UNIVERSITY OF HERTFORDSHIRE

School of Physics, Engineering and Computer Science

MSc Data Science and Analytics with Advanced Research

7COM1039-0109-2023 - Advanced Computer Science Masters Project

19/08/2024

Predicting and Diagnosing Autism Spectrum Disorder (ASD) Using Machine Learning Models

Name: Naga Durga Chaitanya Gulla
Student ID: 21086793
Supervisor: Raghubir Singh

MSc Final Project Declaration

This report is submitted in partial fulfilment of the requirement for the degree of Master of Science in Data Science and Analytics with Advanced Research at the University of Hertfordshire (UH).

It is my own work except where indicated in the report.

I did not use human participants in my MSc project.

I hereby give permission for the report to be made available on the university website provided the source is acknowledged

Naga Durga Chaitanya Gulla

21086793

Abstract

Autism Spectrum Disorder (ASD) is a severe neurological disease characterised by difficulty in social interaction and repetitive activities. Early and accurate diagnosis of ASD is critical for providing effective intervention and better results. This study investigates how machine learning algorithms can enhance accuracy in predicting and diagnosing ASD. The dataset, preprocessed to eliminate outliers and inconsistencies, included 3742 instances and 17 features. It was acquired via Kaggle. The machine learning model to predict and diagnose ASD was built using the Random Forest, K-Nearest Neighbors (KNN), Decision Trees, Support Vector Machines, and Logistic Regression models. To assess the performance of the model metrics like accuracy, precision, recall and f1-score have been used. The study showed that specific machine learning algorithms such as SVM classifiers and Random Forests, were highly accurate in forecasting ASD. These results indicate that machine learning has substantial potential to revolutionise ASD diagnosis by providing a reliable method for early identification and intervention in clinical environments. More research is needed to overcome the limitations related to the size of the dataset and any potential biases. Furthermore, integrating these models into clinical environments necessitates a thoughtful assessment of ethical considerations and practical applicability in real-world scenarios.

Acknowledgement

I am extremely grateful to my supervisor, Raghubir Singh, for their priceless guidance, support, and motivation during this project. Their expert advice and insightful feedback have played a crucial role in influencing the direction and results of my research. I am profoundly thankful for their time and dedication to mentoring me, which has greatly contributed to my professional and personal development. Additionally, I extend my thanks to the faculty and staff at the University of Hertfordshire for creating an optimal learning environment and providing the essential resources for the successful completion of this project.

Table of Contents

Table of Figures

Chapter 1: Introduction

1.1 Background

Autism Spectrum Disorder (ASD) has not always been as well understood as it is now. Early reports from the mid-twentieth century concentrated on certain features discovered in children by Leo Kanner. As the study progressed, the concept expanded to embrace a range of presentations, including characteristics articulated by Hans Asperger about the same period (Alqaysi et al., 2022). Currently, researchers recognise the heterogeneous nature of ASD, with individuals demonstrating varying degrees of difficulties with social communication, restricted interests, and repetitive behaviours. Increased awareness and improved diagnostic tools have resulted in a substantial increase in diagnosed cases. Nonetheless, the intricacy of ASD and the possibility of underdiagnosis in certain populations highlight the need for continued research to develop more objective and accessible diagnostic tools (Ahmed et al., 2022).

Most people are unaware of ASD conditions and therefore can't determine if someone is affected. This often leads to the individual's social isolation rather than their recovery. ASD is a medical disorder that starts during childhood and persists into adolescence and adulthood (Saha et al., 2021). Compared to girls boys are four times more likely to be diagnosed with ASD. Figure 1 shows the global prevalence of autism spectrum disorder. The legend in the figure represents varying levels of ASD with different colours from 0% to 1.2%.

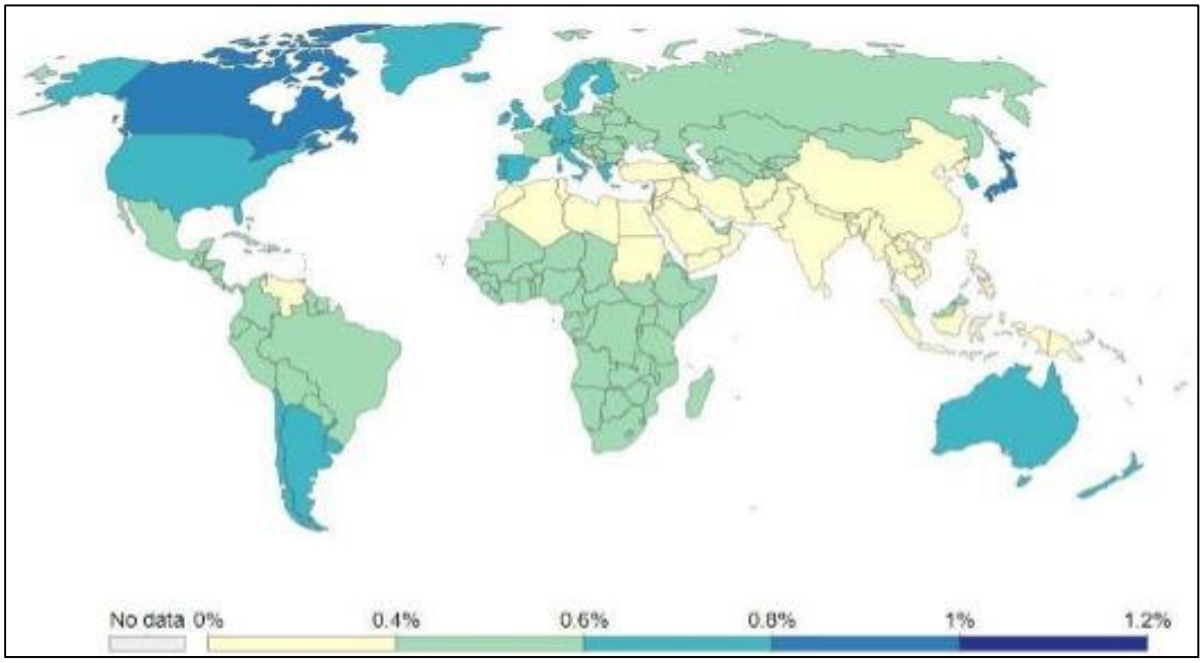

Figure 1: Global Prevalence of Autism Spectrum Disorder (Saha et al., 2021)

According to Doulah et al. (2023) between the ages of 18 and 24 months, ASD can be diagnosed early; during this time its symptoms can be differentiated from other developmental abnormalities and standard developmental delays. About one in every 36 children has been identified with autism spectrum disorder (ASD) according to estimates from the CDC's Autism and Developmental Disabilities Monitoring (ADDM) Network.

However, the rate of occurrence varies according to diagnostic criteria, screening procedures, and the general awareness of ASD. Naik et al. (2023) emphasised the insights of Paul Fergus, who identified three distinct levels of autism severity as "mild," "moderate," and "severe," encompassing Asperger's syndrome, Asperger's disorder, and Pervasive Developmental Disorder (PDD). Individuals with ASD confront various challenges, such as:

1. Lack of concentration.
2. Repetitive use of language, and challenges in interpreting gestures and facial expressions.
3. Delays in speech and motor skills.
4. Sensitive to touch, sound, and smell.
5. Difficulties with body language and vocal intonation.

Despite advances in study and awareness, diagnosing ASD is still difficult due to its broad range of symptoms and reliance on subjective clinical assessments. Traditional clinical procedures, such as the Autism Diagnostic Observation Schedule-Revised (ADOS-R) and the Autism Diagnostic Interview-Revised (ADI-R), can be deemed time-consuming and inconvenient for both patients and clinicians (Sallibi and Alheeti, 2023). Hossain et al. (2021) identified other limitations of traditional clinical procedures in their research. They are children who are too young and have delayed speech impairments, scoring approximately 25% of the total ADI-R items since the patient cannot accurately answer the verbal parts. Furthermore, one of the primary drawbacks of the ADOS-R technique is its tendency to overclassify children with various clinical problems. Every child with ASD is different, so a technological solution that works for one may not work for another. As a result, researchers have begun incorporating some technology to help children with autism to determine which technologies are most suited to everyone.

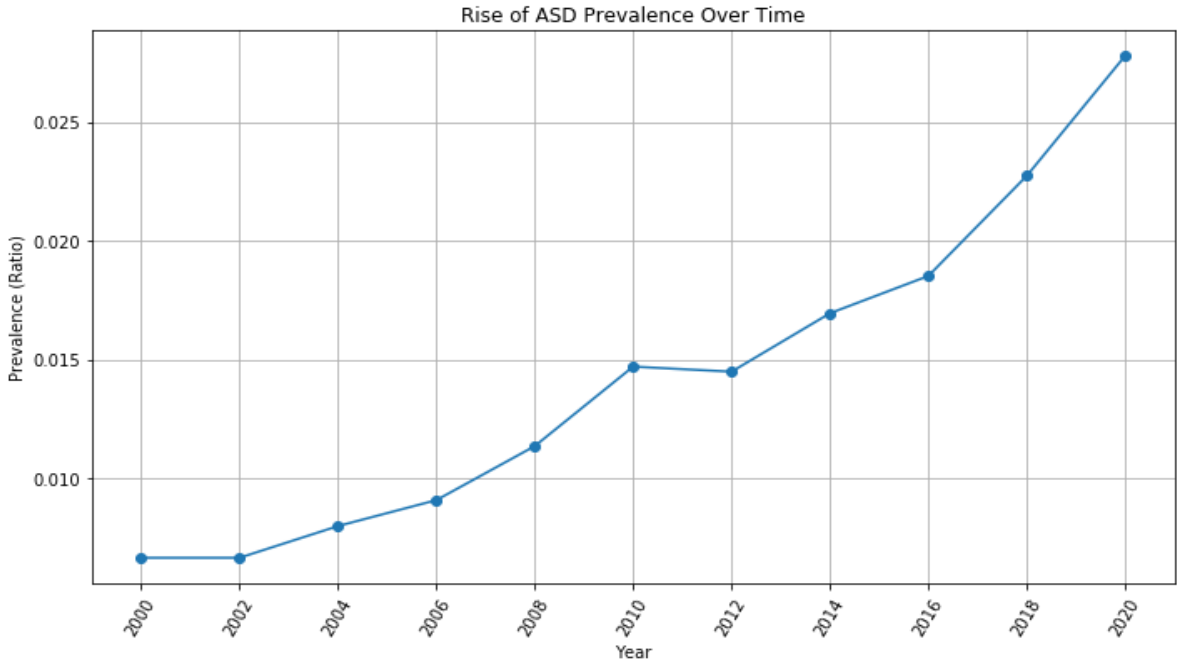

Figure 2: Rise in the ASD prevalence over the years (2000-2020)

(Source: Self-Originated)

According to the findings in Figure 2, there has been a consistent rise in the prevalence of ASD annually. The information for this visual representation was sourced from the Centers for Disease Control and Prevention (CDC) website and includes the prevalence rates of ASD from 2000 to 2020. Using Matplotlib in Python, I generated the line chart.

Healthcare practitioners are in desperate need of time-saving, simple, and accessible screening tools that may effectively predict whether a patient with a specific measurable trait has ASD and recommend individuals on whether they should go for a medical diagnosis or not. With the rapid growth of machine learning technology, there is growing interest in using these tools to enhance the diagnostic process for ASD by providing a more objective and data-driven approach. Machine learning (ML) provides a promising opportunity to improve the diagnosis of autism spectrum disorder (ASD). By processing extensive sets of clinical data, ML models can potentially recognise variations and characteristics that may indicate the presence of ASD. This technology can improve the precision of diagnoses, enable earlier identification, and simplify the diagnostic procedure.

1.2 Problem Statement

Autism Spectrum Disorder (ASD) significantly affects how people think, interact socially, and behave around the world. Individuals, with ASD may show behaviour that can put themselves and others at risk underscoring the importance of evaluations by psychologists and medical experts to diagnose the condition accurately. However traditional diagnostic methods are often subjective and time-consuming causing delays in diagnosis and hindering access, to treatment. This project aims to address these challenges by developing and testing machine learning models using publicly available data to predict and diagnose ASD more objectively and efficiently.

1.3 Research Questions

This project investigates the potential of machine learning for ASD diagnosis by answering the following research questions:

1. “Can machine learning models accurately predict and diagnose autism spectrum disorder based on clinical data and how can this technology be effectively integrated into clinical practice to improve early identification and intervention for individuals with `ASD?”
2. “What are the limitations of using existing datasets for training and evaluating machine learning models for ASD, and how can these limitations be addressed?”

1.4 Objectives

1. To assess the feasibility of using clinical data to predict and diagnose ASD using machine learning methods.
2. To construct and train machine learning models using clinical datasets for ASD diagnosis.
3. Evaluate the performance of the trained classification techniques on unseen data to select the most accurate algorithm for ASD prediction.

4. To identify the potential constraints of the existing datasets for training and testing machine learning techniques to predict and diagnose ASD.
5. To explore the application of machine learning models in clinical environments for early detection and intervention of ASD.

1.5 Legal and Ethical Considerations

According to Deng (2021), ethics is the body of norms that direct our expectations and others' actions. For this project, I have utilised a dataset obtained from the Kaggle platform. The dataset is publicly accessible and free to use and does not contain any private or confidential information. The following are the legal and ethical considerations of this project.

Ethical Considerations:

1. To ensure fairness in the machine learning models, evaluating the Kaggle dataset for potential biases in how different demographics are diagnosed or represented is important. This is because the models can inherit biases from the training data.
2. The model's accuracy might have been restricted to the population the data represents. Therefore, I was transparent about these limitations and the specific target population for which the model was intended.
3. It is imperative to comprehend how the model makes its predictions, especially in a medical context such as diagnosing ASD. Therefore, I investigated ways to improve the model's interpretability through various techniques.

Legal Considerations:

1. Throughout the project, data protection laws and regulations, such as the GDPR in Europe and HIPAA in the United States, were strictly followed to safeguard the privacy and rights of individuals featured in the dataset.
2. To ensure the safe usage of machine learning models for diagnosing ASD, I have taken the necessary steps to identify and minimise potential risks, including data breaches, algorithmic biases, and unintended consequences. I will also implement safeguards to address these risks in case they arise.

1.6 Structure of the Report

The report is structured into several chapters to provide a thorough understanding of the project. The "Introduction" section gives background information on ASD, introduces the research questions, and offers a brief overview. In the "Literature Review" section, recent studies on predicting and diagnosing ASD using machine learning are comprehensively discussed. The "Methodology" section explains the proposed machine learning models' implementation and advantages and disadvantages. "Results and Discussion" discusses the outcomes of the machine learning models, compares their performance, and addresses the research questions. Finally, the "Conclusion and Future Work" section summarises the key findings, discusses the study's limitations, and suggests potential areas for future research. "Appendix" contains code used in the project.

1.7 Chapter Summary

This section examines autism spectrum disorder (ASD), beginning with early studies by Leo Kanner and Hans Asperger and progressing to current understandings that appreciate its diversity. It emphasises the difficulties associated with diagnosing ASD due to its vast range of symptoms and the limits of current approaches, which are time-consuming and subjective. This section highlights the significance of early and correct diagnosis for appropriate treatment. It suggests employing machine learning (ML) to improve ASD diagnosis by providing a more objective and data-driven approach, which could improve accuracy and early detection. Additionally, it delves into the project's problem statement, research questions, and objectives, such as determining the feasibility of using clinical data for ML-based ASD diagnosis, developing, and assessing ML models, and investigating their implementation in clinical practice for early detection and treatment. It highlighted the necessity of considering legal and ethical considerations when utilising machine learning in healthcare, particularly for sensitive diagnoses such as ASD. It briefly outlined the report's structure, which consists of a literature review, methods, results and discussion, inferences, and recommendations for further research.

Chapter 2: Literature Review

Many research papers have employed machine learning techniques in different ways to enhance and expedite the detection of ASD. This chapter comprehensively summarises the existing studies on applying machine learning techniques for predicting and diagnosing ASD.

The study of Islam et al. (2020) used machine-learning models for early diagnosis of ASD in toddlers. The authors have used the Random Forest Classifier, SVM, Naïve Bayes, KNN, Decision Tree Classifier and Logistic Regression models in this study. They excluded the logistic regression and decision tree classifiers from these six classifiers as these two models resulted in overfitting. The authors employed one-hot encoding to convert values in the categorical columns of the dataset to numerical ones. The model's performance has been evaluated with the help of the ROC-AUC curve and confusion matrices. Of the remaining 4 models KNN and random forest models have shown the highest accuracy of 98% and 93% respectively. After these two classifiers, Naïve Bayes was the third-best model in this study with an accuracy of 89%. In their upcoming work, the authors plan to use larger datasets to predict ASD and to build an intuitive mobile application that will help individuals predict ASD by themselves and get medical assistance when necessary.

Mashudi, Ahmad, and Noor (2021) conducted a study using a machine-learning approach to classify adult autistic spectrum disorder. In their research, the authors utilised the data set with 703 instances from the UCL machine learning repository. Thabtah collected this data through his mobile application called ASD test. The authors used Weka software to proceed with the pre-processing and segmentation of ASD for the dataset and filtered the data into nominal features with the help of discretisation. In this research, authors have used KNN, SVM, Naïve Bayes, J48 decision tree, AdaBoost, Bagging and stacking techniques to classify ASD in adults. The authors have used 3, 5, and 10-fold cross-validations to measure the model's performance. The SVM, KNN, bagging, and J48 techniques have shown 100% accuracy in all 3, 5, and 10-fold cross-validation. Similar to this study, the present research removed the non-autistic features to enhance the model's accuracy.

The recent study of Manoj and Praveen (2023) used machine learning and deep learning techniques to support the prediction of ASD in children, toddlers, adolescents, and adults. The authors have identified the best classifier with the help of precision, recall, and F1-score metrics. The authors obtained the children's data from Kaggle and the data for the toddlers, adolescents, and adults from the UCL machine learning repository. This study employed stratified k-fold cross-validation to minimise the class imbalance associated with toddler and adult datasets. Also, One-hot encoding was used for ethnicity and country_of_res features as these features have more than two classes. Logistic regression and Multi-Layer Perceptron models showed the best results for child and toddler datasets, XGBoost and Logistic regression models have given better accuracy on adult and adolescent data.

The study conducted by Vakadkar, Purkayastha, and Krishnan (2021) revealed that it is possible to identify children at risk of ASD at an early age, leading to expedited diagnosis and treatment. Using a variety of parameters, including age, sex, ethnicity, and so on, the researchers employed five machine-learning models in this study to categorise individual

subjects as either having ASD or not. The best-performing model was then identified by evaluating each classifier. One-hot encoding was utilised for multiclass features to prevent the model from sorting the data hierarchically, and label encoding was used to transform the labels into numerical form so that machines could read them. Logistic regression produced the best accuracy among the five models used to analyse the dataset. To improve the system's overall performance and resilience, the researchers planned to apply deep learning algorithms that integrate CNNs with classification in their future work. In the present study, label encoding was utilised to convert categorical columns such as "Sex", "Jaundice", "Family_mem_with_ASD", and "ASD_Traits" into numerical representations, as was done in the work by Vakadkar, Purkayastha, and Krishnan (2021). This process allows the machine learning models to process these categorical features efficiently during training.

Ahmed et al. (2022) employed machine learning and deep learning approaches to introduce a method for early detection and diagnosis of ASD with eye-tracking technology. The authors used an image dataset collected from the Figshare data repository in this study. The dataset consists of 547 pictures, of which 219 images were related to ASD and 328 were related to Typically Developing (TD) Children. As the data size is small for deep learning methods authors have employed a data augmentation technique on training data. After data augmentation, the count of images was increased to 1750 for ASD and 1834 for TD children. In this study, the authors used feed-forward neural networks, artificial neural networks, and convolution neural networks to predict ASD utilising image data. Additionally, they have used a hybrid method called GoogleNet + SVM and ResNet-18 + SVM, combining deep learning and machine learning techniques. The feed-forward and artificial neural networks showed the highest accuracy of 99.8% and AUC of 0.99. The GoogleNet model has an accuracy of 93.6% and the ResNet-18 model has an accuracy of 97.6%.

The study of Zhou, Yu, and Duong (2014) focused on using machine learning algorithms and graph theory to predict clinical outcomes in autism spectrum disorder (ASD) based on multi-parametric MRI data. The research aimed to predict outcomes such as ADI-R and ADOS using multi-parametric MRI data and to compare functional and structural MRI data between ASD and typically developing (TD) children. The authors utilised the ABIDE database for MRI data analysis and implemented principal component analysis (PCA) for feature selection to enhance classifier generalisation. The classifiers were evaluated using the Waikato Environment for Knowledge Analysis (WEKA) software with cross-validation methods, revealing that small-world network analysis, based on graph theory, demonstrated efficiency inequalities in the ASD and TD groups. Furthermore, the study highlighted that essential imaging features identified through advanced machine learning algorithms proved to be highly predictive of phenotypic traits in ASD, including ADOS and ADI-R.

Farooq et al. (2023) used a federated learning technique that was specifically designed to diagnose autism in both children and adults. They used logistic regression and support vector machine classifiers to classify and identify ASD conditions. According to the study, autism develops due to environmental or genetic factors influencing the nervous system and individuals' social and cognitive skills. The results of the logistic regression (LR) and support vector machine (SVM) classifiers were calculated in terms of accuracy, precision, and F1

score, before being sent to a central server for meta-classifier training. The study involved updating models on local devices and sending them again to the central server for aggregation. The dataset used in this research is similar to the one utilised in the present study. Random forest, decision tree, SVM, and KNN classifiers are also used to predict and diagnose ASD. The random forest and SVM classifiers yielded better accuracy than other models in predicting ASD. From their study, an SVM classifier was incorporated in this research to separate the ASD dataset into affected and non-affected classes to predict targets and manage overfitting.

The research conducted by Alteneiji, Alqaydi, and Tariq (2020) looked to employ an advanced machine learning (ML) approach for the identification of autism spectrum disorder (ASD). They applied the knowledge discovery in database (KDD) method to uncover hidden algorithmic patterns important for model development. In addition to characteristics like age, gender, ethnicity, jaundice, and family history of ASD, the study used data from various questionnaire screening methods focusing on key ASD symptoms across different age groups. The specific questionnaires used in the study are detailed in Tables 1, 2, and 3.

Feature	Autism Spectrum Quotient (AQ-10) Children screening features
A1	S/he frequently detects quiet noises that others may overlook
A2	S/he generally focuses on the overall view rather than specific details
A3	In a social setting, s/he can effortlessly monitor multiple conversations happening among different people
A4	S/he has no trouble transitioning between various activities
A5	S/he lacks the proficiency to sustain engaging conversations with peers
A6	S/he excels at casual social conversation
A7	When s/he listens to a story, s/he struggles to understand the character's motivations or emotions
A8	During preschool, s/he used to relish engaging in imaginative play with other children
A9	S/he can easily discern someone's thoughts or emotions by simply observing their facial expressions
A10	S/he struggles to form new friendships

Table 1: Autism Spectrum Quotient (AQ-10) for Children 4-11 (Alteneiji, Alqaydi, and Tariq, 2020)

Feature	Autism Spectrum Quotient (AQ-10) Adolescent screening features
A1	S/he constantly recognises patterns in things
A2	S/he usually concentrates more on the whole picture rather than the small

	details
A3	In a social setting, s/he has no trouble keeping up with multiple conversations among different people
A4	If there's a disruption, s/he can swiftly return to the task at hand
A5	S/he frequently finds it challenging to sustain a conversation
A6	S/he excels at casual social conversation
A7	When s/he was younger, s/he relished participating in imaginative play with other kids
A8	S/he has trouble envisioning what it's like to be in someone else's position
A9	S/he navigates social situations with ease
A10	S/he struggles to form new friendships

Table 2: Autism Spectrum Quotient (AQ-10) for Adolescents 12-17 (Alteneiji, Alqaydi, and Tariq, 2020)

Feature	**Autism Spectrum Quotient (AQ-10) Toddler screening features**
A1	S/he frequently responds by making eye contact when you call his/her name
A2	S/he frequently has no trouble making eye contact with you
A3	S/he can effortlessly use gestures to communicate a desire for something
A4	S/he can readily use pointing to express interest to you
A5	S/he can easily pretend
A6	S/he excels at casual social conversation
A7	S/he can easily follow where you are looking
A8	When you or someone in the family is upset, s/he can show signs of wanting to comfort them
A9	S/he initially spoke typical words
A10	S/he finds it easy to use simple gestures

Table 3: Autism Spectrum Quotient (AQ-10) for Toddlers 1-3 (Alteneiji, Alqaydi, and Tariq, 2020)

A recent study by Naik et al. (2023) developed a machine-learning model to predict the prevalence of autism spectrum disorder (ASD) based on user-provided characteristics. The study employed machine learning algorithms to build the predictive model, including K-Nearest Neighbors (KNN), Logistic regression, Random Forest Classifier, Decision Tree Classifier, Naïve Bayes, and XGB classifier. Except for the Naïve Bayes and XGB classifier,

all the algorithms mentioned above were used in the present research and included a Support Vector Machine (SVM) for model construction. The researchers validated the model by giving a use case in which the system analysed user-supplied characteristics for disease identification. This included preparing the incoming data, developing a predictive model, and evaluating its correctness. Upon implementing their methodology, the study reported an impressive 98% accuracy in predicting ASD using the employed algorithms.

Sallibi and Alheeti's (2023) research investigated the potential of machine learning for the early detection of autism spectrum disorder (ASD) in children. The researchers underscored the limitations of traditional clinical methods such as the Autism Diagnostic Observation Schedule-Revised (ADOS-R) and the Autism Diagnostic Interview-Revised (ADI-R) for ASD diagnosis. They also highlighted risk factors for ASD, including low birth weight, family history of ASD, and advanced parental age. The researchers worked with a dataset comprising 1154 instances and 18 variables, which were preprocessed to remove outliers and missing values. To address the imbalanced target variable in the dataset, they applied the Synthetic Minority Oversampling Technique (SMOTE). After evaluating multiple machine learning models, the researchers discovered that the random forest model had 100% accuracy, average precision, and an AUC of 1.0 for early ASD identification.

In their recent study, Saha et al. (2021) introduced a dashboard with interactive features designed for analysing autism spectrum disorder (ASD) through machine learning. This creative dashboard, built with Tableau software, offers a range of features such as ASD Rates per Ethnicity, Heat Map, Total ASD Counts, Impact of Family Members, ASD Traits vs. No ASD Traits, ASD in Different ethnicities, and ASD based on gender. Within the ASD Rates per Ethnicity section, the researchers presented data on the number of participants who completed the ASD test and the percentage of those who exhibited ASD characteristics. They discovered that support vector machines (SVM) and convolutional neural networks (CNN) delivered more accurate results than other machine learning algorithms. Notably, their analysis revealed a higher prevalence of ASD among males compared to females.

Based on the review of the previous research studies on early identification and diagnosis of ASD using machine learning algorithms, it has been observed that machine learning models are highly effective in predicting ASD with the given data. Most of the studies in the review have shown more than 90% accuracy in detecting ASD. However, all of these studies, except Farooq et al. (2023), used smaller datasets, including nearly 800 – 1000 instances; these models might result in less accuracy when used on larger datasets.

2.1 Research Gap

The research studies disclosed in the “Literature Review” exhibit the capability of machine learning ASD diagnosis, however, there are possibilities to enhance the practicality, accessibility, and generalisation of the model. Due to the use of small datasets in the reviewed studies, the ability of the model to apply to real-world populations has been limited. Furthermore, the emphasis on high accuracy frequently surpasses the need for models that are easy to interpret. It is necessary to assess the potency of machine learning algorithms

including ensemble methods like Ada Boost, and Gradient Boost comprehensively to compare accuracy, generalisability, and interpretability.

This project aims to fill this gap by using a dataset with 3742 instances larger than the datasets used by studies mentioned in the “Literature Review”. Furthermore, this study addresses the limitations of using existing datasets for training and evaluating machine learning models for ASD. With ensemble methods such as Ada Boost, and Gradient Boosting this study will improve the accuracy of models in predicting ASD. Finally, this research project will compare ensemble methods to single classifiers (Random Forest, KNN, SVM, Logistic Regression, and Decision Tree) to determine their effectiveness in terms of accuracy, generalisability, and interpretability for predicting ASD on larger datasets. By filling these research gaps, this project seeks to contribute to developing more reliable, understandable, and generally applicable machine learning models for ASD diagnosis.

2.2 Chapter Summary

This chapter looked into existing research studies that used a machine-learning approach to predict and diagnose ASD. These existing studies showed promising outcomes with more than 90% accuracy for most of the classification algorithms they used. The usage of small datasets is a common limitation in these studies, which may hinder the generalisability to real-world situations. This chapter outlines a research gap in systematically analysing ensemble approaches and leveraging larger datasets to improve model accuracy, generalisability, and interpretability. The study intends to close these gaps by using a larger dataset and comparing ensemble approaches to single classifiers, resulting in more accurate and understandable ASD diagnostic models.

Chapter 3: Research Methodology

This section starts with an explanation of the working model used for predicting and diagnosing autism spectrum disorder using machine learning models. The later part of this section explains the data collection and preprocessing process, a detailed explanation of the machine learning models used, and the process to develop and evaluate machine learning models for predicting ASD.

3.1 Working Model

Figure 3, illustrates the block diagram of the working model for this research study. The process starts with data collection and pre-processing, including handling missing values and outliers, noise removal, and standardisation. Also, the feature engineering technique called label encoding was employed to transform the categorical variables into numerical form, resulting in a fast and accurate training process for the machine learning models. After the pre-processing of data has been completed machine learning algorithms such as Random Forest Classifier, KNN, Decision Tree Classifier, Support Vector Machine and Logistic Regression are used to predict ASD. The performance of each of these algorithms has been assessed and compared using evaluation metrics such as accuracy, precision, recall and f1-score. After comparing evaluation metrics, the model with the highest accuracy and f1-score is selected as the optimal model and will be used for more training and classification. The block diagram of this working model has been drawn by using an online website called Drawio in which it is possible to draw block diagrams, flowcharts, ER diagrams, and sequence diagrams.

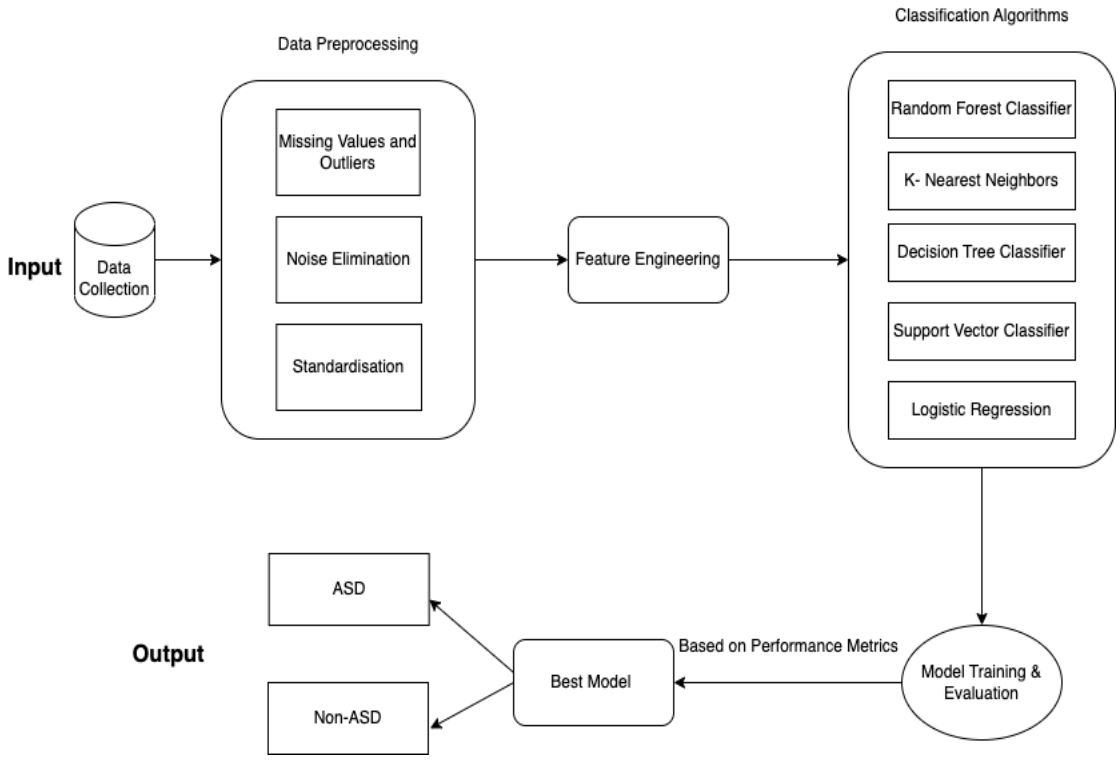

Figure 3: Block diagram of the working model for predicting ASD

(Source: Self-Originated)

3.2 Data Collection and Pre-Processing

The dataset for this research study has been obtained from the Kaggle platform, it is freely available to the public and does not contain any personal or hidden details. Pre-processing transforms a dataset before it is fed into a model (Vakadkar, Purkayastha, and Krishnan, 2021). The dataset contains no missing values except an outlier in the "Age_Years" column with an age of 383 years. This entire row corresponding to that age has been dropped using the df.drop method. After the completion of the data pre-processing the dataset has 3742 rows and 17 columns, including 11 numerical and 6 categorical columns. Figure 4 summarises the dataset, including the mean and standard deviation values in numerical columns. The column name of the "A10_Autism_Spectrum_Quotient" has been renamed to "A10".

```
In [15]: from skimpy import skim
         skim(df1)
```

skimpy summary

Data Summary

dataframe	Values
Number of rows	3742
Number of columns	17

Data Types

Column Type	Count
int64	11
string	6

number

column_name	NA	NA %	mean	sd	p0	p25	p50	p75	p100	hist
A1	0	0	0.453	0.4978	0	0	0	1	1	
A2	0	0	0.3381	0.4731	0	0	0	1	1	
A3	0	0	0.3121	0.4634	0	0	0	1	1	
A4	0	0	0.3819	0.4859	0	0	0	1	1	
A5	0	0	0.3894	0.4877	0	0	0	1	1	
A6	0	0	0.3784	0.4851	0	0	0	1	1	
A7	0	0	0.4447	0.497	0	0	0	1	1	
A8	0	0	0.3808	0.4857	0	0	0	1	1	
A9	0	0	0.3367	0.4727	0	0	0	1	1	
A10	0	0	0.5099	0.5	0	0	1	1	1	
Age_Years	0	0	11.15	10.64	1	3	8	14	64	

string

column_name	NA	NA %	words per row	total words
Sex	0	0	1	3742
Ethnicity	0	0	1.6	5898
Jaundice	0	0	1	3742
Family_mem_with_ASD	0	0	1	3742
Who_completed_the_test	0	0	2.2	8161
ASD_traits	0	0	1	3742

End

Figure 4: Summary of the Dataset

(Source: Self-Originated)

3.3 Feature Engineering

Using categorical variables directly in machine-learning algorithms can make models less interpretable and the complexity of the model might be increased. To address these problems, this project employs label encoding. This technique assigns a unique integer value to each value in categorical columns, making the data machine-readable. The categorical columns "Sex", "Jaundice", "Family_mem_with_ASD", and "ASD_traits" contain Yes and No values and these values are converted to 1s and 0s using label encoding. The specific command for this conversion is "df1["ASD_traits"]= encoder.fit_transform(df1["ASD_traits"]). The Same command is used for the other three columns by just replacing them in place of ASD_traits. The remaining two categorical columns "who_completed_the_test" and "Ethnicity" have been dropped as these two attributes do not have any effect on the output column "ASD_traits". The dataset that has finished the pre-processing and feature engineering methods will be fed to the machine learning algorithms to predict ASD.

3.4 Classification Algorithms

Classification algorithms such as Support Vector Machine (SVM), Random Forest, K-Nearest Neighbors (KNN), Decision Trees, and Logistic Regression have been used in this study. Initially, Stratified K-Fold validation was used to evaluate each model with n_splits=10. The models have been trained and validated in each split on the validation set which contains 674 samples. The metrics of each model have been shown in tabular format for every split along with the best hyperparameters in that particular split. Later to assess each model on test data (unseen data) every algorithm was trained and tested separately to predict ASD. This section will divide the dataset into 80% of the training and 20% of the test sets. The training set consists of 2993 instances from the original dataset, which will be used to train the machine learning algorithms employed for this study. The test data has 749 samples which will be used as unknown data to assess the model's accuracy and other metrics. Additionally, ensemble methods such as AdaBoost, Gradient Boosting and Hist Gradient Boosting were used to compare the effectiveness of these methods over single classifiers.

3.4.1 Support Vector Machine (SVM)

Support Vector Machine (SVM) is the most extensively used supervised machine learning technique introduced by Vladimir N. Vapnik in 1990, and it is used for classification and regression problems. SVM develops a hyperplane that separates the dataset into two groups in the most practical way. SVM is good at pattern recognition and classification, therefore it is perfect for classifying ASD from non-ASD cases. It can handle high-dimensional data which commonly occurs in ASD diagnosis with features such as behaviour, genetics, and brain scan. Additionally, SVMs perform well on complex datasets, particularly when handling imbalanced classes. In this project, SVM conquers the other models except random forest with an accuracy of 92%, an AUC score of 0.97, and an average precision of 0.98. Hyperparameter tuning with GridSearchCV is used to optimise the model performance. Initially, the linear kernel gave less accuracy and later the use of the radial basis function (RBF) kernel improved the model's performance with better accuracy.

```
Accuracy of SVM: 0.9212283044058746
Best Parameters: {'C': 1, 'kernel': 'rbf'}
              precision    recall  f1-score   support

           0       0.89      0.95      0.92       355
           1       0.96      0.89      0.92       394

    accuracy                           0.92       749
   macro avg       0.92      0.92      0.92       749
weighted avg       0.92      0.92      0.92       749
```

Figure 5: Output of SVM Classifier to Predict ASD

3.4.2 Random Forest

The core principle behind a random forest algorithm is to combine many weak learners to yield strong learners. A Random Forest algorithm is a group of decision trees that work

together to classify objects. When a new object must be categorised based on its attributes, each tree in the forest does so individually. The final classification is determined via a democratic process, where the class with the most "votes" from the trees is chosen as the overall classification for the object (Tavasoli, 2023). Using an ensemble of decision trees in a random forest algorithm reduces the chance of overfitting. Furthermore, Random forests are resilient to outliers and perform well with numerical and categorical data, which is common in ASD diagnosis datasets. Like SVM, the random forest algorithm is also used for classification and regression problems. Random forest surpasses the other models with an accuracy of approximately 93%, an AUC score of 0.98, and an average precision of 0.99.

```
Accuracy of RF: 0.9279038718291055
Best Parameters: {'max_depth': 10, 'n_estimators': 100}
              precision    recall  f1-score   support

           0       0.89      0.97      0.93       355
           1       0.97      0.89      0.93       394

    accuracy                           0.93       749
   macro avg       0.93      0.93      0.93       749
weighted avg       0.93      0.93      0.93       749
```

Figure 6: Output of Random Forest Classifier to Predict ASD

3.4.3 Decision Trees

A decision tree classifier is a decision-making tool with a tree structure and a flowchart-like look (Rasul et al., 2024). It primarily focuses on all decisions and possible outcomes. This will be classified both as regression and classification techniques. Also, the tree-like structure of the decision tree makes it simpler to determine which characteristics are most significant. Decision Trees provide an obvious visual representation of decisions, making it simple to see how predictions were formed. Figure 8 shows a decision tree for ASD classification with a maximum branch level of 3 splits. From this figure, it is easy to identify the number of ASD and non-ASD classes. However, the model's accuracy is low, with a maximum depth of 3. After trying different values, the model performed well and yielded an accuracy of 90%, an AUC score of 0.95, and an average precision of 0.94 with a maximum depth of 10.

```
Accuracy of DT: 0.897196261682243
Best Parameters: {'max_depth': 10}
              precision    recall  f1-score   support

           0       0.86      0.94      0.90       355
           1       0.94      0.86      0.90       394

    accuracy                           0.90       749
   macro avg       0.90      0.90      0.90       749
weighted avg       0.90      0.90      0.90       749
```

Figure 7: Output of Decision Tree Classifier to Predict ASD

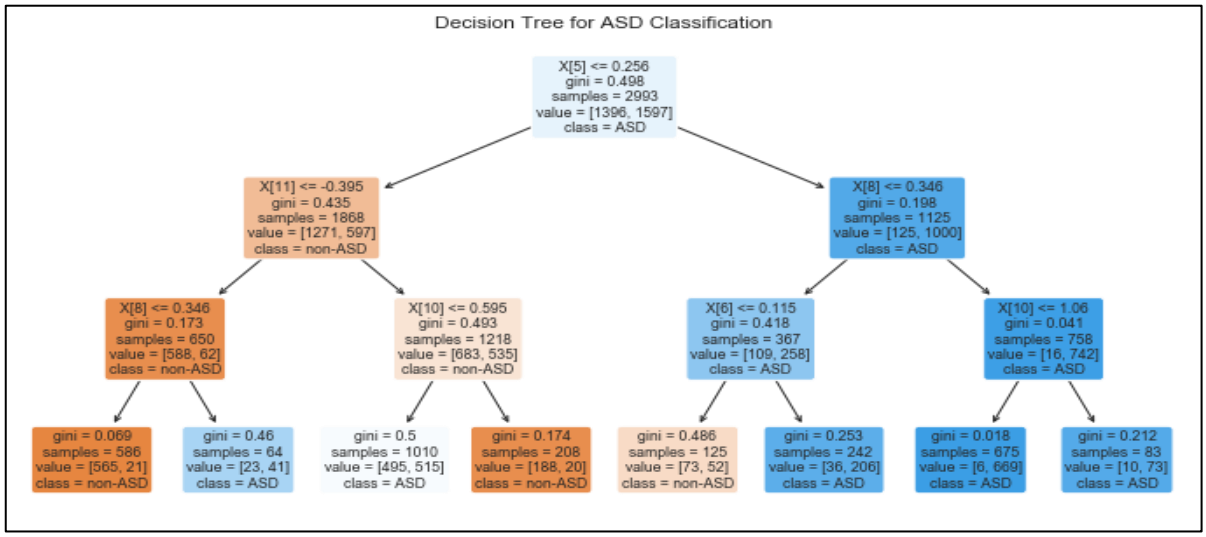

Figure 8: Decision tree for ASD classification with max_depth = 3

3.4.4 K- Nearest Neighbors (KNN)

The k-nearest neighbors (KNN) technique is a simple supervised machine learning approach commonly employed in classification and regression tasks. The KNN algorithm believes that similar things occur close together. In other words, comparable items are located close together (Naik et al., 2023). The input parameter's starting value is the number of classes in the dataset that have a tiny value and are positive integers. The bulk of neighbours are classed as input data. The k-NN method must run numerous times with different K values to find the K that minimises mistakes while maintaining accurate predictions (Mashudi, Ahmad, and Noor, 2021). Because of the stable performance of KNN in clinical studies, it was used in this research study to predict ASD. This study tried different k values to select the right k value to make precise predictions on the unseen data. The other parameter p, calculates the distance between the data points. When the p-value is 1, it uses the Manhattan distance; for p= 2, it uses the Euclidean distance to calculate the distance between points. In this project, there is no difference in using the p-value as either 1 or 2 because the accuracy is almost similar for both values. Figure 9 gives an overview of the output of the KNN algorithm to predict ASD.

```
Accuracy of KNN: 0.8798397863818425
Best Parameters: {'n_neighbors': 9, 'p': 1}
              precision    recall  f1-score   support

           0       0.89      0.86      0.87       355
           1       0.87      0.90      0.89       394

    accuracy                           0.88       749
   macro avg       0.88      0.88      0.88       749
weighted avg       0.88      0.88      0.88       749
```

Figure 9: Output of KNN Classifier to Predict ASD

3.4.5 Logistic Regression

Logistic regression is a supervised learning algorithm used for binary classification which means the target variable contains true and false or yes and no as outputs. It expresses an association between a dependent binary variable and a nominal or ordinary variable (Raj and Masood, 2020). Logistic regression is a statistical method widely used in machine learning for binary classification tasks (Sallibi and Alheeti, 2023). The main reason for using logistic regression in this study is that the dataset's output variable is categorical and has only yes and no as the values. Since logistic regression performs well on binary classification tasks it is suitable for ASD diagnosis to predict the output variable as ASD or non-ASD. In this study, the logistic regression model achieved an accuracy of 85%, an AUC score of 0.89 and an average precision of 0.93.

```
Accuracy of LR: 0.8451268357810414
Best Parameters: {'C': 0.001, 'solver': 'liblinear'}
              precision    recall  f1-score   support

           0       0.81      0.88      0.84       355
           1       0.88      0.81      0.85       394

    accuracy                           0.85       749
   macro avg       0.85      0.85      0.85       749
weighted avg       0.85      0.85      0.85       749
```

Figure 10: Output of Logistic Regression to Predict ASD

3.5 Ensemble Methods

Ensemble methods such as the AdaBoost, Gradient Boosting and Hist Gradient Boosting classifiers were used to compare their performance with single classifiers to predict ASD on large datasets. The AdaBoost classifier is a meta-estimator that begins by training a classification algorithm on the original dataset and then applies more copies of the classifier on the same dataset, adjusting the weights of badly classified instances so that succeeding classifiers focus more on difficult conditions (Mashudi, Ahmad, and Noor, 2021). The Gradient Boosting classifier uses gradient descent to transform several ineffective learners into robust learners by training each new model to minimise the previous model's loss function, such as mean squared error or cross-entropy. The n_estimators parameter was used to specify the number of decision trees to be built in the model. To evaluate the model's performance different values were tried between 100 to 1000, and the model produced the best accuracy at n_estimators value of 500 which is 92%. Hist Gradient Boosting classifier is an advanced version of the gradient boosting classifier with built-in assistance for handling missing values. This classifier works similarly to the gradient-boosting classifier, but it uses a histogram-based binning technique to enhance the efficiency of the learning process. Learning rate and max_iter are the hyperparameters used in this model, learning rate controls the step size taken by the model when updating the weights of its decision trees during the boosting process. The max_iter parameter describes the highest number of iterations that will be built on the Hist Gradient Boosting classifier. Of the three methods, the Hist Gradient

Boosting classifier gave the highest accuracy of 93%, AUC of 0.99 and AP of 0.99 when the hyperparameters learning rate and max_iter were set to 0.05 and 100. Figure 11 shows the output of the Hist Gradient Boosting Classifier along with the classification report. The random forest classifier in the single classifier and the hist gradient boosting classifier in the ensemble methods gave better accuracy in the prediction of ASD.

```
Accuracy of HGBC:0.9292389853137517
              precision    recall  f1-score   support

           0       0.89      0.97      0.93       355
           1       0.97      0.89      0.93       394

    accuracy                           0.93       749
   macro avg       0.93      0.93      0.93       749
weighted avg       0.93      0.93      0.93       749
```

Figure 11: Output of Hist Gradient Boosting Classifier

3.6 Hyperparameter Tuning

In machine learning, hyperparameters are user-defined variables that are crucial in shaping a model's performance, distinct from parameters (Shah, 2024). This study employs GridSearchCV to identify the most effective hyperparameters for enhancing overall performance. GridSearchCV combines grid search with cross-validation, thereby refining the model training process. Within GridSearchCV, hyperparameters are specified in a parameter grid containing different parameters for each model. In the random forest model n_estimators and max_depth were used as hyperparameters, for the KNN classifier n_neighbors and P were used as hyperparameters. Similarly, different hyperparameters are used for the other classification algorithms used in this study. Figure 12 shows the parameter grid used for hyperparameter tuning. GridSearchCV loops over all possibilities in the grid, assessing each model with cross-validation. It chooses the hyperparameter values that yield the best performance measure on the validation set.

```
models = {
    "Logistic Regression": LogisticRegression(solver="liblinear"),
    "SVM": SVC(kernel="rbf"),
    "KNN": KNeighborsClassifier(),
    "Decision Tree": DecisionTreeClassifier(),
    "Random Forest": RandomForestClassifier(),
}

parameter_grid = {
    "Logistic Regression": {"C": [0.01, 0.1, 1, 10]},
    "SVM": {"C": [0.01, 0.1, 1]},
    "KNN": {"n_neighbors": range(1, 21), "p": [1, 2]},
    "Decision Tree": {"max_depth": [4, 8, 10]},
    "Random Forest": {"n_estimators": range(10, 101, 10), "max_depth": [5, 7,10]},
}
```

Figure 12: Parameter grid for hyperparameter tuning

3.7 Chapter Summary

This chapter examined the methods for building a machine-learning model to identify and diagnose autism spectrum disorder (ASD). It involved careful data collecting from Kaggle, thorough preprocessing to manage missing values and outliers, and data standardisation. Label encoding was also used to convert categorical values. The chapter also included classification techniques such as support vector machines (SVM), random forest, KNN, decision trees, and logistic regression. To increase the model's performance, hyperparameters were adjusted. Ensemble techniques, such as AdaBoost, Gradient Boosting, and Hist Gradient Boosting, were also explored, and their performance was compared to individual classifiers. Random Forest and Hist Gradient Boosting are the most accurate models for predicting ASD. This chapter presented a complete approach for creating a machine-learning model to predict ASD and the advantages and disadvantages of the classification methods used for this particular purpose. The Gantt chart given in the chapter depicts the project schedule, which includes the start and end dates of each task specified.

Chapter 4: Results and Discussion

This chapter will explore the findings obtained from data analysis, model building, and evaluation of the model's performance using accuracy, precision, F1-score, and confusion matrix metrics. Also, the ROC curves that were drawn to assess the model's performance will be shown in this section. The answers to the research questions of this study will be provided in the discussion section.

4.1 Importing Libraries and Loading Dataset

The necessary Python libraries such as NumPy, Pandas, Matplotlib, and sci-kit-learn have been imported into this study to load the datasets used for this study, build the machine learning models, and evaluate the performance of the models using performance curves. Figure 13 shows the Python implementation of importing the libraries and loading the dataset. The df. head() command gives the sample of the dataset used in this study before it was pre-processed. Statistics of the autism data have been provided by describe function, including the dataset's mean, standard deviation, and minimum and maximum values. At the same time, the info function will provide information about the dataset, including column names, data types of columns, and the columns' non-null count.

Import necessary libraries

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.model_selection import StratifiedKFold, train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LogisticRegression
from sklearn.svm import SVC
from sklearn.neighbors import KNeighborsClassifier
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.ensemble import AdaBoostClassifier, GradientBoostingClassifier, HistGradientBoostingClassifier
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score
from sklearn.model_selection import GridSearchCV
from tabulate import tabulate
from sklearn.impute import SimpleImputer
from sklearn.metrics import classification_report
from sklearn.metrics import ConfusionMatrixDisplay
from sklearn.metrics import roc_auc_score, roc_curve
from sklearn.metrics import precision_recall_curve, average_precision_score

import warnings
warnings.filterwarnings("ignore")
```

Loading the Data

```
df = pd.read_csv("autism.csv")
df.head()
```

Figure 13: Importing Necessary Libraries and Loading the Dataset

4.2 Data Analysis

To prepare the data for pre-processing the data was analysed to reveal different patterns, and trends and to find the main characteristics of the data. A count plot has been used to visualise the number of males and females, and the number of ASD and non-ASD conditions in the dataset. The count plot and text annotations on the top of each bar help in identifying the pattern of how many data points belong to each category of the specified column. From Figure 14, it can be observed that the dataset has 2549 males and 1193 females. Also, it has been observed from Figure 15 that the dataset has 1991 ASD and 1751 non-ASD cases. A

boxplot was obtained with the help of the seaborn library to determine the outliers present in the Age_Years column.

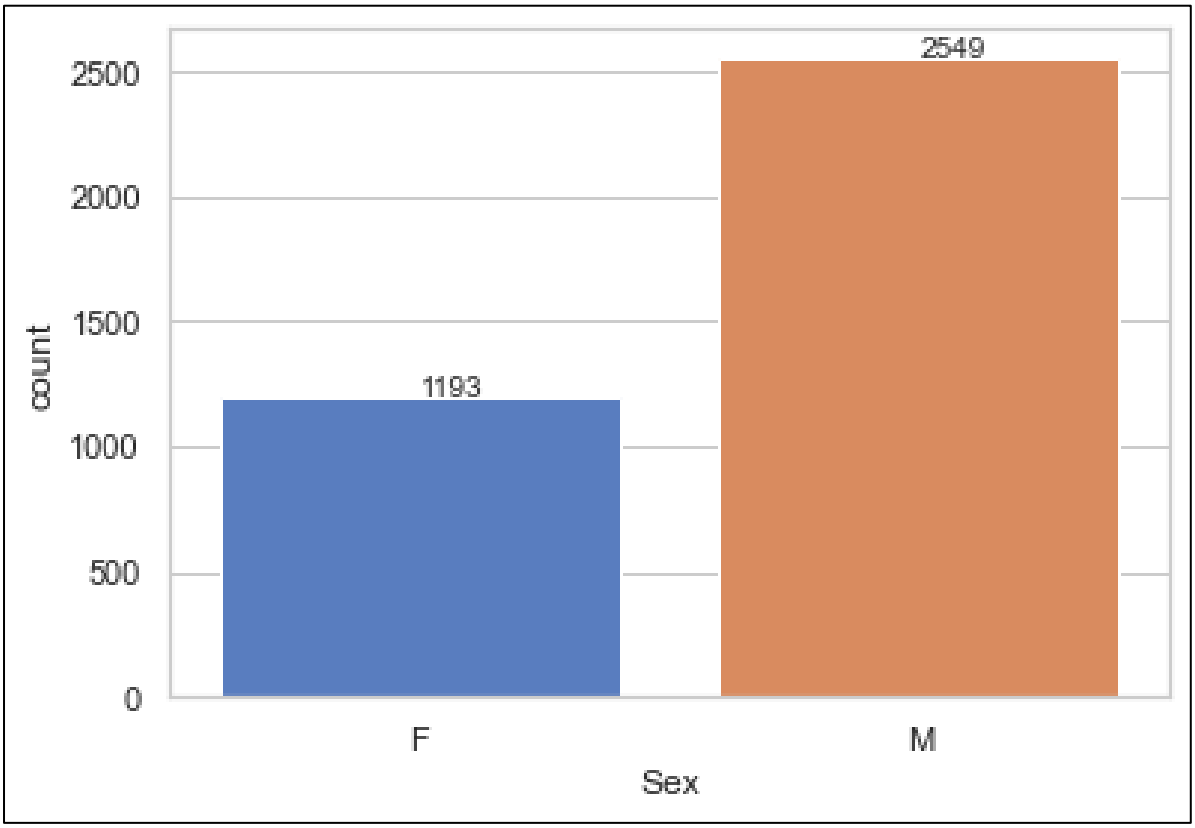

Figure 14: Count plot to visualise the number of males and females in the Data

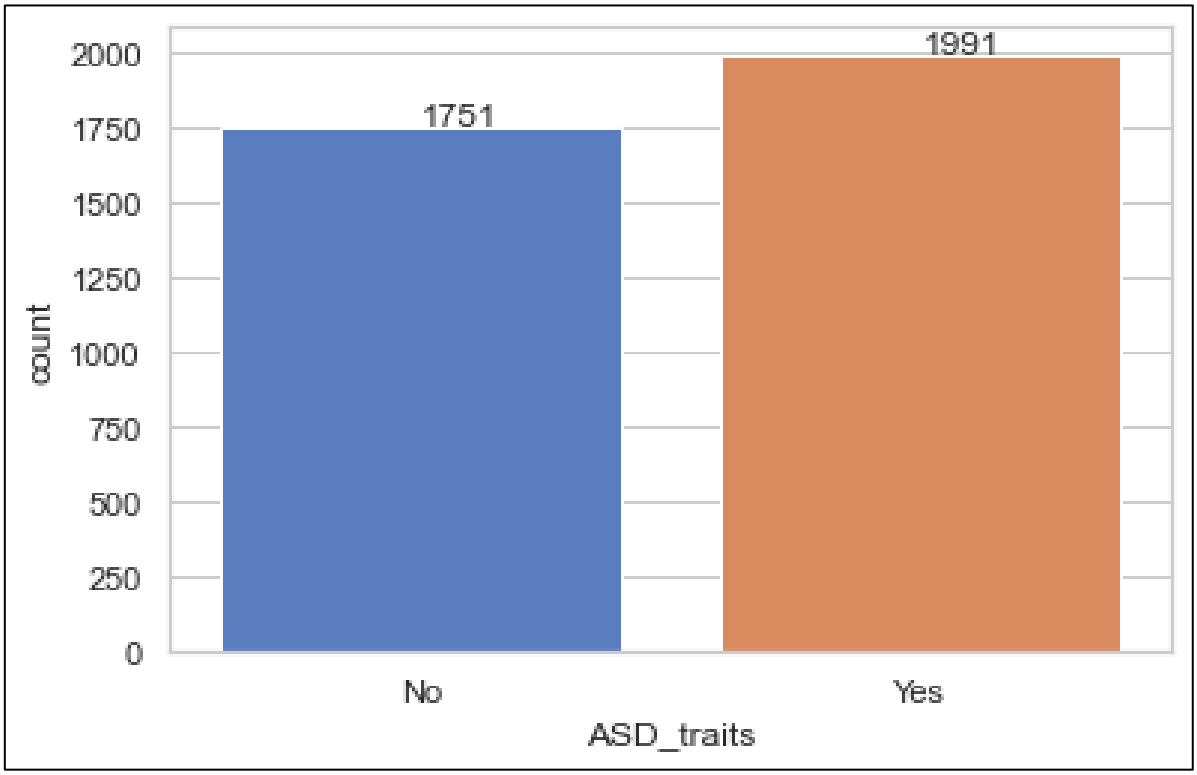

Figure 15: Count plot to visualise ASD and non-ASD instances in the data

Furthermore, the Ethnicity and ASD_traits columns have been visualised to find the number of persons affected by ASD under each ethnicity. From the findings, it has been noticed that White Europeans and Asians are mostly affected by ASD when compared to the other ethnic people in the used dataset. Also, a bar plot has been drawn between the Jaundice and ASD_traits columns to find whether individuals with Jaundice are affected by ASD or not. Finally, a bar plot is used to determine whether males are mostly affected by ASD, or females are affected by ASD. From Figure 16, it is evident that males are mostly affected by ASD compared to females.

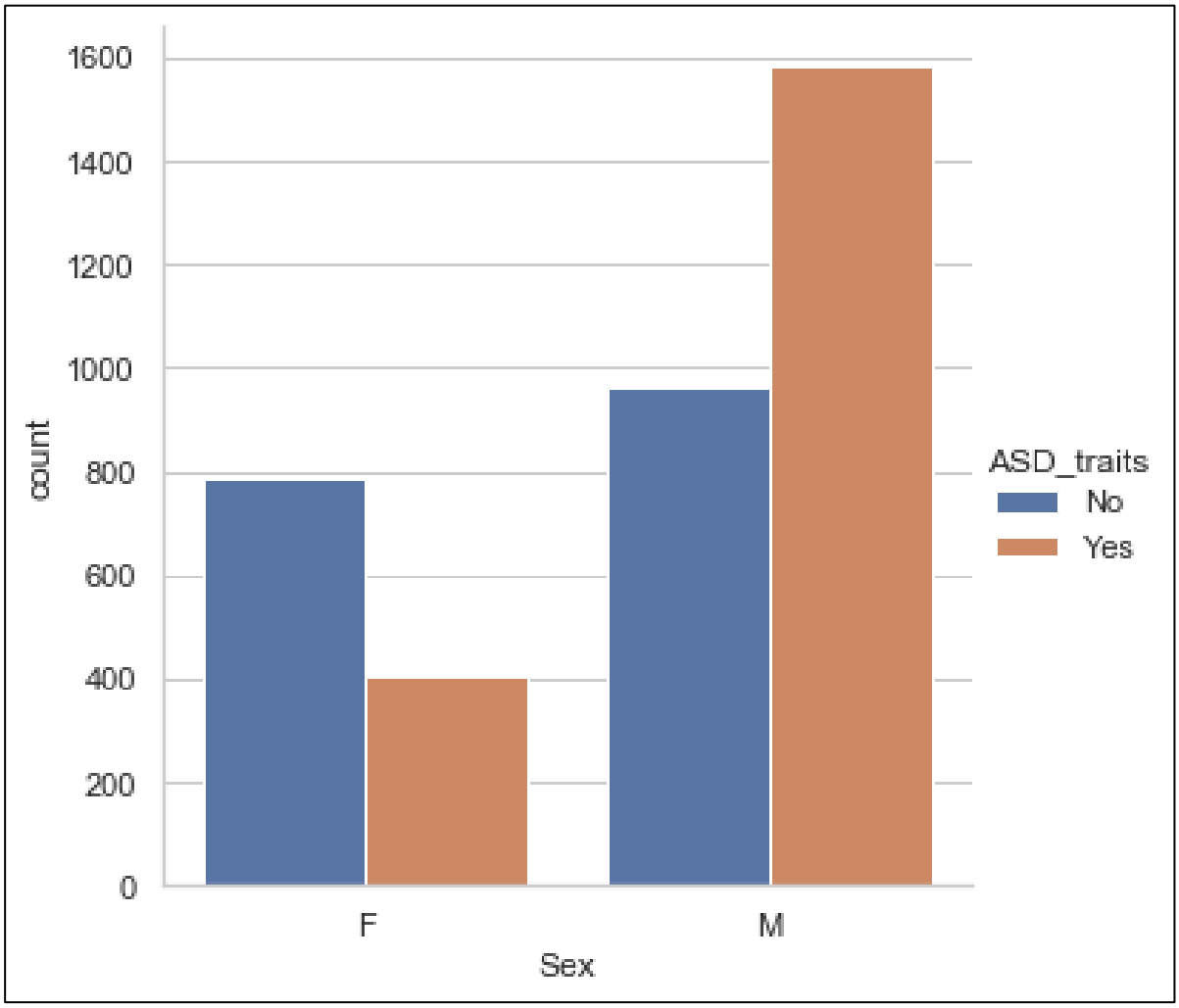

Figure 16: Bar plot to Determine the Prevalence of ASD between Males and Females

4.3 Comparison of Machine Learning Models with Accuracy

In this section, the performance of the machine learning models used in this study has been evaluated using accuracy and the accuracies of these models were compared with the help of a bar plot. The random forest classifier leads the other classifiers with the highest accuracy of 93% and SVM is the second-best model with an accuracy of 92%. The remaining models performed well with an accuracy of 90% for decision tree, 88% for KNN and 85% for logistic regression. The below figure shows the bar plot comparing the accuracy of the models used in this research. The ensembling techniques Ada Boost, Gradient Boosting and Hist Gradient Boosting Classifiers got an accuracy of 84%, 92%, and 93% respectively.

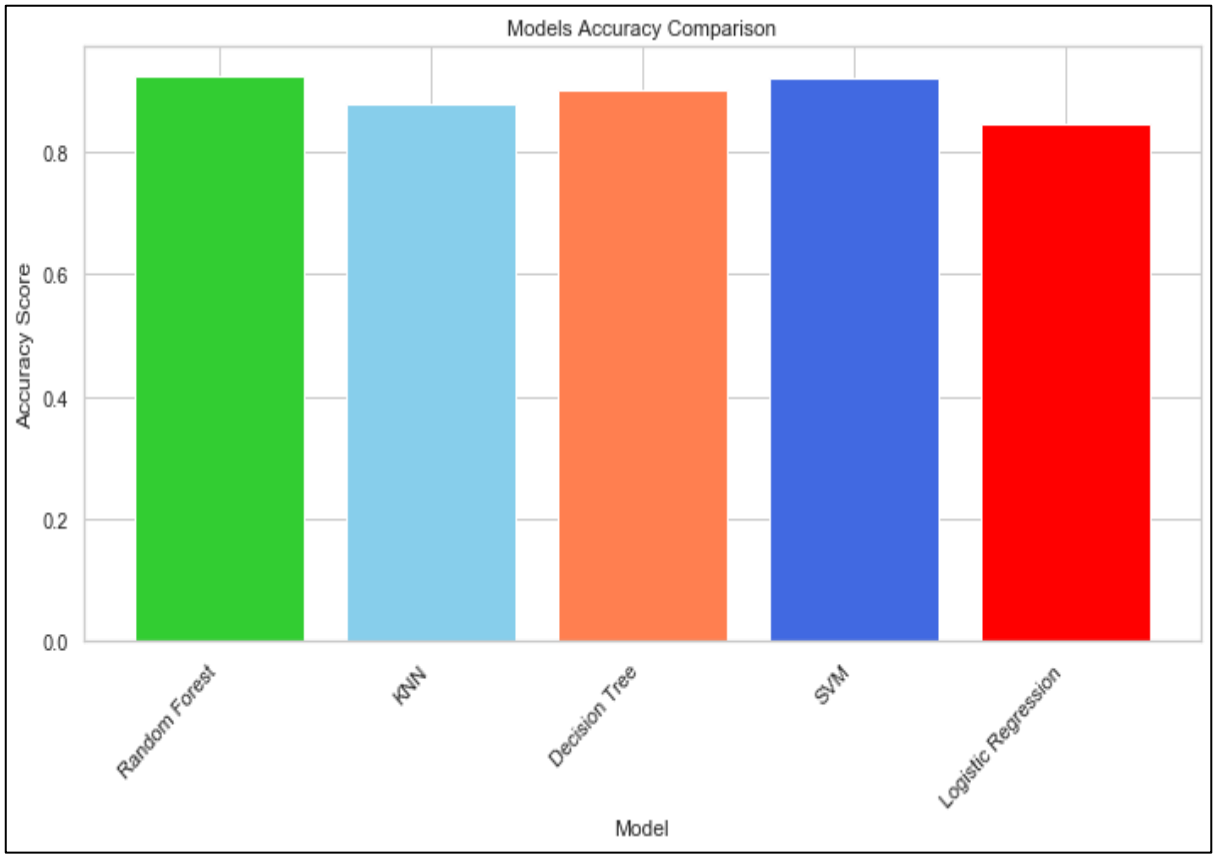

Figure 17: Comparison of Machine Learning Models Accuracy

4.4 Confusion Matrix

A confusion matrix is a significant tool for analysing classification model performance and computing performance measures including accuracy, precision, recall, and F1-score. A 2X2 confusion matrix of 4 values will be used for binary classification problems. The rows of the confusion matrix indicate the true label and the column represents the predicted label of the target variable. The accuracy of a model can be calculated by using the true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) in the confusion matrix. For TP and TN, the true and predicted labels are the same. Whereas for the FP and FN, if the true label is 0, the predicted label is 1 and vice-versa, Also, a false positive is identified as a type I error and a false negative is defined as a type II error. The confusion matrix for all the classification algorithms used in this study has been developed and shown as subplots in Figure 18. These confusion matrices helped calculate the precision and recall theoretically to compare them with practical values. From the figure it has been observed that HistGradientBoostingClassifier has given a greater number of TP and TN.

The accuracy of a classification algorithm is not the only factor defining its performance; it also considers measures such as precision, recall, and F1-score. These metrics can be calculated using TP, TN, FP, and FN values. Furthermore, the true positive rate (tpr), false positive rate (fpr), precision, and recall are important for constructing the ROC-AUC curve and Precision-Recall curve, which will be covered in more detail in this chapter.

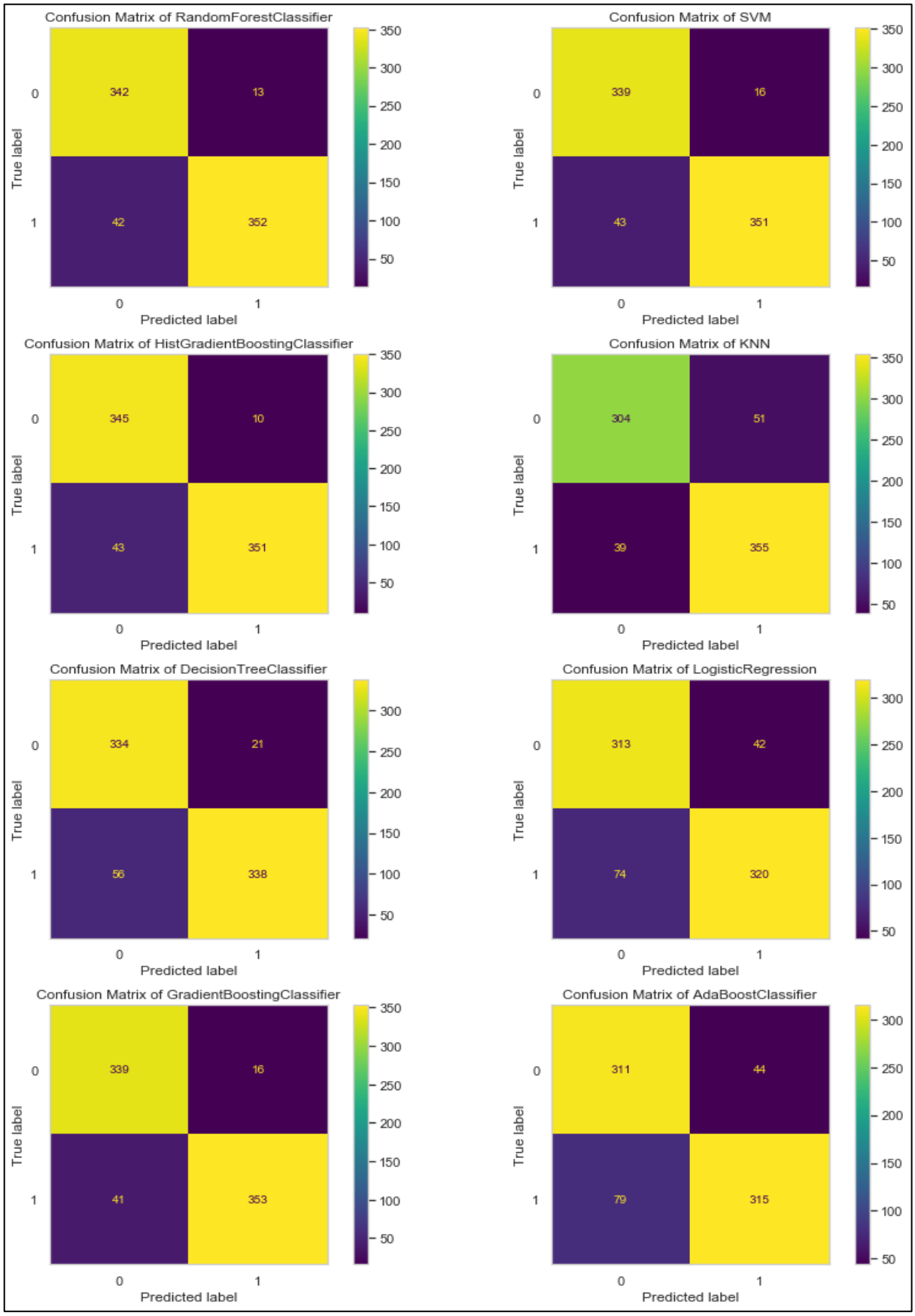

Figure 18: Confusion Matrices

(Source: Self-Originated)

4.5 Correlation Matrix

After Completing the data pre-processing a correlation matrix was developed between the features used for the classification algorithm to find the relation between the input features and target variable ASD_traits. The absolute value of the correlation coefficient represents the strength of the association between two features. From Figure 19, it is evident that weak positive and weak negative correlations are shown in green. In contrast, strong correlations are shown in blue colour.

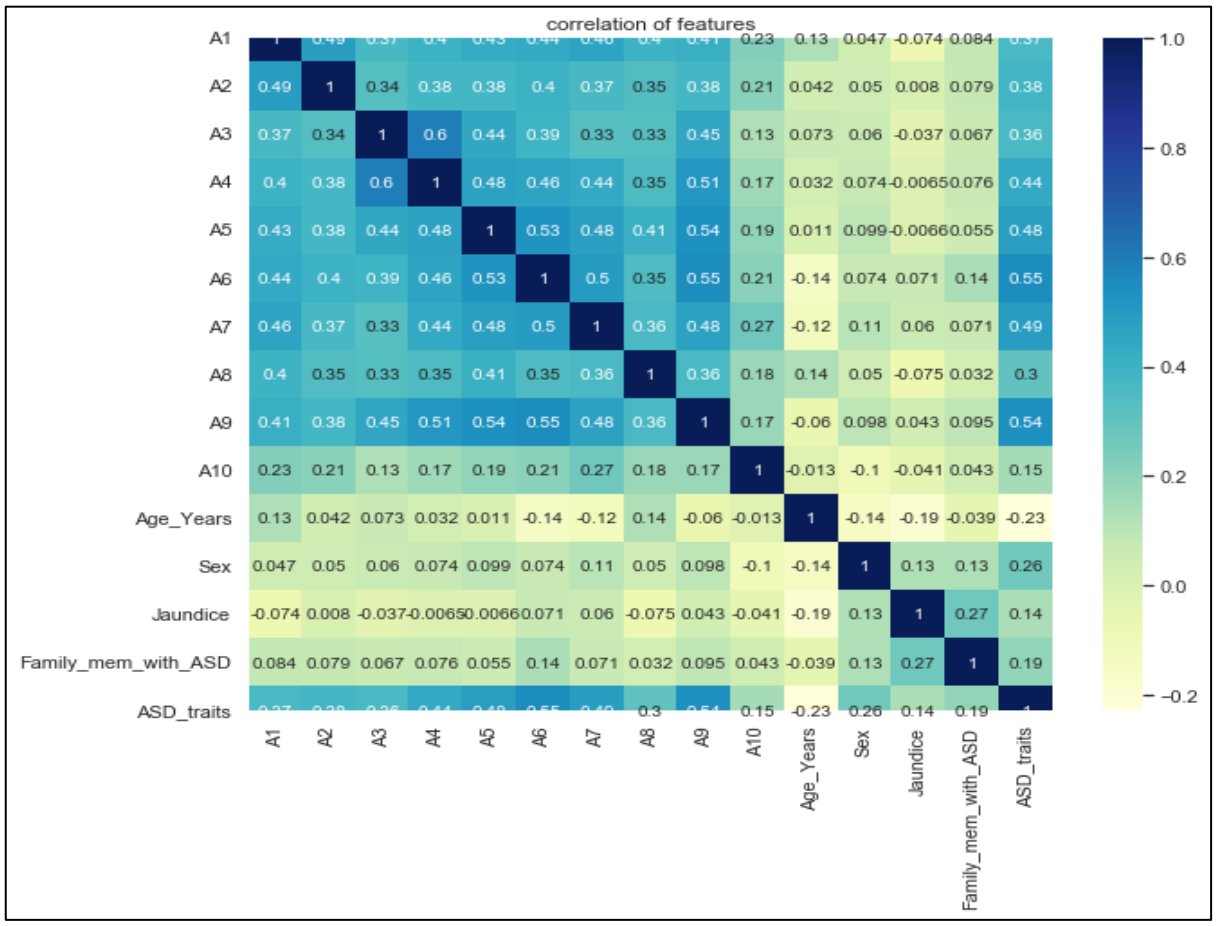

Figure 19: Correlation Matrix Between Input Features and Target Variable

4.6 ROC AUC Curves

A Receiver Operating Characteristic (ROC) curve is a graphical representation that illustrates the effectiveness of classification models at different thresholds. It is a crucial evaluation metric for assessing the performance of any classifier. A ROC curve plots between True Positive Rate (TPR) and False Positive Rate (FPR) at different threshold values for a classification algorithm. AUC is an acronym for "Area Under the ROC Curve." AUC calculates the entire two-dimensional area under the whole ROC curve. The value of AUC varies from 0 to 1. AUC values between 0.7 and 0.8 are acceptable, 0.8 to 0.9 are excellent, and any value beyond 0.9 is exceptional. AUC is a better indicator than accuracy for assessing model performance (Rasul et al., 2024). The greater the AUC value, the more accurate the model can distinguish between patients with and without the disease (Narkhede, 2018). When the value of AUC lies between 0.5 and 1 there is a strong possibility that the

model can differentiate between positive and negative classes. The figure below shows the ROC curves of the classification techniques used in this study and their AUC scores are displayed as a legend in the graph.

Random Forest Classifier is the best model in this study with an AUC score of 0.98 which is near to 1. The AUC score of 0.98 for the random forest classifier indicates that this model can better identify ASD and non-ASD patients effectively. In this work, the other models such as SVM, Decision Tree, KNN and Logistic regression have an AUC of 0.97, 0.95, 0.85, and 0.89. The ensembling methods were used in this study to compare their performance with single classifiers in predicting ASD, and it has been done using ROC-AUC curves. Of the three ensembling techniques, the Hist Gradient Boosting Classifier performed well with an AUC of 0.99. The Ada Boost and Gradient Boosting Classifiers have an AUC of 0.89 and 0.98.

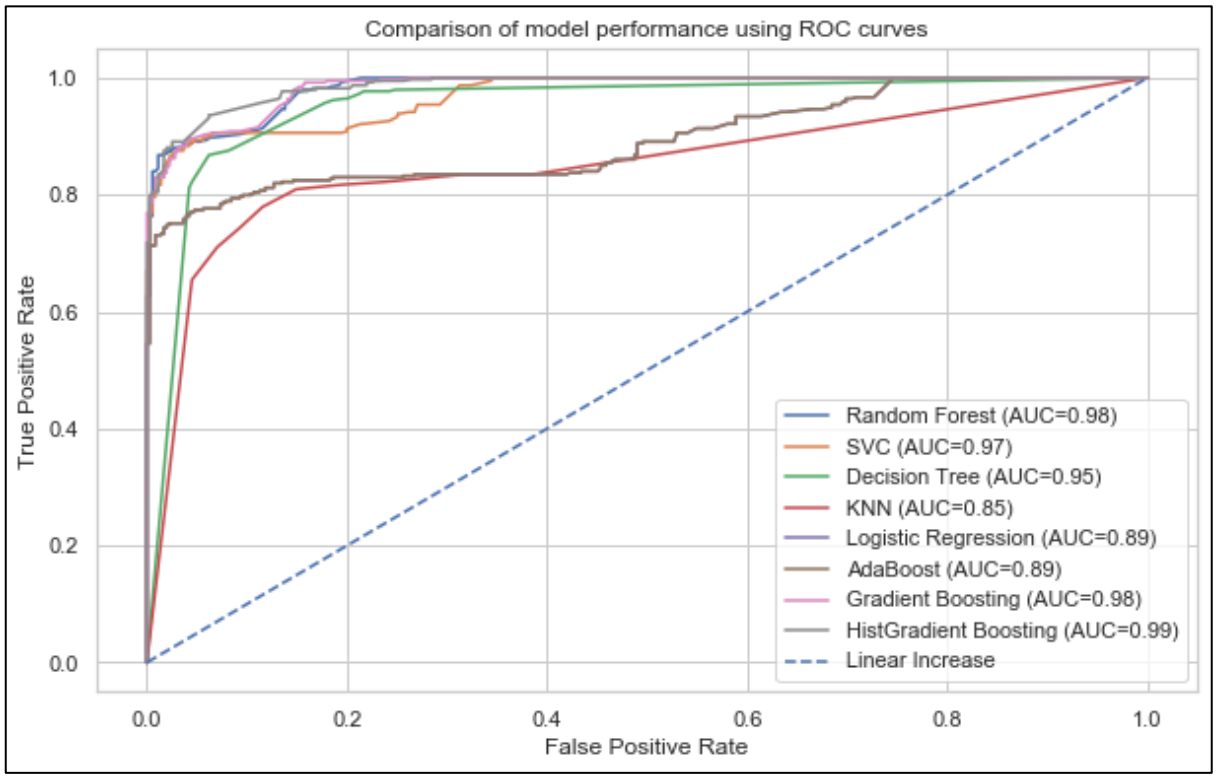

Figure 20: ROC - AUC Curves

(Source: Self-Originated)

4.7 Precision-Recall Curves

A Precision-Recall curve is a graph showing the relation between a classification algorithm's precision and recall at various threshold values. Precision is calculated as the number of true positives separated by the sum of true and false positives. It represents how well a model predicts the positive class. Recall is computed as the number of true positives divided by the total of true positives and false negatives. Sensitivity and recall are equivalent (Brownlee, 2018). When the area under the curve is high it reflects both high recall and high precision, where high precision is related to a low false positive rate and high recall corresponds to a low false negative rate. The precision-recall curve provides greater insights than the ROC

curve when working with imbalanced datasets with a large difference in the proportion of positive to negative classes (Tiwari, 2023). A curve near the top-right corner suggests superior performance, with good precision and recall over an extensive range of thresholds. Figure 21 illustrates the Precision-Recall curves for the machine-learning models employed in this research with average precision presented as a legend in the plot. In this study, the Random Forest classifier is the best model with an average precision (AP) of 0.99. the remaining classification methods SVM, decision tree, logistic regression, and KNN have an AP of 0.98, 0.94, 0.93 and 0.86. In ensembling algorithms, Ada Boost, Gradient Boosting, and Hist Gradient Boosting Classifiers have shown an average precision of 0.93, 0.99, and 0.99. From the figure it has been observed that every model in this study performs better than the baseline.

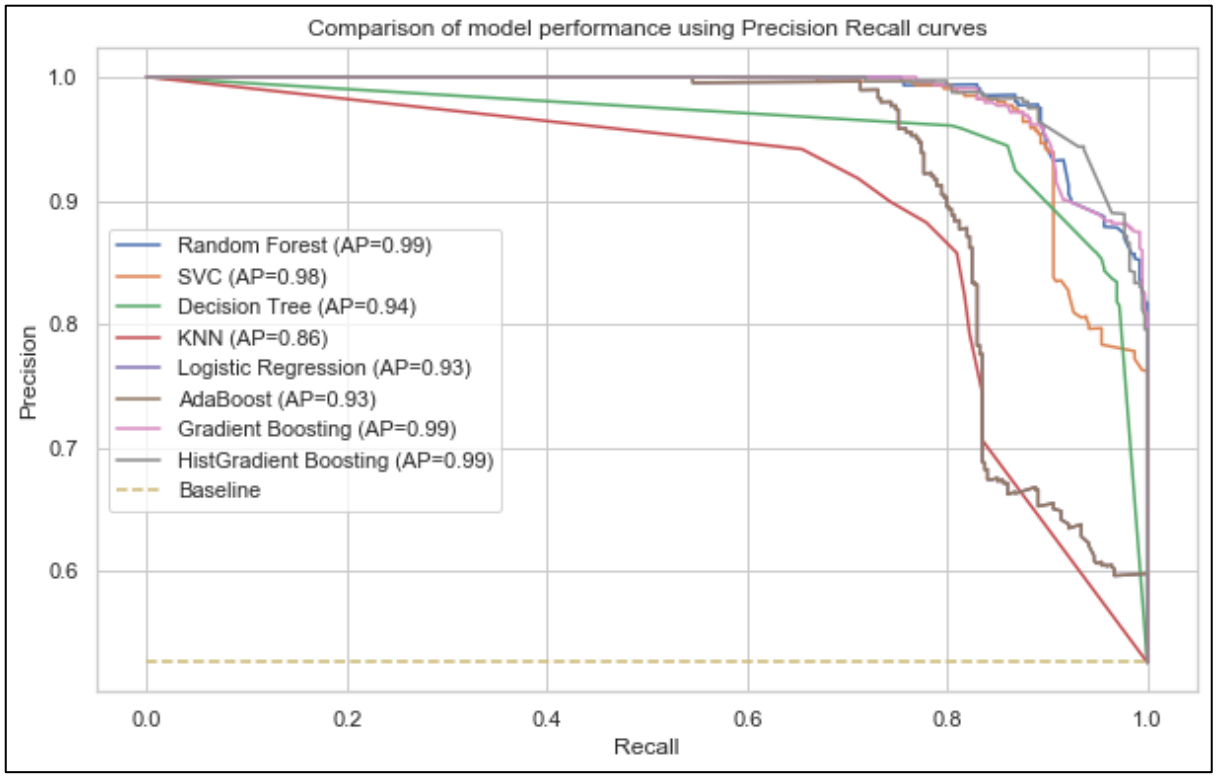

Figure 21: Precision-Recall Curves

(Source: Self-Originated)

4.8 Discussion

The main aim of this project is to build a machine-learning model to predict and diagnose ASD. This study plans to research how accurately these classification techniques will help to predict and diagnose ASD when the models are integrated into clinical practice to improve early identification and early intervention for patients suffering from ASD. Early treatment for ASD will enhance the chances for individuals to interact socially, reduce stress and anger, and increase the sense of well-being. Additionally, this research intends to look into the limitations of using existing datasets for training and assessing machine learning models to predict ASD. These research questions will be answered in detail further in this section. The dataset was pre-processed to handle the missing values, outliers, and features such as

Ethnicity, and Who_completed_the_test that do not show any impact on the target variable, so these variables were removed before being used to train and assess classification models. Removing features that do not affect the target variable dropped the number of features in the dataset from 17 to 15 (Mashudi, Ahmad, and Noor, 2021).

Five different machine-learning models were used in this work to address the research questions and compared to the research studies mentioned in the literature review based on the size of the data set every model used in this study performed well in predicting ASD with an accuracy of more than 85%. Alteneiji, Alqaydi, and Tariq (2020) have applied the Random Forest model to predict ASD in child and adolescent data and got an accuracy of 87% for child data and 88% for teenage data. The Random Forest classifier used in this study has shown an accuracy of 93% after trying different hyperparameter values and cross-fold validations.

The first research question in this study is "Can machine learning models accurately predict and diagnose autism spectrum disorder based on clinical data and how can this technology be effectively integrated into clinical practice to improve early identification and intervention for individuals with `ASD?" The practicality and effectiveness of classification systems in predicting ASD are examined in this question. Regarding the model's performance, this question examines how these machine-learning models might be included in clinical procedures to help with ASD patients' early diagnosis and treatment. From the performance metrics of classification techniques used in this research, it has been observed that machine-learning models can accurately predict and diagnose ASD using clinical data. The ROC-AUC curves and Precision recall curves in Figures 20 and 21 revealed that the Random Forest Classifier and SVM performed well with an AUC of 0.98 and AP of 0.99 for the random forest model, and AUC of 0.97 and AP of 0.98 for SVM. These two models predicted True Positive(ASD) and True Negative (non-ASD) with high precision and recall. The other models used in this study also performed well in predicting ASD. The data patterns analysing ability of machine learning models can identify ASD earlier than traditional methods. This machine-learning technology can be integrated into the clinical process by creating a user-friendly screening tool incorporating the classification model for the initial screening of ASD. The usage of classification models simplifies the initial screening process. It saves more time for doctors, so they can use this time to treat patients who are vulnerable because of ASD.

The second research question of this study is "What are the limitations of using existing datasets for training and evaluating machine learning models for ASD, and how can these limitations be addressed?" This question focused on the drawbacks of using available datasets to develop and assess machine learning models. The size of the dataset, the Imbalanced dataset, missing values in the data, and the presence of outliers are the limitations of using the existing dataset. The usage of data with missing values and outliers will reduce the performance of the model and the model will give biased outputs because of the imbalanced datasets. Using limited-size data will lead the model to overfitting and the generalisability of the model will be affected. Sampling techniques such as under-sampling, and oversampling will reduce the data imbalance problem in existing datasets. Data augmentation will reduce the size of the dataset limitation as this technique intentionally increases the dataset size when

working with limited datasets. Alternatively, employing larger datasets will reduce model overfitting and increase model performance on new data. Cleaning and pre-processing data will remove outliers and reduce missing values, improving data quality and model performance.

4.9 Chapter Summary

This chapter revealed the research results of using machine learning models to predict autism spectrum disorder (ASD). Exploratory data analysis uncovered patterns in the data set, such as the distribution of ASD patients and their gender and ethnicity. The accuracies of the model have been compared using bar plots to compare the performance of each model used in the study. These machine-learning models were also evaluated using ROC-AUC and Precision-Recall curves by showing AUC scores and average precision as legends in the respective curves. The discussion part answered the research questions of this study, which indicates that machine learning models can accurately predict and diagnose ASD using clinical data and can be used in clinical settings for preventive measures. This chapter concluded by discussing the challenges of using existing datasets to assess machine learning models in predicting ASD and how to overcome them.

Chapter 5: Conclusion & Future Work

5.1 Conclusion

This research investigated the feasibility of using machine-learning models in predicting and diagnosing autism spectrum disorder (ASD). This study successfully identified the effectiveness of classification techniques in predicting ASD using clinical datasets. This research revealed that machine learning models can predict and diagnose ASD with high accuracy, with some methods dominating others. The dataset with 3742 instances has been used in this study, this data was analysed and pre-processed before using for the machine-learning algorithms. It has been found that classification algorithms such as the Random Forest Classifier, and SVM showed the highest accuracy of 93% and 92% followed by Decision Tree Classifier with 90%, KNN with 88%, and Logistic regression with 85% accuracy. At present there is no medication to diagnose ASD completely, it can be identified only by the behavioural patterns, social interactivity, and cognitive skills of the individuals. Traditional clinical tests ADOS-R and ADI-R are considered tedious and complicated for patients and practitioners. The machine learning models used in this study will be helpful in the initial screening test for ASD for early intervention and diagnosis of patients and reduce the time taken for screening tests. The model's performance was assessed using evaluation metrics such as accuracy, ROC-AUC curve, and Precision-Recall curve. How these machine-learning methods can be efficiently implemented in medical practice, and the limitations of using existing datasets, including data imbalance, missing values, and dataset size, have been discussed in this research.

5.2 Future Work

This research has laid a solid foundation for examining the usage of machine learning models to predict and diagnose ASD. The future work of this research has many possibilities for improving the use of machine learning models for ASD diagnosis. Using larger and more varied datasets will allow the model to train on more samples and enhance its generalisability. Investigating the use of deep learning models to image data in predicting ASD is another area of future research for this study. Deep learning models are good at recognising patterns in complex data such as images, videos, and audio. Deep learning techniques such as convolution neural networks (CNN) for analysing photos and Recurrent Neural Networks (RNN) for analysing videos may be very effective in predicting and diagnosing ASD as the image data contains the patients' facial expressions, gestures, and eye gaze patterns. These methods can also examine videos of ASD patients' behaviour and cognitive skills. The feedback connections of RNNs make them ideal for speech recognition tasks. Data augmentation is an important approach in the pre-processing of images to increase the training data size by changing the design, shape, and size of images, which helps improve the reliability and performance of deep neural networks. This further research can enhance the application of machine learning models for ASD diagnosis in healthcare settings.

References

Alteneiji, M.R., Alqaydi, L.M. and Tariq, M.U., 2020. Autism spectrum disorder diagnosis using optimal machine learning methods. *International Journal of Advanced Computer Science and Applications*, *11*(9).

Ahmed, I.A., Senan, E.M., Rassem, T.H., Ali, M.A., Shatnawi, H.S.A., Alwazer, S.M. and Alshahrani, M., 2022. Eye tracking-based diagnosis and early detection of autism spectrum disorder using machine learning and deep learning techniques. *Electronics*, *11*(4), p.530.

Alqaysi, M.E., Albahri, A.S. and Hamid, R.A., 2022. Diagnosis-based hybridization of multimedical tests and sociodemographic characteristics of autism spectrum disorder using artificial intelligence and machine learning techniques: a systematic review. *International Journal of Telemedicine and Applications*, *2022*.

Alwidian, J., Elhassan, A. and Ghnemat, R., 2020. Predicting autism spectrum disorder using machine learning technique. *International Journal of Recent Technology and Engineering*, *8*(5), pp.4139-4143.

Brownlee, J., 2018. How to use ROC curves and precision-recall curves for classification in Python. *Machine learning mastery*, *30* (Accessed: 19 July 2024).

Balasubramanian, J., Gururaj, B. and Gayatri, N., 2024. An effective autism spectrum disorder screening method using machine learning classification techniques. *Concurrency and Computation: Practice and Experience*, *36*(2), p.e7898.

Bala, M., Ali, M.H., Satu, M.S., Hasan, K.F. and Moni, M.A., 2022. Efficient machine learning models for early stage detection of autism spectrum disorder. *Algorithms*, *15*(5), p.166.

Bahathiq, R.A., Banjar, H., Bamaga, A.K. and Jarraya, S.K., 2022. Machine learning for autism spectrum disorder diagnosis using structural magnetic resonance imaging: Promising but challenging. *Frontiers in Neuroinformatics*, *16*, p.949926.

Chowdhury, K. and Iraj, M.A., 2020, November. Predicting autism spectrum disorder using machine learning classifiers. In *2020 International conference on recent trends on electronics, information, communication & technology (RTEICT)* (pp. 324-327). IEEE.

Doulah, A.B.M.S.U., Rasheduzzaman, M., Arnob, F.A., Sarker, F., Roy, N., Ullah, M.A. and Mamun, K.A., 2023. Application of Augmented Reality Interventions for Children with Autism Spectrum Disorder (ASD): A Systematic Review. *Computers*, *12*(10), p.215.

Deng, T., 2021, March. Classifying Autism Spectrum Disorder using Machine Learning Models. In *CS & IT Conference Proceedings* (Vol. 11, No. 3). CS & IT Conference Proceedings.

Farooq, M.S., Tehseen, R., Sabir, M. and Atal, Z., 2023. Detection of autism spectrum disorder (ASD) in children and adults using machine learning. *scientific reports*, *13*(1), p.9605.

Farooqui, Q.A. and Rahman, M.A., 2022, February. Autism Spectrum Disorder (ASD) Diagnosis and Reinforcement by Machine Learning and Neural Networks. In *2022 Second International Conference on Artificial Intelligence and Smart Energy (ICAIS)* (pp. 444-451). IEEE.

Hossain, M.D., Kabir, M.A., Anwar, A., and Islam, M.Z., 2021. Detecting autism spectrum disorder using machine learning techniques: An experimental analysis on toddler, child, adolescent, and adult datasets. *Health Information Science and Systems*, *9*, pp.1-13.

Islam, S., Akter, T., Zakir, S., Sabreen, S., and Hossain, M.I., 2020, December. 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)* (pp. 1-6). IEEE.

Karuppasamy, S.G., Muralitharan, D., Gowr, S., Arumugam, S.R., Devi, E.A. and Maharajan, K., 2022, April. Prediction of autism spectrum disorder using convolution neural network. In *2022 6th International Conference on Trends in Electronics and Informatics (ICOEI)* (pp. 1096-1100). IEEE.

Karunakaran, P. and Hamdan, Y.B., 2020. Early prediction of autism spectrum disorder by computational approaches to fMRI analysis with early learning technique. *Journal of Artificial Intelligence*, *2*(04), pp.207-216.

Mashudi, N.A., Ahmad, N. and Noor, N.M., 2021. Classification of adult autistic spectrum disorder using machine learning approach. *IAES International Journal of Artificial Intelligence*, *10*(3), p.743.

Manoj, M., and Praveen, J.I., 2023, August. A Hybrid Approach to Support the Detection of Autism Spectrum Disorder (ASD) through Machine Learning and Deep Learning Techniques. In *2023 12th International Conference on Advanced Computing (ICoAC)* (pp. 1-7). IEEE.

Naik, S.K.R., Deepa, M., Ruhi, P.B., Prakash, S. and Royal, U.J., 2023, June. Determination and Diagnosis of Autism Spectrum Disorder using Efficient Machine Learning Algorithm. In *2023 3rd International Conference on Intelligent Technologies (CONIT)* (pp. 1-5). IEEE.

Narkhede, S., 2018. Understanding AUC-roc curve. *Towards data science*, *26*(1), pp.220-227.

Raj, S. and Masood, S., 2020. Analysis and detection of autism spectrum disorder using machine learning techniques. *Procedia Computer Science*, *167*, pp.994-1004.

Rasul, R.A., Saha, P., Bala, D., Karim, S.R.U., Abdullah, M.I. and Saha, B., 2024. An evaluation of machine learning approaches for early diagnosis of autism spectrum disorder. *Healthcare Analytics*, *5*, p.100293.

Sallibi, A.D. and Alheeti, K.M.A., 2023, September. Detection of Autism Spectrum Disorder in Children Using Efficient Pre-Processing and Machine Learning Techniques. In *2023*

International Conference on Decision Aid Sciences and Applications (DASA) (pp. 505-514). IEEE.

Saha, A., Barua, D., Mohib, Z. and Choya, S.B.Z., 2021. *Development of an Interactive Dashboard for Analyzing Autism Spectrum Disorder (ASD) Data using Machine Learning Techniques* (Doctoral dissertation).

Shah, R., 2024. Tune Hyperparameters with GridSearchCV URL: https://www.analyticsvidhya.com/blog/2021/06/tune-hyperparameters-with-gridsearchcv/ (Accessed: 16 July 2024).

Thabtah, F., 2019. An accessible and efficient autism screening method for behavioural data and predictive analyses. *Health informatics journal*, *25*(4), pp.1739-1755.

Tiwari, R., 2023. 'Advanced evaluation metrics for imbalanced classification models' Medium Blog, 9 February. Available at: https://medium.com/cuenex/advanced-evaluation-metrics-for-imbalanced-classification-models-ee6f248c90ca/. (Accessed: 19 July 2024).

Tavasoli, S., 2023. Top 10 Machine Learning Algorithms For Beginners: Supervised, and More. Available at: https://www.simplilearn.com/10-algorithms-machine-learning-engineers-need-to-know-article/ (Accessed: 09 July 2024).

Vakadkar, K., Purkayastha, D. and Krishnan, D., 2021. Detection of autism spectrum disorder in children using machine learning techniques. *SN computer science*, *2*, pp.1-9.

Zhou, Y., Yu, F. and Duong, T., 2014. Multiparametric MRI characterization and prediction in autism spectrum disorder using graph theory and machine learning. *PloS one*, *9*(6), p.e90405.

Appendices

The data for Figure 2 on the increase in the prevalence of ASD over the years (2000 – 2020) has been obtained from the CDC website (https://www.cdc.gov/autism/data-research/index.html) and the below table shows the data used for that figure.

Years	Prevalence Rate of ASD
2000	1 in 150
2002	1 in 150
2004	1 in 125
2006	1 in 110
2008	1 in 88
2010	1 in 68
2012	1 in 69
2014	1 in 59
2016	1 in 54
2018	1 in 44
2020	1 in 36

Using this data the line chart has been drawn with Python code in jupyter notebook.

```
import matplotlib.pyplot as plt

years = ["2000", "2002", "2004", "2006", "2008", "2010", "2012", "2014", "2016", "2018", "2020"]

ratios = [1/150, 1/150, 1/125, 1/110, 1/88, 1/68, 1/69, 1/59, 1/54, 1/44, 1/36]

plt.figure(figsize=(10, 6))

plt.plot(years, ratios, marker='o', linestyle='-')

plt.xlabel("Year")

plt.ylabel("Prevalence (Ratio)")

plt.title("Rise of ASD Prevalence Over Time")

plt.xticks(rotation=60)

plt.grid(True)

plt.tight_layout()
```

```
plt.show()
```

The coding part for the autism.csv data to predict and diagnose ASD using machine learning models is submitted as a text file. This text file consists of code for data loading, data cleaning, model building, model evaluation using different metrics and loading the model in a pickle file. To run this code Jupyter Notebook needs to be installed with new versions of libraries.