IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 11, Issue 6, June 2022

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.118

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 6, June 2022||

| DOI:10.15680/IJIRSET.2022.1106092|

Autism Diagnosis Using Machine Learning and Neural Networks

A Swapna

Asst Professor, Dept of Computer Science and Engineering, Mahatma Gandhi Institute of Technology, Hyderabad, Telangana, India

ABSTRACT*:* Autism is a type of developmental condition that affects the brain. It creates communication and social interaction issues, as well as limiting physical ability to some level. If it is detected at an early stage, the kid can receive the necessary care, making it easier for them to participate in society. Machine learning is a promising tool for investigating the reliability of patterns across larger, more heterogeneous datasets. We use the personal characteristics, social responsiveness scale data as well as the IQ. Multiple machine learning algorithms are used such as Random Forest, Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Artificial Neural Networks (ANN) and XGBoost to compare the accuracies given by each algorithm and which algorithm gives the best accuracy for the dataset. To help with the project we use the data provided from Autism Brain Image Date Exchange(ABIDE) which is openly available in the internet

KEYWORDS*:* K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Artificial Neural Networks (ANN) ,XGBoost

I. INTRODUCTION

The heart rate (HR) of a person represents the number of heart beats per minute. It is an essential physiological parameter, a source of information related to the entire cardiovascular system and has great importance in diagnosis or assessment of the stress levels experienced by the person. The heart rate normal values differ depending on the age, medical history or usual physical activity. For example, people who are less physically active are expected to have a higher heart rate, as their heart muscle has to work harder to maintain a constant cardiac rhythm. It has been well-known for decades that any unusual variations of the cardiac pulse have to be taken into consideration for further investigation and diagnosis.

As a consequence of society becoming more health conscious, various concepts of remote health monitoring platforms are developed. Among others, these also include supervising elderly people or chronic diseases patients from residential environments. Also, there are situations in which continuous heart rate monitoring is required but skin contact is problematic, and the patient feels uncomfortable to be continuously connected to a pulse measuring apparatus. Research exhibits significant advancements over the last few years and demonstrates that standard video cameras are reliable devices that can be employed to measure a large set of biomedical parameters without any contact with the subject.

Photoplethysmography (PPG) and Ballistocardiography (BCG) are the two main principles for measuring pulse rate in video streams recorded by a camera. Ballistocardiography relates to the observation of small body displacements that appear during systole (cardiac contraction). Photoplethysmography consists in indirect observation of blood volume variations by measuring absorption and reflection of light on skin tissues.These fluctuations in volume are periodic and produced at each heartbeat: the volume of blood increases during systole (cardiac contraction) and decreases during diastole (cardiac relaxation).It must be emphasized that the definition of the principle is still discussed today: light variations that are remotely measured by the camera might be related to elastic deformations of the capillary bed, by a rise of the capillary density that compresses tissues during systole, instead of a direct observation of the changes in sections of the pulsatile arteries.

We present a new approach to contact-free methods for measuring heart rate using video processing. This technique requires the subject to be relaxed and to be placed near a webcam. The distance between the camera and the patient can vary between 75cms to 100 cms, and the illumination conditions have to be constant during the process. In order to measure the heart rate in real time, image processing will be performed using OpenCV and implemented in Python programming language.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 6, June 2022||

| DOI:10.15680/IJIRSET.2022.1106092|

II. PROPOSED METHODOLOGY

The objective is to check which machine learning algorithm gives the most accuracy. For achieving this, the main proposal consists of using a Neural Networks, SVM, Random Forest, K-NN, Logistic Regression and XGBoost. Our approach comprises a sequence of steps for preparing the data by

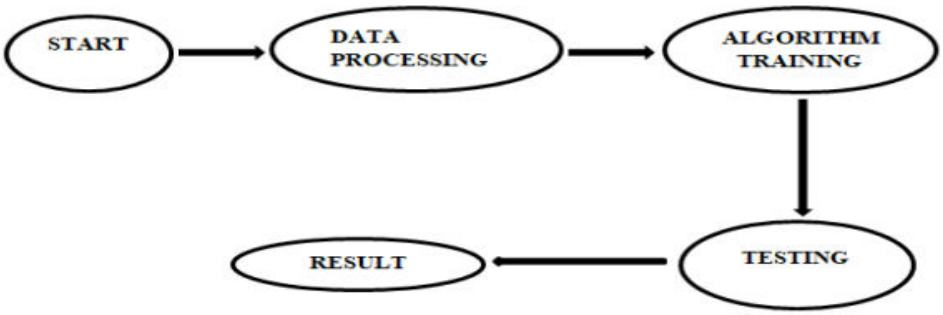

Fig. 1 System Architecture

reducing the dimensionalityof the problem, using a pre-trained model in the initial layers. The general training work-flow can be seen in Figure 1

A. Data Processing Module

a) Loading the Data:

The dataset is downloaded from ABIDE dataset which is available in the internet publicly through COINS. These 1112 datasets are composed of structural and resting state functional MRI data along with an extensive array of phenotypic information

b) Data Preprocessing:

Data Preprocessing is a Data Mining technique that involves transforming raw data into an understandable format. Steps Involved in Data Preprocessing:

1. Data Cleaning: The data can have many irrelevant and missing parts. To handle this part, data cleaning is done. It involves handling missing data, noisy data etc.
2. Data Transformation: his step is taken in order to transform the data in appropriate forms suitable for the mining process. It involves Normalizations, Attribute Selection, Discretization, Concept Hierarchy.
3. Data Reduction:It is a capacity optimization technique in which data is reduced to its simplest possible form to free up capacity on a storage device..It involves Data Cube Aggregation, Attribute Subset Selection, Numerosity Reduction, Dimensionality reduction

c) Feature Extraction:

A dataset has many features which are responsible for autism. So, we select particular features that were considered main in this procedure and diagnose autism by training the dataset. The below mentioned category of features are extracted.

- SUB_ID:This is the identification number for the patient.
- SEX: The Gender of the patient
- HANDEDNESS_CATEGORY This attribute tells us whether the patient is left handed or right handed or ambidextrous.
- AGE_AT_SCAN: The age at which the patient is taking the scan.
- FIQ: This is the full IQ which is the total score comparing VIQ and PIQ.
- VIQ: This tells us the IQ of the person at verbal communication and how good they an hold a conversation.
- PIQ: This is the performance IQ of the person which tells us how good patient is at executing the tasks given to them.
- ADOS_TOTAL: This is sum of communication and social interaction subtotals.
- ADI_R_SOCIAL_TOTAL_A: This attribute tells the social skills of the patient.
- ADI_R_VERBAL_TOTAL_BV: This tells the level of the patient based on his speech and

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 6, June 2022||

| DOI:10.15680/IJIRSET.2022.1106092|

consistency in speaking

- ADI_RRB_TOTAL_C: This attribute tells the total score of the patient in physical activities and the daily situations.
- ADI_R_ONSET_TOTAL_D: This attribute tells us the age at which the symptoms of ASD have started and how fast they are progressing.
- SUB_IN_SMP: The symptoms are mentioned in this attribute and the severity of the symptoms is also shown.
- CURRENT_MED_STATUS: The medical status of the patient is shown, this helps in diagnosing the patient because sometimes Asperger's and ADHD are misdiagnosed as ASD.
- DSM_IV_TR: This is the revised edition of diagnostic manual which tells us how to relate the attributes and which disease the patient is diagnosed with.

B. Random Forest Module:

Random forest (RF) is an ensemble learning classification. In RF, prediction is achieved using decision trees. During the training phase, a number of decision trees are constructed (as defined by the programmer) which are then used for the autism class prediction; this is achieved by considering the voted classes of all the individual trees and the class with the highest vote is considered to be the output.

The following parameters were taken into consideration while creating a Random Forest Classifier:

- *min_samples_split:* This is the minimum number of samples required to split an internal node. The min_samples_split parameter will evaluate the number of samples in the node, and if the number is less than the minimum the split will be avoided and the node will be a leaf. The optimal value for the given model was found to be 3.
- *max_depth:* It represents the depth of each tree in the forest. The deeper the tree, the more splits it has and it captures more information about the data. The maximum depth of the tree, optimal value chosen is 15 here.
- *random_state:* Decision tree do not guarantee the same solution globally. There will bevariations in the tree structure each time you build a model. Passing a specific seed to random_state ensures the same result is generated each time you build the model.

C. Logistic Regression Module:

It is used to understand the relationship between the dependent variable and one or more independent variables by estimating probabilities using a logistic regression equation. This type of analysis can help you predict the likelihood of an event happening or a choice being made. The following parameters were taken into consideration while creating a Logistic Regression Model:

- *solver :* It tries to find the parameter weights that minimize a cost function. The solver that is being used in this model is 'saga' which is an extension of 'sag' solver where the gradient of the loss is estimated each sample at a time and the model is updated along the way with a constant learning rate but in 'saga' training is comparatively faster.
- *max_iter:* This parameter tells us the number of iterations we want for our model to process.

D. K-Nearest Neighbors Module:

KNN is a simple supervised classification algorithm we can use to assign a class to new data point. Larger K value leads to smoother decision boundary (less complex model). Smaller K leads to more complex model (may lead to overfitting). Testing accuracy penalizes models that are too complex (over fitting) or not complex enough (underfit). The following parameters were taken into consideration while creating a K-Nearest Neighbors Model:

- *n_neighbors:* This parameter tells us the number of neighbors that need to be considered by the k-nn queries. By default the value is 5
- *weights:* This parameter is used for the prediction of possible values based on the type of weight. In the model we use weight based on distance which means based on the inverse of distance closer neighbors of a query point will have a greater influence than neighbors which are further away

E. Support Vector Machine Module:

Support vector machines (SVMs) are a set of supervised learning methods used for classification, regression and outliers' detection.The following parameters were taken into consideration while creating a SVM Model:

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 6, June 2022||

| DOI:10.15680/IJIRSET.2022.1106092|

• *kernel:* The function of kernel is to take data as input and transform it into the required form. In the model Radial Basis Function (RBF) kernel is applied on the c-support vector classifier which is the default kernel used.

• *gamma:* Intuitively, the gamma parameter defines how far the influence of a single training example reaches, with low values meaning 'far' and high values meaning 'close'. The gamma parameters can be seen as the inverse of the radius of influence of samples selected by the model as support vectors.

• *C:* The C parameter tells the SVM optimization how much you want to avoid misclassifying each training example. For large values of C, the optimization will choose a smaller-margin hyperplane if that hyperplane does a better job of getting all the training points classified correctly.

F. Artificial Neural Network Module:

Artificial neural networks, usually simply called neural networks, are computing systems inspired by the biological neural networks that constitute animal brains. An ANN is based on a collection of connected units or nodes called artificial neurons, which loosely model the neurons in a biological brain. In the model we create it using 'sequential' which is nothing but a stack of layers.The following parameters were taken into consideration while creating ANN Model:

• *dense:* The dense layer feeds all outputs from the previous layer to all its neurons, each neuron providing one output to the next layer. The layer has the following sub attributes

1. units: Units are one of the most basic and necessary parameters of the Keras dense layer which defines the size of the output from the dense layer

2. activation: Activation function decides, whether a neuron should be activated or not by calculating weighted sum and further adding bias with it. The purpose of the activation function is to introduce non-linearity into the output of a neuron. We use 'relu' and 'softmax' activations

* *dropout*: It reduces the overfitting in the ANN model by preventing co-adaptations on training data

• *optimizer:* It is used to change the attributes of neural networks such as weights and learning rate to reduce losses.

• loss: It is the prediction error for neural networks model that is built.

G. XGBoost Module:

XGBoost (eXtreme Gradient Boosting) is an advanced implementation of gradient boosting algorithm. It's a powerful machine learning algorithm especially where speed and accuracy are concerned. This model requires parameter tuning to improve and fully leverage its advantages over other algorithms. The following parameters were taken into consideration from XGBoost Model:

• *max_depth :* Maximum depth of a tree. Increasing this value will make the model more complex and more likely to overfit.

• *objective :* multi:softmax: set XGBoost to do multiclass classification, also need to set num_class(number of classes)

• *subsample :* corresponds to the fraction of observations (the rows) to subsample at each step.

• *eta :* learning_rate or step size shrinkage used in update to prevents overfitting. After each boosting step, we can directly get the weights of new features, and eta shrinks the feature weights to make the boosting process more conservative.

IV. TESTING AND RESULT

Based on the .training we have given to the dataset using machine learning algorithms, we now test the algorithm by changing the parameters and check which parameters give the best result

We will test each algorithm and the test cases will be shown below.

A. Random Forest Testing

As shown in Fig 4.1, the maximum accuracy is given when the minimum samples split is 3 and the maximum depth is 15.We can observe from the test cases that from lower splits value to higher value the accuracy increase and after reaching the highest accuracy it starts to decrease.If the depth is decreased then we can see that the accuracy slowly reduces

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 6, June 2022||

| DOI:10.15680/IJIRSET.2022.1106092|

Min_samples_split	Max_depth	Accuracy
2	15	83.43
3	15	84.07
5	15	82.80
2	12	82.80
3	12	82.16
5	12	80.89
5	18	81.52

Fig. 2 *Random Forest Testing*

B. Logistic Regression Testing

As we can see in the Fig 3, the best accuracy is given for 'saga' solver. We use the One vs. Rest multi_class for all the test cases because the data is multinomial and one vs. one cannot be used for such a large dataset. Saga solver gave the best result because it even supports L1 regularization

Solver	Multi_class	Accuracy
Newton-cg	One vs. Rest	66.874
lblfgs	One vs. Rest	66.874
sag	One vs. Rest	66.873
saga	One vs. Rest	66.876

Fig. 3 Test cases for Logistic Regression

A. K-Nearest Neighbors Testing

N_neighbors	weights	Accuracy
5	distance	68.78
7	distance	68.15
3	distance	68.15
7	uniform	67.51
3	uniform	67.51

Fig 4: Test cases for KNN

As shown in Fig 4 while testing the k-nearest neighbors, there are only two parameters and we can observe that distance and neighbors of 5 is giving the best accuracy.This is because if the neighbor pool is huge then the comparison of data becomes slow and accuracy reduces. And distance gives better accuracy than uniform because distance gives us the radius in which we can compare the data unlike the uniform parameter.

B. Support Vector Machine Testing

Kernel	C	Accuracy
rbf	0.5	76.42
rbf	0.8	79.61
rbf	1.0	78.37
sigmoid	0.5	66.24
sigmoid	0.8	69.42
sigmoid	1.0	70.06

Fig 5: Test cases for SVM

The test cases in Fig 5 indicate that 'rbf' kernel gives the best accuracy when the value of c is 0.8 The c value tells how much misclassification must be avoided so higher the value the better, but completely avoiding it is not possible hence we get low accuracy for c value 1.The rbf kernel is better because it maps the larger dataset into linear or non-linear hyper planes without overfitting the data and reduces the redundancy thus giving more accuracy

C. *Artificial Neural Networks Testing*

Optimizer	Loss Function	Accuracy
adam	categorical_crossentropy	81.14
sgd	categorical_crossentropy	80.52
adagrad	categorical_crossentropy	78.63
adadelta	categorical_crossentropy	79.81

Fig 6: Test cases for ANN

The test cases in Fig 6 show us using the same loss function, the reason is for such a large dataset which has a lot of categorical data other loss functions would take a lot of time to run and few functions cannot

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 6, June 2022||

| DOI:10.15680/IJIRSET.2022.1106092|

even work our data.Out of all the optimizers, 'adam' gave the highest accuracy because it is invariant to diagonal rescale of the gradients and little memory is required.

D. XGBoost Testing

Depth	Objective	Accuracy
11	Multi_softmax	84.7
7	Multi_softmax	84.07
15	Multi_softmax	83.35

Fig 7: Test cases for XGBoost

As shown in Fig 7 the main reason for change is accuracy in XGBoost model was found to be because of the depth of the model.The accuracy increase from low value to high value of depth till certain point and then decreases, this is because as depth increases the classification becomes easier but after certain limit there tends to be too much time consumption in executing the code and the accuracy tends to decrease.

The result for the work is shown below in Fig.8. The algorithms Random Forest, Logistic Regression, K-Nearest Neighbors, Support Vector Machine, Artificial Neural Networks and XGBoost give the accuracies of 84.07%, 66.87%, 68.78%, 79.61%, 81.14% and 84.7% respectively. Out of the entire algorithms XGBoost algorithm gave the best accuracy

	Classifiers	Accuracy score	AUC-ROC Score
0	Random Forest	0.840764	0.735244
1	Logistic Regression	0.668790	0.599839
2	kNN	0.687898	0.619831
3	SVM_rbf	0.796178	0.694598
4	ANN	0.811441	
5	XGBoost	0.847134	0.747647

Fig 8: Result

V. CONCLUSION

Our proposed model allows us to have a more accurate result in terms of detecting autism. In the ABIDE dataset we are using the questions are provided to the parents to identify if their children are in danger or out of danger is set in a way to maintain their privacy. Using the dataset from ABIDE our proposed model can predict using Random Forest, Logistic Regression, KNN , SVM, ANN and XGBoost with 84%, 66% ,68%,79%,81% and 84% accuracy in case of toddlers. Algorithms which are supervised are selected to run our dataset after preprocessing it. The outcome for XGBoost showed better performance compared to the others.

The limitation for the work is the lack of the correlation between the attributes of the dataset due to which the accuracy is not as high compared to the prediction of other diseases. This is mainly due to the fact that required attention is not being given to ASD compared with other diseases.

Hence we have less knowledge about the disease The purpose of choosing age as attribute is to know at which age people tend to show symptoms of ASD and in which pace the symptoms tend to progress. Taking all these into consideration we have executed the work to give better accuracy.

VI. FUTURE SCOPE

There are many ways in which the performance of this model can be enhanced in the future. Having better correlated attributes would make huge difference for the prediction of ASD. Beyond that, exploration of different types of loss functions could improve the efficiency of producing better results but since our data does not have proper correlated data we are unable to use loss functions other than 'categorical_crossentropy'. Recently research is being done based on the MRI as well as genes of the patient, so if we use data regarding the MRI as well as multiple information regarding the health issues of family and lifestyle of the patients, then we can diagnose the patients more accurately

REFERENCES

[1] Shirajul Islam;Tahmina Akter;Sarah Zakir;Shareea Sabreen;Muhammad Iqbal Hossain 2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)"Autism Spectrum Disorder Detection in Toddlers for Early diagnosis Using Machine Learning"

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 6, June 2022||

| DOI:10.15680/IJIRSET.2022.1106092|

[2] . SARANYA, A., & ANANDAN, "Autism Spectrum Prognosis using Worm Optimized Extreme Learning Machine (WOEM) Technique. 2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)".

[3] Jiao, Z., Li, H., & Fan. "Improving Diagnosis of Autism Spectrum Disorder and Disentangling its Heterogeneous Functional Connectivity Patterns Using Capsule Networks. 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)"

[4] Vabalas, A., Gowen, E., Poliakoff, E., & Casson, A. J. (2019). "Kinematic features of a simple and short movement task to predict autism diagnosis. 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)."

[5] Dutta, S. R., Datta, S., & Roy. "Using Cogency and Machine Learning for Autism Detection from a Preliminary Symptom. 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence)."

[6] Shuaibing Liang, Aznul Qalid Md Sabri; Fady Alnajjar; Chu Kiong Loo. "Autism Spectrum SelfStimulatory Behaviors Classification Using Explainable Temporal Coherency Deep Features and SVM Classifier"

INNO SPACE
SJIF Scientific Journal Impact Factor
Impact Factor: 8.118

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details