update_score_card('Navie Byes',nb,X_test,Y_test)
update_score_card('KNN',knn,X_test,Y_test)
update_score_card('Random Forest',RFmodel,X_test,Y_test)
update_score_card('ADA Boost',ada,X_test,Y_test)
update_score_card('XG Boost',xgb_clf,X_test,Y_test)
update_score_card('Decisiontree',tree_clf,X_test,Y_test)

In [65]: score_card
```

Out[65]:

	Model Name	AUC Score	Precision Score	Recall Score	Accuracy Score	Kappa Score	f1-score
0	Navie Byes	0.612918	0.433735	0.972973	0.500000	0.143176	0.600000
1	KNN	0.877233	0.766667	0.932432	0.864583	0.725231	0.841463
2	Random Forest	0.988204	0.945946	0.945946	0.958333	0.912048	0.945946
3	ADA Boost	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
4	XG Boost	0.990266	0.935065	0.972973	0.963542	0.923619	0.953642
5	Decisiontree	0.892235	0.797619	0.905405	0.875000	0.742627	0.848101

In [ ]:

Figure 4.1 Scorecard of various algorithms

A simple user interface is created where the user can login to the portal and undertake the screening. For the first time user the registration is mandatory. For the returning user should login using the credentials. The series of questions are displayed, the user can answer those questions and depending upon the trained datasets , the system predicts whether the user is autistic or not.

The chatbot can be used for patients to interact and to provide a live human kind of interface. The user can answer the questionnaires which is asked by the chatbot. Based on the way chatbot is trained and its experience based on interactions, it tries to give the approriate response.

The chabot can provide 2 kinds of interface-one is questions asked for parents related to the behaviour of their kids(recommened for kids below 15 years of age) and also the person can answer on their own behalf.

Figure 4.2 Chatbot UI

The user can also answer the list of questionnaires'' and the classification is made based on the data that us trained to the model.

S.no	Attributes	Values
1	Often notice noise when others not	0-1
2	give attention to the picture rather than small details	0-1
3	Able to perform more than one task at once	0-1
4	Can switch between the activities easily	0-1
5	Can concentrate on study while talking with someone	0-1
6	If someone bored with me then I am able to tell him	0-1
7	Difficult to make out characters while reading the book	0-1
8	Collect the information's about things	0-1
9	By look at someone face can easily make out what he is thinking	0-1

Figure4.2 Question set 1

10	Difficult to trust people	0-1
11	Age	17 or above
12	Gender	1-2
13	Ethnicity	1-11
14	Born with jaundice	1-2
15	Who completes the test	1-5
16	Used screening app before	1-2
17	Screening method type	1-4
18	country	1-52
19	Score	0-10

Fig 4.3 Question set 2

V. CONCLUSION

This analysis provides a threefold outcome: firstly, a prediction model was developed to predict syndrome traits with 97% accuracy by using Ensemble method. This analysis provides a comparative read among different machine learning approaches in terms of their performance. Finally, an easy user interface has been developed appliance to predict the syndrome traits. Since designation of the syndrome traits is kind of an expensive and lengthy method, it's usually delayed attribute to the difficulty of detecting syndrome in kids and adolescents. With the assistance of autism screening application, a private will be guided at an early stage that may forestall the case from obtaining associate degree worse and scale back prices related to delayed diagnosing.

Our future work will focus on collecting more data from various sources and to improve the proposed machine learning classifier to enhance its accuracy. A user study will also be conducted to evaluate the usability and user experience (UX) of the mobile application.

VI. REFERENCES

[1] U. Frith and F. Happé, "Autism spectrum disorder," *Current biology*, vol. 15, no. 19, pp. R786–R790, 2005.

[2] WHO, *Autism spectrum disorders*, 2017 [Accessed August 22, 2018].[Online]. Available: http://www.who.int/news-room/fact-sheets/detail/ autism-spectrum-disorders

[3] J. A. Cruz and D. S. Wishart, "Applications of machine learning in cancer prediction and prognosis," *Cancer informatics*, vol. 2, 2006.

[4] N. S. Khan, M. H. Muaz, A. Kabir, and M. N. Islam, "Diabetes predicting mhealth application using machine learning," in *2017 IEEE International WIE Conference on Electrical and Computer Engineering (WIECON-ECE)*. IEEE, 2017, pp. 237–240.

[5] 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.

[6] D. Bone, S. L. Bishop, M. P. Black, M. S. Goodwin, C. Lord, and S. S. Narayanan, "Use of machine learning to improve autism screening and diagnostic instruments: effectiveness, effificiency, and multi-instrument fusion," *Journal of Child Psychology and Psychiatry*, vol. 57, 2016.

[7] 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.

[8] 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.

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

[10] B. van den Bekerom, "Using machine learning for detection of autism spectrum disorder," 2017.

[11] 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.

[12] A. S. Heinsfeld, A. R. Franco, R. C. Craddock, A. Buchweitz, and F. Meneguzzi, "Identifification of autism spectrum disorder using deep learning and the abide dataset," *NeuroImage: Clinical*, vol. 17, 2018.

[13] 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.