IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 26, Issue 5, Ser. 1 (Sept. – Oct. 2024), PP 37-44
www.iosrjournals.org

A Review Of Autism Spectrum Disorder Detection Using Machine Learning

Rubeena Khan
Modern Education Society's Wadia College Of Engineering.
Savitribai Phule Pune University Pune, India

Shaikh Umme Hanni Aiyaz Hussain
Modern Education Society's Wadia College Of Engineering.
Savitri Bai Phule Pune University Pune, India

Devesh Sandeep Singh
Modern Education Society's Wadia College Of Engineering.
Savitribai Phule Pune University Pune, India

Sakshee Vijay Phadtare
Modern Education Society's Wadia College Of Engineering.
Savitri Bai Phule Pune University Pune, India

Mohammed Aslaan
Modern Education Society's Wadia College Of Engineering.
Savitribai Phule Pune University Pune, India

Abstract
A neurological condition known as an autism spectrum disorder (ASD) profoundly impacts an individual's lifelong ability to engage and interact with others. ASD is an illness that can be recognized at early stages in a person's life and is often classified as a "behavioral disease" due to the presence of numerous symptoms that frequently manifest within the first two years of life, as per what most people believe about autism theory, these challenges typically emerge during childhood and endure into adolescence and adulthood [1].
In response to the growing popularity of medical diagnosis, machine learning techniques [2] are used to help doctors make better decisions about a person's health, extensive efforts have been made to leverage various algorithms, including Naive Bayes, Support Vector Machine, Logistic Regression, K-Nearest Neighbors (KNN), Neural Networks, and Convolutional Neural Networks (CNNs) [3], to predict and analyze ASD-related issues across different age groups—children, teenagers, and adults. These endeavors have been detailed in a research survey paper.
These predictive models were assessed using the ABIDE dataset, which is openly accessible for research purposes. The study's findings underscore the effectiveness of CNN-based prediction models, which consistently outperformed other machine-learning techniques [4]. Notably, these models achieved impressive accuracy rates of 99.53%, 98.30%, and 96.88% for screening and diagnosing Autistic Spectrum Disorder in datasets representing adults, children, and adolescents, respectively.

Keywords: *Autism, Convolutional neural network (CNN); Artificial Neural Network (ANN); K- Nearest Neighbors (KNN); Logistic Regression (LR); Support Vector Machine (SVM) [5].*

Date of Submission: 03-09-2024 Date of Acceptance: 13-09-2024

I. Introduction

In recent years, the field of medicine has adopted a range of techniques and strategies to detect and predict various disorders. To make things more accurate, scientists are now working on creating computer programs that can learn and make predictions in the field of medicine [6].

In particular, there has been a growing interest among researchers in the field of neuroscience regarding brain disorders such as Autism-Spectrum Disorder (ASD). Despite the ongoing search for definitive biomarkers associated with ASD, neuroscience studies have shed light on the potential significance of the

corpus callosum and intracranial brain volume in its identification. Building upon these insights, authors have proposed machine-learning programs created to automatically find signs of ASD [7].

These algorithms have been subjected to rigorous evaluation and analysis using data sourced from the ABIDE Datasets. Our approach leverages functional Magnetic Resonance Imaging (fMRI), which offers enhanced insights into autism due to its ability to reveal aberrations in brain function.

This research is trying to help and add something valuable to the field advancement of ASD identification and understanding with the help of computer smarts, authors can do the thing mentioned to neuroimaging data [8].

A. Symptoms Of The Disease

- Individuals with ASD may exhibit a lack of sensitivity to pain.
- They might struggle to establish proper eye contact.
- Responses to auditory stimuli may be limited or inappropriate.
- Some individuals with ASD may not seek or desire cuddling or physical contact.
- Expressing themselves through gestures can be challenging for them.
- Interactions with others may be notably absent or difficult.
- Inappropriate attachment to objects can be observed.
- A preference for solitude may be evident.
- Echolalia, or the repetition of words or phrases, is a common trait.

People with ASD may also display a tendency for repetitive behaviors, including:

- Repetition of specific actions repeating the same words or phrases many times[9].
- Becoming upset when daily habits or schedules are disrupted [10].
- Developing a strong interest as they might really like certain things, like numbers or facts [11].
- Exhibiting lower awareness of environmental factors like light or noise [12].

Early detection and intervention are crucial in mitigating the signs of autism spectrum disorder and enhancing the goodness of life for individuals affected by Autism Spectrum disorder [13]. It is important to note that there is currently no medical test available for the direct detection of autism.

II. Dataset Description

Autism spectrum disorder (ASD) is described by certain differences in how a person thinks, feels, and behaves impairments in being social and friendly, as well as doing the same things over and over, sticking to specific patterns, and having strong interests in certain topics[14]. While before, ASD was thought to be uncommon, but now authors know it impacts more than 1 in 100 children[15]. In spite of continuous advancements in study, progress in identifying early-age diagnoses, optimal treatments, and outcome predictions for ASD has not kept pace with the urgency of the situation. This challenge primarily arises from the complications and diversity of ASD[16].

Addressing these problems necessitates the use of large- scale datasets[17]. However, individual research laboratories are often unable to acquire datasets of sufficient size to uncover the underlying neural mechanisms of ASD. To address this, the ABIDE is a place where they collect and share brain images to help study autism[23]. The initiative has undertaken the task of consolidating data from brain scans that show both how the brain works and its physical structure, collected from labs worldwide. The main goal is to speed up our understanding of the brain basis of autism[22]. The ABIDE project has two big collections, ABIDE I and ABIDE II[14]. They focus on helping scientists discover things and compare different samples of data.

Both of these collections have been assembled by aggregating datasets independently collected across more than 24 international brain imaging laboratories[11]. These datasets are provided to researchers across the globe, aligning following the principles of open science, that underlie initiatives like the International Neuroimaging Data-sharing Initiative[15]. For further details about these initiatives, please visit the collection-specific pages: ABIDE I and ABIDE II[16].

III. Methodology

A. Algorithms

Random Forest (RF)

Random Forest (RF) is an ensemble classification method followed by decision trees that employ a divide-and-conquer approach on several decision trees created using the supplied dataset, collectively referred to as a "forest" [8]. The RF algorithm operates in two main phases, as outlined below:

1. Selection of Random Samples: Initially, the algorithm randomly selects a few examples from the training dataset [13].

2. Construction of Decision Trees: For every training sample, the algorithm constructs individual decision trees [16].
3. Determination of Tree Count: The magnitude 'N' is chosen to specify the count of decision trees to be created in the forest [19].
4. Repetition of Sampling and Tree Construction: Steps 1 and 2 are repeated as needed to build the designated number of decision trees.
5. Classification of Test Samples: For each test sample, predictions are generated using each decision tree. The final class value for the test sample is determined through a majority voting process, where the most frequently predicted class among the decision trees is selected.

This approach allows Random Forest to harness the power of multiple decision trees, making it a robust and effective method for classification tasks.

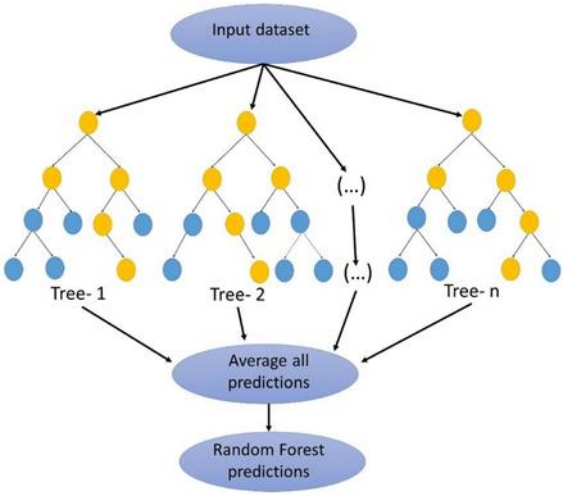


Fig. 1. Random Forest

Decision Tree (DT)

The decision tree is a straightforward and easily interpretable method[15]. This method's parameters are adjusted to yield three distinct types of trees:

1) Coarse Tree: This type of tree contains only a few leaves, allowing for coarse distinctions. While it enhances prediction robustness, it typically does not achieve high training accuracy.
2) Medium Tree: The medium tree strikes a balance with a moderate number of leaves.
3) Fine Tree: The fine tree, on the other hand, contains numerous leaves, enabling it to make finely detailed distinctions. However, this approach can be prone to overfitting.

These variations in decision tree structures serve different purposes in the context of machine learning, offering a range of trade-offs between prediction robustness and training accuracy. Noteworthy for its shorter training time in comparison to SVM and ME models.

Logistic Regression (LR)

Logistic regression operates on a dataset of independent variables, assessing the probability of a target variable, such as voting or not voting [1]. The outcome is in the form of probability, and the dependent variable's range spans from 0 to 1[4]. In logistic regression, odds, which represent the probability of success divided by the probability of failure, undergo a transformation using the logit formula [5]. The given formulas show the logistic function, also known as the log odds or the natural logarithm of odds [26]:

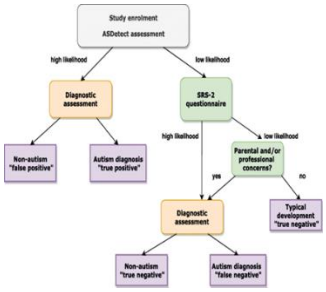


Fig 2: Decision Tree

Here, 'p' denotes the probability of instance 'x'[26]. During model training, for each instance ' x1, x2, x3, ... xn, ' the logistic coefficients are ' b0, b1, b2, ... bn[26].' The stochastic K-NN algorithm classifies data points based on their similarity to other data points (neighbors) within a training dataset[21]. It is a straightforward yet efficient method for the classification of new data data points[23]. After assessing the impact of various parameter configurations, such as the count of neighbors, distance weight, and distance measurement method, on the efficiency of k-NN classifiers, the authors gradient descent method is employed to estimate and update these coefficient values [26].

The coefficient values are now updated using the equation below:

$v = b_0x_0 + b_1x_1 + \cdots\cdots\cdots\cdots + b_nx_n$

Selected five k-NN variations for classifying our Self- stimulatory Behavior Dataset[2].

1. Fine k-NN: This variant achieves highly detailed separable distinctions between classes When the value of N is set to 1. It employs the Euclidean distance metric with equal distance weights[25].
2. Medium k-NN: The medium k-NN configuration provides a moderately separable group of distinct classes, utilizing 10 neighbors[25].
3. Coarse k-NN: Coarse k-NN produces great separation between the distinct classes, employing 100 neighbors [25].

Additionally, authors explored variations using different distance metrics:

1. Cosine k-NN: When using the cosine distance metric, this method yields medium distinctions between classes, with 10 neighbors and equal distance weights.
2. Weighted k-NN: With equal distance weights and 10 neighbors, the weighted k-NN approach also produces normal distinctions between classes, using Euclidean distance metric[16].

These variations in k-NN methods offer different levels of discrimination between classes, providing flexibility in addressing various classification tasks.

Naïve Bays (NB)

This approach falls under supervised machine learning and relies on the principles of probability. It is known for its efficient computation speed and predicting capabilities[4]. Naive Bayes (NB) is rooted in statistical concepts, where it predicts the possibility of a particular outcome[2],[3]. It is Now, the following equation is used to update the values of the coefficients:

$b = b + l \times (y - p) \times (1 - p) \times p \times x$

Initially, all coefficient values are set to 0, and 'y' represents the target variable for each training sample [2]. 'l' signifies the learning rate, and 'x' denotes the biased input for 'b0,' which is invariably set to 1. This process continues to update the coefficient values up to the point that it correctly forecasts the output during the training phase [1].

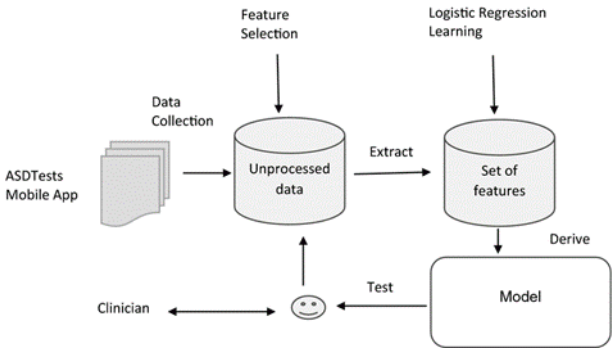


Fig 3: Flow of Logistic Regression

Support Vector Machine (SVM)

Support Vector Machine (SVM) is commonly employed in classification problems, SVM identifies the hyperplane that best distinguishes a given dataset into two classes [27]. The margin, which shows the separation of the hyperplane and the nearest training data point, is a key factor. [27] SVM's objective is to maximize this margin within the training data by identifying the most optimal separating hyperplane. Initially, the authors initiated our training using a linear Radial Basis Function (RBF) kernel and observed its effectiveness compared to a non-linear kernel.

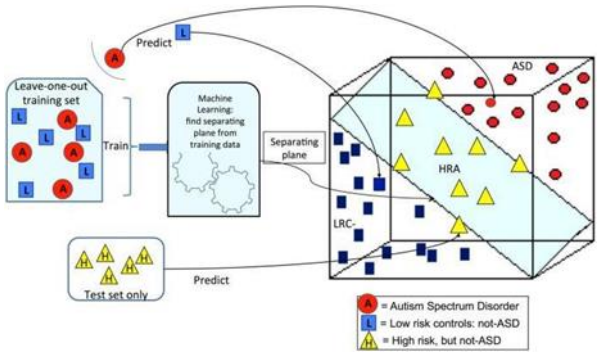


Fig 4: Flow of Logistic Regression

Convolutional Neural Network (CNN)

Convolutional neural network (CNN/ConvNet) represents a class of deep neural networks, primarily utilized while analyzing visual imagery [21]. While conventional neural networks often evoke thoughts of matrix multiplications, this isn't the case with ConvNet. Instead, ConvNet utilizes a unique method called convolution [20]. In mathematical terms, Convolution is a process that combines two functions to produce a third function. elucidates how the shape of one function is changed by the other [18].

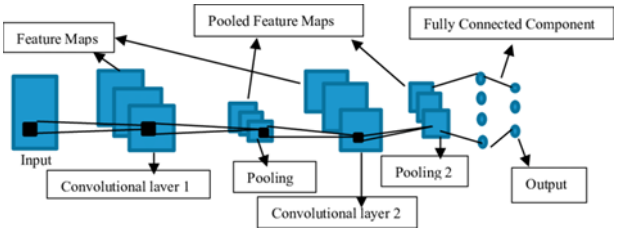


Fig 5: Convolutional Neural Network

IV. Comparison And Analysis

In this research, authors detected that they used computer programs (machine learning) to study Autism-Disorder. models on four distinct non-clinical ASD screening datasets. These datasets are publicly available from Kaggle, the UCI machine learning repository, and ABIDE I Datasets, covering various age groups: toddlers, children, adolescents, and adults. Authors evaluated the model performance using various assessment criteria for ASD identification. Comparing our results to recent studies in this field, our models outperformed other classifiers, particularly after addressing missing values in the Toddler's autism spectrum dataset.

Analysis on accuracy

Accuracy tells us how well a classifier can predict things in reality. A higher accuracy number means it's doing a better job of predicting with fewer misclassifications. Here are the accuracy scores for different classifiers on various datasets.

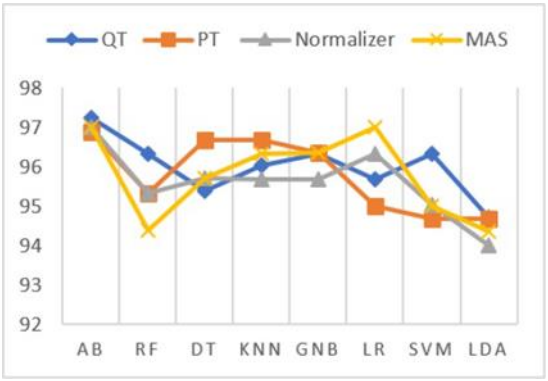


Fig 6: Comparison of accuracy

Accuracy of various ML algorithms on autism datasets are given below:

Algorithms	**Feature Scaling Techniques**			
	QT	***PT***	***Normalizer***	MAS
AB	97.95	96.89	97.02	97.02
RF	96.35	95.35	95.35	94.4
DT	95.39	96.68	95.71	95.71
KNN	93.03	96.88	95.70	96.35
GNB	98.34	96.37	95.70	96.37
LR	95.68	95.02	96.33	97.02
SVM	96.35	94.69	95.03	95.01
LDA	94.72	94.72	94.02	94.37

Analysis on Precision

Precision is a measure that tells us how often a classifier is correct when it predicts something as positive, and a higher precision value it shows that when something is actually true (true positives), it's good at not wrongly saying it's true when it's not (false positives). The values of precision for different classifiers on various he adjusted datasets are shown in the table below:

Algorithms	**Feature Scaling Techniques**			
	QT	***PT***	***Normalizer***	MAS
AB	94.02	93.86	94.57	94.23
RF	96.63	92.37	93.94	94.2
DT	94.60	92.71	92.71	92.31
KNN	93.39	92.17	91.06	93.77
GNB	95.68	93.41	94.73	94.88
LR	95.45	93.31	95.86	96.16
SVM	93.73	93.00	92.10	96.01
LDA	94.66	92.69	92.55	94.22

Precision of the different ML classifiers on ASD datasets:

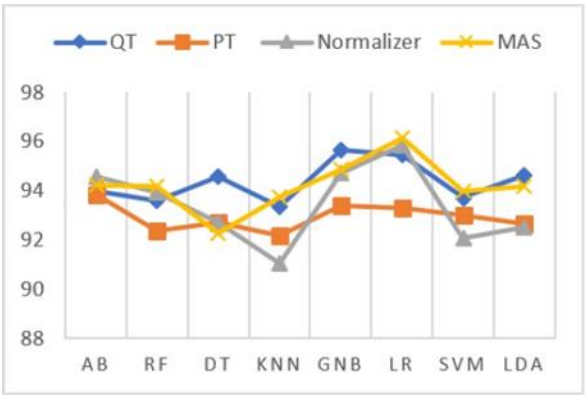


Fig 7: Comparison of precision

Analysis on Recall

Recall signifies the true positive rate, and when it's higher. the value indicates a high true positive count and having few incorrect rejections (false negatives) is good. A higher recall value means better results. Prediction. Below is the table presenting the recall values of various ML methods used on various datasets with adjusted features.

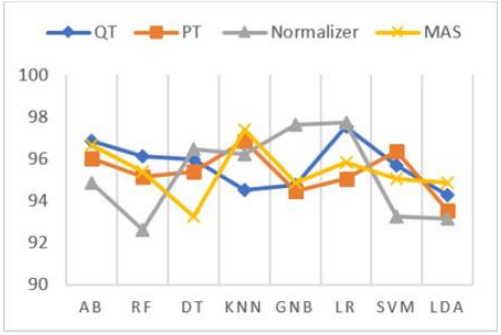


Fig 8: Comparison of Recall

Recall of the different ML classifiers on ASD datasets:

Algorithms	Feature Scaling Techniques			
	QT	***PT***	***Normalizer***	MAS
AB	98.84	96.05	94.88	96.64
RF	96.11	95.17	92.63	95.4
DT	96.00	95.39	96.48	93.25
KNN	94.51	96.87	96.21	97.38
GNB	94.78	94.49	97.63	94.86
LR	97.54	95.03	97.72	95.81
SVM	95.70	96.39	93.25	95.03
LDA	94.24	93.56	93.14	94.84

Analysis on ROC

Calculating the ROC value serves as an indicator of a classifier's effectiveness in distinguishing between positive and negative classes. Below, you'll find the ROC values for various machine learning classifiers diverse datasets that have been scaled to include specific.

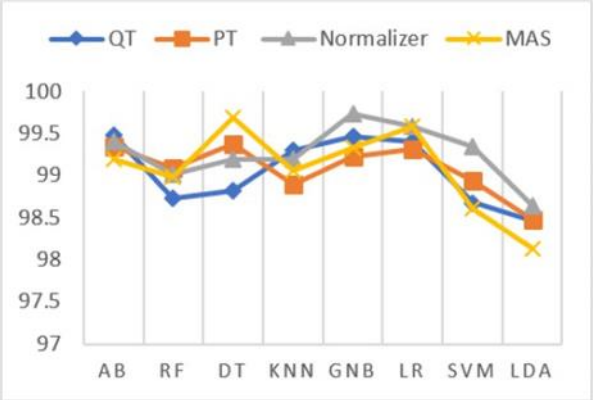


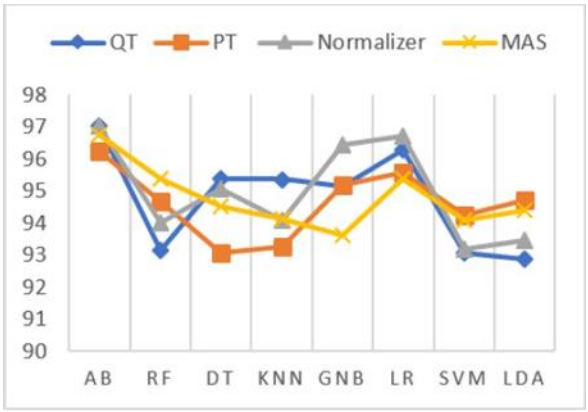
Fig 9: Comparison of ROC

ROC of various ML algorithms on autism datasets are given below:

Algorithms	Feature Scaling Techniques			
	QT	***PT***	***Normalizer***	MAS
AB	99.48	99.35	99.41	99.19
RF	98.73	99.09	99.02	98.99
DT	98.82	99.38	99.20	99.69
KNN	99.30	98.9	99.19	99.06
GNB	99.46	99.23	99.73	99.33
LR	99.41	99.31	99.59	99.58
SVM	98.68	98.94	99.34	98.61
LDA	98.47	98.48	98.64	99.13

Analysis on F1 – Score

The F1-score is like a report card that looks at both precision and recall together to give a single score, where a higher answer indicates better predictive performance. Here, you can find the F1-score values for various machine learning classifiers on a range of datasets that have been scaled to include specific features.



F1-score of various ML algorithms on autism datasets are given below:

Algorithms	**Feature Scaling Techniques**			
	QT	***PT***	***Normalizer***	MAS
AB	97.02	96.27	97.02	96.78
RF	93.15	94.89	94.00	95.38
DT	95.38	93.06	95.07	94.51
KNN	95.37	93.26	94.10	94.12
GNB	95.17	95.18	96.43	93.64
LR	96.29	95.58	96.74	95.38
SVM	93.06	94.27	93.21	94.11
LDA	92.90	94.72	93.45	94.41

V. Conclusion

Identifying autism disorder is a type of neurological state, is notably challenging, particularly in young patients. Magnetic Resonance Imaging (MRI) scans offer a way to detect significant changes in brain structure associated with ASD. In this research, authors explored three methods for building a model to identify ASD using image classification tools: Convolutional Neural Network (CN), and many such algorithms. Authors evaluated the success of these techniques for automated ASD detection on the ABIDE dataset, using accuracy as the performance metric.

The findings indicated that the CNN method worked the best, with a 95% accuracy rate. It's worth noting that exploring other modalities, such as EEG, speech analysis, or kinesthetic data, concurrently could further enhance ASD detection.