

**PREDICTION OF AUTISM SPECTRUM DISORDER IN THE EARLY STAGE BASED
ON MACHINE LEARNING APPROACH**

BY

ASIFA AFSARI HEMU

181-15-1716

AND

SABBIR AHAMED

181-15-1996

AND

MD. NAIM JANNAT NIPU

181-15-1747

This Report Presented in Partial Fulfillment of the Requirements for the Degree of
Bachelor of Science in Computer Science and Engineering

Supervised By

Md. Mahfujur Rahman

Sr. Lecturer

Department of CSE

Daffodil International University

Co-Supervised By

Tania Khatun

Sr. Lecturer

Department of CSE

Daffodil International University

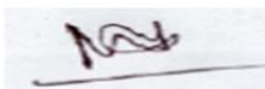


**DAFFODIL INTERNATIONAL UNIVERSITY
DHAKA, BANGLADESH
JANUARY 17, 2022**

APPROVAL

This Project titled “**Prediction of Autism Spectrum Disorder in The Early Stage based on Machine Learning Approach**”, submitted by “**Asifa Afsari Hemu**”, Id No: 181-15-1716, “**Sabbir Ahamed**”, Id No: 181-15-1996 and “**Md. Naim Jannat Nipu**”, Id No: 181-15-1747 to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 17 January, 2022

BOARD OF EXAMINERS



Dr. Md. Ismail Jabiullah

Professor

Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University



Internal Examiner

Narayan Ranjan Chakraborty

Assistant Professor

Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University



Internal Examiner

Dr. Mohammad Shorif Uddin

Professor


Department of Computer Science and Engineering
Jahangirnagar University

External Examiner

DECLARATION

We hereby declare that, this project has been done by us under the supervision of **Md. Mahfujur Rahman, Sr.Lecturer & Tania Khatun, Sr.Lecturer Department of CSE** Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

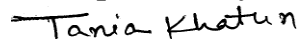
Supervised by:


15.1.2022

Md. Mahfujur Rahman

Sr.Lecturer
Department of CSE
Daffodil International University

Co-Supervised by:



Tania Khatun

Sr.Lecturer
Department of CSE
Daffodil International University

Submitted by:



Asfia Afsari Hemu

ID: 181-15-1716
Department of CSE
Daffodil International University



Sabbir Ahamed

ID: 181-15-1996
Department of CSE
Daffodil International University



Md. Naim Jannat Nipu

ID: 181-15-1747
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First, we express our heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the final year project/internship successfully.

We really grateful and wish our profound our indebtedness to **Supervisor of Md. Mahfujur Rahman, Sr.Lecturer** Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of “*Machine Learning*” to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts and correcting them at all stage have made it possible to complete this project.

We would like to express our heartiest gratitude to **Md. Mahfujur Rahman, Sr.Lecturer** and Head, Department of CSE, for his kind help to finish our project and also to other faculty member and the staff of CSE department of Daffodil International University.

We would like to thank our entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, we must acknowledge with due respect the constant support and patients of our parents.

ABSTRACT

Autism Spectrum Disorder (ASD) is a set of neurological impairments which are incurable but can be improved with early treatment. We obtained slightly earlier detected ASD datasets pertaining to children and highly processed the dataset as needed. Various ML approaches were applied to the collected dataset and compared their performance based on accuracy, precision, recall, f-measure, log loss, kappa statistics, and MCC. We found that DT provides the best performance with 100% accuracy. Then different FSTs methods were applied to the dataset to show the importance and identify the significant features responsible for ASD. The study's findings indicate that, when properly tuned, machine learning approaches can offer accurate forecasts of ASD status. According to the findings, the suggested model has the ability to diagnose ASD in its early phases.

Keywords: *ASD, Neurodevelopmental, Classifier, Logistic Regression, Random Forest.*

TABLE OF CONTENTS

CONTENTS	PAGE
Board of examiners	ii
Declaration	iii
Acknowledgements	iv
Abstract	v
Table of Contents	vi-vii
CHAPTER 1: Introduction	1-7
1.1 Problem Outline	1-6
1.2 Motivation	6
1.3 Objective	7
1.4 Contribution	7
CHAPTER 2: Background Study	8-9
CHAPTER 3: Materials & Methods	10-17
3.1 Data Collection	10-11
3.2 Methods	12-17
3.2.1 Data processing	12
3.2.2 Supervised Machine Learning Analysis	12-15
3.2.3 Performance Evaluation Criteria	16
3.2.4 FST Methods	17

CHAPTER 4: Results & Discussion	18-26
1.5 Statistical & Exploratory Data Analysis	18-21
1.6 Performance Analysis	22-26
CHAPTER 5: Conclusion & Future Scope	27
REFERANCE	28-30

LIST OF FIGURES

FIGURES	PAGE NO
Figure 1: visualization of ASD increasing density	2
Figure 2: Type of ASD	4
Figure3: Workflow of the Research	10
Figure 4: Heat map to show the hypothesis test based on p value.	21
Figure 5: Visualization of Accuracy	22
Figure 6: Visualization of Performance analysis	23
Figure 7: Area under Receiving Operating Characteristics (ROC) Curve. & Area under Precision – Recall (PRC) Curve (A) ROC Curve, (B) PRC Curve	24

LIST OF TABLES

TABLES	PAGE NO
Table 1: Features description	11
Table 2: The explanation of all the performance evaluation metrics.	16
Table 3: The explanation of all the performance evaluation metrics.	17
Table 4: Statically & Exploratory Data Analysis	18-20
Table 5: Performance Measurement Of Different ML Approaches	22
Table 6: Results of FST approach	25

CHAPTER 1

Introduction

1.1 Problem Outline

Autism spectrum disorder (ASD) is a neurodevelopmental illness characterized by communication and behavioral impairments [1]. Autism spectrum disorder affects around one out of every 54 children, according to the CDC's Autism and Developmental Disabilities Monitoring (ADDM) Network (ASD). ASD has been reported in people of various races, ethnicities, and socioeconomic backgrounds. Boys are four times as likely as girls to have ASD. According to parent reports, from 2009 to 2017, almost one in every six (17%) children aged 3–17 years were diagnosed with a developmental disability. Among these were autism, attention deficit/hyperactivity disorder, blindness, and cerebral palsy. In 2000 year, there are 6.7% children are caused by ASD which is about 1 in 150 children. In 2016 year, there are 18.5% children are caused by ASD which is about 1 in 54 children. So, it is increased day by day [2].

While ASD is a lifelong condition, the severity of functional impairment caused by these problems differs across children with ASD. Before a kid turns one year old, parents or physicians can detect early indications of this condition. Symptoms, on the other hand, generally become more constant by the time a kid is 2 or 3 years old. In other circumstances, the functional impairment caused by ASD may be moderate and not noticeable until the kid enters school, at which point their impairments may become evident when they are among their companions [3].

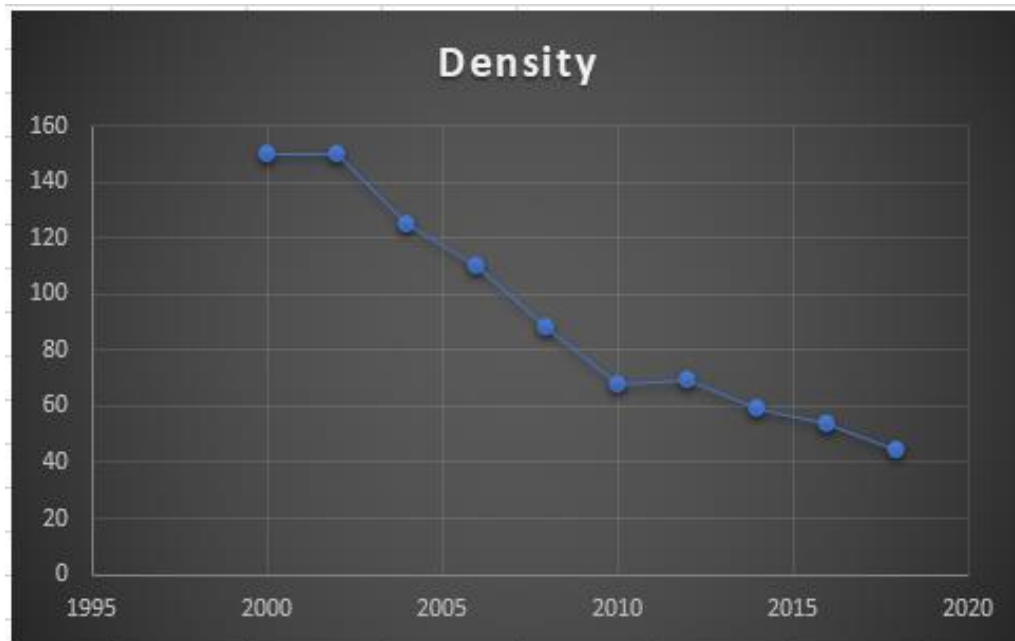


Figure 1: visualization of ASD increasing density

Children with ASD exhibit difficulty with social contact and consultation, have restricted interests, and prefer repetitive activities [1]. Not all people with ASD display all of these behaviors, but the vast majorities do. Sharing gratitude for items or activities with others by pointing out or showing them things is uncommon. Making infrequent or erratic eye contact. Not looking at or listening to others. Having trouble with the back and forth of communication. Frequently chatting for an extended period of time on a preferred subject without realizing that others are not engaged or without giving others an opportunity to react. Exhibiting odd habits or repeating particular traits echolalia is the practice of repeating words and expressions [1]. Getting a long strong interest in specific things, such as statistics, details, or facts. Having too restricted interests, such as moving things or parts of objects. Being upset by little changes in a routine. Being more or less sensitive to sensory input such as light, noise, clothing, or temperature than others. According to research, genes can interact with environmental factors to alter development in ways that contribute to ASD. Having a sibling with ASD, older parents, specific genetic disorders, and persons with illnesses such as Down syndrome, birth weight is quite low are all risk factors of ASD [1].

Having a long-term, great interest in a certain subject, such as data, details, or facts. Excessively limited interests, like as moving things or portions of objects. Distress caused by little changes in one's routine. Being more or less sensitive to sensory input than other people, such as light, noise, clothing, or temperature. Genes can combine with environmental variables to alter development in ways that contribute to ASD, according to research. Having an ASD sibling, older parents, particular genetic flaws, persons with conditions such as Intellectual impairment, and having a poor body weight increase are all risk factors for ASD [1].

Because there is no diagnostic method for autism spectrum disorder (ASD), including a blood test, detecting the condition might be challenging. When diagnosing a youngster, doctors mainly consider their cognitive history and behavior [35].

Early indications of ASD, based on the National Institute of Mental Health, are including [11]:

- a. There seems to be little eye contact.
- b. Not correctly pointing out or showing others what you're enjoying about goods or activities.
- c. Adults have a hard time getting their attention, and children have a hard time responding to them.
- d. Conversing for a long time without evaluating other people's interest
- e. A voice with a monotone voice.
- f. A habit of recurring certain actions, words, or sentiments.
- g. A strong desire for some things.
- h. Irritated by variations in regularity.
- i. Sleeping difficulties.

The majority of people diagnosed ASD have other traits. These could include the following [35]:

- a. Linguistic skills that are behind schedule.
- b. Movement's abilities that are delayed.

- c. Impaired cognitive and learning abilities.
- d. Seizures are a type of epilepsy.
- e. Irregular sleeping as well as having to eat patterns.
- f. Irritable bowel syndrome (IBS) is a condition that affects the digestive
- g. Mood swings or emotional responses that are out of the ordinary.
- h. Anxiety, tension, or excessive worry is all examples of anxiety.
- i. Fearlessness or a higher level of fear than predicted.

Since 1994 till 2013, autism has been divided into five distinct groups [36]. Asperger's syndrome, Childhood Disintegrative Disorder, Autistic Disorder, Pervasive Developmental Disorder, Ret Syndrome.

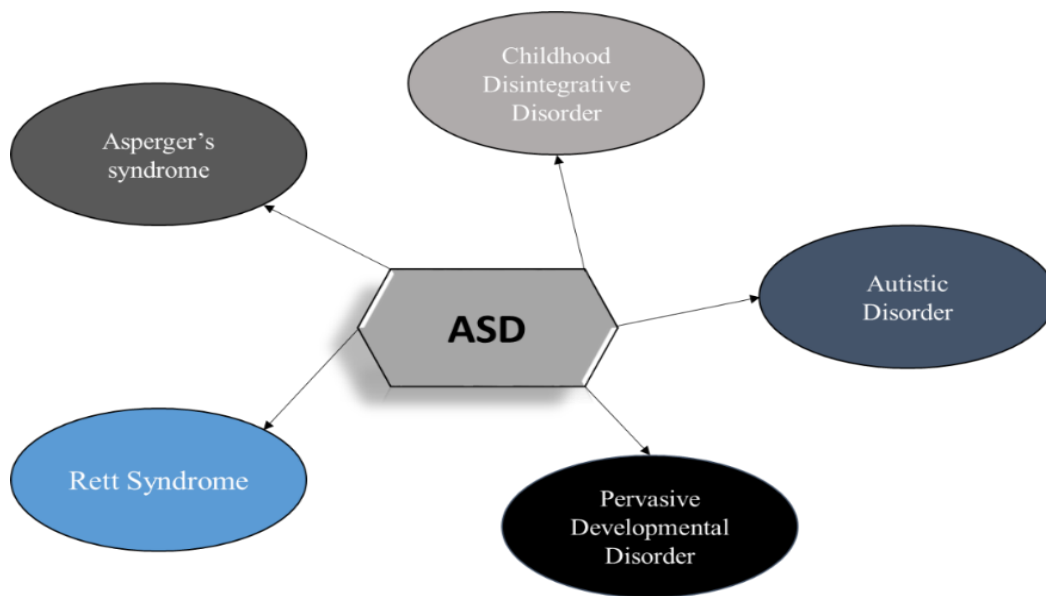


Figure 2: Type of ASD

Summary of five different types of autism in brief:

A. Asperger's syndrome

Asperger's syndrome belongs to the autism spectrum disorder category of neurodevelopmental diseases, despite the fact that it is no longer a recognized diagnosis (ASD). According to doctors, what was once known as Asperger's syndrome is now

classified as a moderate form of autism. You've probably seen that this is known as level 1 ASD. The primary difference between Asperger's syndrome and autism spectrum disorder is that persons with Asperger's are more likely to [37]:

- a. Display milder autistic symptoms.
- b. Possess great linguistic abilities and have no speech impediments.

B. Childhood Disintegrative Disorder

Children with childhood disintegrative disorder develop normally until they are three or four years old. The youngsters then lose linguistic, physical, social, and other previously gained skills over the period of several months. Childhood disintegrative disorder is a developmental impairment that is classified as part of the autism spectrum [38].

C. Autistic Disorder

Autism spectrum, this word sits between Asperger's and PDD-NOS. The symptoms are the same as previously, but on a much larger scale [39].

D. Pervasive Developmental Disorder not otherwise specified PDD(NOS)

PDD is also known as Unusual Autism since it is diagnosed when a child exhibits a number of signs of autism but not all of them. PDD (NOS) is most frequently diagnosed if a child has speech difficulties and demonstrates specific repetitive actions [4].

E. Rett Syndrome

Rett Syndrome is a rare and dangerous disorder caused by a chromosomal X deficit that affects mostly women. Rett Syndrome is characterized by normal periods of development followed by a progressive loss of capacities, most notably speech and intentional hand motions [4].

As children with autism develop into teens and young adults, they may find it difficult to build and maintain friendships, participate in discussions with adults, and understand

what is expected of them at school or at work. If they exhibit founder disorders including cognitive impairment, attention deficit disorder, feelings of despair, or behavior disorder, they may be sent to a doctor [2]. Children with ASD should be monitored, screened, evaluated, and diagnosed as soon as possible to ensure that they receive the help and support they need to reach their full potential [40].

The American Academy of Pediatrics (AAP) recommends psychological and developmental assessments for all children at certain ages at routine well-child visits [2]:

- A. Nine-month period.
- B. Eighteen-month period.
- C. Thirty months.

There are three degrees of autism, which help with diagnosis and indicate how much support the individuals will require [36].

Autism's three levels:

- a. A level one diagnostic is the highest cognitive and requires minimal amount of assistance.
- b. Level two is severe and necessitates a significant amount of assistance.
- c. Third level is perhaps the most difficult and will almost certainly necessitate extensive assistance.

1.2 Motivation

There are a handful of youngsters with ASD in our immediate vicinity. We can't take effective action or care of them until we get confirmation that they have ASD. Because the symptoms are less severe in the early stages, it is difficult to detect. When it's a serious matter, it typically draws my attention. It is really difficult to reduce ASD at the time. Although ASD cannot be cured, it is feasible to moderate and control it if it is discovered early on. Clinical diagnosis, on the other hand, is too expensive for the majority of Bangladeshis. From this standpoint, my goal is to create a Machine Learning model that can identify ASD at an early stage with high accuracy. It will also be cost-effective for everyone.

1.3 Objective

In today's environment, autism spectrum disorder (ASD) is a dangerous disorder. It is a type of neurodevelopmental condition that impairs a person's capacity to engage with others. Autism is a long-standing issue in our country, and we've been striving to address it for a long time. The suggested technology will be able to confirm whether or not a patient has ASD at an early stage. To achieve the optimum efficiency and accuracy, an effective machine learning approach will be developed. Patients can use the method to identify ASD at a cheaper cost.

1.4 Contribution

From the foregoing explanation, it is clear that the model's efficacy in predicting ASD at an early stage may be improved. Even yet, by enhancing the present methods, it is feasible to achieve more accuracy. In that light, the study's goal is to present a newly designed machine learning model that can diagnose ASD in children at an early stage. In comparison to the previous offered models in this part, our proposed model delivered more accurate and efficient results.

CHAPTER 2

Background Study

A number of studies have been conducted in the last ten years to develop a model to detect ASD using classification and other ML methods for children. Vakadkar et al. 2021 [5] employed five machine learning models, "Random Forest Classifier, Logistic Regression, Naive Bayes, Support Vector Machines, and KNN," to categorize individual participants as having ASD or not having ASD based on age, gender, ethnicity, and other variables. According to their findings, Logistic Regression delivered the highest accuracy for the dataset they used. Thabtah et al. 2020 [4] proposed Rules Machine Learning (RML), a novel machine learning approach that gives users a knowledge foundation of rules for understanding the basic reasons of categorization and diagnosing ASD symptoms. Li et al. 2019 [6] used the ABIDE database to identify 6 personality characteristics in 851 people and trained and evaluated Machine Learning models using a cross-validation technique. This was utilized to distinguish between ASD patients. T. Akter et al [9] proposed a machine learning-based technique for the early detection of ASD. They used nine machine learning classifiers on four datasets based on age category, such as toddler, child, adult, and adolescent. They proposed four classifiers for four age groups. Hyde et al. 2019 [8] conduct a literature study on supervised categorization in autism spectrum disorder. The author examined 45 different papers. Naive Bayes, SVM, ADtree, and Random Forest were the algorithms used. SVM and ADtree techniques for data mining for ASD were the most commonly utilized algorithms. Stevens et al. 2019 [7] employ Gaussian Mixture Models and Hierarchical Clustering to detect ASD behavioral features. The advantage of utilizing machine learning is that behavioral subtypes and their connections may be discovered. Moon et al. 2019 [10] intend to conduct a systematic review and a meta-analysis to synthesize the data on the performance of machine learning algorithms in diagnosing ASD. They used a subgroup meta-analysis of structural magnetic resonance imaging (sMRI) that yielded an 83 percent sensitivity and an 84 percent specificity. Abbas et al. 2018 [11] presented a machine learning technique for early autism prediction by integrating a survey and home video screening.

They used two previously trained algorithms to detect autism. They demonstrate a significant improvement in accuracy over conventional screening approaches in terms of AUC, sensitivity, and specificity. Duda et al. 2016 [12] utilized forward feature selection and under sampling to discriminate between autism and ADHD using a 65-item Social Responsiveness Scale. Bone et al. 2016 [13] used a support vector machine for the same purpose and achieved 89.2 percent sensitivity and 59 percent specificity (SVM). In their study, 1264 people with ASD and 462 people without ASD participated. However, because of the wide age range, their study was not approved as a screening approach for persons of different ages. The Alternating Decision Tree (AD Tree) was utilized by Wall et al. 2012 [14] to reduce screening time and quickly find ASD features. Using data from 891 people, they employed the Autism Diagnostic Question and Answer Session, revised (ADI-R) approach and obtained a high degree of accuracy. The exam, on the other hand, was restricted to children aged 5 to 17, and it failed to predict ASD for various age groups.

CHAPTER 3

Materials & Methods

To conduct this study, we used Python (Version 3.8.5) for machine learning, statistical and exploratory data analysis. The study was conducted in several steps, which are mentioned in figure 3.

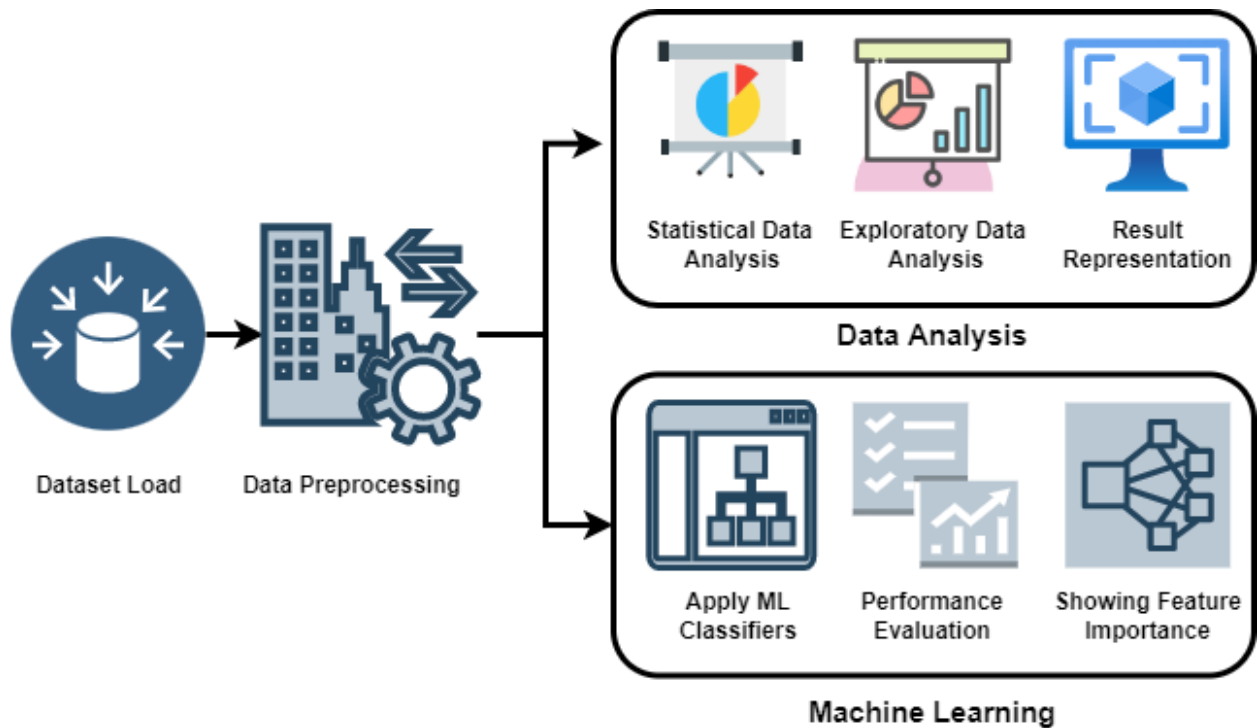


Figure 3: Workflow of the Research

The figure 3 represents all the steps undertaken to conduct the study sequentially. The overall workflow and steps are found in this figure.

3.1 Data Collection

In this work, we used a ASD dataset to build our expected model. The ASD dataset was gathered from Kaggle. This dataset has 21 characteristics. The dataset comprises 292 patient records, comprising 208 men and 84 females of various ages, with 141 patients being positive and 151 patients being negative. There are 103 male patients and 38

female patients among the ASD patients. A brief overview of all the features is represented in Table 1.

TABLE 3.: FEATURES DESCRIPTIONS

Feature	Type	Description
Age	Numeric	Age of child
Gender	String	Male(M) and Female(F)
Ethnicity	String	List of common ethnicities in text format
Born with jaundice	Boolean	Child was born with jaundice or not
Family member with PDD	Boolean	If any immediate family member has a PDD
Relation	String	Parent, medical staff, self, caregiver, clinician, etc.
Country of residence	String	List of countries in text format
Used the screening app before	Boolean	If the user has used a screening app
Screening Method Type	Numeric	Type of screening methods chosen based on age category
A1_Score: Answer of Q1	Boolean	Mention in the Table 2 for details Q1
A2_Score: Answer of Q2	Boolean	Mention in the Table 2 for details Q2
A3_Score: Answer of Q3	Boolean	Mention in the Table 2 for details Q3
A4_Score: Answer of Q4	Boolean	Mention in the Table 2 for details Q4
A5_Score: Answer of Q5	Boolean	Mention in the Table 2 for details Q5
A6_Score: Answer of Q6	Boolean	Mention in the Table 2 for details Q6
A7_Score: Answer of Q7	Boolean	Mention in the Table 2 for details Q7
A8_Score: Answer of Q8	Boolean	Mention in the Table 2 for details Q8
A9_Score: Answer of Q9	Boolean	Mention in the Table 2 for details Q9
A10_Score: Answer of Q10	Boolean	Mention in the Table 2 for details Q10
Scoring Result	Numeric	Mention in the Table 2 for details
ASD	Boolean	Child with ASD

3.2 Methods

3.2.1 Data processing

Data preprocessing is required for any machine learning or data mining strategy, because the effectiveness of a machine learning methodology is dependent on how effectively the dataset is prepared and structured. We detected missing value of the collected dataset and dropped them. We employed Interquartile Range (IQR) to detect outliers and extreme values. The IQR is a way for measuring a dataset's variability around the median. The outlier is a data point that goes outside of the expected range of the data and may be identified. presumed to be attributable to recording mistakes or other irrelevant occurrences for the purposes of the investigation [15]. To obtain a better analytical or statistical outcome, such outliers must be removed using machine learning (ML) or data mining approaches [16]. Data is divided into three quartiles for outlier detection: Q_3 , Q_2 , and Q_1 . The data boundaries in this case are Q_1 and Q_3 . $IQR = Q_3 - Q_1$ was used to determine the value of IQR . Then, using the following equations [17], the lower boundary B_l and upper boundary B_u were calculated:

$$B_l = Q_1 - 1.5 * IQR$$

$$B_u = Q_3 + 1.5 * IQR$$

An outlier is defined as a result that is less than B_l but larger than B_u . To balance the unbalanced dataset, the synthetic minority oversampling method (SMOTE) was used.

3.2.2 Supervised Machine Learning Analysis

In this work, four (04) supervised learning methods were used and compared to select a proper ML classification method to build the desired model. The identified training dataset is used first and foremost in supervised machine learning algorithms to train the ML model. This approved model is subsequently placed into a non-labeled testing dataset in order to test and evaluate the most perfect ML methods for the expected model [25]. The corresponding subsection provides a brief summary of these suggested supervised machine learning techniques for disease diagnosis.

A. K-nearest neighbor (KNN)

One of the most fundamental and simplest classification techniques [21,22] or statistical teaching approaches [23] is K closest neighbors. K is the number of nearest neighbors used, which can be explicitly set in the objects constructor or estimated using the upper limit offered by the stated value [24]. As a result, related examples are categorized similarly [25], and a new instance is classified by assessing its similarity to each of the existing cases [26]. When an unknown sample is received, the closest neighbor algorithm searches the pattern space for the k training examples that are next to the unknown substance. Predictions from numerous neighbors based on their distance may be generated from the test instance, and two distinct strategies are offered to convert the distance into a weight [23,27]. The approach has a lot of advantages, including being analytically manageable and easy to execute [23]. The algorithm performs based on different distance functions like: Minkowski Distance, Manhattan Distance, Euclidean Distance. In this study, Minkowski Distance function has been used. The Minkowski Distance for two points $U (u_1, u_2, \dots, u_n)$ and $V (v_1, v_2, \dots, v_n)$ can be represented by the following equation, here q demonstrates the order of the Minkowski Distance.

$$distance(U, V) = \left(\sum_{i=1}^n (|u_i - v_i|)^q \right)^{(1/q)}$$

B. Multilayer Perception (MLP)

A multilayer perceptron is a well-known neural network-based classification technique that consists of three or more layers: an input layer, an output layer, and one or more hidden layers between the input and output layers [28]. Each layer has a number of 'neurons' that connect all of the layers. MLP is a global multivariate non-linear mappings calculator derived from training data's capacity to learn and generalize [31] via the use of backpropagation learning methods [29]. MLP classifiers are built by specifying appropriate input variables and network types, relevant data pre-processing and partitioning, network infrastructure configuration, success parameter specification, training algorithm specification (optimization of relation weights), and finally model

evaluation [30]. In this study, the default configuration produced the best findings from this predictor. The error of the k^{th} output node in the data point n can be represented by the equation below where d and c represent the actual and predicted values respectively.

$$e_k(n) = d_k(n) - c_k(n)$$

C. Random Forest (RF)

Random Forest is one of the most well-known and effective machine learning techniques. Bagging or Bootstrap Aggregation is a form of machine learning algorithm. The bootstrap is a powerful statistical approach for estimating a value from a sample of data, such as the mean. A large number of data samples are collected, the mean is established, and then all of the mean values are averaged to offer a more accurate prediction of the real mean value [18]. Bagging uses a similar method, but instead of computing the mean of each data sample, decision trees are frequently used. Several training data samples are analyzed here, and models are developed for each data sample. When a forecast is required for any data, each model generates a prediction, which is then averaged to offer a more accurate approximation of the real output value [18]. The algorithm works based on following steps:

Step-1: Select random K data points from the training set.

Step-2: Build the decision trees associated with the selected data points (Subsets).

Step-3: Choose the number N for decision trees that you want to build.

Step-4: Repeat Step 1 & 2.

Step-5: For new data points, find the predictions of each decision tree, and assign the new data points to the category that wins the majority votes.

D. Decision Tree (DT)

One of the first and most well-known machine learning techniques is the decision tree (DT). A decision tree depicts the decision logics, or the tests and results for classifying data items into a tree-like structure. The nodes of a DT tree may contain several levels,

with the first or top-most node referred to as the root node. Each internal node denotes a test on one or more input variables or features. Based on the test result, the classification algorithm branches to the appropriate child node, and the process of testing and branching is continued until it reaches the leaf node [19]. The leaf or terminal nodes reflect the possible outcomes. DTs have been demonstrated to be simple to read and understand, and they are being used in a variety of medical diagnostic processes. When traversing the tree to categorize a sample, the results of all tests at each node along the route will provide enough information to make a prediction about its class [20]. The decision of splitting the data is controlled by entropy and which can be defined by the equation below, where p_j is the probability of the j^{th} class.

$$E(S) = \sum_{j=1}^n -p_j \log_2 p_j$$

Different versions of Decision Tree are available as for instance: ID3, C4.5, CART, etc.

3.2.3 Performance Evaluation Criteria

TABLE 2: THE EXPLANATION OF ALL THE PERFORMANCE EVALUATION METRICS.

Evaluation Criteria	Explanation	Formula
Accuracy	Accuracy is defined as the proportion of properly categorized cases. [19,20].	$Acc = \frac{TP + TN}{TP + TN + FP + FN}$
Precision	When we need to be really sure in our forecast, precision is an acceptable evaluation parameter finding. The ratio of True Positives to all Positives is defined as precision. [19,20].	$Precision = \frac{TP}{TP + FP}$
Recall	The recall measures how successfully our model detects True Positives.	$Recall = \frac{TP}{TP + FN}$
F-Measure	The F1 Score is calculated as the weighted average of Precision and Recall. [19,20].	$F = \frac{2 * Precision * Recall}{Precision + Recall}$
Log Loss	Log-loss is a useful performance indicator where the model output represents the likelihood of a binary result. The log-loss measure considers estimate trust while choosing how to penalize incorrect categorization. [20].	$L_{(log)}(y, p) = (y \log(p) + (1 - y) \log(1 - p))$

All of the performance evaluation criteria are listed in Table 2 together with their functional mathematical equation. Table 2 also includes a definition and a description of the definitions and their objectives. These were the key criteria utilized in this study to assess all of the classification algorithms and choose the best performing algorithm for early detection of ASD.

3.2.4 FST Methods

TABLE 3: THE EXPLANATION OF ALL THE PERFORMANCE EVALUATION METRICS.

CFST	Abbreviation	Explanation	Formula
Correlation based Feature Subset Selection	CFSSE	It assesses the value of a subset of characteristics by taking into account each feature's unique predictive capacity as well as the degree of redundancy between them.[32]	F_s $= \frac{N * ra}{N + N(N - 1)r_n}$
Gain Ratio based Attribute Evaluation	GRAE	It validates the value of a feature by calculating the gain ratio in relation to the class.[33]	$GR(C, A)$ $= (H(C)_H(C A)) / H(A)$
Info Gain based Attribute Evaluation	IGAE	It analyzes the value of a feature by calculating the information gain in relation to the class.[33]	$IG(C, A)$ $= (H(C)_H(C A))$
ReliefF based Attribute Evaluation	RFAE	It determines the value of an attribute by repeatedly sampling an instance and comparing the value of the provided attribute to the value of the nearest instance of the same and other classes.[34]	R_x $= P(\text{diff } X \text{diff class}) - P(\text{diff } X \text{same class})$

Table 3 shows all of the feature selection techniques (FST) utilized in this study to rank and highlight the value of all of the aspects used to diagnose ASD in children at an early stage. In this study, four different types of FST approaches were used. Table 3 includes a list of all of them, as well as their functional and mathematical equations, definitions, and descriptions.

CHAPTER 4

Results & Discussion

4.1 Statistical & Exploratory Data Analysis

This statistical study identifies the number of positive and negative patients for each binary feature, as well as the number of normal and ASD positive patients for each group. For numerical characteristics, the mean and standard deviation (STD) are determined. In addition, the p value for each feature is determined in relation to the target feature. Table 4 shows the statistical outcome of all the characteristics in relation to the goal feature.

TABLE 4: STASTICAL & EXPLORATORY DATA ANALYSIS

Categorical Feature					
Features	Category	All Patients N=248(%)	Patient's condition		P Value
			Positive	Negative	
Gender					<0.001
	Male	208 (71.23%)	105(50.48%)	103(49.52%)	
	Female	84 (28.77%)	46(54.76%)	38(45.24%)	
Born with Jaundice					<0.001
	Yes	292(100.00%)	151(50.71%)	141(48.29%)	
	No	0(0)	0(0)	0(0)	
A1					<0.001
	Yes	185 (63.36%)	68(36.76%)	117 (63.24%)	
	No	107 (36.64%)	83(77.57%)	24(22.43%)	
A2					0.215
	Yes	156(53.42%)	64(41.03%)	92(58.97%)	

	No	136 (46.58%)	87(63.97%)	49(36.03%)	
A3					5.282
	Yes	217(74.32%)	87(40.09%)	130(59.91%)	
	No	75 (25.68%)	64(85.33%)	11(14.67%)	
A4					0.098
	Yes	161(55.14%)	42(26.09%)	119(73.91%)	
	No	131 (44.86%)	109(83.21%)	22(16.79%)	
A5					5.282
	Yes	217(74.32%)	88(40.55%)	129(59.45%)	
	No	75 (25.68%)	63(84.00%)	12(16.00%)	
A6					1.058
	Yes	208(71.23%)	80(38.46%)	128(61.54%)	
	No	84 (28.77%)	71(84.52%)	13(15.48%)	
A7					0.003
	Yes	177(60.62%)	72(40.68%)	105(59.32%)	
	No	115(39.38%)	79(68.70%)	36(31.30%)	
A8					0.741
	Yes	145(49.66%)	43(29.66%)	102(70.34%)	
	No	147 (50.34%)	108(73.47%)	39(26.53%)	
A9					0.804
	Yes	144(49.32%)	39(27.08%)	105(72.92%)	
	No	148 (50.68%)	112(75.68%)	36(24.32%)	
A10					1.118
	Yes	212(72.60%)	81(38.21%)	131(61.79%)	
	No	80 (27.40%)	70(87.50%)	10(12.50%)	
Family member with PDD					<0.001
	Yes	292(100.00%)	151(51.71%)	141(48.29%)	
	No	0(0)	0(0)	0(0)	

Used the screening app before					<0.001
	Yes	292(100.00%)	151(51.71%)	141(48.29%)	
	No	0(0)	0(0)	0(0)	
Numerical Feature					
Features	All Patients		Patient's Condition		P Value
			Positive	Negative	
	Mean (STD)	Mean (STD)	Mean (STD)		
Age	6.354167 (2.365456)	6.539568 (2.494243)	6.181208 (2.233207)	0	

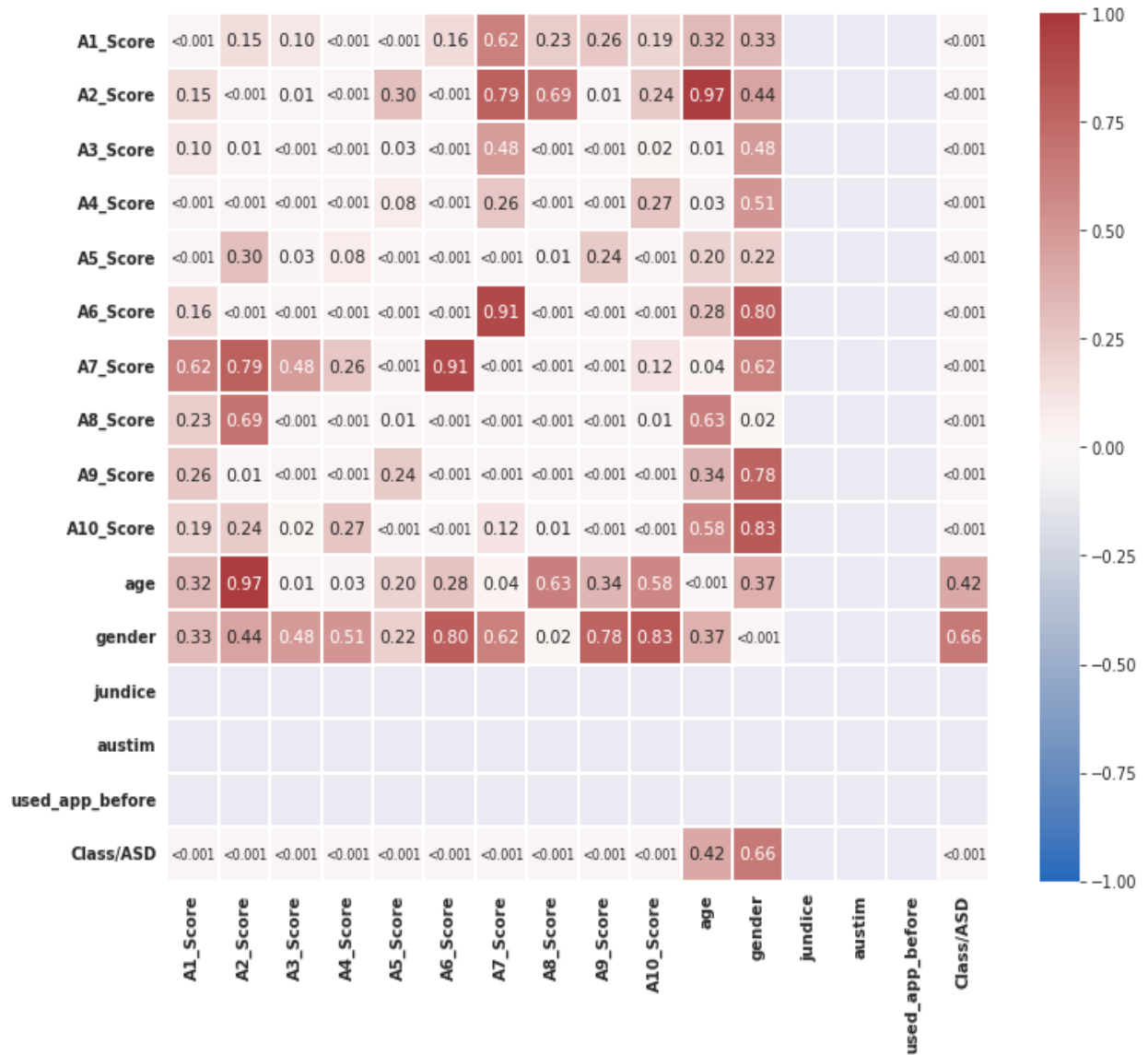


Figure 4: Heatmap to show the hypothesis test based on p value.

The hypothesis testing result is depicted in Figure 4 among all the attributes. The attributes jaundice, autism, and used app before have only positive values in this dataset; there is no negative value. As a result, these characteristics have no p value and their cell is empty. The characteristics with a p value less than 0.001 were statistically significant. Significant characteristics can also be defined as a p value less than 0.05.

4.2 Performance Analysis

TABLE 5: PERFORMANCE MEASUREMENT OF DIFFERENT ML APPROACHES

Algorithms	Accuracy	Precision	Recall	F-measure	Log-Loss	Kappa Statistics	MCC
KNN	0.92	0.92	0.92	0.92	2.76	0.84	0.84
RF	0.90	0.90	0.90	0.90	3.22	0.81	0.81
MLP	0.92	0.92	0.92	0.92	2.76	0.84	0.84
DT	0.95	0.95	0.95	0.95	1.84	0.89	0.89

Table 5 demonstrates the performance result of all the applied classifiers. The performance result indicates that RF provides the least performance with 90% accuracy, where DT provides the best result with 95% accuracy. In terms of all the parameters, DT provides the best performance, which indicates that DT is potential classifier for the detection of ASD among children in early stage.

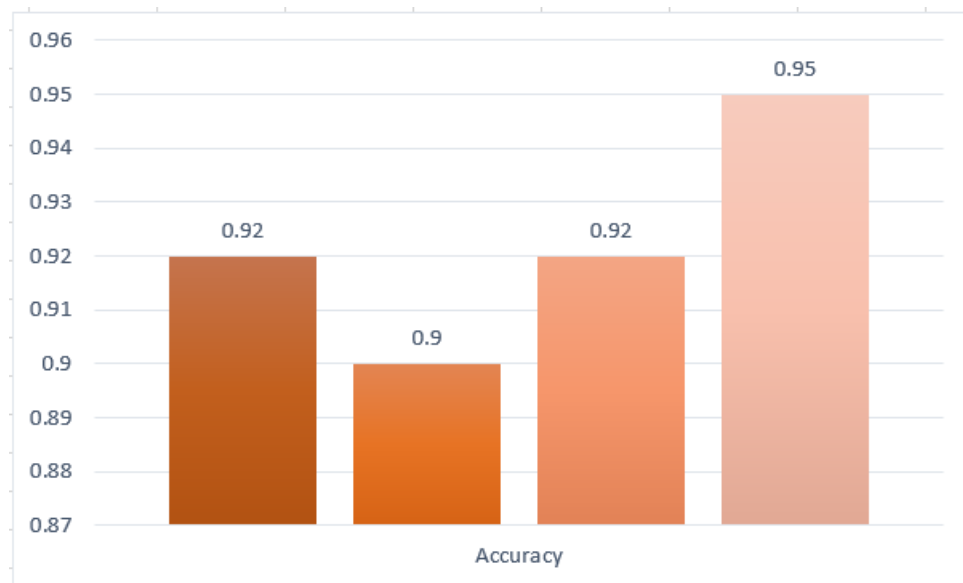


Figure 5: Visualization of Accuracy

Figure 5 illustrates the graphical visualization of accuracy comparison. The accuracy of all the applied classifiers is compared in the figure 4 and found that Random Forest generated the best accuracy.

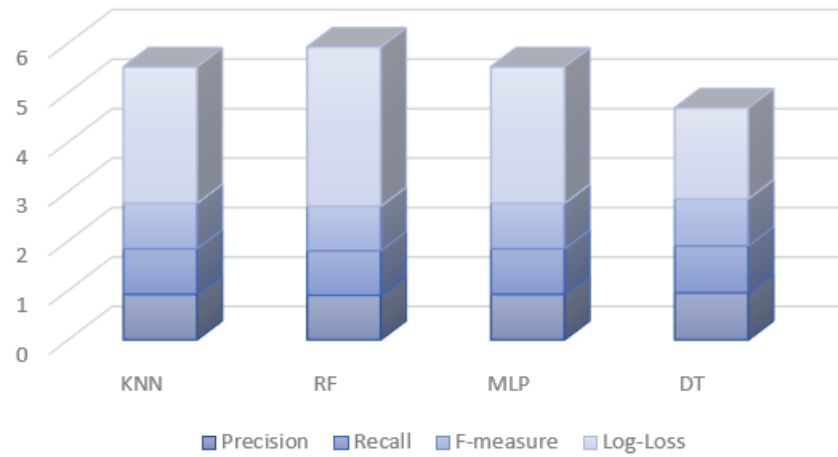


Figure 6: Visualization of Performance analysis

Figure 6 visualizes the other considered performance evaluation metrics to get a clear idea about the comparison of all the applied classifier's performance. The figure also supports that Random Forest is the best performing classifier compare to all other applied classification algorithms.

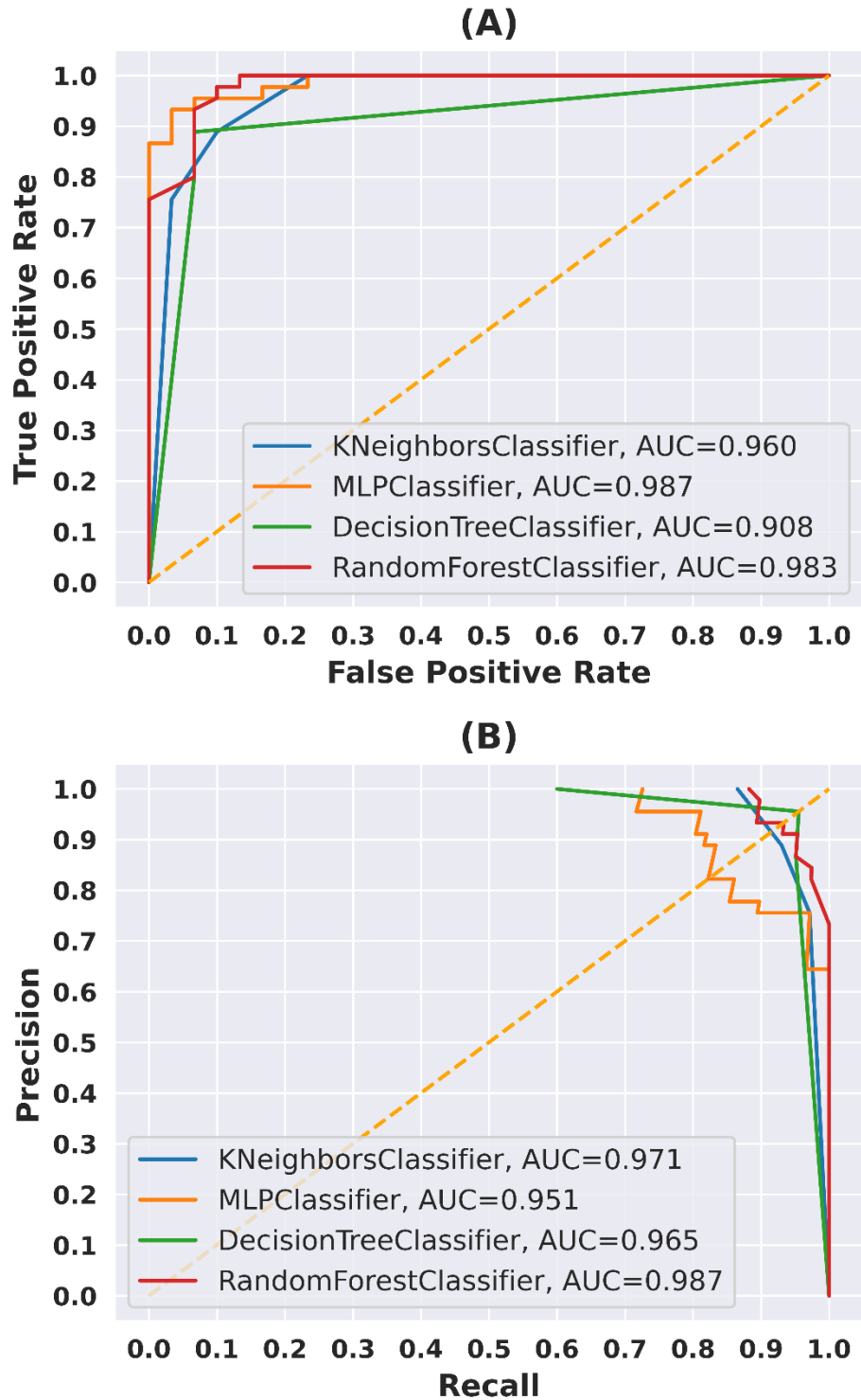


Figure 7: Area under Receiving Operating Characteristics (ROC) Curve. & Area under Precision – Recall (PRC) Curve (A) ROC Curve, (B) PRC Curve

Figure 7 depicts the result of area under the ROC (AUROC) curve at figure 7 (A) and are under the precision-recall (PRC) curve at figure 7(B). AUROC curve and PRC curve shows the efficiency of a model based on the area covered by the learning figure. Based on the figure 7, DT is the least performer classifier since it has covered 0.926 and 0.942 in AUROC and AUPRC respectively. However, the most efficient classifier is MLP and RF since they have covered 100% area in both of AUROC and AUPRC curve.

TABLE 6: RESULTS OF FST APPROACH

Feature Name	IGAE	GRAE	CFSSE	RFAE
Age	0	0	0.0753	0.00871
Jaundice	0.000453	0.000535	0.025	-0.00308
Gender	0.001086	0.001255	0.0388	0.00274
Use app before	0.001632	0.007048	0.0472	-0.00205
Autism	0.001724	0.002642	0.0488	-0.00822
Relation	0.011968	0.014917	0.0614	0.00671
Ethnicity	0.036504	0.014577	0.0556	0.01757
A1_score	0.116637	0.123048	0.3935	0.19007
A2_score	0.038218	0.038348	0.229	0.07637
A3_score	0.122714	0.149296	0.3955	0.11062
A4_score	0.249648	0.251567	0.5685	0.28185
A5_score	0.112333	0.136666	0.3799	0.15582
A6_score	0.135612	0.156652	0.4173	0.12295
A7_score	0.055143	0.057011	0.2739	0.07945
A8_score	0.143457	0.143462	0.4384	0.18459
A9_score	0.177918	0.177942	0.4862	0.16575
A10_score	0.153613	0.181338	0.4399	0.18596
Country of res	0.235509	0.054281	0.0752	0.02705
Result	0.999154	1	0.8359	0.2024

The table 6 describes the analysis result of FST methods. Four FST methods such as CFSSE, IGAE, GRAE, and RFAE are applied and their result is represented in the table. This table enables us to identify the most important risk factor associated with ASD.

To summarize, we gathered an ASD dataset from Kaggle for building our expected model, Then the collected dataset was highly processed as needed. Then we performed exploratory data analysis for better understanding the source data. Then we employed five ML techniques such as KNN, MLP, DT, and RF, and assessed their results based on accuracy, sensitivity, specificity, kappa statistics, precision, recall, F-Measure, log loss, and MCC. We discovered that all of the applied algorithms performed well, where DT generated the greatest performance with 95% accuracy. It indicates that it is the most accurate at predicting ASD in the early stages. In addition to that, we applied FST methods to show the feature importance employing four techniques such as CFSSE, IGAE, GRAE, and RFAE and the result is represented in table 6. The results of FST methods help us to identify the most important factors associated with ASD. It should be noted, however, that the quantity of data on ASD provided by this dataset was inadequate to address these problems adequately, and that more data analytic approaches are necessary to construct an useful prediction model. Nonetheless, we expect that in the future, we will be capable of understanding the limitations of this approach and that more data analysis will allow for very precise forecasts of ASD and associated illnesses utilizing ML methodologies.

CHAPTER 5

Conclusions & Future Scope

ASD is a type of disease related to mental growth and development, which leads to potentially fatal complications. Because of the possibility for precise illness prediction rate, ML approaches might be utilized to anticipate their incidence. We employed an ASD dataset to investigate the applicability of ML techniques to ASD identification, and discovered that DT functioned exceedingly well with 95% correctness. In addition to that we applied FST methods to identify most important risk factor, which are associated with ASD. The research attempted to uncover the best ML approaches among a range of widely-accepted and simple-to-implement methods, and discovered that, at least for this dataset, they performed effectively. This is an early stage of employing ML techniques, but it implies that it might be a valuable addition to clinical outcomes.

In the future, we intend to focus on additional research on early detection of autism spectrum disorders using machine learning approaches, either by developing a new algorithm or tweaking an existing algorithm, in order to achieve high levels of accuracy when applied to such an important subject.

Reference:

1. Eggett, A. (2018). What is Autism Spectrum Disorder? In *Groupwork for Children with Autism Spectrum Disorder* (pp. 1–18). Routledge.[Accessed on 9-10-2021]
2. CDC. (2021, December 13). *Data & statistics on autism spectrum disorder*. Centers for Disease Control and Prevention.
<https://www.cdc.gov/ncbddd/autism/data.html?fbclid=IwAR3lXYyBrqmwCPZ6WIOgqx67FcFpa6liVA0P5Knek0eZIFapZJvcInMc9IU> [Accessed on 10-10-2021]
3. Anderson, M. P. (2018). Autism spectrum disorders. In *Developmental Neuropathology* (pp. 477–495). John Wiley & Sons, Ltd.[Accessed on 10-10-2021]
4. Thabtah, F. and Peebles, D., 2020. A new machine learning model based on induction of rules for autism detection. *Health informatics journal*, 26(1), pp.264-286.
5. Vakadkar, K., Purkayastha, D. and Krishnan, D., 2021. Detection of Autism Spectrum Disorder in Children Using Machine Learning Techniques. *SN Computer Science*, 2(5), pp.1-9.
6. Parikh, M.N., Li, H. and He, L., 2019. Enhancing diagnosis of autism with optimized machine learning models and personal characteristic data. *Frontiers in computational neuroscience*, 13, p.9.
7. Stevens, E., Dixon, D.R., Novack, M.N., Granpeesheh, D., Smith, T. and Linstead, E., 2019. Identification and analysis of behavioral phenotypes in autism spectrum disorder via unsupervised machine learning. *International journal of medical informatics*, 129, pp.29-36.
8. Hyde, K.K., Novack, M.N., LaHaye, N., Parlett-Pelleriti, C., Anden, R., Dixon, D.R. and Linstead, E., 2019. Applications of supervised machine learning in autism spectrum disorder research: a review. *Review Journal of Autism and Developmental Disorders*, 6(2), pp.128-146.
9. Akter, T., Satu, M.S., Khan, M.I., Ali, M.H., Uddin, S., Lio, P., Quinn, J.M. and Moni, M.A., 2019. Machine learning-based models for early stage detection of autism spectrum disorders. *IEEE Access*, 7, pp.166509- 166527.
10. Moon, S.J., Hwang, J., Kana, R., Torous, J. and Kim, J.W., 2019. Accuracy of machine learning algorithms for the diagnosis of autism spectrum disorder: Systematic review and meta-analysis of brain magnetic resonance imaging studies. *JMIR mental health*, 6(12), p.e14108.
11. Abbas, H., Garberson, F., Glover, E. and Wall, D.P., 2018. Machine learning approach for early detection of autism by combining questionnaire and home video screening. *Journal of the American Medical Informatics Association*, 25(8), pp.1000-1007.
12. Duda, M., Ma, R., Haber, N. and Wall, D.P., 2016. Use of machine learning for behavioral distinction of autism and ADHD. *Translational psychiatry*, 6(2), pp.e732-e732.
13. Bone, D., Bishop, S.L., Black, M.P., Goodwin, M.S., Lord, C. and Narayanan, S.S., 2016. Use of machine learning to improve autism screening and diagnostic instruments: effectiveness, efficiency, and multi-instrument fusion. *Journal of Child Psychology and Psychiatry*, 57(8), pp.927-937.

14. Wall, D.P., Dally, R., Luyster, R., Jung, J.Y. and DeLuca, T.F., 2012. Use of artificial intelligence to shorten the behavioral diagnosis of autism.
15. M.R. Rahman, T. Islam, T. Zaman, M. Shahjaman, M.R. Karim, F. Huq, J.M. Quinn, R.D. Holsinger, E. Gov, M.A. Moni, Identification of molecular signatures and pathways to identify novel therapeutic targets in alzheimer's disease: insights from a systems biomedicine perspective, *Genomics* 112 (2) (2019) 1290–1299.
16. Four Techniques for Outlier Detection, [https://www.kdnuggets.com/2018/12 /four-techniques-outlier-detection.html](https://www.kdnuggets.com/2018/12/four-techniques-outlier-detection.html).
17. Md Satu, Syeda Atik, Mohammad Moni, A Novel Hybrid Machine Learning Model to Predict Diabetes Mellitus, 2019.
18. Nashif, S., Raihan, M.R., Islam, M.R. and Imam, M.H., 2018. Heart disease detection by using machine learning algorithms and a real-time cardiovascular health monitoring system. *World Journal of Engineering and Technology*, 6(4), pp.854-873.
19. Quinlan JR. Induction of decision trees. *Mach Learn.* 1986;1(1):81–106.
20. Cruz JA, Wishart DS. Applications of machine learning in cancer prediction and prognosis. *Cancer Informat.* 2006;2:59–77.
21. T. Cover, P. Hart, Nearest neighbor pattern classification, *IEEE Trans. Inf. Theor.* 13 (1) (1967) 21–27.
22. B.V. Dasarathy, Nearest neighbor (NN) norms: NN pattern classification techniques, *IEEE Comput. Soc. Tutorial* (1991), 10012834200.
23. K.H. Raviya, B. Gajjar, Performance Evaluation of different data mining classification algorithm using WEKA, *Indian J. Research* 2 (1) (2013) 19–21.
24. X. Luo, F. Lin, Y. Chen, S. Zhu, Z. Xu, Z. Huo, M. Yu, J. Peng, Coupling logistic model tree and random subspace to predict the landslide susceptibility areas with considering the uncertainty of environmental features, *Sci. Rep.* 9 (1) (2019) 1–13.
25. S.B. Kotsiantis, I. Zaharakis, P. Pintelas, Supervised machine learning: a review of classification techniques, *Emerg. Artif. Intel. Appl. Comput. Eng.* 160 (2007) 3–24.
26. R.L. De Mantaras, E. Armengol, Machine learning from examples: inductive and Lazy methods, *Data Knowl. Eng.* 25 (1–2) (1998) 99–123.
27. S. Vijayarani, S. Sudha, Comparative analysis of classification function techniques for heart disease prediction, *Int. J. Innov. Resear. Compute. Commun. Eng.* 1 (3) (2013) 735–741.
28. K. Kwon, D. Kim, H. Park, A parallel MR imaging method using multilayer perceptron, *Med. Phys.* 44 (12) (2017) 6209–6224.
29. S. Tajmiri, E. Azimi, M.R. Hosseini, Y. Azimi, Evolving multilayer perceptron, and factorial design for modelling and optimization of dye decomposition by biosynthesized nano CdS-diatomite composite, *Environ. Res.* 182 (2020) 108997.

30. Y. Azimi, Prediction of seismic wave intensity generated by bench blasting using intelligence committee machines, *Int. J. Eng.* 32 (4) (2019) 617–627.
31. R.D. Canlas, *Data Mining in Healthcare: Current Applications and Issues*, School of Information Systems & Management, Carnegie Mellon University, Australia, 2009.
32. M. S. Satu, S. Ahamed, F. Hossain, T. Akter, and D. M. Farid, “Mining traffic accident data of N5 national highway in bangladesh employing decision trees,” in *Proc. IEEE Region 10 Humanitarian Technol. Conf. (R10-HTC)*, Dec. 2017, pp. 722–725.
33. M. S. Satu, F. Tasnim, T. Akter, and S. Halder, “Exploring significant heart disease factors based on semi supervised learning algorithms,” in *Proc. Int. Conf. Comput., Commun., Chem., Mater. Electron. Eng. (IC4ME2)*, Feb. 2018, pp. 1–4.
34. R. J. Urbanowicz, M. Meeker, W. La Cava, R. S. Olson, and J. H. Moore, “Relief-based feature selection: Introduction and review,” *J. Biomed. Inform.*, vol. 85, pp. 189–203, Sep. 2018.
35. Hyman, S.L., Levy, S.E., Myers, S.M., Kuo, D.Z., Apkon, S., Davidson, L.F., Ellerbeck, K.A., Foster, J.E., Noritz, G.H., Leppert, M.O.C. and Saunders, B.S., 2020. Identification, evaluation, and management of children with autism spectrum disorder. *Pediatrics*, 145(1).
36. Rowden, A. (2021, November 3). *What are the types of autism?* Medicalnewstoday.Com. <https://www.medicalnewstoday.com/articles/types-of-autism> [Access on 15-12-21 at 6.20pm]
37. *Autism spectrum disorder - childhood disintegrative disorder.* (n.d.). Medlineplus.Gov. Retrieved January 16, 2022, from <https://medlineplus.gov/ency/article/001535.htm> [Access on 16-12-21 at 12.10am]
38. *WebMD - Better information. Better health.* (n.d.). WebMD. Retrieved January 16, 2022, from <https://www.webmd.com/brain/autism/autism-spectrum-disorders> [Access on 16-12-21 at 12.20am]
39. White Swan Foundation. (2016, September 23). *Autism spectrum disorder.* White Swan Foundation; White Swan Foundation. https://www.whiteswanfoundation.org/disorders/neurodevelopmental-disorders/autism-spectrum-disorder?gclid=Cj0KCQiAweaNBhDEARIsAJ5hwbeqAJiVdDH-2xxqpkMZ-4r5A-DDPbEY-DU5ma8N4HZvonsSmvLGbTkaArM2EALw_wcB. [Access on 16-12-21 at 12.25am]
40. Burke, D. (2021, November 9). *What is Asperger syndrome? Causes and symptoms.* Healthline. <https://www.healthline.com/health/asperger-syndrome> [Access on 16-12-21 at 12.05am]