Enhancing Autism Detection through Machine Learning Models Focusing on Behavioral Analysis

by

Tania Sultana Tamanna
20101384
Mahmudul Hassan
24141223
Razin Sumyta Monsoor
20101529
Shehrin Hoque
20101148
Rageeb Mohammad Ridwan
23241073

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

Department of Computer Science and Engineering
Brac University
May 2024

****

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at BRAC University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Tania Sultana Tamanna

20101384

Mahmudul Hassan

24141223

Razin Sumyta Monsoor

20101529

Shehrin Hoque

20101148

Rageeb Mohammad Ridwan

23241073

Approval

The thesis/project titled "Enhancing Autism Detection Through Machine Learning Models Focusing on Behavioral Analysis " submitted by

1. Tania Sultana Tamanna (20101384)
2. Mahmudul Hassan (24141223)
3. Razin Sumyta Monsoor (20101529)
4. Shehrin Hoque (20101148)
5. Rageeb Mohammad Ridwan (23241073)

Of Summer, 2024 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on May 21, 2024.

Examining Committee:

Supervisor:
(Member)

Dr. Md. Ashraful Alam, PhD

Associate Professor
Department of Computer Science and Engineering
Brac University

Program Coordinator:
(Member)

Dr. Md. Golam Rabiul Alam, PhD

Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Sadia Hamid Kazi, PhD

Chairperson and Associate Professor
Department of Computer Science and Engineering
Brac University

Abstract

Autism Spectrum Disorder (ASD) is a complex and enduring condition characterized by challenges related to communication and behavior. It is a complex neurological disorder related to an individual's psychological difficulties, which eventually impact their behavior or reactions to the outside world. While it is feasible to detect autism symptoms at any stage of an individual's life, there is a greater likelihood of detection within the first two years after birth, as differences in typical activities, communication gaps, or a lack of understanding typically become more noticeable during this early developmental period. The paper suggests a deep learning-based method that makes use of behavior to identify autism in both adults and children by analyzing their behavioral characteristics through machine learning approaches and determining a process that makes autism detection easier and cost-effective. The recommended approach works by behavioral monitoring of children and adult datasets that were collected from online platforms and went through successive processing and finally, those datasets were applied to different models with the help of binary classification towards the end to determine autism detection correctly. Behavioral data includes a range of indicators, including patterns of social interaction, communication abilities, and recurrent actions. For behavior analysis, we implemented specific models like KNN, Random Forest, CatBoost, SVM, GradientBoost, and Logistic Regression and also ensembled models by incorporating a few of our pre-trained models together to give better accuracy rates. We have also integrated different confusion matrices in our paper. This will help in evaluating and fine-tuning the performance of our detection model. We have acquired behavioral datasets from publicly available platforms called UC Irvine Machine Learning Repository and Kaggle. Our primary goal is to improve the accuracy of autism detection or contribute to research by developing a comprehensive research paper. This paper aims to facilitate model comparisons and streamline the autism detection process using advanced machine learning techniques available today.

Keywords: Autism detection, Machine learning, KNN, Logistic Regression, Random Forest, CatBoost, GradientBoost, SVM, behavior analysis.

Acknowledgement

First of all, all praise to Allah that our thesis have been completed without any major interruption and without changing any members.
Secondly, to our supervisor Dr. Md. Ashraful Alam sir for his kind support and RA Efaz sir for their constant support and advice in our work. Both of them helped us whenever we needed help.
Lastly to our parents, who supported us throughout our university life.

Table of Contents

List of Figures

List of Tables

Nomenclature

The next list describes several symbols & abbreviation that will be later used within the body of the document

ANN Artificial Neural Network

ASD Autism Spectrum Disorder

KNN K-Nearest Neighbours

LR Logistic Regression

RF Random Forest

SVM Support Vector Machine

VGG Visual Geometry Group

Chapter 1

Introduction

1.1 Exploring Autism: Understanding its Complexities and Nuances

Autism Spectrum Disorder (ASD) represents a lifelong condition that is associated with the development of the human brain. It's a rapidly growing category of disabilities characterized by repetitive behavior patterns, specific interests, and challenges in social interactions. Autistic individuals often encounter difficulty in expressing themselves, whether through verbal communication or non-verbal means such as gestures, facial expressions, and physical touch. Individuals with autism may experience difficulties in the learning process, and their skill development may occur unevenly, with some areas progressing at a different pace than others. This variability is a characteristic feature of autism. Now, individuals with autism can also exhibit remarkable abilities in various memory-intensive activities, excelling in areas such as mathematics, art, and more. These specialized areas of strength are sometimes referred to as "splinter skills" or "special talents" and can be a notable aspect of autism. The classification of autism as high-functioning or low-functioning is contingent upon the individual's level of severity. The high-functioning autistic individuals may excel in communication, have higher IQ levels, or demonstrate slightly more proficiency in typical daily activities compared to their low-functioning counterparts. Autistic individuals typically exhibit symptoms that differ from those seen in neurotypical individuals. The symptoms of autism can include lack of eye contact, a limited range of interests or intense focus on specific topics, repetitive behaviors, heightened sensitivity to sensory stimuli, difficulty in social engagement, aversion to physical touch, and challenges in adapting to changes in routines. Occasionally, individuals with autism can become overwhelmed in certain situations, leading to what is known as a meltdown, they may express their distress by crying, screaming, or engaging in physical behaviors or they might completely withdraw and become unresponsive. A considerable amount of research has been dedicated to enhancing the accuracy of autism detection through the utilization of various algorithms.

1.2 Challenges

Autism detection is a multifaceted sphere that involves not only physical characteristics but also hugely relies on behavioral tendencies. In our research paper, we

aim to underscore this comprehensive approach by focusing on behavior analysis of adults and children. Among the challenges that we faced, the hardest one has to be data collection. Since autism is a sensitive topic and also involves privacy issues of patients and different individuals we had a hard time collecting the data both primarily and secondarily. Among the available secondary data sources, many suffer from authenticity issues, which was a primary reason we could not utilize image datasets in our thesis. Additionally, autism-related centers and hospitals refused to provide patient information due to privacy concerns, and personal data collection was infeasible due to the insufficient quantity of data for meaningful research. We encountered challenges in identifying consistent behavioral tendencies across the same autistic patients sourced from different platforms, which, combined with the collected images, complicated our efforts to conduct a comparative study using exact patient datasets. Nonetheless, we proceeded by implementing the models separately with different datasets to individually analyze accuracy rates based on behavior datasets.

1.3 Problem Statement

Our paper aims to explore behavioral characteristics using conventional models to further refine autism detection methodologies. We collected behavioral data from both autistic patients and neurotypical individuals from the available online source called UC Irvine Machine Learning Repository and Kaggle, then we conducted preprocessing and testing of the data and then applied machine learning algorithms like Logistics Regression, KNN, GradientBoost, SVM, CatBoost, and RandomForest to analyze and extract valuable insights from the dataset. We will obtain our predictions by training the models, which will enable us to identify and ascertain symptoms of autism based on the analysis of behavioral feature data. Here we have employed different algorithms to train datasets, facilitating the extraction of components associated with human behaviour and expressions. The behavioral dataset included different observations of characteristics and everyday behavior that were analyzed through the different models. Recently, there has been a focus on analyzing data related to physical biomarkers and assessing clinical data through the application of machine-learning approaches. In our paper, we focused on detecting autism during infanthood and adulthood focusing on behavioral tendencies as we cannot confine the detection of autism only to one limited sphere.

1.4 Research objective

The goal of detecting autism is to identify if someone has autism by looking at their behavior patterns and characteristics. This research on autism detection will help in:

Early Identification : We can find out if a child has autism when they are young, which helps them get special help early on.

Understanding Behaviors: We can learn more about how people with autism behave and express themselves through their expressions and actions.

Objective Diagnosis: It helps doctors make a clear and certain diagnosis based on behaviors, which is more accurate to identify if someone has autism or not.

Enhancing Accuracy rate: We want to enhance the accuracy rates of different models through our research work to contribute to the medical diagnosis sphere.
Equality: We can make sure that everyone, no matter where they come from or who they are, gets a fair chance of being diagnosed with autism.
Helpful Tools: We can create tools and technology that make it easier and less scary for children to go through the diagnosis process.
Tracking Progress: We can see how a person's expressions and behaviors change over time. It will help us understand how well interventions are working.
Privacy and Safety: We can learn how to protect people's privacy and keep their behavior data safe while still using them to help with autism diagnosis.
Training Models: These pre-trained models can be used by others in the future.
Raising Awareness: Detecting autism early can help people understand the victim better and be kinder to those who have it. It will make the world a better place.
Inexpensive diagnosis: Autism diagnosis is usually very expensive and time-consuming. We want to make it easier for general people.
Exploring Machine Learning Algorithms: Study and explore machine learning algorithms like KNN, Random Forest, Logistic Regression, and so on.
The main goal of this research is to detect autism with the help of behavioral data to alert the patient and their families to early treatment and make sure that they get special care from society. They should get all the opportunities to learn and grow like a normal child and as a person.

1.5 Thesis Structure

The remaining sections of this paper are structured as follows:

- The literature review for autism detection is in section 2.
- The work plan, data collection and data pre-processing of the research are in section 3.
- The description of the used models is presented in section 4.
- The results are analyzed in section 5.
- The results are compared in section 6.
- Conclusion and future works are presented in section 7.

Chapter 2

Literature Review

In paper [7], the differences between kids in Bangladesh who had been diagnosed with autism spectrum disorder (ASD) and those who didn't are examined. With the help of the Autism Barta App and fieldwork in Savar, researchers used a variety of approaches to figure out the characteristics of autism. They discovered that a number of variables including birth time and gender can also greatly affect the chance of developing autism. By examining more than 600 records, scientists were able to show how certain traits could be used to identify autism early on. The study also identified regional differences in autism frequency throughout Bangladesh providing insight into a variety of reasons. These discoveries offer insightful information to researchers and medical experts. On top of that, it also promotes better comprehension and treatment of autism spectrum illnesses.

In another paper [13], Convolutional Neural Network (CNN) classifier with transfer learning is used to identify autistic children. CNN is a core deep learning algorithm that extracts key characteristics from images for classification using pooling and convolution processes. Besides, transfer learning imitates how people apply previously acquired information to new tasks by employing a pre-trained model for a secondary task. It also gives better accuracy. To categorize photos into different categories including autistic and non-autistic the study used a pre-trained VGG-19 model, which is a 19-layer deep convolutional neural network trained on over a million images from the ImageNet database. Confusion matrices and classification reports were used to calculate metrics like specificity, sensitivity, and accuracy. The study acquired an accuracy of 84.67% on the validation data.

In another research [4], the Binary Firefly algorithm was used to eliminate the noisy behavioral features from the behavior dataset that was used to classify autism. They tried to show that the Binary Firefly algorithm can be a better option for optimum feature selection than ranking-based feature reduction algorithms. After the feature reduction (from 21 to 10), they found a significant change in accuracy for the KNN model. Before the feature reduction, KNN had an accuracy of 87.67% in detecting ASD, which increased to 93.84% afterward. Other models like the J48 decision tree and Naïve Bayes also slightly improved and their final accuracies in detecting ASD were 92.12% and 95.55% respectively. However, the accuracies of their two best-performing models, SVM and MLP, decreased slightly after the feature reduction.

Another research [15], developed a web app using MobileNet, which is a CNN architecture-based deep learning model. They used image data from online sources to detect facial features that resemble those of ASD patients. First, they fine-tuned the layers and optimized the models (using the RMSprop optimizer for output error reduction) according to their dataset. Three deep learning models, namely MobileNet, Xception, and InceptionV3 were applied to the dataset and the accuracies in detecting ASD patients were pretty high. Although the accuracy for MobileNet was the highest in their research, a more complex model might have been a better choice for building an ASD detection application, as it would reduce the chance of missing the incredibly intricate facial features.

In another research paper [21], we can see the implementation of pre-trained deep learning models like Xception and VGG16 (based on Transfer Learning) to diagnose autism. They used facial images of children for their dataset. After preprocessing the data and training and testing the models, they found that the Xception model performed the best, topping VGG16. It had 91% accuracy whereas VGG16 had 78% in successfully detecting ASD in children. Xception, as a DL model, has a high model complexity, which can lead to finding better results since intricate facial features cannot be easily missed by it. However, Xception, due to its complex design, can require more computational resources to train and infer, making it less suitable to run on applications with constrained hardware capabilities. There is also a risk of overfitting the data while training the model, due to the larger capacity and complex structure of Xception. This can lead to decreased accuracy for unseen data.

Another research [17] worked on behavioral datasets of toddlers, children, adolescents, and adults to detect ASD. They used different types of feature scaling techniques like QT, and normalizer, for their datasets. They used 8 different machine learning algorithms, such as Ada Boost (AD), Linear Discriminant Analysis (LDA), Random Forest (RF), and K-Nearest Neighbors (KNN), to all 4 feature-scaled datasets. For the toddler and children datasets, AD had the highest accuracy among all the classifiers with a mean of 98.6%. For the adolescent and adult datasets, LDA had the highest accuracy with a mean of 98.075%. They tried to emphasize the claim that ASD detection is easiest during the early stages. They claimed that the limitation of their work was that they could not gather more data to build a generalized model that could be used by people of all age-groups. Lack of enough data can cause issues like overfitting, bias, etc.

Another research [23] also worked with behavior data for children. They conducted a survey that collected answers to critical questions about behavior of children. Their research targeted early detection of ASD, which is why they chose a single age group for their dataset. They chose to run KNN, Random Forest, SVM and Naïve Bayes algorithms on their data to accurately detect ASD in children. The highest accuracy was found for KNN (98%) and the lowest was found for SVM (83%). For RF and Naïve Bayes, their accuracies were 93% and 89%, respectively. They also mentioned the lack of large data which resulted in two of the algorithms they had tested overfitting. Consequently, the model might perform poorly on new, unseen data. Additionally, relying on subjective survey data might introduce biases which can affect the model's performance.

Another research [12] gathered datasets for toddlers, children, adolescents, and adults from the Kaggle and UCI ML repository. During the data preprocessing, they discarded highly co-linear features in the datasets and used methods like standardization and normalization to transform the features. Finally, to detect ASD, they used different algorithms like KNN, Logistic Regression (LR), Decision Tree, and so on. They also used models like ANN and MLP. LR gave the highest accuracy for all the datasets. The mean accuracy was 98.59% for LR. However, there might be some issues with choosing LR for their model, as LR is not designed to capture interactions between features. In real-world scenarios, interactions between variables might be essential, as this is often the case for ASD detection.

Looking into another paper [20], the researchers worked with a dataset that consisted of dialogs from actual parents of autistic children who had been undergoing communication, behavior, and speech therapy. They analyzed each sentence of the dialogs to identify potential ASD symptoms, extracted the relevant features, and then proceeded to train machine learning models to diagnose ASD with new unseen data. They used models like KNN, SVM, LR, and RF. The highest accuracy was found with SVM and LR (71%). KNN had an accuracy of 62% and RF had that of 69%. The data was collected through an online survey on social media, which might have contributed to inaccuracies in data and introduced bias. This likely resulted in the relatively low accuracy of the models.

Another research [19] worked with eye movement data and tried to implement machine learning algorithms to predict ASD. They designed a VR scene that included social stimuli like real-life objects, figures, landscapes, and recorded the eye movement to enrich their dataset. In their preliminary experiment, they recruited 107 adults, recorded their eye movements while watching the VR scene, and then had them complete a questionnaire that recorded their statements about imagination, social skills, behavior, and other important characteristics. All of these were converted into features, and then models like SVM, DT, RF, KNN, Naive Bayes, and Ensemble were tested on the gathered data for ASD detection. Ensemble had the highest accuracy at 73% whereas KNN had the lowest at 55%.

According to Rad and Furnanello, their research was based on the communication and social problems of young teens having ASD. Besides, while doing their research the characteristic movement of the young patients stood out noticeably. Children with autism have SMMs as an important part of their life. So, it was very important for the researchers to develop a technique that would be effective for locating the changes of young patients with abnormal behavior [5].

In another case, researchers utilized AI to develop a sequence of repetitive examples of strolling. This research was established by applying the motor and kinematic characteristics of strolling. To detect the problem, they used linear analysis of learning classification. The negative and positive parts of linear analysis helped in regulating the research. DSM-IV was used as an ASD screening tool and it was consistent in its result. With the help of tomography of the cerebrum, they observed significant control of autism [2]. Data sets from the ASD Tests program were used as a source

for another Artificial Neural Network in other research. So, 10 questions were used for collecting information. With those questions including the age and gender of the member, research was conducted.

In [3], a prediction system was created by using the previous records of long-time patients to distinguish the signs of ASD. Both traditional and rare types of ASD were detected using this prediction system. For creating a matrix, machine learning and colloquy theory were applied together to establish the system. It utilized all sets of associations to create rules for better results. It must be noted that their ASD system consumed more limited memory for conducting the research. Also, because of single-time database access, the system was faster [6].

For another work, K Nearest Neighbors (KNN) and Linear Discriminant Analysis (LDA) were utilized for the detection of autism in kids aged between 4-11. So, they used 70 percent of the data for the training, and the rest 30 percent were used for testing. As a result, those algorithms helped in getting more accurate results for prediction and reduced predictions that were biased [11].

Another research [18] suggests a deep learning model for early ASD identification using eye-tracking scan path data. The model was trained and assessed based on eye-tracking data from kids with and without ASD. However, the authors acknowledged that more study is required to verify the model on a bigger, more varied dataset since they had a relatively small sample size. Since this is such an important tool for early ASD diagnosis, we can substantially improve the detection accuracy through longitudinal analysis, using more advanced machine learning models and a more diverse and larger dataset.

In [14], the authors used VGG16 as their model for facial recognition of autism in children. Thus, they applied VGG16 for transfer learning as a pre-trained model. In the case of training and testing, they divided the dataset into 80 percent and 20 percent respectively.Moreover, the dataset was composed of East Asian children's facial features. As a result, after training they got a 93 percent accuracy rate which was a good percentage. Both models, after going through training and testing, classified children as autistic or normal. The accuracy rate for the Xception and VGG16 models was 91 percent and 78 percent respectively. Xception had the highest accuracy rate while VGG16 had the lowest accuracy rate So, Xception performed better than VGG16 in terms of performance level.

In another work [16], VGG19, Xception, and NASNETMobile were enforced for classifying autism and non-autism in children.These three models were trained and tested for extracting traits from facial images of children for classification purposes. As a result, the Xception model had the highest accuracy among the three models which was 91 percent. On the other hand, VGG19 and NASNETMobile had an accuracy rate of 80 percent and 78 percent respectively.So, the accuracy rate of NASNETMobile and VGG19 was too low compared to the Xception model.

The authors in [22] proposed using MobileNet, InceptionV3, Xception, EfficientNetB0, EfficientNetB7, and VGG16 for their research. Models were trained and

tested using a dataset from Kaggle. After testing, the accuracy rate of MobileNet, Xception, InceptionV3, VGG16, EfficientNetB0, and EfficientNetB7 was 88 percent, 87.7 percent, 86.1 percent, 86.3 percent, 85.6 percent, and 82.6 percent respectively. Therefore, MobileNet performed better than other models because of having a higher accuracy rate than other models. Conversely, the performance level of EfficientNetB7 was low.

In [9], the authors only implemented a deep learning model called MobileNet for autism detection. For extracting features and classification of images, they applied two dense layers of MobileNet. With the help of 3014 images, the model was trained and tested by splitting 90 percent of the data for training and 10 percent of the data for testing. As a result, they got 94.6 percent accuracy in the detection of autism which is a good accuracy rate. Autism can also be detected in people using behavioral features in the dataset.

Naive Bayes [1] and Logistic Regression [8] were used in the detection of autism in people. It was applied for all three datasets- adult, child, and adolescent. The accuracy rate of adults was better than that of adolescents and children because it had more instances. Between Logistic Regression and Naive Bayes, Logistic Regression performed better in terms of accuracy by 4.12 percent. In the case of sensitivity and specificity, logistic Regression achieved better results by 4.2 percent and 3.01 percent respectively.

Authors in [10], applied Naive Bayes, Logistic Regression, K Nearest Neighbor, Support Vector Machine, Artificial Neural Network and Convolutional Neural Network for autism detection of children, adolescents and adults. They got the datasets from the UCI ML repository which had 21 features except for toddlers and all the datasets had 1100 instances. For the adult dataset, the performance of the Convolutional Neural Network was better than that of other models because it had an accuracy rate of 99 percent. In the adolescent dataset, again Convolutional Neural Network performed better since the accuracy rate of the model was 95 percent but the rate of specificity was a bit low. Similarly, the accuracy of CNN was the highest compared to the other models in the child dataset. Although the rate of sensitivity of the Convolutional Neural Network was a bit low in the child dataset, the rate was still acceptable. From the result obtained, the accuracy rate of the Convolutional Neural Network was higher in all three datasets than in other models.

Chapter 3

Methodology

3.1 Workplan

This work plan detail outlines the procedures for utilizing machine learning to identify autism using behavior analysis. We have used some pre-processing techniques such as merging the data, handling missing values, and imputing them, handling categorical value, feature importance, and principal component Analysis. Then, we used machine learning models such as KNN, Logistic regression, SVM, Random forest, GradientBoosting and CatBoost. Then these models are tested and validated. The best model was chosen according to their performance. The work plan's final stage entails a comprehensive review of all models used in this research.

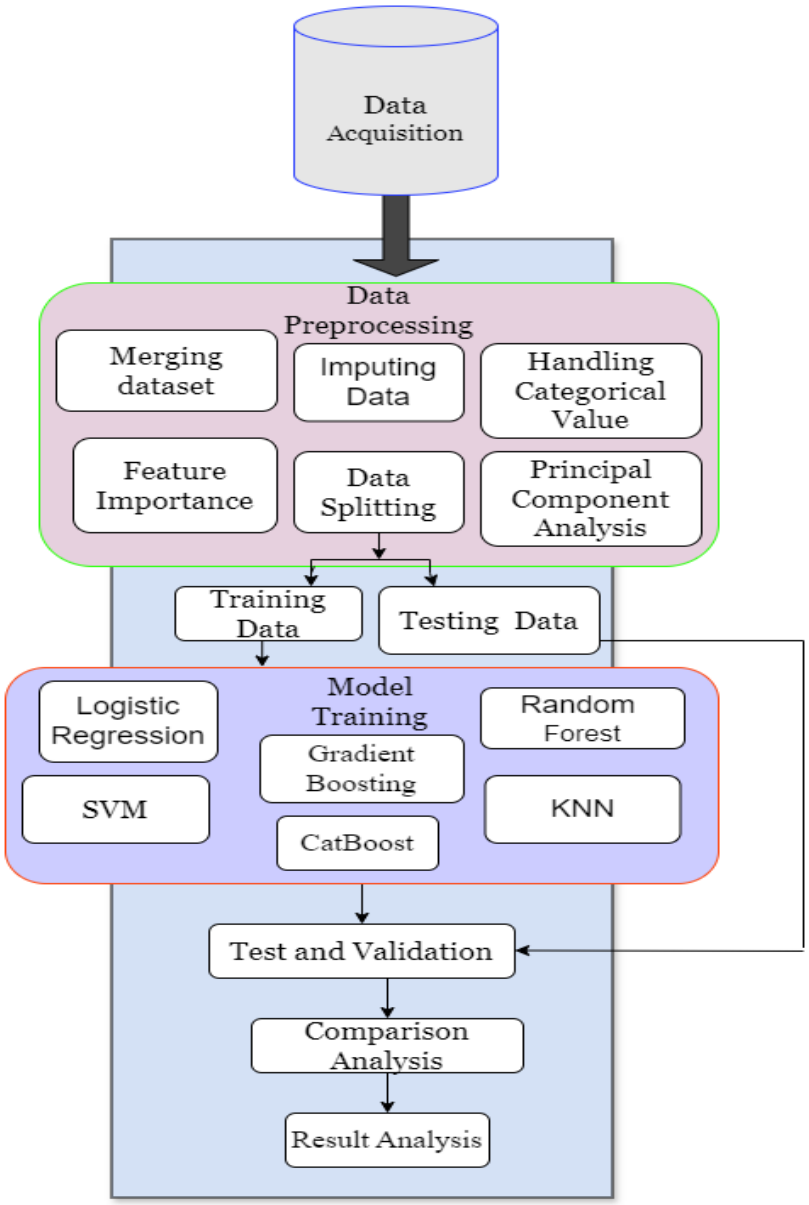

Figure 3.1: Step by Step Workflow of Our Research

3.2 Datasets Description

The dataset for autism screening in Adults has also been collected from Kaggle.In the Kaggle dataset, there are also 704 instances and the attribute type is categorical, continuous, and binary. There are two datasets used for autism screening in toddlers. The one Autism Screening of Toddlers dataset has 1054 cases and 18 attributes—Q-Chat-10 items (A1-A10), age, gender, ethnicity, jaundice, family history of ASD, country of residence, previously used app, and daily screen usage. The Q-Chat-10 score identifies ASD traits. Another dataset is collected from AUTISM RESEARCH: UNIVERSITY OF ARKANSAS Computer Science. In this dataset, there are 1985 instances and 28 attributes. The features of autism spectrum disorder include the Autism Spectrum Quotient, Social Responsiveness Scale, Age, Q-CHAT-10 Score, Genetic Disorders, Depressive Symptoms, Learning Disorder,

1	I often notice small sounds when others do not
2	I usually concentrate more on the whole picture, rather than the small details
3	I find it easy to do more than one thing at once
4	If there is an interruption, I can switch back to what I was doing very quickly
5	I find it easy to 'read between the lines' when someone is talking to me
6	I know how to tell if someone listening to me is getting bored
7	When I'm reading a story, I find it difficult to work out the characters' intentions
8	I like to collect information about categories of things (e.g. types of car, types of bird, types of train, types of plant etc)
9	I find it easy to work out what someone is thinking or feeling just by looking at their face
10	I find it difficult to work out people's intentions

Figure 3.2: Ten ASD Question

Speech Delay/Language Disorder, Global Growth Delay/Intellectual Disabilities, Social/Behavioral Issues, Childhood Autism Rating Scale, Anxiety Syndrome, Sex, Ethnicity, History of Jaundice, and Family Members with ASD One dataset for autism screening in adults has also been collected from Kaggle. Both dataset has some missing values that need to be added.

3.3 Pre-Processing of Data

We have used three datasets for our research paper. We know the raw data set has missing values, inconsistent data, and errors. So, we must preprocess the raw data to clean it and standardize it to maintain reliability and consistency. The step-by-step preprocessing is as follows:

3.3.1 Identify Common Columns and Merge them

Our three datasets contain different rows and columns. So, Our first task was to identify common columns across the datasets. We have found 17 columns that are similar in each dataset. We renamed all the common columns using the same name. Next, We Combine all the datasets into a single dataset. Some features engineering was done such as deriving age in years from months to maintain consistency. After combining all the data we got 3743 instances and 17 attributes.

3.3.2 Standardize the Categorical Value

For standardizing the categorical value we first Identify the unique values present in each column to understand the inconsistency and variations. We detected variances

	Objects	Unique values	number of unique values
0	Sex	[f, m, F, M]	4
1	Ethnicity	[White-European, Latino, ?, Others, Black, Asi...	23
2	Jaundice	[no, yes, Yes, No]	4
3	Family_mem_with_ASD	[no, yes, No, Yes]	4
4	Who_completed_the_test	[Self, Parent, ?, Health care professional, Re...	11
5	ASD_traits	[NO, YES, No, Yes]	4

Figure 3.3: Values Variations

in data entry procedures by printing unique values. For example, We standardize "Yes" and "No" in place of "Yes," "yes," "YES," "No," and "no."Many more columns are required to standardize it as well. This method ensures consistency, improves data quality, and reduces errors.

3.3.3 Handling Missing Values

	Missing Values
A1	0
A2	0
A3	0
A4	0
A5	0
A6	0
A7	0
A8	0
A9	0
A10_Autism_Spectrum_Quotient	0
Age_Years	2
Sex	0
Ethnicity	95
Jaundice	0
Family_mem_with_ASD	0
Who_completed_the_test	95
ASD_traits	0

Figure 3.4: Missing Values

Our dataset contains missing values which can negatively affect our result analysis part. That's why we have determined missing values in our column and impute them. we denoted missing values in our dataset using "NaN" which is the standard way. Then we impute missing values using the most frequent strategy. This process helps to retain data integrity, maintain dataset size, and prevent data loss.

	Missing Values
A1	0
A2	0
A3	0
A4	0
A5	0
A6	0
A7	0
A8	0
A9	0
A10_Autism_Spectrum_Quotient	0
Age_Years	0
Sex	0
Ethnicity	0
Jaundice	0
Family_mem_with_ASD	0
Who_completed_the_test	0
ASD_traits	0

Figure 3.5: Impute Values

3.3.4 Convert Categorical to Numeric

Our dataset has some columns with categorical values. However, many ML algorithms need to convert categorical values to numerical values as they only understand numerical formats. We have used the replace method to convert binary categorical values to numerical format. For instance, we converted 'No' to '0', 'Yes' to 1, M' to 1, and 'F' to 0. We have used one hot-coded encoding for multiclass categorical values. Binary columns are created for each column present in a specified column such as the 'Ethnicity' column has categories like 'Asian', 'Black', 'White', etc., One-hot encoding will produce new columns called "Ethnicity-Asian," "Ethnicity-Black," and so on, with binary values indicating the presence or absence of each category. After, one-hot encoding we have a total of 34 columns and 3743 instances. It enhances model accuracy and is compatible with machine learning algorithms.

3.3.5 Correlation Matrix and Feature Importance

The correlation matrix is a strong tool for understanding the relationship between different features present in our dataset. It used Pearson correlation by default which measures linear relationship between all the features. The table displays the correlation between two variables in each cell. The value ranges from -1 to 1, where 1 (red color) indicates that the two variables have a perfect, positive link and -1(blue color) denotes a perfect negative correlation between the two variables. A value of 0 (lighter color) indicates that the two variables have no relationship.

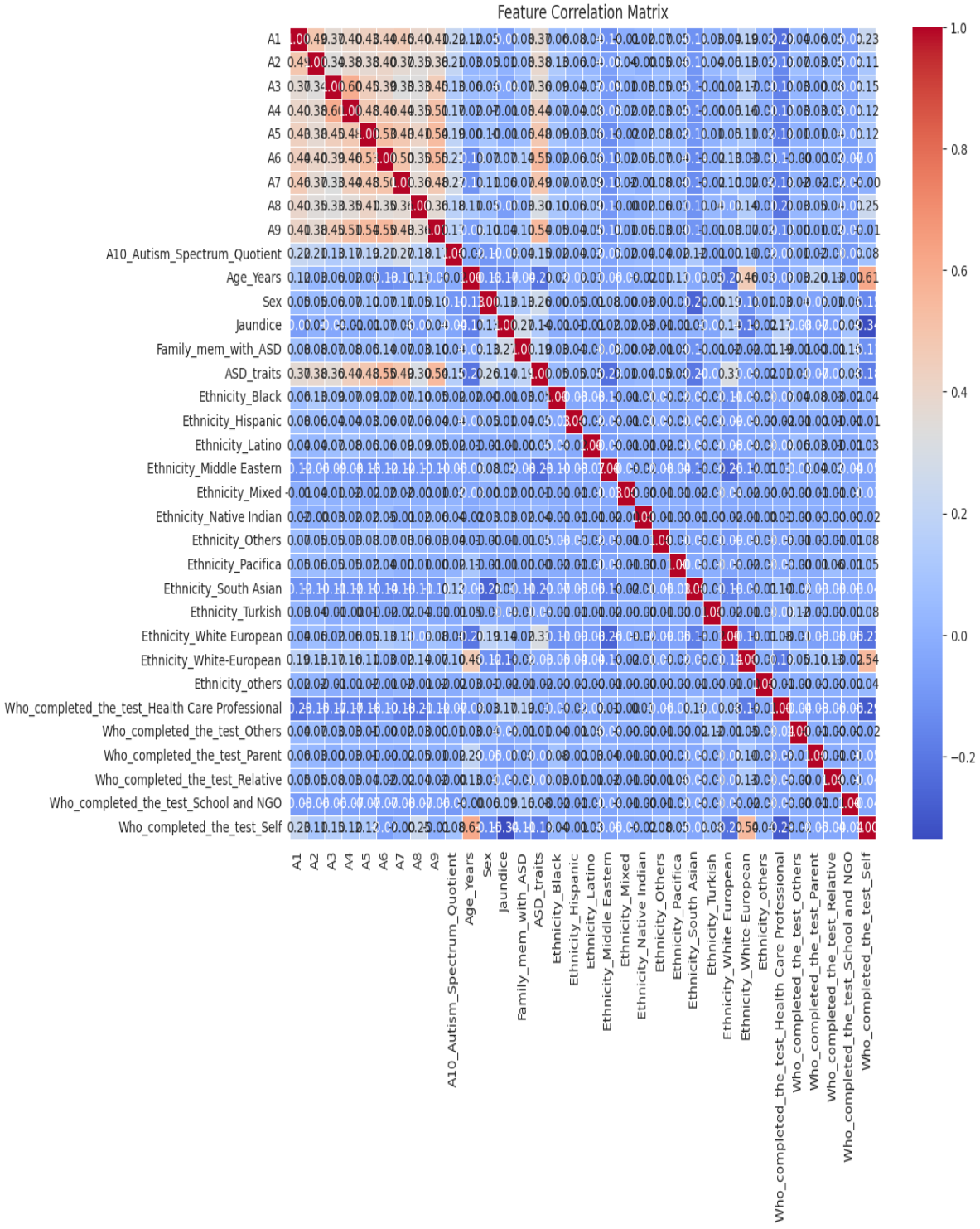

Figure 3.6: Correlation Matrix

From the correlation matrix we can see A1 to A9 have higher inter-correlation with each other and Autism Spectrum Quotient (A10) and Family-mem-with-ASD have moderate correlations with A1 to A9. So, these features are connected to the Autism spectrum quotient. However, other columns show a low correlation with each other and are not strongly related to other features. Ethnicity and sex show a low correlation with autism-related features. Then we derive the feature's importance from our model to highlight the contribution of each feature to predict the model performance.

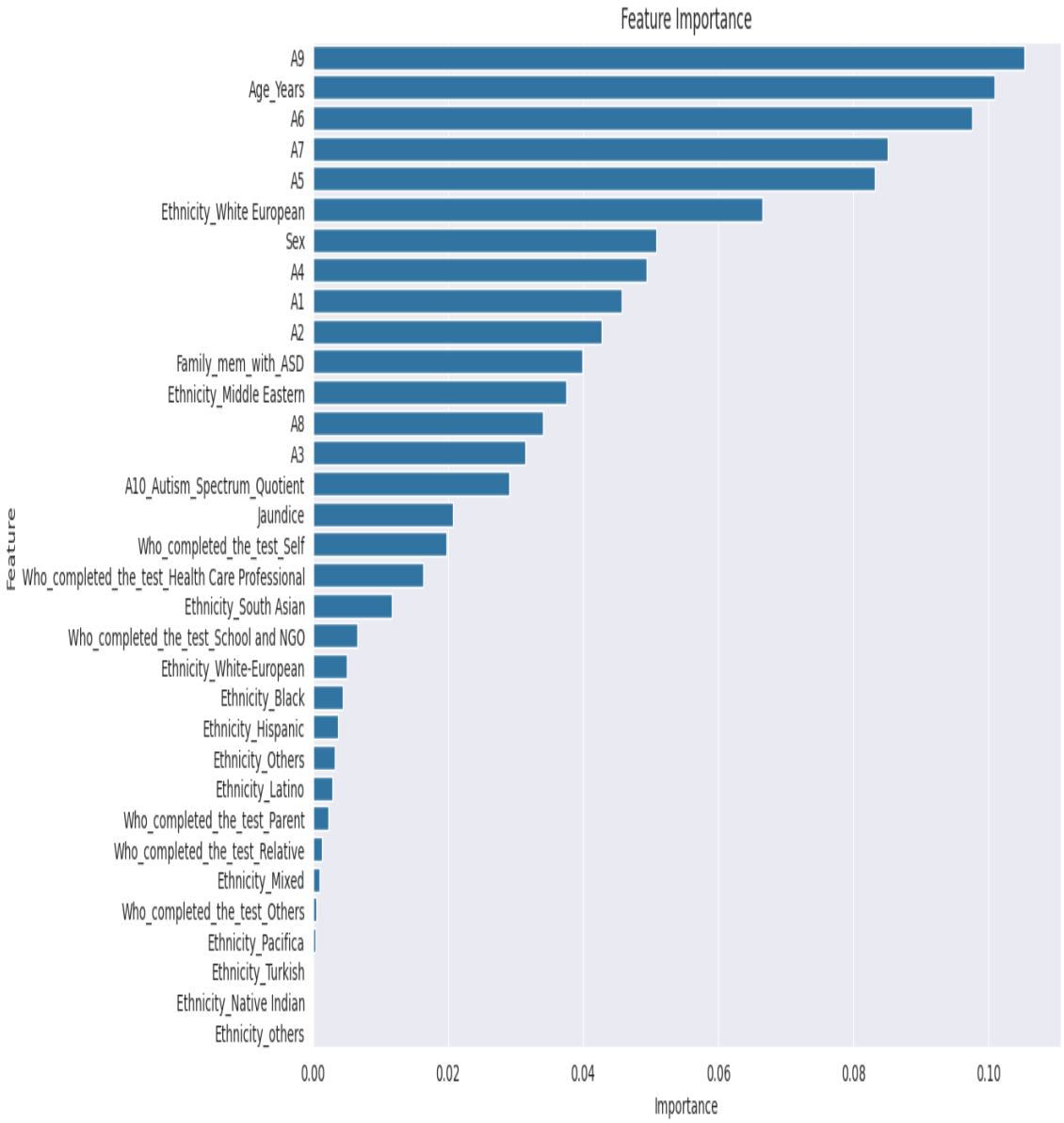

Figure 3.7: Feature Importance

From the graph, We can see the top features of our model which shows the highest scores are A9, age-years, A6, A7, and A5., on the other hand, Ethnicity-Turkish, Ethnicity-Native Indian, and Ethnicity-others show little significance, indicating that they have little bearing on prediction. The correlation matrix and feature importance give a thorough understanding of the relationships and contributions of the variables in our study and the underlying data structure and model dynamics. Then we dropped some columns that have less contribution to our dataset.

3.3.6 Principal component Analysis

Principal Component Analysis is a machine-learning technique that uses the dimensionality reduction method to reduce the size of a big data set to a small one while preserving the majority of the data's variance. It makes the process of processing data points for machine learning algorithms much quicker and easier. For applying PCA, we first standardized the data by arranging the data such that its mean is 0 and its standard deviation is 1 for every feature. After that, we use Principal Component Analysis (PCA), which keeps enough components to account for 95 percent of the variation. Our graph shows the explained variance ratio by principal components.

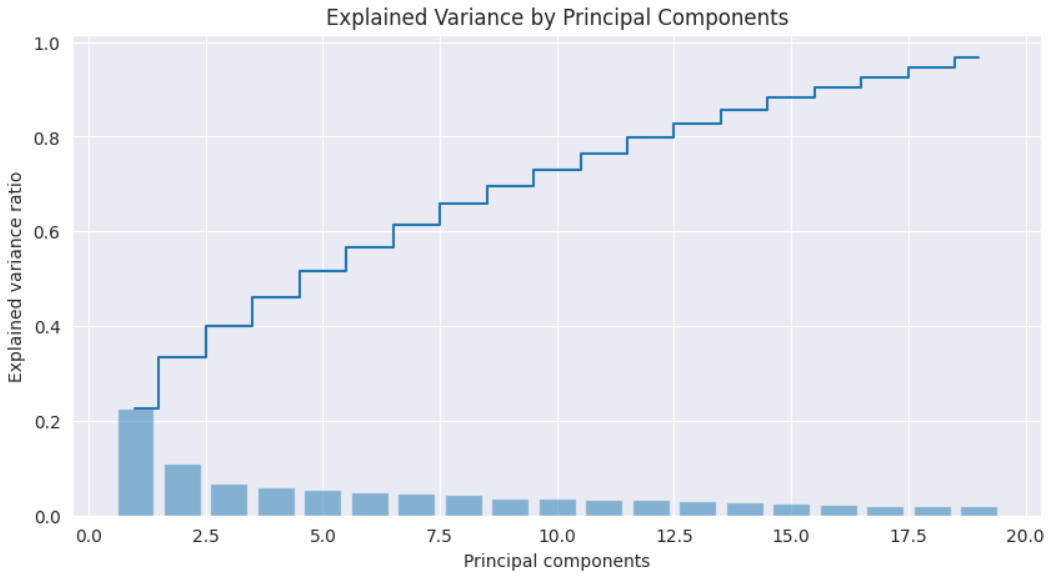

Figure 3.8: PC Analysis

The X-axis shows the principal components arranged from the first principal component (PC1) to the twentieth principal component (PC20). The Y-axis represents the ratio of the overall variance explained by every principal component. Each component's explained variance ratio denoted how much of the variance in the entire dataset it accounts for. The line plot displayed the ratio of cumulative explained variance. It is calculated by adding the principal component explained variance ratios consecutively. From the graph, we can see that the first component captures the most variance which is around 20 percent. The bar height is dropped moving from the first to the latter primary components. This indicates that the variance captured by each following principle component is smaller than that of the previous one. The

cumulative explained variance line starts at the first principal component's variance and increases as more components are added. After 10 components it might capture around 75 percent variance. As the right-sided components capture less variance, the line levels decrease gradually.

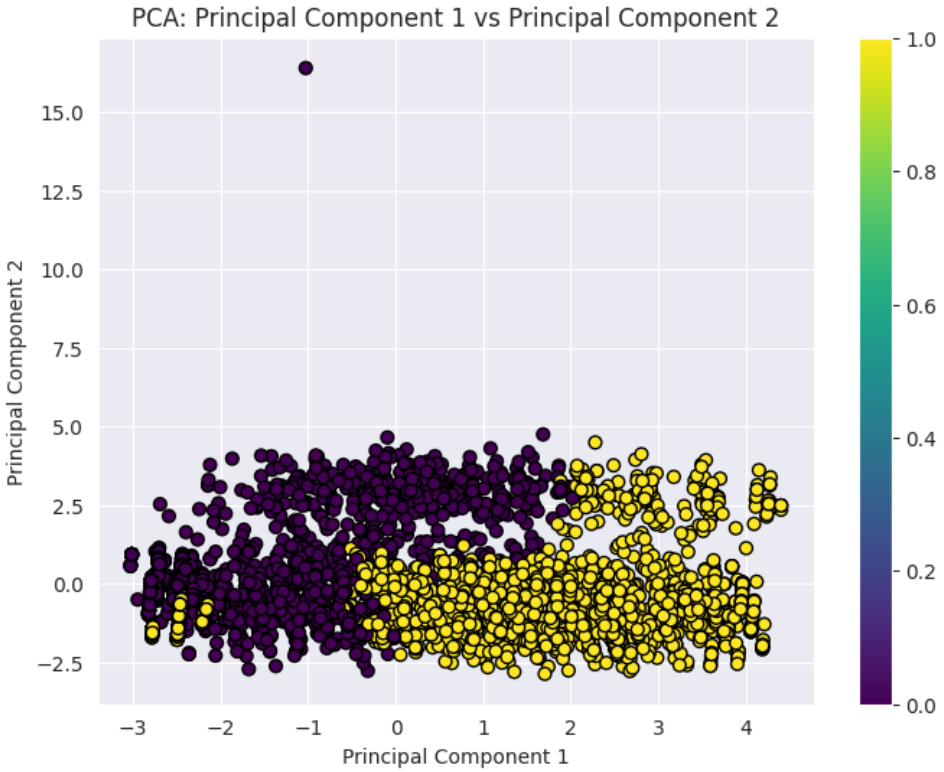

Figure 3.9: Relationship Between Two Components

The scatter plot represented the relation between the two components (PC1 and PC2). The X-axis represents the values of the first principal component and Y axis represents the values of the second principal component. PC1 separates the two groups (yellow and purple) with a spread from -3 to +4, suggesting that the data differs significantly in this direction. PC2's spread is less for the bulk of points showing that while it is less significant than PC1 for most points.it does capture variance in a way that makes certain data points stand out. The sample regarding the second principal component that stands out considered unique and distinct from the others is the one with a high PC2 value.

3.3.7 Data Splitting

We have a total of 3743 instances in our dataset. Our dataset is divided into testing and training sets. We have set 80 percent data for the training set and 20 percent data for the testing set.

Chapter 4

Description of Models

In the next subsections we described about KNN, Random Forest, Logistic Regression, SVM, Gradient Boosting and CatBoost Model.

4.1 KNN (K Nearest Neighbors)

KNN is a prevalent machine learning algorithm that operates by assessing proximity to make predictions for various points within datasets. This algorithm majorly depends on the labeled training datasets to make predictions and is widely known for its service as a non-parametric supervised learning classifier as it does not focus on the underlying datasets. This algorithm emphasizes similarity rates to predict outcomes. By comparing new input data with available trained datasets, it seeks similarities among different cases to make predictions. The algorithm assigns new points based on their similarity to the stored data, ensuring reliability by not solely relying on the training dataset but actively seeking similar actions to make predictions. The way this algorithm works is it selects neighbors (K), calculates the Euclidean distances, assigns new data points to the category with the majority of nearest neighbors, everytime the dataset is updated.

Advantages

KNN is one of the simplest functioning algorithms which gives pretty straightforward outputs. The algorithm's ease of implementation makes it conducive to research, while its emphasis on finding similarities and recognizing new input points, rather than relying solely on trained datasets for immediate predictions, contributes to an overall increase in accuracy rates for different kinds of new data sets. When dealing with large datasets in research, the algorithm tends to perform more accurately and yield better results, rendering it well-suited for handling extensive data volumes.

4.2 Random Forest

Random forest is a machine learning algorithm that operates by creating an ensemble of decision trees at training time for generating a singular accurate prediction. It is a type of supervised machine-learning algorithm which is very simple. By merging

multiple decision trees, random forest can construct a robust model. The bagging method is generally used for combining decision trees at the time of training. With the help of a bagging method, random forest combines decision trees flexibly and simply for more accuracy. The diverse dataset is generated through bagging which helps in training each tree. Then it composes multiple sets of training data by sampling the original dataset arbitrarily. Subsets of the feature are chosen at the end of each split during the formation of a tree. This technique is called the Random Subspace method. As a result, this method helps in improving the robustness of random forest models. So, the random forest model can be applied for classification and regression. During the expansion of trees, additional information can be added to the model. The best feature is selected from the subset of features as an alternative to the most important feature in the subset. So, a diversified version of the model is achieved as an outcome which is good for increasing the strength of the model. At the time of separating a node, a subset of the feature is selected for using the random forest classifier. Since multiple trees are combined, high accuracy can be obtained in a random forest model. For high accuracy, we can apply this model.

Advantages

The chances of overfit occurring are less likely due to the average of the errors of each tree.So, accuracy will be high because of random forest assembling the prediction. Instead of a single decision tree, this model uses an ensemble method which helps in improving the rate of accuracy. Moreover, noisy data does not affect the performance of the model because of its adaptability. Also, it can manage missing values better because of having substituted splits. It reduces the dimensionality as important features can be selected. So, random forest can be chosen to achieve a higher accuracy rate as it is flexible to use.

4.3 Logistic Regression

The logistic regression model is a type of supervised machine learning algorithm that predicts the probability of an event or outcome by performing binary classification. Only two possible outcomes can be generated in the logistic regression model. The outcome can either be binary or dichotomous. Relation between one or more independent variables can be examined in a logistic regression model and classification can take place. It is simply used for prediction purposes. Whether an instance falls under a certain category or not can be determined by this model with the help of mathematical probability. If the result is one, it means it is in the positive class. Otherwise, it is in the negative class for zero. Moreover, it uses a sigmoid function for the prediction of an event or outcome. An S-curve will be generated by applying the sigmoid function. So, this model can be used as it is very simple.

Advantages

The logistic regression model can be easily implemented. It requires less computational power for training. Also, it gives information about how important a coefficient size is and whether the direction of the outcome is positive or negative. Besides, the classification of unrevealed records is very fast in this model. If the

dataset is linearly separable then the performance level is very high. It does not need much time for training. Assumptions are not made about features for not following specific rules of distribution. So, different types of data can be managed by this model. This model can be used for binary classification purposes. Therefore, a random forest model can be chosen.

4.4 SVM

SVM is a supervised learning model that can be used for solving both classification and regression problems. It is mainly used for solving tasks of binary classification. For resolving the classification task, it needs to sort out the elements to classify into two groups of a dataset. Data points are called support vector which are crucial for finding out the decision boundary. Also, it helps in finding out the position and orientation of a hyperplane. In the building of a decision boundary, the support vector plays an important role. Different classes have data points that are divided by decision boundary. This model can easily separate data classes by maximizing the margin. Here, the margin is called the distance between data points closer to each class and the hyperplane. The decision boundary becomes a hyperplane when it is in a dimension higher than 3D. So, SVM has to discover the best hyperplane that can easily distinguish the classes. In the case of data that can not be divided into a single straight line, SVM works the best. As a result, it can handle complex datasets. Besides, SVM can be called a non-linear SVMs. For non-linear SVMs, a mathematical formula is used to convert data into higher dimensional space. So, it is reliable to discover the boundary easily. The SVM model helps in figuring out a linear separation with the help of the kernel function. The kernel function plays an important role in mapping the data into kernel space from the original input space. As a result, the SVM model can be called a reliable model.

Advantages

Both linearly separable and non-linearly separable data can be managed by the SVM model with the help of a kernel. Classification errors are reduced in this model. This model can handle high-dimensional data. If the number of features is large, the SVM model can handle it efficiently. This model also reduces overfitting. Since this model maximizes the margin, the generalization of unseen data is handled effectively. Complex relationships can be captured very well with the help of this model. Different kernel functions make the SVM model flexible. So, different types of problems can be solved. If the training dataset is small, the performance of the SVM model becomes very high.

4.5 Gradient Boosting

Gradient Boosting is an effective machine learning technique. It is mainly used for regression and classification tasks. It forms a powerful ensemble model by iteratively combining the strengths of several weak learners, usually decision trees. The process begins with creating an initial basic predictive model. Then, a new tree is added to it, which is trained to predict the loss function of the existing ensemble model. By

concentrating on the errors of the previous models, every new tree helps minimize the overall error of the ensemble model. This procedure is repeated for a certain number of iterations or until the model's performance stops getting better.

Gradient Boosting reduces the loss function by calculating the gradient of the loss according to the model's predictions. The loss function can be customized for the specific task at hand. For example, mean squared error for regression and log loss for classification can be used as loss function. The overall model's performance depends on hyperparameters like learning rate, number of trees, and depth of each tree, as they control the speed and complexity of the model.

Advantages

Gradient Boosting can manage the complexity of the behavioral data by analyzing correlations (non-linear) and feature interactions. The overall predictive accuracy is improved by its iterative approach, which fixes errors from earlier models. The model's adaptability in utilizing various loss functions and its capability to deal with subtle and intricate criteria makes it a good fit for ASD detection.

4.6 CatBoost

CatBoost is a supervised machine learning method. It is applied by the Train Using AutoML tool. Decision trees are used for classification and regression purposes in this model. Two important features of CatBoost are managing categorical data and applying gradient boosting. In the gradient boosting technique, many decision trees are built at each iteration. The result of the previous trees can be enhanced by sequential trees. So, CatBoost refined the gradient boosting technique for better results. Also, categorical features are encoded with the help of ordered encoding. For replacing the categorical features, target statistics are used from the rows in order to calculate the value. One important feature of CatBoost is the usage of symmetric trees. Split conditions remain the same for all the decision nodes at every level. Scaling for the columns is done internally in this method. Also, the best hyperparameters are selected by using cross-validation internally. Besides, CatBoost can also handle numerical features. So, CatBoost can be a suitable model that is easy to implement for categorical features.

Advantages

CatBoost is very easy to use since there is no need for preprocessing extensively. Because it has a built-in technique that helps in handling the relationship of all categories effectively. The chances of overfitting occurring in this model are reduced due to ordered boosting. So, it makes sure that overfitting does not occur on the training data. Also, it can handle both small and large datasets. In the case of a large dataset, this model enhances the speed of training. Imputation is not needed since CatBoost can internally manage missing data. So, this model is recommended because it is very effective and fast.

Chapter 5

Result Analysis

In the next subsections, the results are analyzed of the models.

5.1 Performance Analysis of Logistic Regression

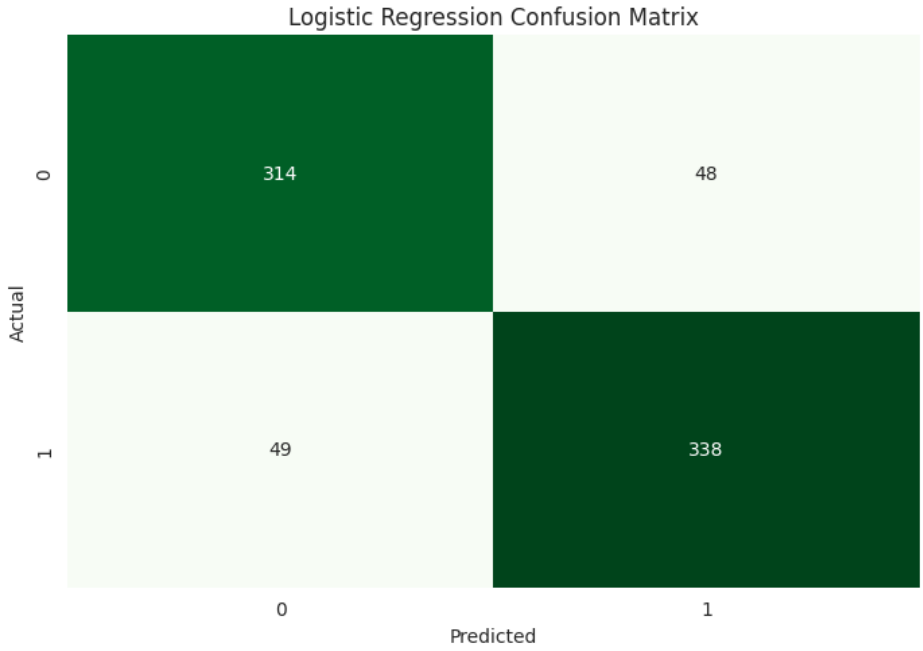

Figure 5.1: Logistic Regression Confusion Matrix

Our dataset contains 3743 data points. After splitting the data as training and testing datasets as 2994 and 749 data points respectively we start with training our datasets with different models. In our processed dataset, we used the logistic regression model on the test dataset and it gave us an accuracy of 87.05%. For a binary classification task, this model actually gave us a decent accuracy. Out of 749 test data, the model predicted true positive 338 instances and true negative 314 instances. False positive and false negative instances were 48 and 49 respectively. For class 0 (Non-Autistic), among 362 instances, precision, recall and F1 scores are 87% each. For class 1 (Autistic), among 387 instances, precision, recall, F1

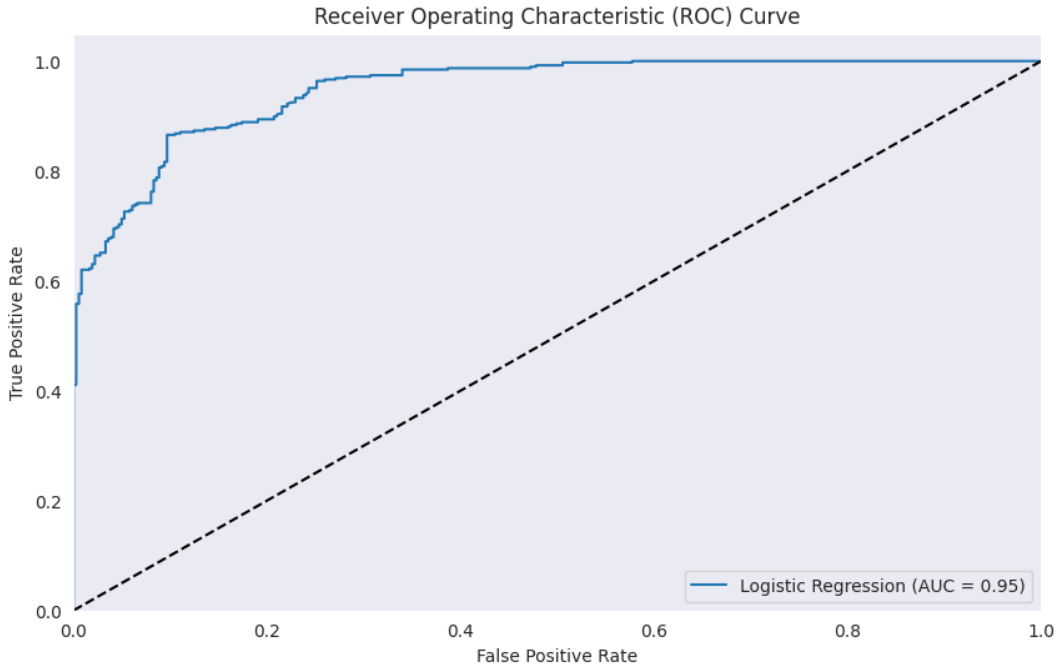

Figure 5.2: ROC Curve of Logistic Regression

scores are 88%, 87%, 87% respectively. The macro average and weighted average for precision, recall, and F1 score are all 0.87, indicating balanced performance across both classes. Overall, the confusion matrix shows a balanced distribution of errors between false positives and false negatives.The model's effectiveness can be seen from the classification report. The accuracy we got is decent but not outstanding.

5.2 Performance Analysis of KNN

K-Nearest Neighbors (KNN) is the next model we used on the test dataset . We achieved an accuracy of 93.59%. For a binary classification task this model provided a high level of accuracy and proved to be good. Among 749 test data instances, the model successfully predicted 336 true positives and 365 true negatives. The number of false positives and false negatives were 26 and 22 respectively. For class 0 (Non-Autistic), among 362 instances, the precision, recall, and F1 scores were 94%, 93%, and 93% respectively. For class 1 (Autistic), among 387 instances, the precision, recall and F1 scores were 93%, 94%, and 94% respectively. The macro average and weighted average for precision, recall, and F1 score were all 94%, indicating a well-balanced performance across both classes. The confusion matrix shows a slightly lower number of false negatives compared to false positives, reflecting the model's strong predictive capability. The classification report underscores the model's effectiveness, with high precision, recall, and F1 scores. AUC value was found for the model is 98%. Overall, the accuracy and detailed performance metrics suggest that the KNN model is highly effective for this binary classification task.

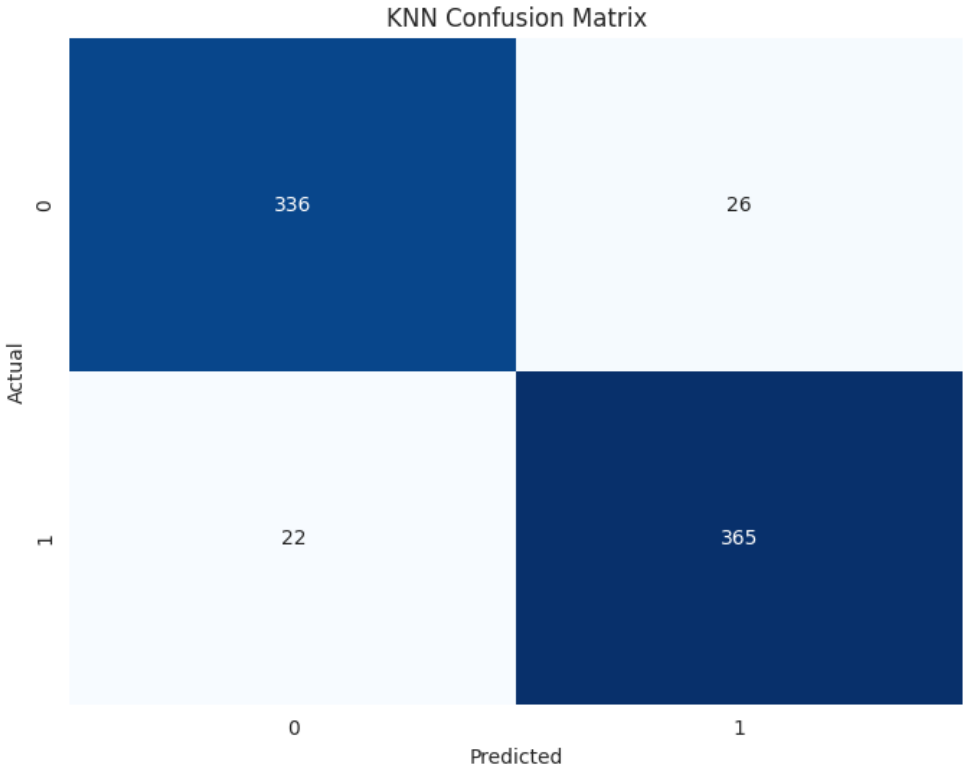

Figure 5.3: Confusion Matrix of KNN

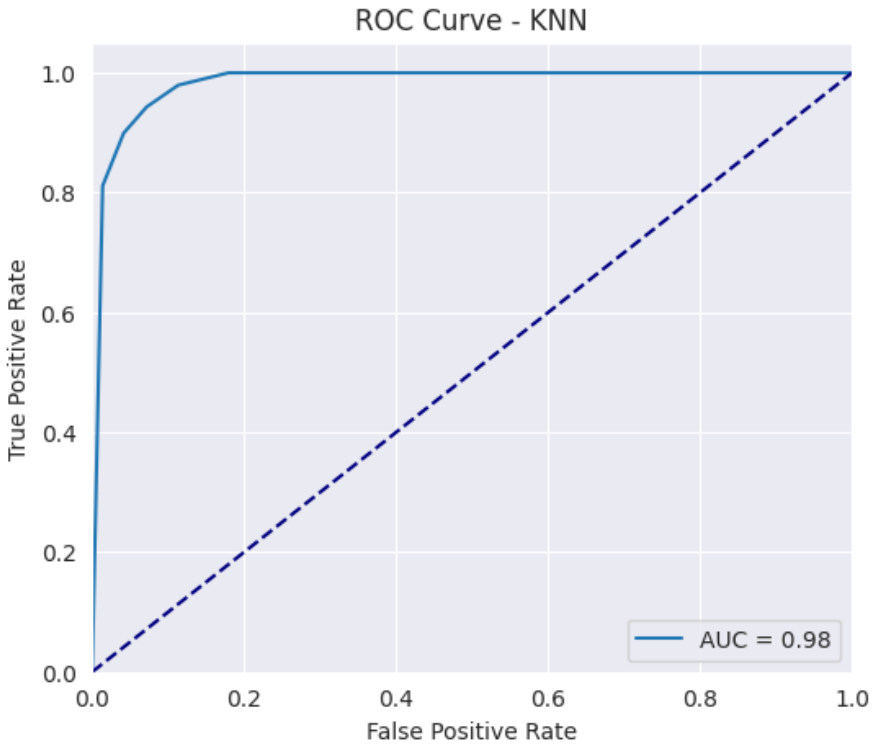

Figure 5.4: ROC Curve of KNN

5.3 Performance Analysis of Gradient Boosting

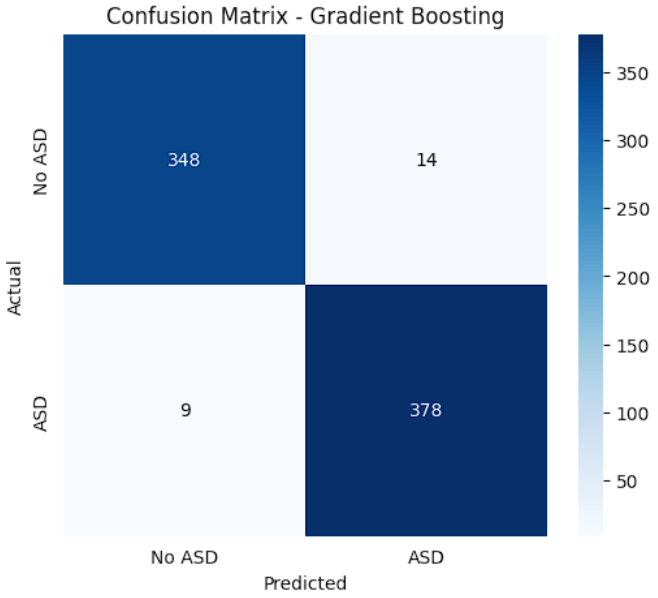

Figure 5.5: Gradient Boosting Confusion Matrix

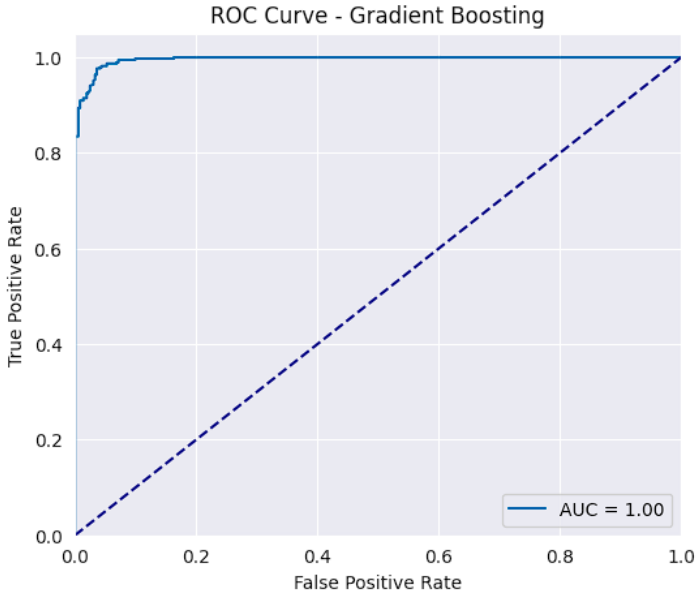

Figure 5.6: ROC Curve of Gradient Boosting

An overall accuracy of 96.93% using Gradient Boosting was achieved, which indicates the model's strong performance. From the confusion matrix, we found out that the model correctly identified 348 out of 362 instances of the Non-autistic class and 378 out of 387 instances of the Autistic class. Both classes having low false positive rates and high true positive rates indicate a balanced performance across classes. Class 0 (Non-autistic) had precision and recall of 0.97 and 0.96, respectively. Class 1 (Autistic) had precision and recall of 0.96 and 0.98, respectively. An average score of 0.97 was found for the precision, recall and f1-scores of both classes. This high

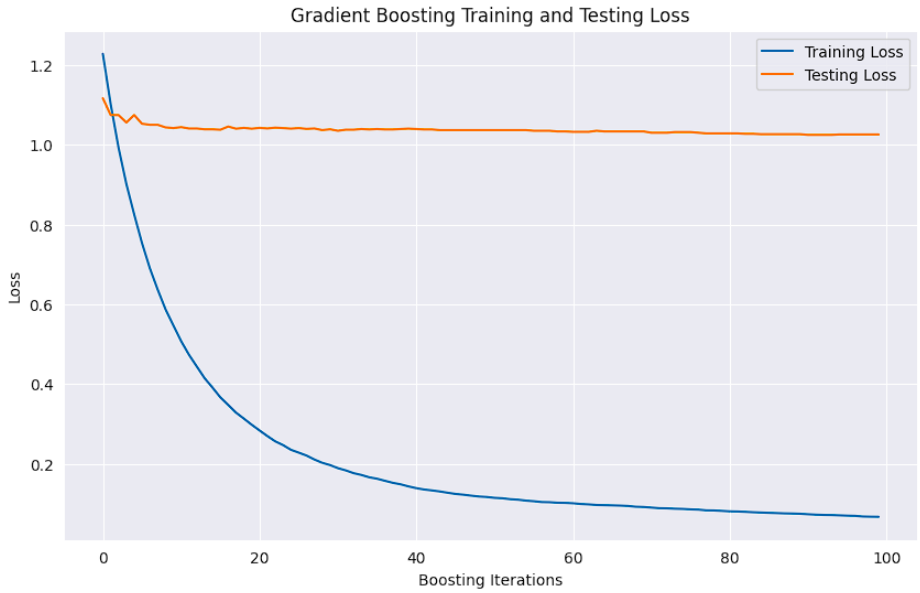

Figure 5.7: Training and Testing Loss of Gradient Boosting

score suggests that the model was able to minimize the false positives and false negatives quite effectively. The macro and weighted averages of these scores were 0.97 as well. This indicates consistent and robust performance of the Gradient Boosting model across classes, making it a reliable option for our binary classification task. Furthermore, the AUC score was found to be 1.00 which points to the perfect discrimination of the model between classes.

5.4 Performance Analysis of Random Forest

We applied the Random Forest model to the test dataset and got an impressive accuracy of 97.73%. For a binary classification task, this accuracy is exceptionally high. The result also indicates the model's robust performance. The model correctly predicted 353 true positives and 379 true negatives cases. The false positives and false negatives were 9 and 8 respectively. For class 0 (Non-Autistic), which comprised 362 instances, the model achieved precision, recall, and F1 scores of 98% each. For class 1 (Autistic), with 387 instances, the model also achieved precision, recall, and F1 scores of 98% each. The macro average and weighted average for precision, recall, and F1 score were all 98% which demonstrated balanced and exceptional performance across both classes. The confusion matrix indicates a very low error rate. The number of false positives and false negatives are very low. It reflects the model's high accuracy. The ROC curve further underscores the model's effectiveness, with an area under the curve (AUC) of 0.99, highlighting its excellent discriminative ability between the two classes. The training and validation loss curves show the model's learning progress over different training set sizes. The training error rate decreases consistently with more data, while the validation error rate remains low, indicating the model's strong generalization capability. Overall,

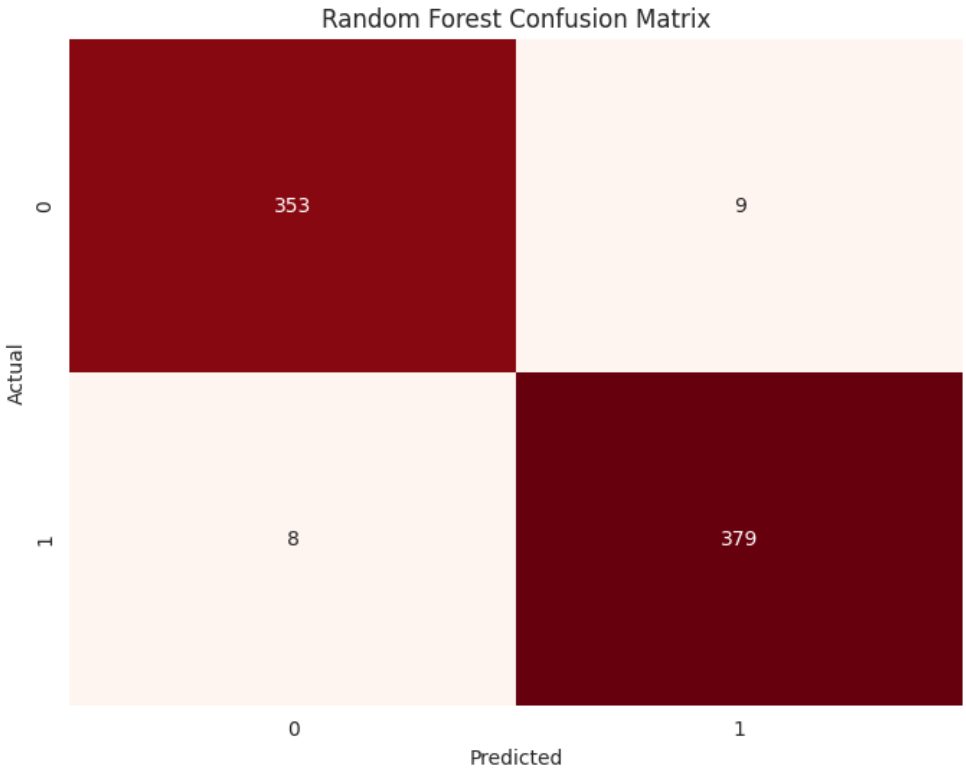

Figure 5.8: Random Forest Confusion Matrix

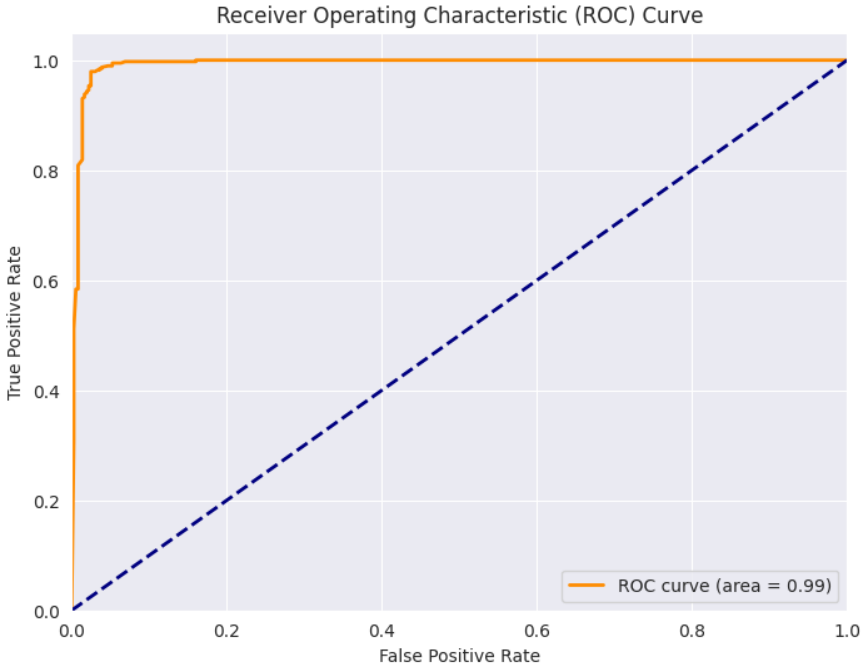

Figure 5.9: ROC Curve of Random Forest

the Random Forest model demonstrates outstanding accuracy and balanced performance across all metrics. It proves Random Forest as a highly reliable classifier for this binary classification task.

5.5 Performance Analysis of SVM

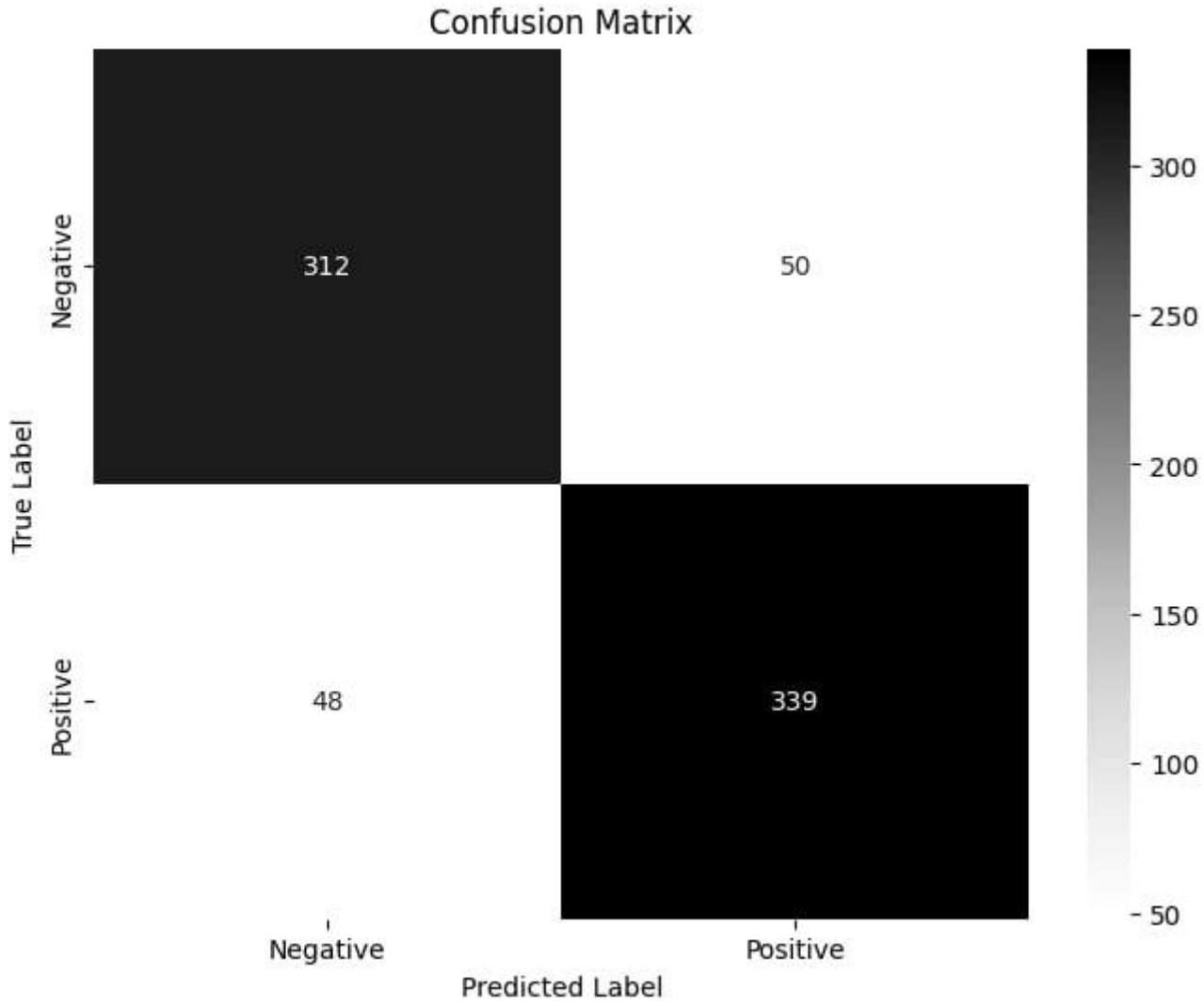

Figure 5.10: SVM Confusion Matrix

In our paper, we utilized the SVM model with a linear kernel to categorize autism traits, evaluating its effectiveness on a test dataset. The SVM model showcased an accuracy of 0.89, signifying that 89 percent of its predictions accurately distinguished individuals with and without ASD traits based on the provided features. Analysis of the confusion matrix revealed that the SVM model correctly identified 70 instances of ASD traits and 90 instances of non-ASD traits. However, it also misclassified 10 instances of non-ASD as ASD and 8 instances of ASD as non-ASD. The comprehensive classification report provided further insights, indicating a precision of 0.87 for the ASD class and 0.92 for the non-ASD class, along with respective recalls of 0.90 and 0.89. Furthermore, the F1 scores, reflecting a balance between

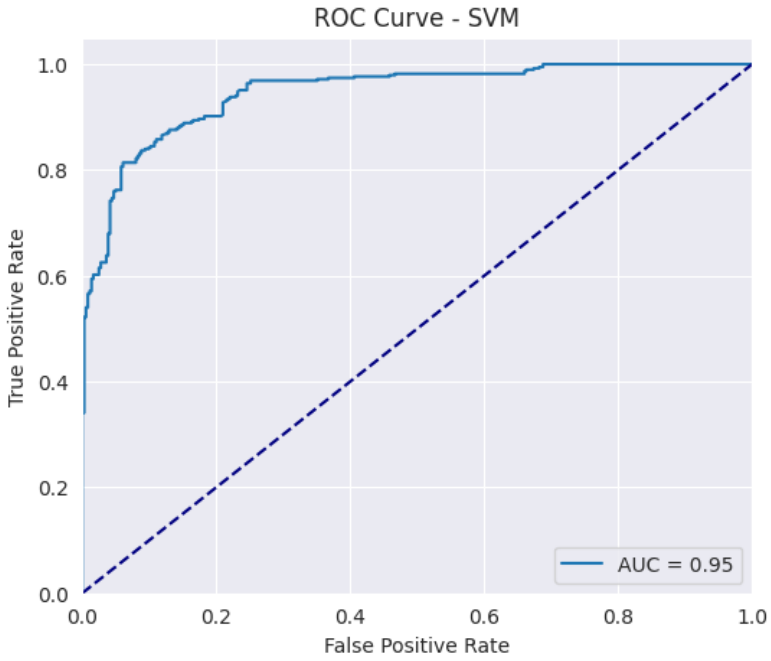

Figure 5.11: ROC Curve of SVM

precision and recall, were 0.88 for the ASD class and 0.90 for the non-ASD class. The Receiver Operating Characteristic (ROC) curve underscored the model's robust discriminatory capability, boasting an Area Under the Curve (AUC) of 0.92. Overall, the SVM model demonstrated commendable performance with good accuracy, precision, recall, and AUC, establishing its credibility as a dependable tool for ASD trait classification while leaving room for future improvement.

5.6 Performance Analysis of CatBoost

In our study, we employed the CatBoost classifier, an advanced gradient-boosting model renowned for its efficient handling of categorical features and robust performance with minimal hyperparameter tuning. The model underwent evaluation using the same test dataset, yielding noteworthy results. With an accuracy of 0.91, the CatBoost model demonstrated its effectiveness in classifying ASD traits, with 72 instances of ASD traits and 92 instances of non-ASD traits correctly identified. However, it misclassified 8 instances of non-ASD as ASD and 6 instances of ASD as non-ASD, providing comprehensive insights into its performance. The precision for the ASD class was 0.90, indicating that 90 percent of instances predicted as ASD were correct, while for the non-ASD class, it was 0.93. The model exhibited a recall of 0.92 for the ASD class and 0.90 for the non-ASD class, reflecting its ability to identify actual ASD instances accurately. The F1 scores were 0.91 for the ASD class and 0.92 for the non-ASD class, offering a balanced performance measure. Furthermore, the ROC curve displayed an excellent discriminatory capability with an AUC of 0.95, confirming the model's superior ability to distinguish between ASD and non-ASD traits. The CatBoost model's training and testing loss curves illustrated a stable learning process, indicating its capacity to generalize well to unseen data. Overall, the CatBoost classifier demonstrated good performance, highlighting its

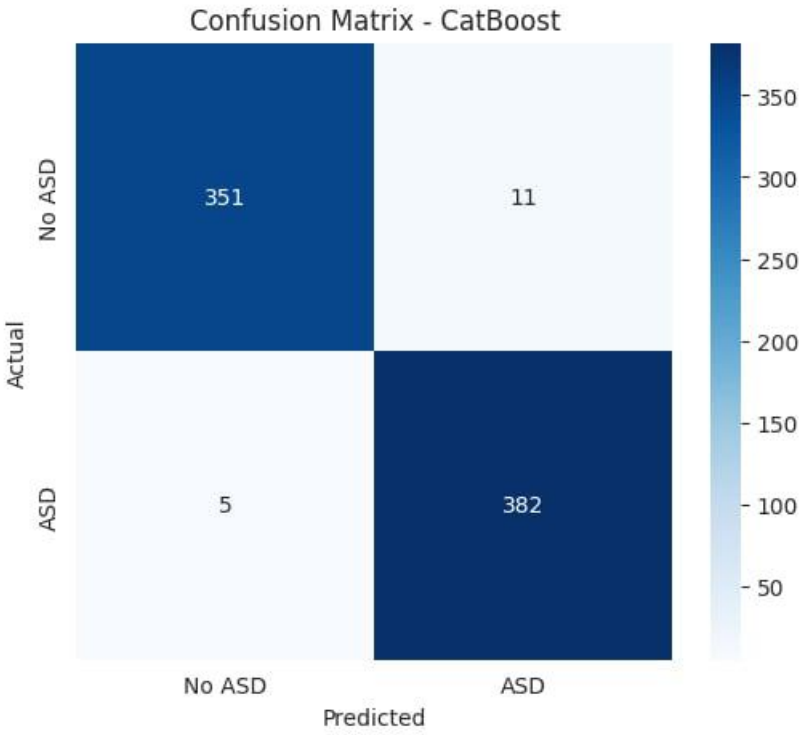

Figure 5.12: CatBoost Confusion Matrix

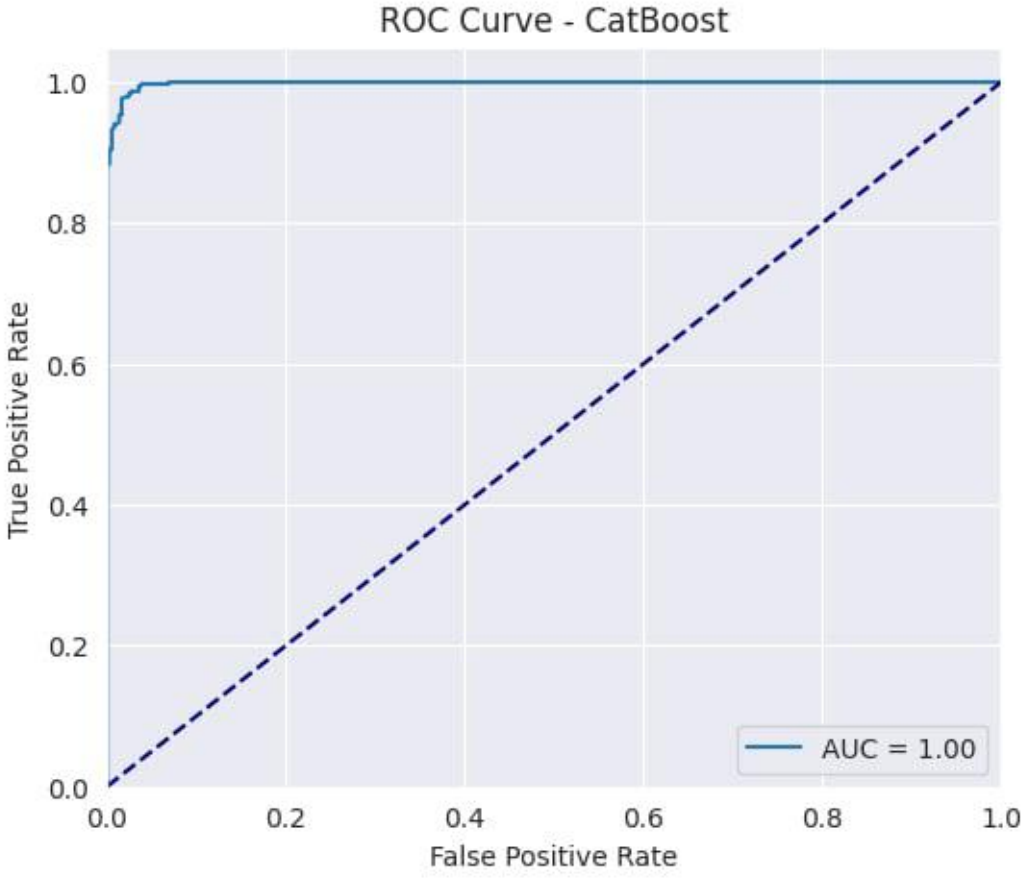

Figure 5.13: ROC Curve of CatBoost

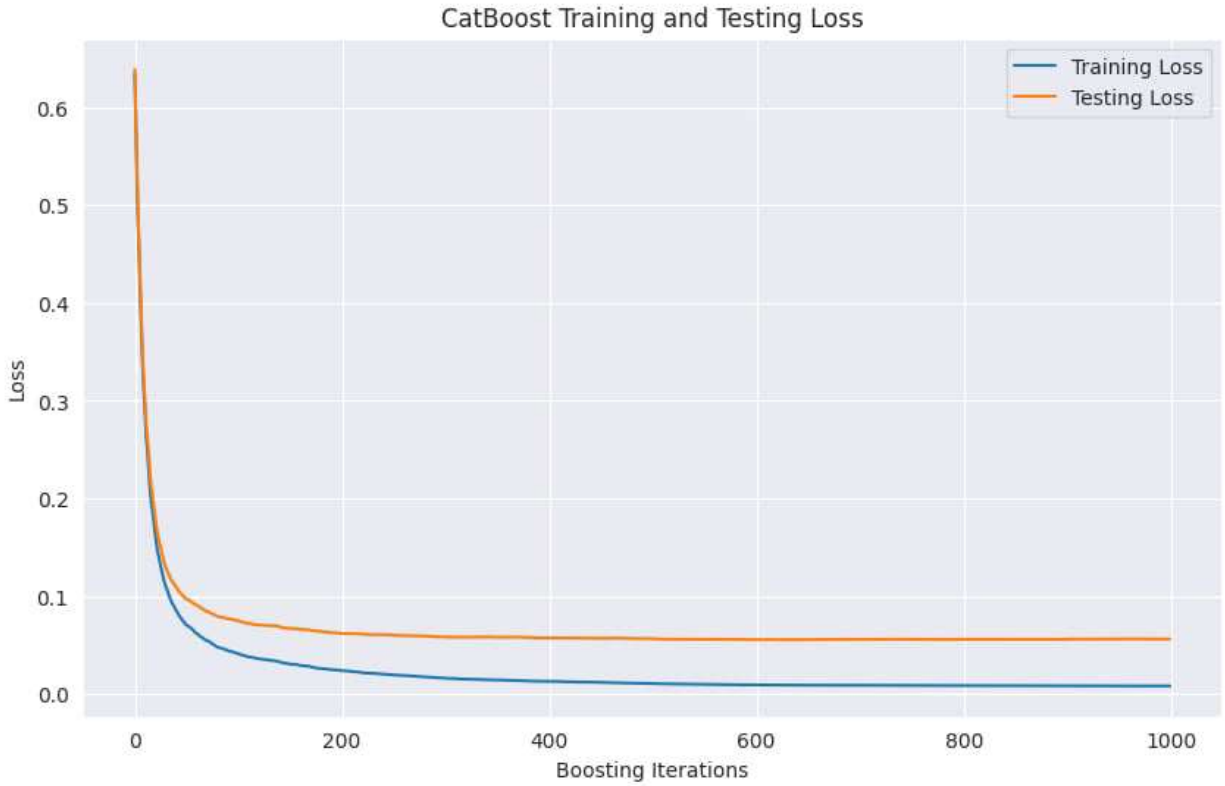

Figure 5.14: Training and Testing Loss of CatBoost

potential as a powerful model for autism detection.

Chapter 6

Result Comparison

We used Logistic Regression, KNN, Gradient Boosting, Random Forest, SVM and CatBoost model to detect autism from the behavior dataset. We got some good numbers in some of these models. The accuracy with the CatBoost model was the highest among all. We got the accuracy of 97.86% and for class-1 (autistic) precision, recall and F1 score of 97%, 98% and 98% with the CatBoost model. With Random Forest we got the second highest accuracy of 97.73%. The precision, recall and F1 scores for class-1 with the Random Forest are 98%. Gradient Boosting also gave us an excellent result with an accuracy of 96.92%. The Precision, recall and F1 score of Gradient Boosting are 96%, 98% and 97%. With KNN we got the accuracy of 93.59% and the precision, recall and F1 score are 93%, 94% and 94% respectively. SVM model gave us an accuracy of 86.91% and precision, recall and F1 score of 87%, 88% and 88%. Logistic Regression gave us an accuracy of 87.04%. Random Forest and CatBoost model was the most successful among these models. KNN and Gradient Boosting also performed very well with our dataset having accuracy above 95%. CatBoost is the best model for detecting autism from the behavior dataset from our research.

In Table 6.1, a comparison between these machine learning is presented. Precision, recall and F1 scores are for class-1(Autistic)

Models	Accuracy	Precision	Recall	F1 Score
Logistic Regression	**87.04%**	**88%**	**87%**	**87%**
KNN	**93.59%**	**93%**	**94%**	**94%**
Gradient Boosting	**96.92%**	**96%**	**98%**	**97%**
Random Forest	**97.73%**	**98%**	**98%**	**98%**
SVM	**86.91%**	**87%**	**88%**	**87%**
CatBoost	**97.86%**	**97%**	**98%**	**98%**

Table 6.1: Models Comparison

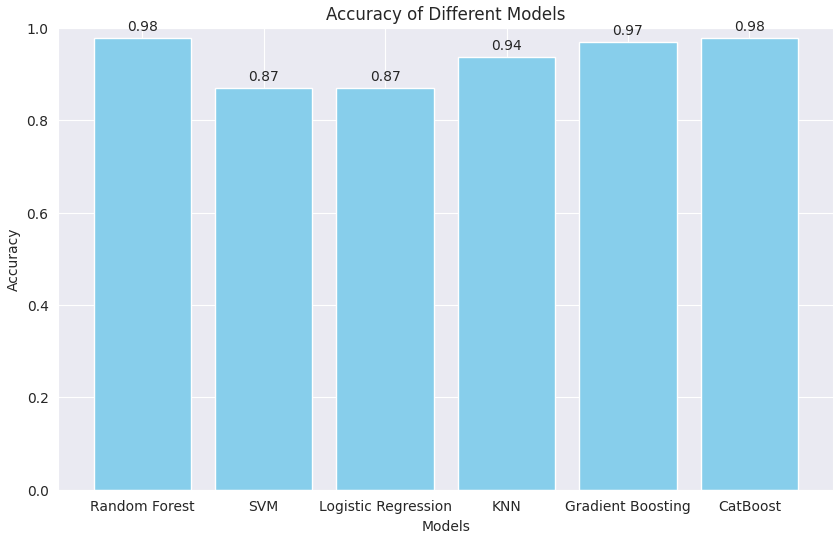

Figure 6.1: Comparing Accuracy of Different Models

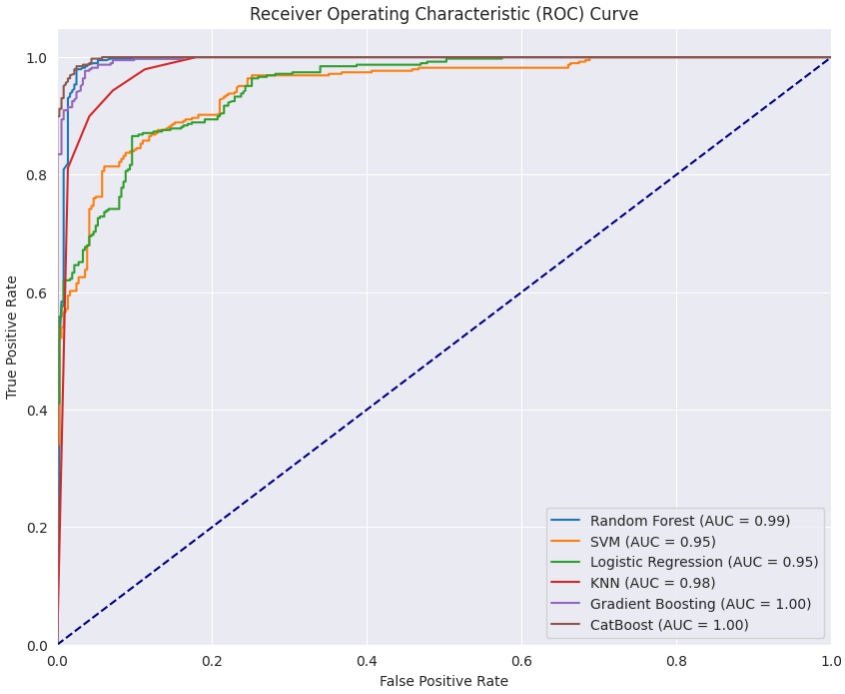

Figure 6.2: ROC Curve of Different Models

Chapter 7

Future work and Conclusion

7.1 Conclusion

In today's world autism is a widely known neurological condition that requires more focus to be shed on its complexities and different spectrums and how to be addressed by others. One approach to address such conditions involves detecting autism, which can sometimes be accomplished during early childhood by examining the facial features or physical attributes of the individual. Alternatively, it may involve identifying behavioral patterns that align with various points on the autism spectrum. In our paper, we aimed to introduce diversity by exploring behavioral datasets through machine learning algorithms. These methods were designed to identify autism from b behavioral datasets. We tried to ensure that there were fewer chances of misdiagnosis by proper training of machine learning algorithms. The confusion matrix assesses the performance of models that we are using in this research which helps in evaluating the negative and positive classes. Application of different machine learning models like Logistic Regression, Random Forest, CatBoost, GradientBoost, SVM, and KNN helped us detect the characteristics of autism beneficially from the behavioral dataset. Furthermore, leveraging autism features extracted from individual modalities of data will aid in enhancing the accuracy rate. Our result analysis revealed that CatBoost performed way better than other models and yielded a higher accuracy of 97.86 percent following the CatBoost model we have our Random Forest model with an accuracy rate of 97.73 percent, Gradient Boosting model showed a 96.92 percent accuracy rate while KNN yielded a 93.59 percent accuracy. Logistic Regression and SVM had slightly lower accuracy rates where SVM achieved 86.91 percent accuracy and Random Forest had 87.04 percentage of accuracy. Overall, our paper demonstrates superior performance with the behavioral dataset and machine learning models.

7.2 Future Work

For the future we want our data collection process to involve acquiring image datasets, predominantly comprising images of children from diverse Asian countries, because when we looked for the required dataset we observed that, the dataset lacked a representation of Asian facial features. Also, we could not find valid image datasets from online platforms hence we want to focus on that sphere. In our study involving behavioral datasets, we observed a lack of corresponding images for individuals whose behavioral data were provided. Consequently, we could not conduct a comparative analysis on the same individual using their images and behavior together while applying different deep learning models. Due to previous circumstances in the country, and also due to privacy issues of patients and most importantly us being undergraduate students, we were unable to personally collect additional data within the given timeframe. In the future, we want to collect image and behavioral datasets from the same individuals so that we can merge the two results and show a comparative study as well as focus on Asian features to enhance our dataset. Our future plans include gathering more data as primary sources to address this gap and enhance the accuracy of our behavior-based models while we work on image-based models on the sideline. We would like to make a web interface that can detect autism through image or behavior datasets. We can make that free and people can get highly benefited from this. We want to create an impact even if it's small in the autism detection sphere of the medical sector.

References

[1] G. H. John and P. Langley, "Estimating continuous distributions in bayesian classifiers," *arXiv preprint arXiv:1302.4964*, 2013.

[2] O. Altay and M. Ulas, "Prediction of the autism spectrum disorder diagnosis with linear discriminant analysis classifier and k-nearest neighbor in children," in *2018 6th international symposium on digital forensic and security (ISDFS)*, IEEE, 2018, pp. 1–4.

[3] A. S. Heinsfeld, A. R. Franco, R. C. Craddock, A. Buchweitz, and F. Meneguzzi, "Identification of autism spectrum disorder using deep learning and the abide dataset," *NeuroImage: Clinical*, vol. 17, pp. 16–23, 2018.

[4] R. Vaishali and R. Sasikala, "A machine learning based approach to classify autism with optimum behaviour sets," *International Journal of Engineering & Technology*, vol. 7, no. 4, p. 18, 2018.

[5] S. R. Dutta, S. Datta, and M. Roy, "Using cogency and machine learning for autism detection from a preliminary symptom," in *2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, IEEE, 2019, pp. 331–336.

[6] B. Marr, "Amazing examples of computer and machine vision in practice," *Forbes.[online]¡ 11nq. com/RqPRP*, 2019.

[7] M. S. Satu, F. F. Sathi, M. S. Arifen, M. H. Ali, and M. A. Moni, "Early detection of autism by extracting features: A case study in bangladesh," in *2019 international conference on robotics, electrical and signal processing techniques (ICREST)*, IEEE, 2019, pp. 400–405.

[8] F. Thabtah, N. Abdelhamid, and D. Peebles, "A machine learning autism classification based on logistic regression analysis," *Health information science and systems*, vol. 7, no. 1, p. 12, 2019.

[9] M. Beary, A. Hadsell, R. Messersmith, and M.-P. Hosseini, "Diagnosis of autism in children using facial analysis and deep learning," *arXiv preprint arXiv:2008.02890*, 2020.

[10] S. Raj and S. Masood, "Analysis and detection of autism spectrum disorder using machine learning techniques," *Procedia Computer Science*, vol. 167, pp. 994–1004, 2020.

[11] F. C. Tamilarasi and J. Shanmugam, "Convolutional neural network based autism classification," in *2020 5th International Conference on Communication and Electronics Systems (ICCES)*, IEEE, 2020, pp. 1208–1212.

[12] T. Akter, M. I. Khan, M. H. Ali, M. S. Satu, M. J. Uddin, and M. A. Moni, "Improved machine learning based classification model for early autism detection," in *2021 2nd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST)*, IEEE, 2021, pp. 742–747.

[13] S. Jahanara and S. Padmanabhan, "Detecting autism from facial image," *International Journal of Advance Research, Ideas and Innovations in Technology*, vol. 7, no. 2, pp. 219–225, 2021.

[14] A. Lu and M. Perkowski, "Deep learning approach for screening autism spectrum disorder in children with facial images and analysis of ethnoracial factors in model development and application," *Brain Sciences*, vol. 11, no. 11, p. 1446, 2021.

[15] Z. A. Ahmed, T. H. Aldhyani, M. E. Jadhav, *et al.*, "Facial features detection system to identify children with autism spectrum disorder: Deep learning models," *Computational and Mathematical Methods in Medicine*, vol. 2022, 2022.

[16] F. W. Alsaade, M. S. Alzahrani, *et al.*, "Classification and detection of autism spectrum disorder based on deep learning algorithms," *Computational Intelligence and Neuroscience*, vol. 2022, 2022.

[17] S. M. Hasan, M. P. Uddin, M. Al Mamun, M. I. Sharif, A. Ulhaq, and G. Krishnamoorthy, "A machine learning framework for early-stage detection of autism spectrum disorders," *IEEE Access*, vol. 11, pp. 15 038–15 057, 2022.

[18] M. R. Kanhirakadavath and M. S. M. Chandran, "Investigation of eye-tracking scan path as a biomarker for autism screening using machine learning algorithms," *Diagnostics*, vol. 12, no. 2, p. 518, 2022.

[19] Y. Lin, Y. Gu, Y. Xu, S. Hou, R. Ding, and S. Ni, "Autistic spectrum traits detection and early screening: A machine learning based eye movement study," *Journal of Child and Adolescent Psychiatric Nursing*, vol. 35, no. 1, pp. 83–92, 2022.

[20] P. Mukherjee, S. Sadhukhan, M. Godse, and B. Chakraborty, "Early detection of autism spectrum disorder (asd) using traditional machine learning models," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 6, 2023.

[21] A. Rashid and S. Shaker, "Autism spectrum disorder detection using face features based on deep neural network," *Wasit Journal of Computer and Mathematics Sciences*, vol. 2, no. 1, 2023.

[22] A. Singh, M. Laroia, A. Rawat, and K. Seeja, "Facial feature analysis for autism detection using deep learning," in *International Conference On Innovative Computing And Communication*, Springer, 2023, pp. 539–551.

[23] S. Islam, T. Akter, S. Zakir, S. Sabreen, and M. Hossain, "Autism spectrum disorder detection in toddlers for early diagnosis using machine learning. in 2020 ieee asia-pacific conference on computer science and data engineering (csde). 2020," *Gold Coast, Australia*, pp. 1–6,