Development of an Interactive Dashboard for Analyzing Autism Spectrum Disorder (ASD) Data using Machine Learning Techniques

Thesis

Submitted By

18-36064-1	Avishek Saha
18-37405-1	Dibakar Barua
18-36041-1	Ziad Mohib
18-36235-1	Sumaya Binte Zilani Choya

Department of Computer Science

Faculty of Science & IT

American International University Bangladesh

September, 2021

Declaration

We declare that this thesis is our original work and has not been submitted in any form for another degree or diploma at any university or other institute of tertiary education. Information derived from the published and unpublished work of others has been acknowledged in the text and a list of references is given.

Avishek Saha
18-36064-1
Faculty of Science & Technology

Dibakar Barua
18-37405-1
Faculty of Science & Technology

Ziad Mohib
18-36041-1
Faculty of Science & Technology

Sumaya Binte Zilani Choya
18-36235-1
Faculty of Science & Technology

Approval

The thesis titled “Development of an Interactive Dashboard for Analyzing Autism Spectrum Disorder (ASD) Data using Machine Learning Techniques” has been submitted to the following respected members of the board of examiners of the department of computer science in partial fulfilment of the requirements for the degree of Bachelor of Science in Computer Science on (September 24, 2021) and has been accepted as satisfactory.

_____

Dr. Md. Mahbub Chowdhury Mishu
Assistant Professor, Head (Undergraduate) & Supervisor
Faculty of Science & Technology
Department of Computer Science
American International University-Bangladesh

_____

Dr. Dip Nandi
Professor and Director & External
Faculty of Science & Technology
Department of Computer Science
American International University-Bangladesh

_____

Professor Dr. Tafazzal Hossain
Dean
Faculty of Science & Technology
American International University-Bangladesh

_____

Dr. Carmen Z. Lamagna
Vice Chancellor
American International University-Bangladesh

Acknowledgment

We are pleased that, despite many limitations in this pandemic circumstance, we were able to finish our thesis.

First and foremost, we want to express our thanks to the Almighty for allowing us to complete our Thesis on time.

We do not want to pass up the chance to thank the Faculty of Science and Technology for keeping thesis credit in the graduate program's curriculum and allowing us to sample the taste of research work that we are interested in.

We would like to thank our supervisor, Dr. Md. Mahbub Chowdhury Mishu, for his consistent advice and assistance during our thesis study. Furthermore, we would like to express our gratitude to, Dr. Carmen Z. Lamagna, our honored Vice-Chancellor, who has been a source of inspiration for us.

Abstract

Autism Spectrum Disease (ASD) is a lifelong neurodevelopmental disorder that impairs a person's capacity to communicate and connect with others. It has an impact on a person's understanding and social relationships. People with ASD also have a wide range of symptoms, such as difficulty communicating with others, repetitive habits, and an inability to operate well in other aspects of daily life. Autism is a "behavioral condition" that can be diagnosed at any age. Symptoms generally develop in the first two years of life. The majority of people are unaware of the disease and so have no way of knowing whether or not someone is disordered. Rather than helping the patient, this usually results in his or her social isolation. ASD is a condition that begins in childhood and continues through adolescence and maturity.

About 25 publications were examined in this study on autism spectrum disorder (ASD) prediction using machine learning or data mining. In this thesis, the approaches or algorithms utilized in those works are explained. Furthermore, the data and findings of those articles are evaluated utilizing various techniques and algorithms. Four publicly available non-clinically ASD datasets are used to evaluate the techniques described in those papers. There are 292 instances and 21 attributes in the ASD screening in children dataset. There are 704 instances and 21 attributes in the ASD screening Adult dataset. There are 104 instances and 21 attributes in the ASD screening in teenagers dataset. With 1054 instances and 19 attributes, Toddler is the fourth dataset. An automated dashboard based on the Toddler dataset was built to analyze it and identify insights. Because the focus was on early ASD prediction, this dataset was chosen.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Autism Spectrum Disorder (ASD) is a neurodevelopmental disorder characterized by social communication and interaction problems [1]. ASD is most commonly identified in the first two years of life, but it can also develop later. Autism is not difficult to identify, but it does need extensive learning and training on the side of doctors. “Autism is not a genetic brain illness, but a systemic body disorder that affects the brain,” explains Dr. Mark Hyman [2]. People of any age, socioeconomic position, ethnicity, or race might be affected by ASD. Adults might experience this disease in many ways, ranging from severe symptoms to minor problems. The autism spectrum disease is one of five childhood-onset illnesses referred to as Pervasive Developmental Disorders (PDDs) (PDD). This ASD is categorized as a complex neurological disorder [3,4]. It's a difficulty with communication that also affects one's behavior. Figure 1.1 depicts the global prevalence of ASD.

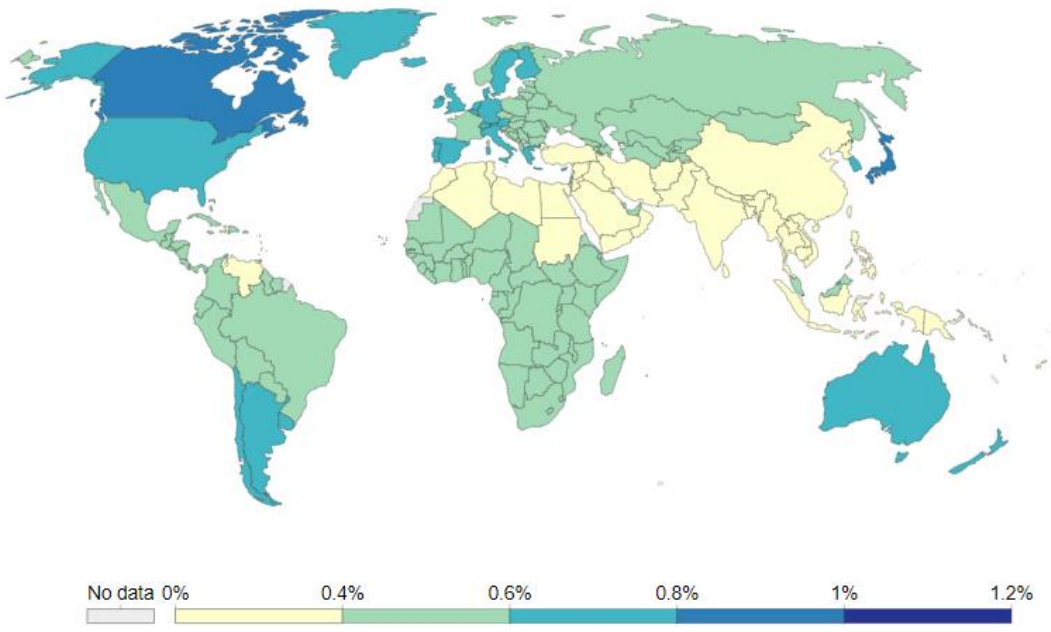

Fig. 1-1: Autism Prevalence Around the World [5]

A person with ASD might have a wide range of symptoms. Depending on the symptoms, ASD can range in severity from moderate to severe [6]. However, several diseases are similar to ASD. One of them is attention deficit hyperactivity disorder (ADHD) [7,8]. For a child with ADHD, social interaction might be challenging [7]. Males are more afflicted by Autism Spectrum Disorder (ASD) than females [9]. ASD symptoms can be noticed as early as the first two years of life. ASD affects people for the rest of their lives [8]. No

therapy is a complete cure for ASD [8]. People with ASD have been found to benefit from medication [9]. Environmental and genetic variables might have a role in the disease's progression [8]. Scientists, on the other hand, have yet to identify the actual causes of ASD. The intellect of someone with Asperger's syndrome, which is comparable to autism, is ordinary to above-average [3]. ASD is influenced by several variables, including low birth weight, having an ASD sibling, and having old parents [8].

Every disease has its own set of signs and symptoms. The following are some frequent ASD synonyms:

- Facial expressions, gestures, and visual communication are inappropriate. For example, avoiding eye contact or employing facial expressions that do not match the message he or she is attempting to convey. [3,4,8,9,10]
- There is no desire to share similar interests or achievements with others. In communication with others, hesitancy or impediment.
- Hesitates approaching people or engaging in social interactions; looks distant and reserved; prefers isolation.
- Difficulties in interpreting their emotions, how they should respond, and nonverbal cues.
- Become enraged when someone touches them, do not want to curdle
- One of the challenges is the inability to make friends with other teenagers of the same age. The emergence of social ties is difficult or delayed. The inability or unwillingness to learn how to communicate.
- Using a strange rhythm, a high pitch, or an unusual tone of voice. Multiple uses of the same words or phrases, each with a distinct meaning. Inappropriate laughing and giggling.
- Impairment in starting or maintaining a spoken language. Difficulty expressing wants or requirements
- Inability to grasp basic inquiries or assertions.
- Taking things too seriously and ignoring wit, irony, and satire.
- Repetitive physical motions that are always moving.
- Obsession with a certain subject, generally involving numbers or symbols.

- Consistency, order, and traditions are important to a strong. Agitated by changes in their routine or environment.
- Clumsiness, abnormal posture, or odd movement patterns are all examples of clumsiness.
- Interested in spinning, moving, or toy characteristics of items.
- Sensory issues and a lack of pain sensitivity are also present.
- Psychological talents that are inconsistent
- Spending a substantial amount of time putting things back in order.
- In various conditions, such as light, noise, and so on, one person is less sensitive than another.

1.1 Research Question

How can an automated dashboard for visualizing ASD traits among people anticipate future trends?

- What effect does machine learning have on ASD detection?
- What were the results of the existing models in terms of detecting ASD traits?
- How can data analysis be used to spot trends and patterns?

1.2 Aims and Objectives

We sought to use machine learning techniques to create an interactive dashboard for displaying ASD traits across individuals all around the world in this study.

- Machine learning's impact on ASD detection.
- Performance of existing machine learning models that detect ASD traits.
- Identifying trends and patterns by data analysis.

1.3 Contribution

In a clinic, ASD diagnosis is costly and time-consuming [9,11]. It is, however, not done by a medical test; rather, it is done through observation, which takes a long time, which is why early diagnosis is so important. As a consequence, machine learning algorithms can help with the diagnosis while simultaneously reducing the cost and inconvenience [9]. It's also possible that a diagnosis isn't correct. Data analytics and machine learning algorithms can assist a healthcare practitioner in quickly and accurately detecting Autism. We employed multiple machine learning algorithms and demonstrated accuracy calculations to diagnose Autism in this paper. We have also created an interactive analytical tool (dashboard) for analyzing and visualizing data. This is a dynamic dashboard. Any changes to the variable will have an impact on the dashboard's appearance. Anyone from anywhere in the world may observe the global ASD situation by going through this. We attempted to study several papers before making any choices. This aided us in our decision-making and paved the way for our study. It will be addressed in-depth in the next chapter.

Chapter 2: Background

ASD symptoms have been identified by several researchers [3,7,8]. The primary symptoms of ASD include inappropriate facial expressions, gestures, and visual communication, as well as avoiding eye contact or employing facial expressions that do not match the message he or she is trying to convey [8]. People with ASD are rarely interested in discussing their common interests or achievements with others. They are hesitant to converse with people. Other symptoms include odd facial expressions, irritability, and a delay in forming social bonds. Researchers are trying hard to develop a computer-assisted detection system that can diagnose ASD at an early stage, but no such system has yet been developed. We compared data from prior studies on computer-aided identification of ASD in this section." Autism Spectrum Disorder"," ASD"," ASD diagnosis"," ASD detection using machine learning"," Data mining approaches to diagnose ASD"," Supervised machine models to identify ASD" were some of the keywords we used to find publications.

After sorting the articles, 24 were chosen for evaluation. The usage of models and accuracy were assessed in the abstracts, methods, and results from parts of these papers. Following the selection of articles, each member of our team creates a summary of the articles. The publications were then discussed by the researchers. During the discussion period, researchers shared their thoughts on publications. We created this review article based on these viewpoints.

The following table is a list of chosen publications with descriptions of the data sets referenced.

Table 2-A: Number of instances and attributes in reviewed papers.

Paper	Instances	Attributes	Technique
[3]	100		Neural Network, SVM, Fuzzy logic.
[4]	609	19	KNN, Random Forest, Logistic Regression
[6]	704	21	Naïve Bayes, J45, Bayesian Network, Majority Model.
[7]	2775	65	LDA, SVM, Decision Tree, Random Forest, Logistic Regression,

			Categorical Lasso.
[8]	292(Child)704(Adult)104(Adolescent)	212121	KNN, SVM, CNN, Logistic Regression, Naïve Bayes, ANN.
[9]	699	19	Random Forest.
[10]	292	19	KNN, LDA.
[11]	851	6	KNN, SVM, Decision Tree, Random Forest, Logistic Regression, Majority Model, Confidence Model, LSA.
[12]	292(Child)704(Adult)104(Adolescent)	212121	Decision Tree, Random Forest.
[13]	704	Not Mentioned	SVM, Decision Tree, Naïve Bayes.
[14]	1565	17	SVM, Decision Tree, Random Forest, Naïve Bayes.
[15] (data set 1)	2009	21	Adaboost, FDA, C5.0, LDA, MDA, PDA, SVM and CART
[15] (data set 2)	248 (Child)609 (Adolescent)98 (Adult)	21	Adaboost, FDA, C5.0, LDA, MDA, PDA, SVM and CART
[16]	702(Adult)	19	KNN, SVM, Naïve Bayes, Linear Regression, Linear Discriminant Analysis, Classification Regression Tree
[17]	515(Not ASD)189(ASD)	21	CBA, CMAR, MCAR, FACA, FCBA, ECBA, WCBA
[18]	21 ASD21 Non ASD	Not Mentioned	SVM, KNN, LDA

In [1], data was gathered from toddlers, children, adolescents, and adults with ASD. Different types of techniques were applied to the ASD datasets and results were evaluated. According to this research, SVM outperformed Adaboost for the toddler dataset, whereas Adaboost outperformed SVM for the children dataset. For the teenage dataset, Glmboost was used, whereas Adaboost was used for the adult dataset. The feature changes that provided the best classifications were the sine function for toddlers and the Z-score for children and teenage datasets. Following these investigations, different feature selection techniques were used to establish the key ASD risk factors for toddlers, children, adolescents, and adults using Z-score-transformed datasets.

In [3], it offers a thorough understanding of the many kinds of ASD. It discussed how tools and approaches are beneficial to ASD children. 100 samples were collected to conduct their research. Weka was used in this case. The decision rules, as well as a decision tree, have been used for visualization.

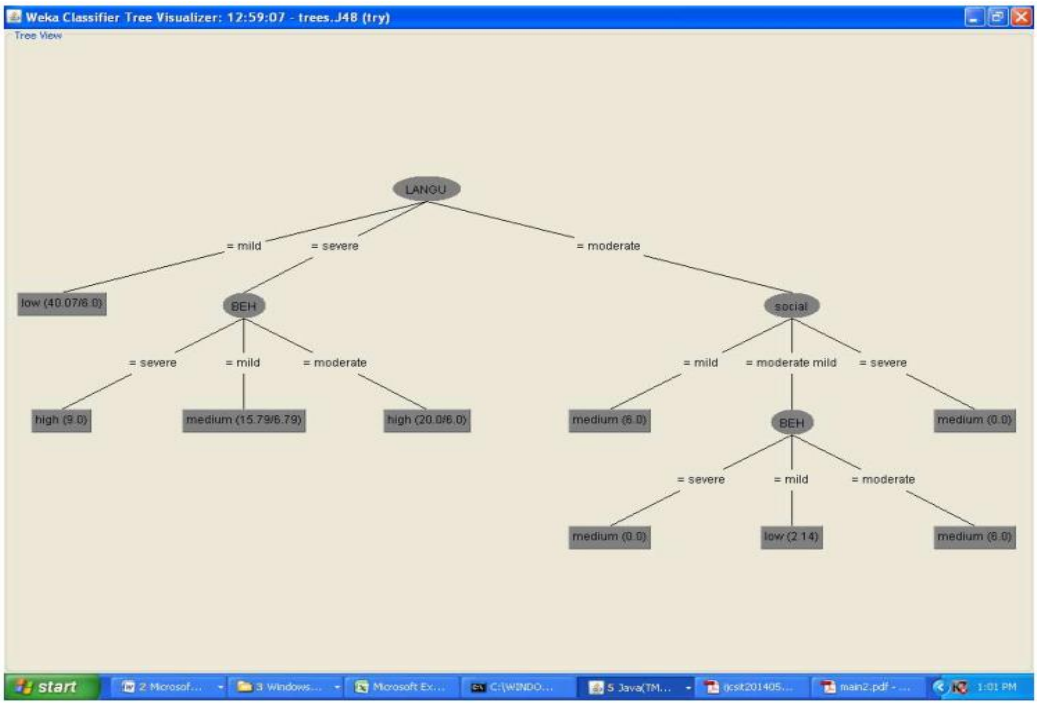

Fig. 2-1: Decision Tree [3].

However, it only included one classification approach when it was meant to discuss several others. How many attributes had been utilized was not mentioned here, and the performance measurements have also been left out.

In [4], There were 704 samples with 20 attributes and one output utilized. Python was used to implement the model. Because the dataset had some missing values, some of the samples were deleted, leaving 609 samples. This study, utilized logistic regression, Random Forest, and K Nearest Neighbors. The hamming distance was utilized for KNN. The 80:20 rule was used to evaluate performance, with 5-fold cross-validation. The patient's age, as well as a graphical representation, were not given. There were no model diagrams. In addition, the workflow diagram is missing. The symptoms of ASD aren't discussed at all, and neither are the different forms of ASD.

In [6], The J48, Multilayer Perception, Nave Bayes, and Bayesian Network algorithms are discussed. For the analysis, the Weka tool was utilized. A total of 704 instances were utilized, each with 21 attributes from the adult dataset. Diagrams and graphs are used to depict the workflow and conclusions. Others were outperformed by Multilayer Perception.

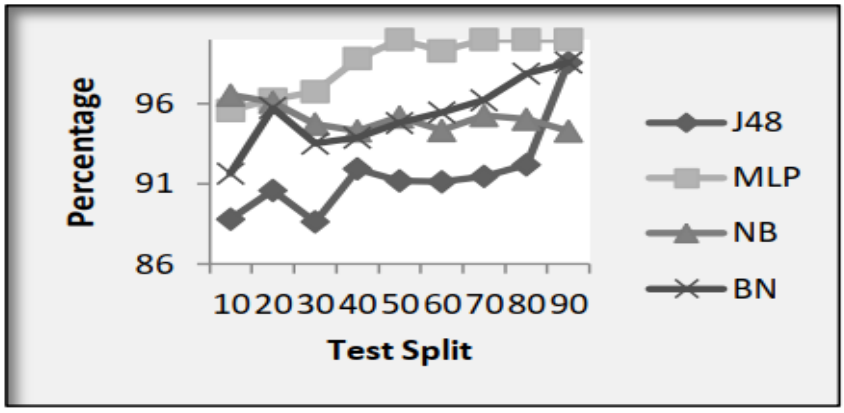

Fig. 2-2: Accuracy of MLP, J48, NB, and BN Classifiers [6].

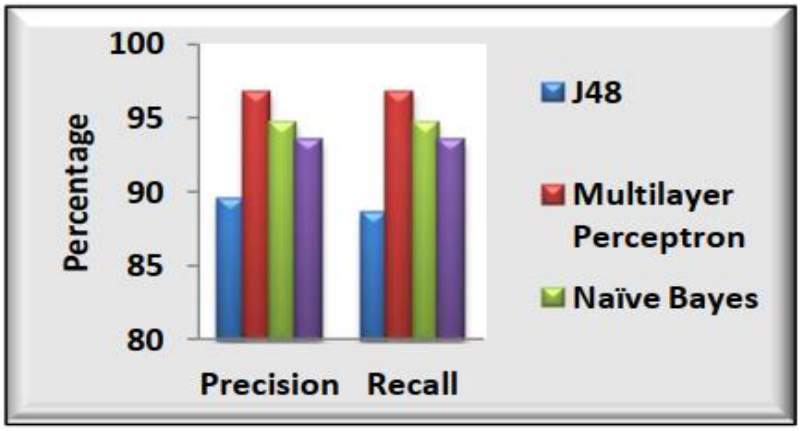

Fig. 2-3: Precision and Recall Rate of 30% Test Split Samples [6].

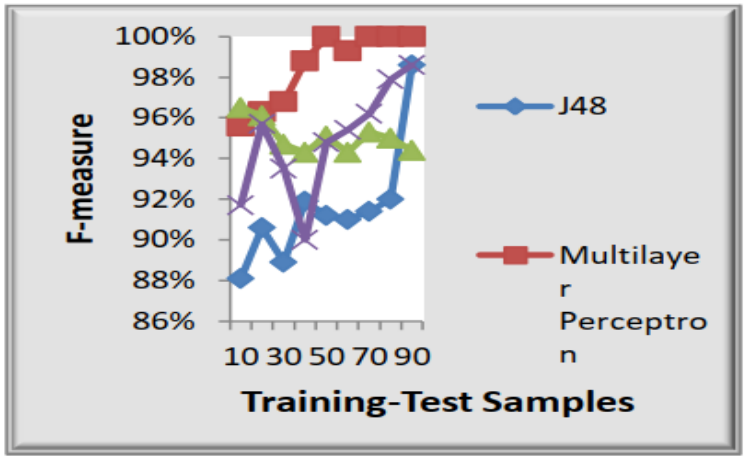

Fig. 2-4: F measures Values of Four Classifiers [6].

Only J48 and Multilayer Perceptron were labeled in this study, the other two were not. The paper's motivation was unclear. ADHD was not discussed, even though it shares several characteristics with ASD. The different forms of ASD were also absent.

In [7], the study makes a distinction between ASD and ADHD. Both of them have comparable symptoms. A total of 2975 instances with 65 attributes were used in this study. There are 2275 ASD cases and 150 ADHD cases among them. The following methods were utilized in this paper: decision tree, random forest, logistic regression, support vector classification, linear discriminant analysis, and categorical lasso. Python was used to carry out the model implementation. Others were outperformed by Support Vector Classification. In this case, 10-fold cross-validation was employed.

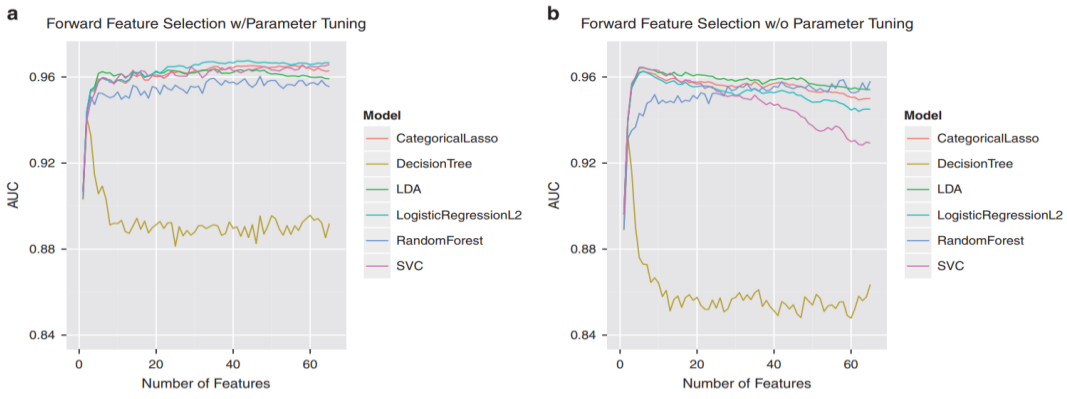

Fig. 2-5: Forward Feature Selections with Parameter Tuning (a) without Parameter Tuning(b) [7]

The age of the ASD patients was not stated in this study. Only accuracy and the ROC curve were used to gauge performance, if other approaches were employed, it would be

apparent which one was the best. There was not a model diagram or a process diagram here. ASD types are not addressed in this article.

In [8], The algorithms Naive Bayes, Support Vector Machine, logistic regression, KNN, Neural Network, and Convolutional Neural Network were used. This study employed 292 instances with 21 attributes from the child dataset, 704 instances with 21 attributes from the adult dataset, and 104 instances with 21 attributes from the adolescence dataset. The data had been analyzed using Python. A schematic of the process flow was provided. The visualization of the discovery is well done. SVM and CNN outperformed the other algorithms in terms of accuracy.

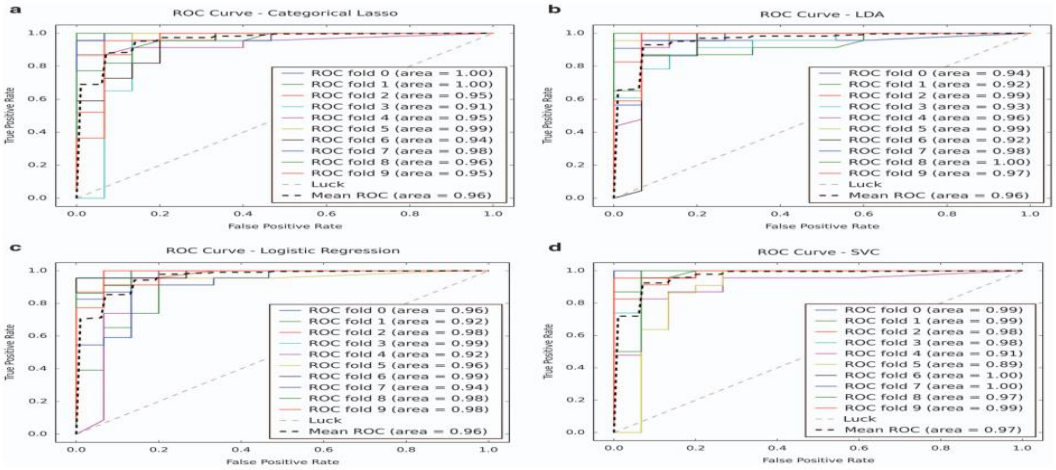

Fig. 2-6: ROC Curves of Top-performed Algorithms [8].

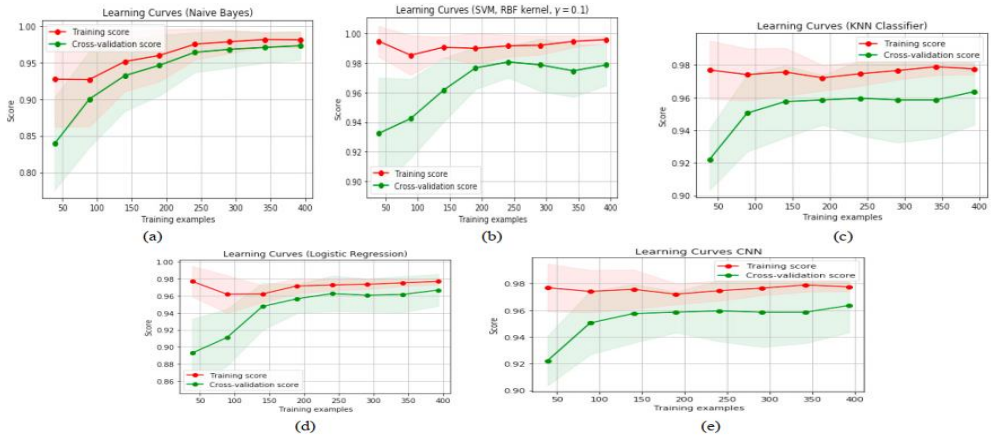

Fig. 2-7: Learning Curve of Naïve Bayes(a), SVM(b), KNN(c), Logistic Regression(d), CNN(e) Algorithm for Adult's Dataset [8]

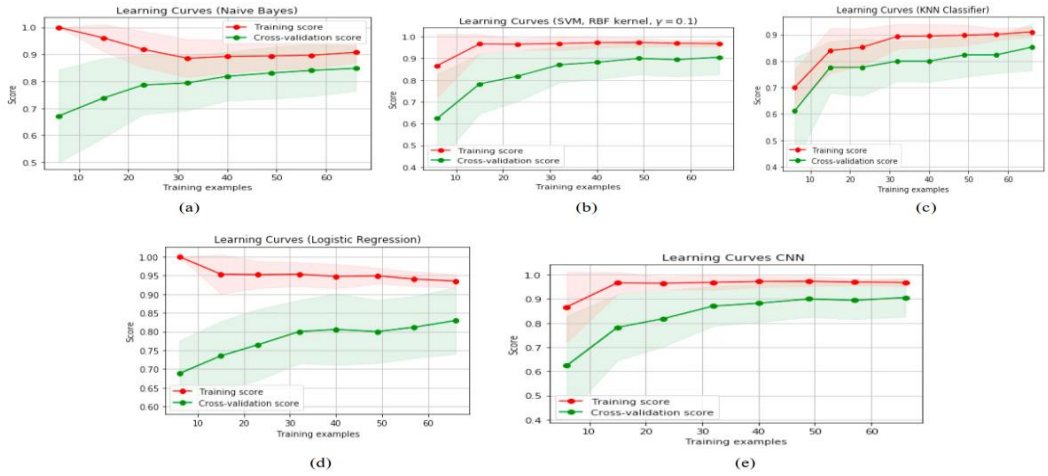

Fig. 2-8: Learning Curve of Naïve Bayes(a), SVM(b), KNN(c), Logistic Regression(d), CNN(e) Algorithm for Adolescent's Dataset [8].

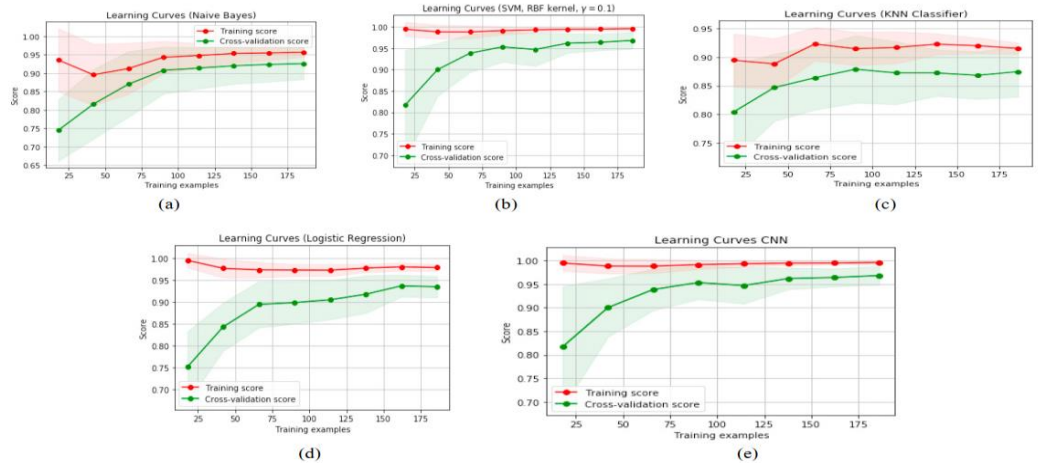

Fig. 2-9: Learning Curve of Naïve Bayes(a), SVM(b), KNN(c), Logistic Regression(d), CNN(e) Algorithm for Children's Dataset [8].

In the figures learning curves were shown. The red line refers to training data and the green line refers to test data. In every graph initially, the training data was far away from the test dataset. But in the end, they are almost overlapping. Overall, in terms of accuracy SVM and CNN outperformed others in each dataset. But there was no mention of ADHD, which has several of the same symptoms as ASD. The different forms of ASD are also absent.

In [9], the Random Forest algorithm was utilized. A dataset of 704 samples with 21 attributes was utilized, which was a dataset of adults. The number of samples in this dataset was reduced to 699 instances with 19 attributes after the duplicate data was removed. The data was gathered from the machine learning repository at UCI. The 80:20

rule was used in this case. This publication included a diagram of the random forest method. Sensitivity, specificity, and the ROC curve were used to assess performance accuracy. Graphs were also used to show the performance metrics.

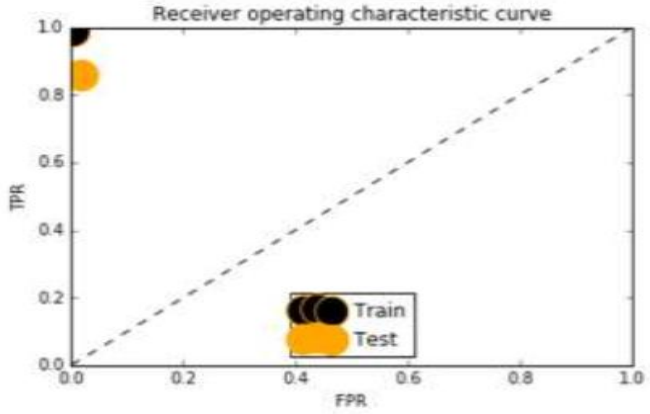

Fig. 2-10: ROC Curve [9].

However, there was no workflow diagram here. Though the goal was to reduce the cost of ASD diagnosis, the rate of reduction was not specified. It was claimed that medicine can be utilized to help ASD patients better their condition. However, no mention was made of how it would be applied to them or what would encourage them to do so. The many forms of ASD were also not highlighted.

In [10], it was included LDA and KNN models for ASD prediction. The dataset in this dataset comes from the University of California, Irvine. There were 292 samples and 19 attributes in all. A total of 141 samples were tested for ASD. As stated in the title, it was a child dataset. For performance testing, they used the 70:30 rule. The Euclidian distance formula was utilized for KKN. Here the confusion matrix was given. LDA outperformed the other two algorithms. However, no visual depiction was provided, and it was also unclear how they are encouraged to complete the task. The flowchart is missing. ASD types were not addressed in this article.

In [11], the researcher used an 851-sample with six attributes. There were 430 non-ASD patients and 421 ASD patients among them. The following algorithms were used: Decision tree, Majority model, Random Forest, SVM(Linear), SVM(Non-Linear), Confidence model, Logistic regression, K-Nearest neighbor, and Neural Network. K-fold cross-validation was used in this case. However, there is no visual representation of the findings. The study was based on an adult dataset that was never revealed. The flowchart for the workflow is missing. The many forms of ASD are also not discussed. There had not been much published about the symptoms.

In [12], machine learning methods were used to predict ASD in different age groups. A smartphone application was developed by researchers that utilize machine learning to predict ASD. The major focus was on creating an autism screening app for adults aged 4 to 11, 12 to 17, and 18 and above to predict ASD. Data gathering, data synthetization, building the prediction model, evaluating the prediction model, and developing a mobile application were the five steps of this study. The optimal algorithm for building the mobile app was utilized after data collecting and synthetization. After obtaining results from several supervised learning methods such as Linear Regression, SVM, Naive Bayes, and Random Forest, it was discovered that the Random Forest was extremely viable and accurate. But the issue is that they did not demonstrate any results and did not explain why the random forest is the best option for their mobile app. They utilized 248 samples for the (4-11) year group, 98 instances for the (12-16) year group, and 608 instances for the 18 and more year group after cleaning the data to train the machine. To train a machine, this data set is relatively little. A computer cannot adequately learn with such a little dataset. The application, on the other hand, was not created for children under the age of four. The app may be more useful if it included contained statistics for children under the age of four.

In [13], Researchers used a data set from the University of California at Irvine's repository to try to predict ASD using WEKA tools' supervised learning algorithms. It did not adequately describe the data set. As a result, it was unclear which age groups or types of people this research was successful for. The accuracy and some mistakes were listed, but no mention was made of where the test set was obtained. Is this a result of unknown data or data that hasn't been mentioned?

In [14], the algorithms utilized in the study were Decision Tree, Support Vector Machine, Random Forest, and Naive Bayes. The data was divided into two sections. 70% of the data is used for training, whereas 30% is used for testing. The proposed algorithms were tested on the data set and proved to be 100 percent accurate in every case. However, the models were not validated using unknown data. If the recommended accuracy for the models were measured using unknown data, this research might be more successful.

In [16], To predict ASD, researchers utilized machine learning techniques such as LDA (Linear Discriminate Analysis), Naive Bayes (NB), Regression Trees (Cart), K – Nearest-Neighbor (KNN), Linear Regression (LR), and Support Vector Machine (SVM). The Euclidean distance calculation formula was chosen for KKN. Here, a dataset of 702 instances with 19attributes from an adult dataset was employed. It was retrieved from the University of California at Irvine's repository. For assessing the performance of the models, the 70:30 rule was used. Among the algorithms used, LDA outperformed the

others. Each model's correctness had been provided for explanation. It would be easier to determine which algorithm was the best if certain additional factors were taken into account. The workflow diagram was not mentioned. Types of ASD were also absent. It should have been indicated earlier that an adult dataset was utilized, which was mentioned last in the conclusion section.

In [19], Only the words in the children's assessments were utilized to predict whether they met the case criteria for ASD using eight supervised learning algorithms. The algorithms' performance was evaluated using classification accuracy, F1 score, and the number of positive calls over ten random train-test splits of the data, evaluating their potential applicability for surveillance. Two of the algorithms that obtained above 87 percent accuracy were random forest and support vector machine with Naive Bayes features (NB-SVM). The number of false negatives produced by this NB-SVM was significantly greater (P = 0.027) than the number of false positives. The random forest outperformed more recently developed models like the NB-SVM and neural network, as well as providing reliable prevalence estimates. NB-SVM may not be a viable choice for usage in a fully automated surveillance workflow due to the increased false negatives. This type of machine learning method, in general, produces varied outcomes and accuracy for persons of various ages. Only children's data from the Georgia ADDM site was used in this study. Using various data sets for different age groups might make this research more successful. A total of ten train test cycles were used in this study. Where the total dataset was randomly split into 57 percent training, 13 percent validation, and 30 percent test sets for each cycle. These test sets were used to assess its performance once it had been trained. They may utilize a test set from a variety of data sets to see how well these algorithms work when dealing with unknown data.

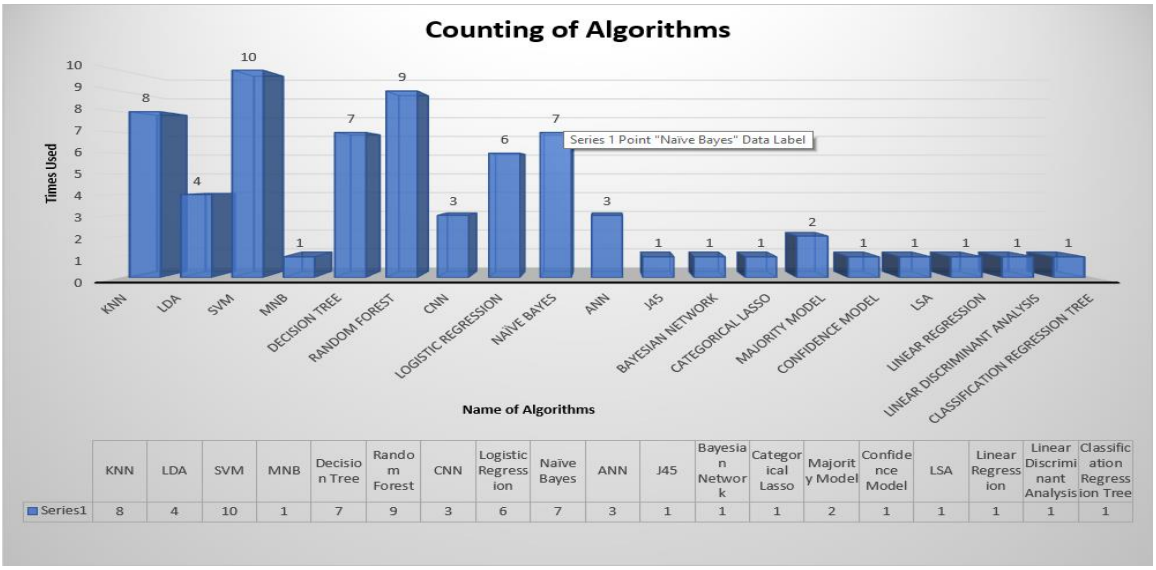

Fig. 2-11: Counting of Algorithms Used in The Selected Papers.

The machine learning algorithms applied in the selected papers were shown in this figure. A total of 19 algorithms were applied. Except for the paper [9], every publication utilized more than one algorithm. SVM was the most applied algorithm. It was used 10 times. The second place went to Random Forest, and the third place went to KNN based on the utilization in papers. Descriptions of some algorithms are given below,

- KNN

The k Nearest Neighbour (kNN) approach [20] [21] is an instance-based learning method that has been used in a variety of data mining applications. Assume you've been given a set of T such vectors, together with their corresponding classes: **x(i),y(i)fori=1,2, T.**

This set is referred known as the training set. Assume we've been handed a fresh sample with x=u. The class to which this sample belongs is what we're looking for. The simplest case is k = 1, in which we identify the sample in the training set that is closest to u and set v = y, where y is the nearest class of the nearest neighbor sample. For k-NN, the 1 NN principle is expanded in the following way. Find u's closest k neighbors, then use a majority decision rule to categorize the new sample. Higher k values smooth the data, which reduces the risk of overfitting caused by noise in the training data. We're talking about neighbors here, which implies that the independent variables are used to compute a distance or dissimilarity measure between samples. In general, we may look at the most frequently used distance measurement: Euclid's distance. [22].

- LDA

A probabilistic generative model of a corpus is LDA (Latent Dirichlet Allocation). Documents are represented as random mixes of latent themes, each of which is characterized by a word distribution. LDA assumes the following generating process for each document w in a corpus D:

1. Select $N \sim \text{Poisson}(\xi)$.

2. Select $\theta \sim \text{Dir}(\alpha)$.

3. Write the following for each of the N-words w_n:

(a) Select a topic $z_n \sim \text{Multinomial}(\theta)$.

(b) Select a word w_n from $p(w_n | z_n, \beta)$, a multinomial probability conditioned on the topic z_n. [23]

Some simplifying assumptions are used in this basic model, some of which will be eliminated in later sections. First, the dimensionality k of the Dirichlet distribution (and hence the dimensionality of the topic variable z) is assumed to be known and fixed. Second, the word probabilities are parameterized by a (k X V) matrix, with I j = p(wj = 1|zi = 1) as a fixed quantity to be estimated for the time being. Finally, for the rest of the method, the Poisson assumption is not essential, and more realistic document length distributions can be used if necessary. Keep in mind that the other data-generating variables (θ and z) do not affect N. As a result, it is treated as an auxiliary variable, and its unpredictability will be ignored in the subsequent development. [24]

A k-dimensional Dirichlet random variable can take values in the (k - 1)-simplex (a k-vector θ lies in the (k−1)-simplex if $\theta_i \geq 0$, $\sum_{i=1}^{k} \theta i = 1$), and has the probability density on this simplex as follows:

$$p(\theta|\alpha) = \left(\left(\Gamma \sum_{i=1}^{k} \alpha_i\right) / \prod_{i=1}^{k} \Gamma(\alpha_i) \right) \theta 1_1^{(\alpha 1-1)} \cdots \theta k_k^{(\alpha k-1)}$$

where α parameter is a k-vector with components $\alpha_i > 0$ and $\Gamma(x)$ is the Gamma function. The Dirichlet is a simplex distribution that belongs to the exponential family, has sufficient statistics in limited dimensions, and is conjugate to the multinomial distribution. These qualities will aid the development of LDA inference and parameters.

Here α and β are given parameters, The joint distribution of a topic mixture θ, a set of N subjects z, and a set of N words w is given by:

$$\mathbf{p(\theta, z, w|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(zn\ |\theta)\ p\ (wn\ |zn, \beta)}$$

- **SVM**

Support Vector Machine (SVM) is based on the notion of structural risk reduction rather than the conventional empirical risk minimization principle, the support vector machine is a breakthrough small-sample learning technique that surpasses prior methods in many ways. The support vector machine (SVM) is a two-dimensional representation of the optimum surface that arises from a linearly separable scenario; Figure 1 illustrates the fundamental concept. Figure 1 depicts the most important idea. H distinguishes between two kinds without producing any errors. H1 and H2 are points that intersect with H at their most recent point. The class interval was defined as the distance between H1 and H2. The goal of an optimum separating surface, also known as the largest class interval, is to allow error-free separation of two classes of data.

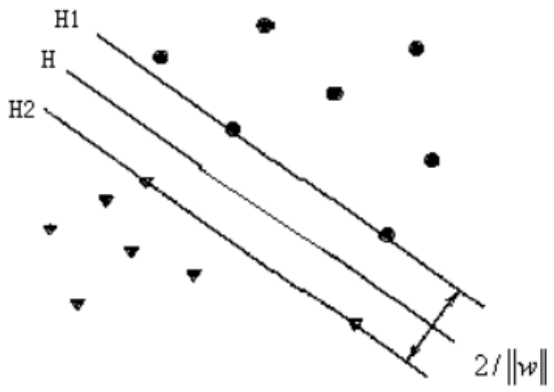

Fig. 2-12: Optimal Separating Surface [24]

The linear discriminate function in n-dimensional space is used in pattern recognition: **g(x) =ω · x + b**, The classification hyperplane equation **(ω · x) + b = 0** can be written. The discriminating function **g(x)** was normalized in the linear separable scenario. So that all training samples are met **|g(x) >=1|**, even if they are met away from the surface classification of the sample **g(x)= 1.** As a result, the class interval is $\frac{2}{||w||}$, making the interval on the equivalent to **||w||** or **|| w||²**. To appropriately classify the surface of all samples, it is necessary to meet the following requirements:

$$y_i [(\omega . x_i) + b]-1 >= 0, i=1,2,\ldots\ldots,n$$

Make **|| w||²** the smallest classification surface the ideal classification surface by satisfying the preceding equation. Support vectors (support vectors) are points on the hyperplane that help to support the best classification surface [25].

- Decision Tree

A decision is the recursive split of the instance space is used to categorize data. The nodes in the decision tree form a rooted tree, which is a directed tree with no incoming edges and a "root" node. There is one incoming edge for each of the other nodes. A node having outgoing edges is known as an internal or test node. The other nodes are leaves (also known as terminal or decision nodes). Based on a discrete function of the input attribute values, each internal node in a decision tree divides the instance space into two or more sub-spaces. In the simplest and most typical scenario, each test analyzes a single attribute, with the instance space partitioned according to the attribute's value. In the case of numeric characteristics, the condition refers to a range. A class is assigned to each leaf based on the best target value. Alternatively,

the leaf might include a probability vector indicating the possibility of the target feature having a particular value. By moving from the root of the tree down to a leaf, instances are categorized based on the outcomes of the tests along the way. To evaluate whether or not a potential client will respond to a direct mailing, a decision tree like the one illustrated in the accompanying image is employed. Internal nodes are represented by circles, whilst leaves are represented by triangles. This decision tree has both nominal and numerical characteristics, which is worth mentioning. When it comes to direct mailing, the analyst may use this classifier to predict a potential customer's reaction (by sorting it along the tree) and determine the behavioral traits of the whole possible customer group. Each node is labeled with the property it is testing, and the values that correspond to that attribute are labeled on its branches. [26]

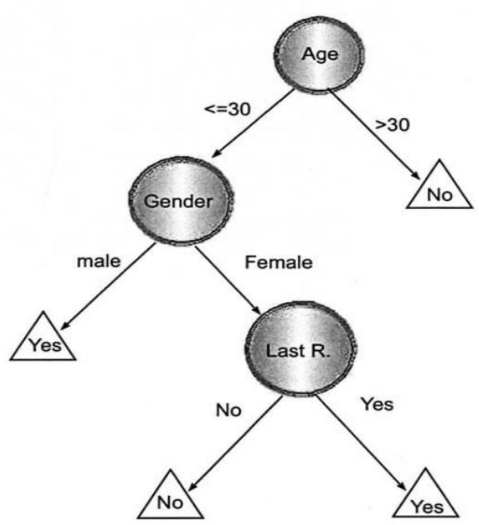

Fig. 2-13: Decision Tree Presenting Response to Direct Mailing. [26]

- Random Forest

To classify an input vector, the random forest classifier is made up of several tree classifiers, each of which is created using a random vector sampled independently from the input vector, and each tree casts a unit vote for the most popular class [27]. The random forest classifier used in this work employs randomly selected attributes or a combination of features at each node to construct a tree. Bagging, a method of creating a training data set by randomly drawing with replacement N samples, where N is the size of the original training set, was used for each feature/feature combination chosen. The most commonly selected class among all the forest tree predictors is used to categorize any instances (pixels). The construction of a decision tree requires choosing an attribute selection measure and a pruning method. There are numerous approaches to selecting characteristics for decision tree induction, and the

majority of methods explicitly give a quality measure to the attribute. The most often used attribute selection measures in decision tree induction are the Information Gain Ratio criteria and the Gini Index. The Gini Index is a measure of an attribute's impurity in respect to the classes that are utilized by the random forest classifier as an attribute selection measure.

The Gini index may be stated as choosing one instance (pixel) at random from a specified training set T and asserting that it belongs to some class

$$\sum_{j \neq i} \sum (f(C_i, \mathrm{T})/|\mathrm{T}|)(f(C_j, \mathrm{T})/|\mathrm{T}|)$$

where f (C_i , T)/ |T| is the probability that the selected case belongs to class C_i . [28]

- Naïve Bayes

The Naive Bayes classifier is based on the notion that, given a goal value, attribute values are conditionally independent [29]. In other words, the probability of observing the conjunction a_1, a_2,.....a_n is just the sum of the probabilities for the component characteristics, given the instance's goal value:

$$\mathbf{P(a_1, a_2,.....a_n | v_j) = \arg\max_{v_j \varepsilon V} a \prod_i P(a_i | v_j)}$$

$$\mathbf{V_{NB} = \arg\max_{v_j \varepsilon V} P(v_j) \prod_i P(a_i | v_j)}$$

In essence, the Naive Bayes Classifier eliminates any input dependencies, such as correlations, and reduces a multivariate issue to a set of univariate problems. In a Naive Bayes classifier, the number of distinct P(a_i| v_j) terms that must be estimated from the training data is simply the number of distinct attribute values multiplied by the number of distinct target values—a much smaller number than if we estimate the P(a_1, a_2,.....a_n| v_j) terms as Bayesian theory requires. [30]

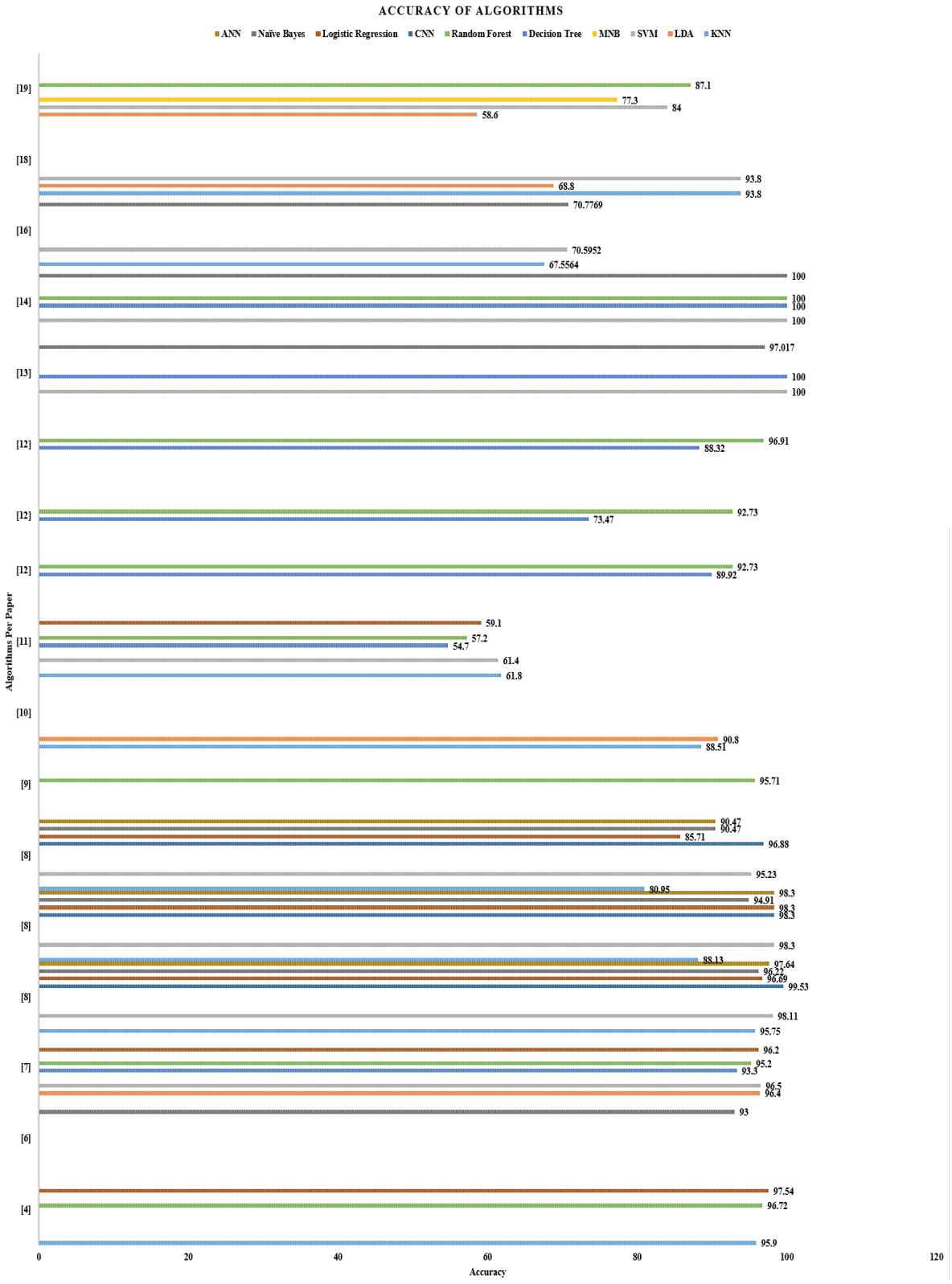

Fig. 2-14: Accuracy of Algorithms Used in The Papers.

The accuracy of the selected papers in detecting ASD is displayed in the figure. We can observe that four algorithms, SVM, Decision Tree, Random Forest, and Majority Model, all had 100% accuracy. Researchers used SVM, Decision Tree, and Random Forest algorithms and achieve 100% accuracy in [14].

Following our review of all of the documents, we have chosen to build an automatic dashboard. For this, a toddler dataset with 1054 occurrences and 19 characteristics was used [31]. Our research was conducted using Tableau. The procedures followed for our research will be detailed in full in the next part.

Chapter 3: Research Methodology

In this research, our goal was to diagnose autism through machine learning and make an interactive dashboard so that researchers can easily analyze the world's autism condition. According to it, we started to find the related papers. After that, we have selected the most informative papers for review. After reading those papers we started to write a literature review. Where we discussed how they conduct their research and what was their defects and also how to overcome those defects. After finishing our literature review, we technically started our research implementation. So that we started to find a good dataset to conduct our research.

As the determination of ASD and making a dashboard is the goal of this paper. A toddler's secondary data was collected from Kaggle. This dataset has consisted of 1054 instances with 19 attributes including the class variable. It was surveyed by Dr. Fadi Thabtah on July 22, 2018. A1-A10 attributes were some questions asked in the Toddler Application. The answer to this question was mapped to "1" or "0". The summation of the answered questions was stored inside another attribute which is Scored by Q-chat-10. The remaining attributes were Age, Sex, Ethnicity, jaundice, Family member with ASD history, who are taken the screening, and Class variable. The questions of this data set are given below.

Table 3-A: Questions of the selected dataset.

Variable	QCHAT-10 Features (18-36 months)	AQ-10-Child Features (4-11 years)	AQ-10-Adolescent (12-15 years)	AQ-10-Adult Features (16 and older)
Q1	Does your child look at you when you call his/her name?	S/he often notices small sounds when others do not	S/he notices patterns in things all the time	I often notice small sounds when others do not
Q2	How easy is it for you to get eye contact with your child?	S/he usually concentrates more on the whole picture rather than the small details	S/he usually concentrates more on the whole picture rather than the small details	I usually concentrate more on the whole picture rather than the small details
Q3	Does your child point to indicate that s/he wants something?	In a social group, s/he can easily keep track of several different peoples conversation	In a social group, s/he can easily keep track of several different peoples conversations	I find it easy to do more than one thing at once

Q4	Does your child point to share an interest with you?	S/he finds it easy to go back and forth between different activities	If there is an interruption, s/he can switch back to what s/he was doing very quickly	If there is an interruption, I can switch back to what I was doing very quickly
Q5	Does your child pretend?	S/he does not know how to keep a conversation going with his/her peers	S/he frequently finds that s/he does not know how to keep a conversation going	I find it easy to read between the lines when someone is talking to me
Q6	Does your child follow where you are looking?	S/he is good at social chit-chat	S/he is good at social chit-chat	I know how to tell if someone listening to me is getting bored
Q7	If you or someone else in the family is visibly upset, does your child show signs of wanting to comfort them?	When s/he is reading a story, s/he fonds it difficult to work out the characters intentions or feelings	When s/he was younger, s/he used to enjoy playing games involving pretending with other children	When I am reading a story I find it difficult to work out the characters intentions
Q8	Would you describe your child's first word as:	When s/he was in preschool, s/he used to enjoy playing pretending games with other children	S/he finds it difficult to imagine what it would be like to be someone else	I like to collect information about categories of things
Q9	Does your child use simple gestures?	S/he finds it easy to work out what someone is thinking or feeling just by looking at their face.	S/he finds social situations easy	I find it easy to work out what someone is thinking or feeling just by looking at their face
Q10	Does your child stare at nothing with no apparent purpose?	S/he finds it hard to make new friends	S/he finds it hard to make new friends	I find it difficult to work out peoples intentions

Before going to analysis, the dataset checked for missing data. Then we have considered "Age", "Scored by Q-chat-10"," Sex", "Ethnicity" and "Family member with ASD history" attributes as the others data do not have an impact on our analysis. We have converted Family members with ASD, Ethnicity, Sex, and Class attribute values to numeric data. We have created new datasets by separating the dataset based on the

Ethnicity attribute. Weka tool is used for analyzing the data and tableau is used for visualization. The full process of our research is shown in the following flow chart.

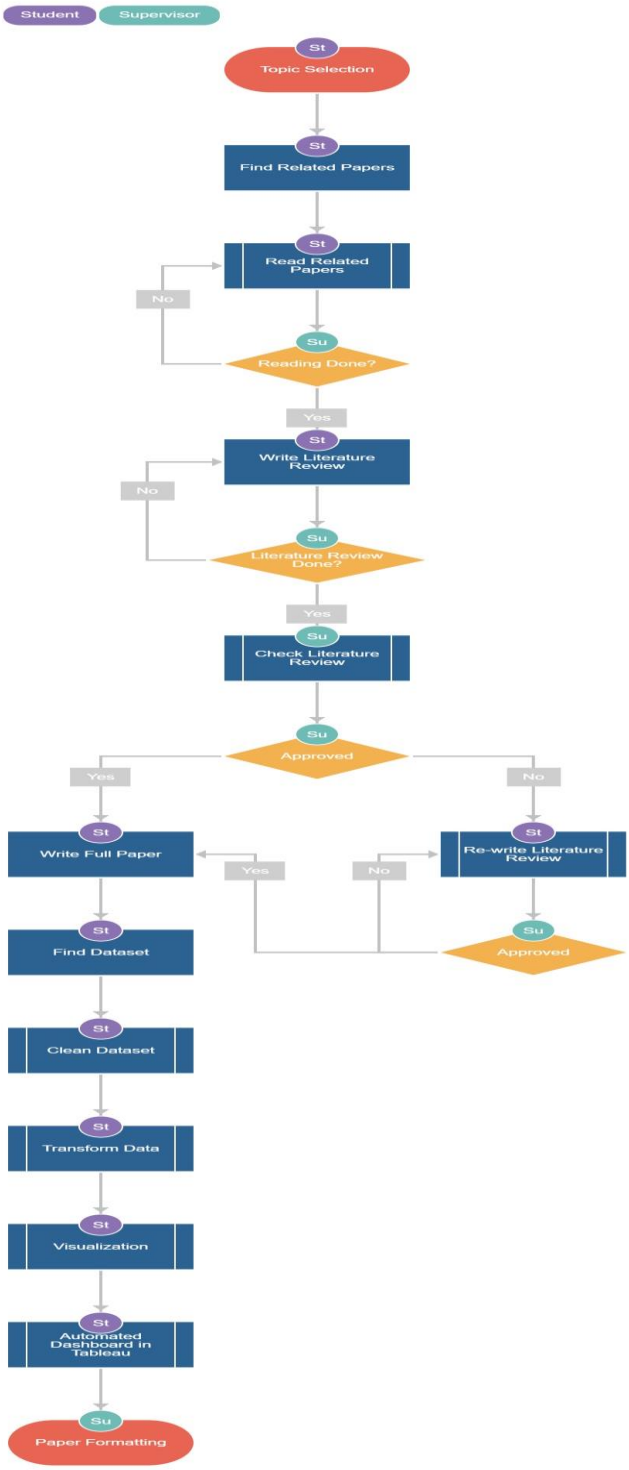

Fig. 3-1: Flow Chart of The Research.

Chapter 4: Development of Automated Dashboard to Detect ASD

We've demonstrated our created dashboard for ASD data analysis in this part. This dashboard makes it simple to evaluate data. The present design allows users to see findings both numerically and graphically. This is an automated dashboard. The dashboard created to evaluate ASD data is shown in Figure 4-1.

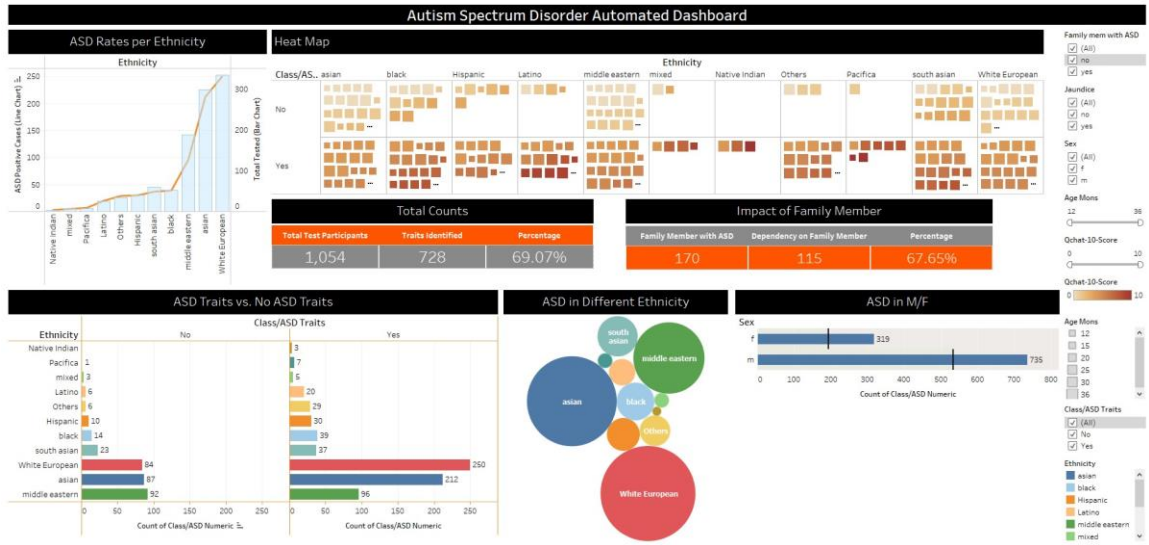

Fig. 4-1: Automated Dashboard for ASD Analysis

Chapter 5: Result Analysis

5.1 Analysis of The Dashboard

In this section different dashboard parts and analyses utilizing those parts are shown. This dashboard provides an interactive summary of ASD characteristics among people all across the world. Our dashboard includes features such as ASD Rates by Ethnicity, Heat Map, Total ASD Counts, Impact of Family Member, ASD Traits vs. No ASD Traits, ASD in Different Ethnicities, and ASD by Gender.

ASD Rates per Ethnicity

One can see how many participants in the toddler dataset took the ASD test and how many of them showed ASD traits in this area of the dashboard. The bars in this graph reflect the overall number of test cases by ethnicity, whereas the line shows the number of people who took part in this test and had ASD symptoms. Figure 5-1 indicates that White Europeans account for the bulk of test instances, whereas Native Americans account for the least. This graph also shows that when the number of test cases is high, the number of ASD Traits Identified people is usually high as well. For South Asians, though, the situation is different. While there were 53 black test participants, 39 of whom had ASD characteristics, there were 60 South Asian test participants, 37 of whom had ASD characteristics. Hovering the mouse over any individual bar or line point displays information about that bar/line point.

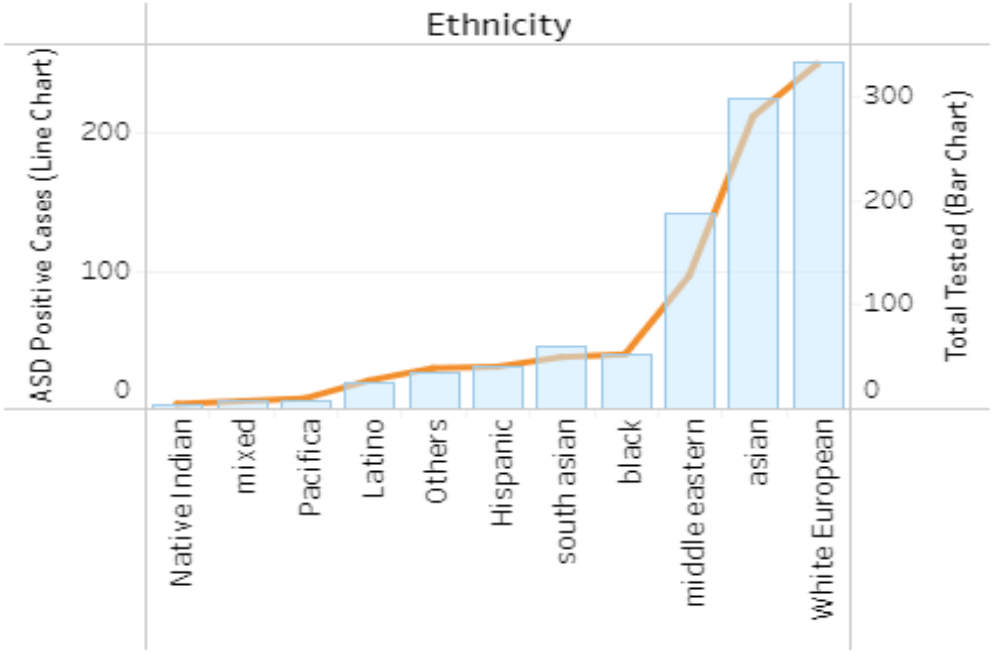

Fig. 5-1: ASD Rates Per Ethnicity

Below is a table created from this graph.

Table 5-A: Ethnicity wise identified ASD Traits

Ethnicity	Test Participants	Traits Identified	Percentage
White European	334	250	74.85%
Asian	299	212	70.90%
Middle Eastern	188	96	51.06%
Black	53	39	73.58%
South Asian	60	37	67.67%
Hispanic	40	30	75%
Others	35	29	82.86%
Latino	26	20	76.92%
Pacifica	8	7	87.5%
Mixed	8	5	62.5%
Native Indian	3	3	100%

According to the table above, around 334 White Europeans were tested, and about 250 of them were ASD positive, which is about 75 percent when compared to tasted vs positives. Native Indians made up the smallest percentage of individuals examined, however those who were tested all exhibited ASD symptoms. There may be some difference observable if the number of tasted patients is raised.

Heat Map

The QChat-10-Score, ethnicity, and age are shown on this graph. The QChat-10-Score ranges from 0 to 10. Figure 5-2 demonstrates that a higher QChat-10-Score indicates a deeper color, and the size of the square indicates age. A bigger square represents a greater age. Each square also represents a different personality. In this Heat Map, we've included various criteria such as ASD Family Member, Jaundice, Sex, Age, and QChat-10-Score. One may experiment with these filters to see some particular results. If we change the QChat-10-Score filter range from 0 to 3 on the dashboard, we can observe that there are people with ASD characteristics. Hovering the mouse over each individual square shows information such as the person's age in months, ethnicity, QChat-10-Score, and whether or not he or she has ASD.

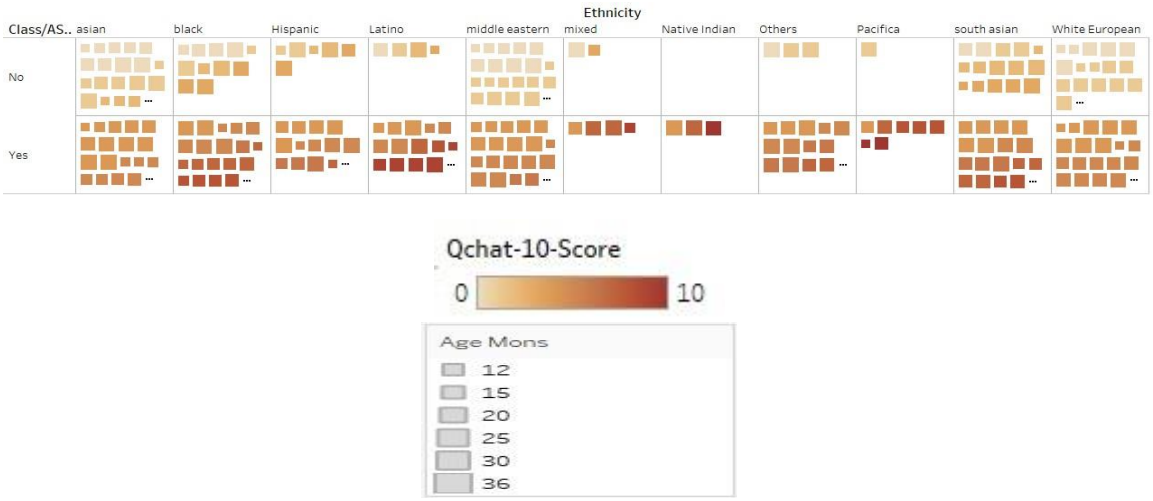

Fig. 5-2: Heat Map

Total Counts

In this portion of the dashboard, it can be seen that the total number of participants all over the world in this dataset [31] was 1,054 and from them, 728 persons had ASD Traits. So, the percentage of ASD Traits identified out of total test participants is 69.07%.

Impact of Family Member

We looked for test participants who had a family member with ASD and counted how many of them showed ASD characteristics. 67.65 percent is the percentage.

Chances of ASD based on Survey

In this survey, each participant was given ten questions. The questions in the dataset [31] were labeled from A1 to A10. In this section, we determined how many participants exhibited ASD characteristics if they replied "Yes" to the questions and how many did not. Among 1054 participants, 594 people said yes to question A1. Among these 594 people, ASD characteristics were found in 532 of the individuals, whereas ASD traits were not found in the rest 62 of the people.

From this graph, we generated a table that shows the percentage of having ASD traits if an individual replied “Yes” to a question.

Table 5-B: Chances of Having ASD

Question Label	Found ASD Traits	Without ASD Traits	% of having ASD Traits
A1	532	62	89.56%
A2	439	34	92.81%
A3	390	33	32.20%
A4	496	44	91.85%
A5	519	34	93.85%
A6	557	51	91.61%
A7	604	81	88.18%
A8	438	46	90.50%
A9	497	19	96.32%
A10	470	148	76.05%

From this table, it can be seen that if a participant replied “Yes” to question A9 has the highest chance of having ASD traits in them.

ASD Traits vs. No ASD Traits

We divide the total number of participants’ ethnicity wise into two sections based on ASD Traits. In one section those who had ASD Traits among total participants and in the other section who had not.

ASD in Different Ethnicity

This graph aids in determining which ethnicity has the most participation. The wider the ethnicity circle, the greater the number of participants. In this graph, we also include filters such as ASD Traits and Ethnicity. With the ASD Traits filter, you can only see people who have ASD traits or don't have ASD traits. The ethnicity filter will spotlight

the ethnicities you choose. Hovering the mouse over any circle reveals the details of that specific circle like the circle is for which ethnicity and how many participants have ASD traits on that ethnicity.

Gender-Based ASD

The number of males and females among the participants is shown below. Show how many of them have ASD traits as well. Figure 5-3 depicts 735 male individuals, 534 of whom had ASD features which were about 73% compared with the total number of males, and 319 female participants, 194 about 61% compared with the total number of females of whom had ASD traits. The black line in the bars represents the number of persons having ASD Traits.

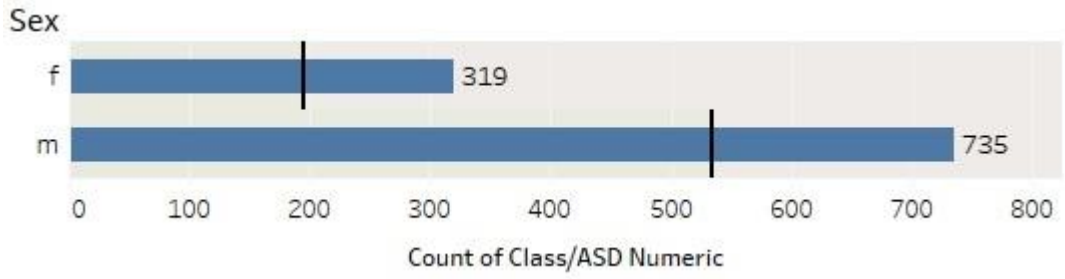

Fig 5-3: Gender-Based ASD

5.2 Analysis of Implemented Models

Based on the dataset K-Nearest Neighbor classifier score is 0.978, Logistic Regression classifier and SVM classifier score is 1.0.

Here's the AURIC scores of different classifiers-

Random Prediction: AUROC = 0.500
Random Forest: AUROC = 1.000
Naive Bayes: AUROC = 1.000
Decision Tree Prediction: AUROC = 1.000
Gradient Boosting Prediction: AUROC = 1.000
K-Nearest Neighbors Prediction: AUROC = 0.998
Logistic Regression: AUROC = 1.000
Support Vector Machine: AUROC = 1.000

After calculating ROC, here's the ROC Curve

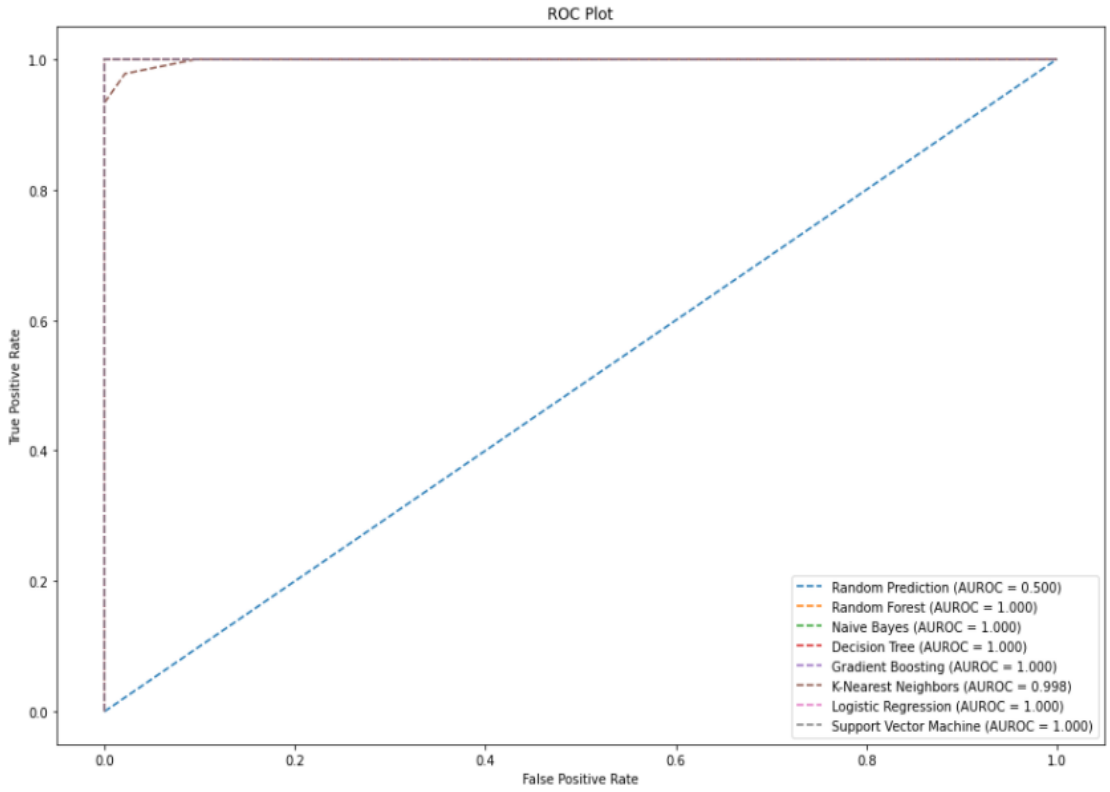

Fig. 5-4: ROC Curve.

Chapter 6: Conclusion

6.1 General Discussion

In this study, we created an interactive dashboard for analyzing ASD data, and also, we have reviewed some research articles based on supervised machine learning. A full background study of the topic follows the development. About 25 research articles are reviewed based on the background. During the review, the most often utilized approaches, guidelines, and tools by researchers are prioritized. In this review, we have also discussed how they conduct their research and what was their defects.

We've also included a table based on the accuracy, sensitivity, and specificity of supervised machine learning algorithms. The definitions and working principles of various basic supervise learning algorithms are presented in this article. How the researchers using these algorithms to detect autism we have also been discussed. Signs and symptoms of ASD are mentioned in this paper. This paper will be very helpful for those who want to research in this field and it will also give a direction to detect ASD using machine learning.

Finally, we made an interactive data visualization dashboard to analyze ASD data. According to our findings, the rate of positives in 1054 cases is about 70%. Males have a higher prevalence of ASD than females. When we look at the number of people who have ASD, we can see that White Europeans are the most afflicted. Early detection of ASD patients can help them become more self-sufficient in the future, and machine learning's ability to interpret data can greatly aid researchers. The created dashboard can be used as a standalone tool by loading pre-processed data and selecting an algorithm to enable machine learning. Users will be able to generate reports and interactively visualize data based on algorithms and data properties.

Early detection is critical since it allows for treatment to begin much sooner. For young children with autism, this implies that the skills they need to realize their full potential are taught at a young age when brain plasticity is much more prominent and intervention's influence is much broader. The traditional diagnosis of ASD is very costly and it is very difficult for some families to bear this cost. Using the machine learning technique, we can get rid of this problem. Machine Learning can help to diagnose Autism Spectrum Disorder easily. Such information, if predicted in advance, can provide valuable insights to clinicians, allowing them to tailor their diagnosis and treatment to each patient. It can be very cost-friendly and also it will save lots of time.
So before researching autism spectrum disorders (ASD) or try to predict ASD using machine learning this study can support other researchers in this field and also it can help them to know about how they can use supervised learning classification algorithms to predict ASD and also which algorithm and tools will suitable for their research. By using our dashboard, they can also get a clear view of the world’s Autism conditions and they

can easily analyze ASD data based on “Age”,” Sex”, “Ethnicity”, “Family member with ASD history” and “some important questionnaires”.

6.2 Future Work

We hope to develop an application for identifying ASD in the future. We will use machine learning techniques to implement this application. In this application users only need to answer some simple questions. Based on those answers the system will show that whether the person has autism or not.

If the user has autism what is the level or type of it so that the user can easily understand what kind of treatment do, they need. By using this application users can save their money from traditional medical diagnosis techniques. So, this will be very cost-friendly and also it will be very user-friendly so that those people who cannot bear the high cost of medical diagnose can easily check whether they or their child has autism or not.

We will also try to make our survey by going to the autistic schools or organizations. So that our dashboard can be up to date and also it will increase our dashboard’s efficiency which can be very beneficial for ASD data analysis and researches. It will also be very effective for our system as we will be able to implement our machine learning model with our real-life data and it will also increase the predictive accuracy of our system.

References

[1] G. Powell, S. Wass, J.Erichsen and S. Leekam, " First Evidence of the Feasibility of Gaze-Contingent attention training for school children with autism", Autism, 2016 in SAGE Journel.

[2] Kundu, Rakhee & Das, Mr. (2019). Predicting Autism Spectrum Disorder in Infants Using Machine Learning. Journal of Physics: Conference Series. 1362. 012018. 10.1088/1742-6596/1362/1/012018.

[3] M. Mythili and A. Shanavas, "A study on Autism Spectrum disorders using classification techniques," 2014.

[4] Abdullah, Azian & Rijal, Saroja & Dash, Satya. (2019). Evaluation on Machine Learning Algorithms for Classification of Autism Spectrum Disorder (ASD). Journal of Physics Conference Series. 1372. 012052. 10.1088/1742-6596/1372/1/012052.

[5] Ritchie, H., 2017. Neurodevelopmental disorders. [online] Our World in Data. Available at: <https://ourworldindata.org/neurodevelopmental-disorders#prevalence-of-autistic-spectrum-disorders> [Accessed 11 August 2021].

[6] V. J. GeethaR. Vivek, "Classification ofAutism Spectrum Disorder Data using Machine Learning Techniques," *International Journal of Engineering and Advanced Technology (IJEAT)*, vol. 8, no. 6S, pp. 365–369, Aug. 2019.

[7] Duda, M., Ma, R., Haber, N., & Wall, D. P. (2016). Use of machine learning for behavioral distinction of autism and ADHD. Translational Psychiatry, 6(2), e732–e732. doi:10.1038/tp.2015.221.

[8] Raj, S., & Masood, S. (2020). Analysis and Detection of Autism Spectrum Disorder Using Machine Learning Techniques. Procedia Computer Science, 167, 994–1004. doi: 10.1016/j.procs.2020.03.399.

[9] S. B. Shuvo, J. Ghosh and A. S. Oyshi, "A Data Mining Based Approach to Predict Autism Spectrum Disorder Considering Behavioral Attributes," *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, Kanpur, India, 2019, pp. 1-5, doi: 10.1109/ICCCNT45670.2019.8944905.

[10] Altay, O., & Ulas, M. (2018). Prediction of the autism spectrum disorder diagnosis with linear discriminant analysis classifier and K-nearest neighbor in children. 2018 6th International Symposium on Digital Forensic and Security (ISDFS). doi:10.1109/isdfs.2018.8355354.

[11] Parikh, M. N., Li, H., & He, L. (2019). Enhancing Diagnosis of Autism with Optimized Machine Learning Models and Personal Characteristic Data. Frontiers in Computational Neuroscience, 13. doi:10.3389/fncom.2019.00009.

[12] K. S. Omar, P. Mondal, N. S. Khan, M. R. K. Rizvi, and M. N. Islam, "A machine learning approach to predict autism spectrum disorder," in *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, 2019.

[13] B. Deepa and K. S. Jeen Marseline, "Exploration of autism spectrum disorder using classification algorithms," *Procedia Comput. Sci.*, vol. 165, pp. 143–150, 2019.

[14] R. A. Musa, M. E. Manaa, and G. Abdul-Majeed, "Predicting autism spectrum disorder (ASD) for toddlers and children using data mining techniques," *J. Phys. Conf. Ser.*, vol. 1804, no. 1, p. 012089, 2021.

[15] T. Akter *et al.*, "Machine learning-based models for early stage detection of autism spectrum disorders," *IEEE Access*, vol. 7, pp. 166509–166527, 2019.

[16] B. Tyagi, R. Mishra and N. Bajpai, "Machine Learning Techniques to Predict Autism Spectrum Disorder," 2018 IEEE Punecon, 2018, pp. 1-5, doi: 10.1109/PUNECON.2018.8745405.

[17] Alwidian, Jaber & Elhassan, Ammar & Rawan, Ghnemat. (2020). Predicting Autism Spectrum Disorder using Machine Learning Technique. 2277-3878. 10.35940/ijrte.E6016.018520.

[18] D. H. Oh, I. B. Kim, S. H. Kim, and D. H. Ahn, "Predicting autism spectrum disorder using blood-based gene expression signatures and machine learning," *Clin. Psychopharmacol. Neurosci.*, vol. 15, no. 1, pp. 47–52, 2017.

[19] S. H. Lee, M. J. Maenner, and C. M. Heilig, "A comparison of machine learning algorithms for the surveillance of autism spectrum disorder," *PLoS One*, vol. 14, no. 9, p. e0222907, 2019.

[20] S. B. Imandoust And M. Bolandraftar. "Application of K-Nearest Neighbour(KNN)Approach for predicting Economic Events:Theoretical Background"S B Imandoust et al. Int. Journal of Engineering Research and Application,vol. 3,Issue 5,pp.605-610,2013.

[21] S. Thirumuraganathan, “A Detailed Introduction to K-Nearest Neighbour(KNN) Algorithm”.” 2010.

[22] Begum, S., Chakraborty, D., & Sarkar, R. (2015). *Data Classification Using Feature Selection and kNN Machine Learning Approach. 2015 International Conference on Computational Intelligence and Communication Networks (CICN).* doi:10.1109/cicn.2015.165

[23] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” *the Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003.

[24] Y. Zhang, “Support vector machine classification algorithm and its application,” in *Communications in Computer and Information Science*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 179–186.

[25] N. Deng, : “Support vector Machine Theory, algorithms and Development.” Science Press, Beijing, p. 176, 2009.

[26] L. Rokach and O. Maimon, *Decision Trees. Data Mining and Knowledge Discovery Handbook*, vol. 165–192, pp. 1007 0–387–25465– 9.

[27] L. Breiman, “RANDOM FORESTS--RANDOM FEATURES,” Sep-1999.

[28] M. Pal, “Random forest classifier for remote sensing classification,” *Int. J. Remote Sens.*, vol. 26, no. 1, pp. 217–222, 2005.

[29] T. M. Mitchell, “Artificial neural networks,” *Machine learning*, vol. 45, pp. 81–127, 1997.

[30] D. Soria, J. M. Garibaldi, F. Ambrogi, E. M. Biganzoli, and I. O. Ellis, “A ‘non-parametric’ version of the naive Bayes classifier,” *Knowl. Based Syst.*, vol. 24, no. 6, pp. 775–784, 2011.

[31] “Autism screening data for toddlers,” https://www.kaggle.com/fabdelja/autism-screening-for-toddlers

[32] D. Eman and A. W. R. Emanuel, "Machine Learning Classifiers for Autism Spectrum Disorder: A Review," *2019 4th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, 2019, pp. 255-260, doi: 10.1109/ICITISEE48480.2019.9003807.