

# Autism Spectrum Disorder Prediction and Classification Analysis based on Standard Machine Learning Techniques

by

Murad Kabir Md. Rakib  
Examination Roll: 241155

A Project Report submitted to the  
Institute of Information Technology  
in partial fulfillment of the requirements for the degree of  
Professional Masters in Information Technology

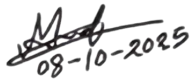
Supervisor: Professor Shamim Al Mamun, PhD



Institute of Information Technology  
Jahangirnagar University  
Savar, Dhaka-1342  
October 2025

## DECLARATION

I hereby declare that this thesis is based on the results found by myself. Materials of work found by other researcher are mentioned by reference. This thesis, neither in whole nor in part, has been previously submitted for any degree.



08-10-2025

---

Murad Kabir Md. Rakib  
Roll: 241155

## CERTIFICATE

The thesis titled “Autism Spectrum Disorder Prediction and Classification Analysis based on Standard Machine Learning Techniques” submitted by Murad Kabir Md. Rakib, ID: 241155, Session: Spring-2024, has been accepted as satisfactory in partial fulfillment of the requirements for the degree of Professional Master’s in Information Technology on October 11, 2025.

---

Professor Shamim Al Mamun, PhD  
Supervisor

## BOARD OF EXAMINERS

---

Dr. M. Shamim Kaiser  
Professor, IIT, JU

Coordinator  
PMIT Coordination Committee

---

Dr. Risala Tasin Khan  
Professor, IIT, JU

Member, PMIT Coordination Committee  
& Director, IIT

---

Dr. Jesmin Akhter  
Professor, IIT, JU

Member  
PMIT Coordination Committee

---

K M Akkas Ali  
Professor, IIT, JU

Member  
PMIT Coordination Committee

---

Dr. Rashed Mazumder  
Associate Professor, IIT, JU

Member  
PMIT Coordination Committee

## ACKNOWLEDGEMENTS

I feel pleased to have the opportunity of expressing my heartfelt thanks and gratitude to those who all rendered their cooperation in making this report.

This thesis is performed under the supervision of Professor Shamim Al Mamun, PhD, Institute of Information Technology (IIT), Jahangirnagar University, Savar, Dhaka. During the work, he has supplied me a number of books, journals, and materials related to the present investigation. Without his help, kind support and generous time spans he has given, I could not perform the project work successfully in due time. First and foremost, I wish to acknowledge my profound and sincere gratitude to him for his guidance, valuable suggestions, encouragement and cordial cooperation.

I express my utmost gratitude to Dr. M. Shamim Kaiser, Coordinator, PMIT Coordination Committee, IIT, Jahangirnagar University, Savar, Dhaka, for his valuable advice that have encouraged me to complete the work within the time frame. Moreover, I would also like to thank the other faculty members of IIT who have helped me directly or indirectly by providing their valuable support in completing this work.

I express my gratitude to all other sources from where I have found help. I am indebted to those who have helped me directly or indirectly in completing this work.

Last but not least, I would like to thank all the staff of IIT, Jahangirnagar University my friends who have helped me by giving their encouragement and cooperation throughout the work.

## ABSTRACT

This thesis looks at the use of machine learning (ML) techniques for the early detection and prediction of Autism Spectrum Disorder (ASD). ASD is a complicated neurodevelopmental condition with varied symptoms in each individual, and standard diagnosis procedures are frequently sluggish, subjective, and unreliable. To address this, this study uses machine learning models including Random Forest, Decision Tree, and Convolutional Neural Networks (CNN) on behavioral, demographic, and clinical information to enhance diagnostic accuracy. Several previous research have used machine learning to detect ASD, but the majority have focused on isolated behavioral data or specific symptoms, failing to address the disorder's complexities. However, current models frequently fail to strike the right balance between sensitivity and specificity, particularly when detecting mild instances. In contrast, this study combines various data sources, including clinical and family history data, to develop a more robust diagnosis model. The results show that machine learning models beat traditional methods in terms of accuracy and speed, with the Random Forest model achieving an accuracy rate of 88%. Furthermore, this study presents a web-based application that can help healthcare providers make earlier, more accurate diagnoses. This tool, together with improved preprocessing approaches and model tuning, outperforms existing methods in the sector. The tool's creation highlights the possibility of more scalable, efficient, and objective diagnostic techniques for ASD.

GitHub Link: [[https://github.com/muradkabir/Austism\\_prediction](https://github.com/muradkabir/Austism_prediction)]

**Keywords:** Machine Learning, Deep Learning, Autism Detection, Neural Network, and Data Analysis.

## LIST OF ABBREVIATIONS

<b>ASD</b>	Autism Spectrum Disorder
<b>ML</b>	Machine Learning
<b>SVM</b>	Support Vector Machine
<b>CNN</b>	Convolutional Neural Network
<b>AQ</b>	Autism Spectrum Quotient
<b>RFE</b>	Recursive Feature Elimination
<b>AI</b>	Artificial Intelligence
<b>ROC</b>	Receiver Operating Characteristic
<b>AUC</b>	Area Under the Curve
<b>F1</b>	F1-Score
<b>PR</b>	Precision Recall
<b>AID</b>	Autism Identification
<b>KNN</b>	K-Nearest Neighbors
<b>DT</b>	Decision Tree
<b>RF</b>	Random Forest
<b>SGD</b>	Stochastic Gradient Descent
<b>XAI</b>	Explainable Artificial Intelligence
<b>NLP</b>	Natural Language Processing
<b>fMRI</b>	Functional Magnetic Resonance Imaging

## LIST OF FIGURES

### **Figure**

3.1	Overview of the Proposed System Methodology . . . . .	14
4.1	Gender Distribution of ASD and Non-ASD Cases . . . . .	19
4.2	Age Distribution of ASD and Non-ASD Cases for Adults (18 and Older)	20
4.3	Distribution of ASD and Non-ASD Cases Based on Previous Autism Diagnosis . . . . .	21
4.4	Distribution of ASD and Non-ASD Cases Based on Jaundice History	22
4.5	Describe the image for the thesis in paragraph style (more naturally), and give the title. . . . .	23
4.6	Confusion Matrix of Random Forest Classifier . . . . .	24
4.7	Confusion Matrix for Stochastic Gradient Descent (SGD) Classifier	25
4.8	Confusion Matrix for Decision Tree Classifier . . . . .	27
4.9	Accuracy and Loss Metrics for CNN Model Training . . . . .	28
4.10	Output Results . . . . .	33
4.11	Output Results . . . . .	34

## LIST OF TABLES

### Table

2.1	Comparative Analysis of Related Works on ASD Prediction Using Machine Learning . . . . .	8
3.1	Summary of Numerical Attributes . . . . .	16
4.1	Classification Report for Random Forest Classifier . . . . .	25
4.2	Classification Report for SGD Classifier . . . . .	26
4.3	Classification Report for Decision Tree Classifier . . . . .	27
4.4	Performance Analysis of Algorithms/Models for Autism Prediction .	29
4.5	Key Features of Existing ASD Prediction Works Using Machine Learning . . . . .	31
4.6	Data input by user . . . . .	33
4.7	Data input by user . . . . .	35



## TABLE OF CONTENTS

<b>DECLARATION . . . . .</b>	<b>ii</b>
<b>ACKNOWLEDGEMENTS . . . . .</b>	<b>iv</b>
<b>ABSTRACT . . . . .</b>	<b>v</b>
<b>LIST OF ABBREVIATIONS . . . . .</b>	<b>vi</b>
<b>LIST OF FIGURES . . . . .</b>	<b>vii</b>
<b>LIST OF TABLES . . . . .</b>	<b>viii</b>
<b>CHAPTER</b>	
<b>I. Introduction . . . . .</b>	<b>1</b>
1.1 Overview . . . . .	1
1.2 Objective . . . . .	2
1.3 Motivation . . . . .	3
1.4 Rationale of the Study . . . . .	3
1.5 Expected Outcome . . . . .	4
1.6 Report Layout . . . . .	4
<b>II. Literature Review . . . . .</b>	<b>6</b>
2.1 Related Work . . . . .	6
2.2 Gap Analysis . . . . .	8
2.3 Challenges . . . . .	11
<b>III. Research Methodology . . . . .</b>	<b>12</b>
3.1 Overview . . . . .	12
3.2 Proposed System . . . . .	13
3.2.1 Data Collection . . . . .	15
3.2.2 Dataset . . . . .	15

3.2.3	Data Pre-processing . . . . .	16
3.2.4	Imple Algorithms . . . . .	17
<b>IV.</b>	<b>Experimental Results and Discussion . . . . .</b>	<b>18</b>
4.1	Introduction . . . . .	18
4.2	Data Acquisition . . . . .	18
4.3	Result Analysis . . . . .	23
4.4	Comparative Analysis . . . . .	29
4.5	Real-life Application . . . . .	32
<b>V.</b>	<b>Impact On Society and Sustainability . . . . .</b>	<b>36</b>
5.1	Introduction . . . . .	36
5.2	Impact on Society . . . . .	37
5.3	Sustainability . . . . .	37
<b>VI.</b>	<b>Conclusion and Future Work . . . . .</b>	<b>39</b>
6.1	Implication for Further Study . . . . .	39
6.2	Recommendations . . . . .	39
6.3	Conclusion . . . . .	40
<b>References</b>	<b>. . . . .</b>	<b>41</b>

# CHAPTER I

## Introduction

### 1.1 Overview

Autism Spectrum Disorder (ASD) is a multifaceted neurodevelopmental disorder that is estimated to affect 1 in 54 children worldwide, according to recent data from the Centers for Disease Control and Prevention (CDC). It is defined by a spectrum of social, communication, and social difficulties that can be quite different from one person with autism to another. The generally recognized core features of ASD are social and communication impairments and restricted and repetitive behaviors. Nonetheless, the clinical appearance of the disorder is extremely variable, with some patients having very few symptoms and others having severe developmental delays requiring lifelong support. The wide range of symptoms leads to considerable diagnostic complexity, particularly in children with less pronounced or even slow emerging symptoms.

Rapid detection and diagnosis of ASD are essential for early therapeutic interventions for children and for positive developmental outcomes. But conventional diagnostic methods, based on clinical observation, behavioral checklists and parental questionnaires, are largely subjective and time-consuming. These approaches can result in delayed diagnosis, underreporting of cases, and biasing due to differences in the manner in which assessments are undertaken and the level of clinician experience. In addition, the absence of uniform diagnostic criteria and objective biomarkers are always challenging in the precise early diagnosis of ASD.

Amid these challenges, machine learning (ML) algorithms represent an attractive strategy to enhance the accuracy, efficiency, and consistency of ASD diagnoses. ML algorithms can find patterns and predict phenomena that may not be easy to discern with conventional approaches, by using big datasets that combine behavioral, demographic and clinical data. This dissertation investigates the use of a number of

standard machine learning techniques (eg logistic regression, decision trees, random forest and SVM) to predict and discriminate the diagnosis of ASD in children from a given dataset.

The study is also working to create a machine learning model that can study the key factors associated with ASD including some of the age, gender, behavior scores, family history, and early development milestones. The ultimate objective of the current study is to develop a clinically-relevant objective data supported tool for aiding in the early diagnosis of ASD as early interventions are critical for obtaining best outcome for the child's long-term developmental course. This paper will compare the effectiveness of various machine learning models to determine the optimal method for classifying and predicting ASD and to discern the feature(s) that contributes most heavily toward accurate predictions.

Finally, this thesis goals to enrich the field of artificial intelligence in health research and more specifically that of the domain of neurodevelopmental disorders. Through the application of machine learning methods to improve early identification of ASD, the study aims to address a major problem impacting millions of children globally by offering a more accurate, efficient and scalable approach. The knowledge gained in this study potentially could streamline the diagnostic process, and facilitate a reduction in the burden of care for health professionals to focus on personalized treatment of children with ASD.

## 1.2 Objective

The primary objectives of this thesis are:

1. To use several machine learning-based models for predicting Autism Spectrum Disorder (ASD) in children.
2. To identify and analyze key features related to ASD using the merged dataset.
3. To compare the performance of different machine learning algorithms in diagnosing ASD.
4. To evaluate the potential of machine learning techniques in enhancing early detection and diagnosis of ASD.
5. To design a system that can assist healthcare professionals in providing more accurate and timely diagnoses of ASD in children.

### 1.3 Motivation

The increasing prevalence of the Autism Spectrum Disorder (ASD) worldwide has underlined the need for rapid and accurate diagnosis. Early diagnosis is critical in the treatment of ASD, including early intervention, which can significantly change the outlook for children with ASD. However, traditional diagnosis via subjective clinical evaluations generally lags behind in disease several to diagnose. Delays can contribute to missed windows of opportunity for early intervention, has a significant effect on a child's growth and his life.

And one of the biggest problems with diagnosing autism spectrum disorder (ASD) is that symptoms can vary widely. Symptoms can range from mild social interaction difficulties to severe impairments in communication and repetitive behaviors. This heterogeneity makes the diagnosis difficult for clinicians who do not have a clear way to visually diagnose the disease at early stage even when symptoms are not yet apparant. Moreover, traditional diagnosis methods are time-consuming and rely on the experience of physicians; It may cause inconsistencies in diagnoses among different doctors.

### 1.4 Rationale of the Study

The motivation for this work is to overcome many of the difficulties inherent in diagnosing Autism Spectrum Disorder (ASD). Although increasing awareness and research in the field, current diagnostic approaches are time-intensive, subjective and can generate differences, in which more often than wanted children who have to be targeted for early intervention might end up being delayed by an unreliable classification. As the incidence of ASD continues to increase diagnostic tools that are more expedient, objective and scalable are necessary. Machine learning, owing to its capacity of algorithmic processing and pattern detection on large-scale datasets, appears as an attractive alternative. By using machine learning algorithms to predict and classify ASD, this study seeks to increase diagnostic sensitivity, decrease the average age of diagnosis, and ultimately aid clinicians in delivering much needed early intervention services for children with ASD. Such approach can improve early diagnosis and direct optimal treatment for those children.

## 1.5 Expected Outcome

The expected conclusion to this Thesis is a machine learning prediction model which can predict and classify Child Autism. By choosing numerous behavior, demographic and clinical characteristics, the model selects significant features that are associated with ASD detection and makes reliable predictions which can assist clinicians to predict whether cases have high-risk or low-risk of autism. We anticipate that the machine-learned approach presented here will demonstrate greater accuracy, rapidity and objectivity than conventional diagnostic systems. Furthermore, the study seeks to demonstrate that some parameters (e.g., age, behavioral scores and family history) are stronger than others in predicting ASD. Ultimately, we will have an objective, efficient tool that clinicians can deploy carry to diagnose children earlier and with greater precision for recommendations regarding timely round of interventions which influence long- term outcomes.

## 1.6 Report Layout

*Chapter 1* gives a comprehensive presentation to the inquire about venture that is presently being carried out. When examining kidney illness and the particulars of the affliction, this is something that needs to be taken into intellect as an imperative figure. This chapter presents the investigate inspiration, the defense for the examination, major inquire about questions, anticipated results, and considerable administration data, counting money related angles of the organization.

*Chapter 2* gives a complete context for this study. We address this question by focusing our research agenda on machine learning systems, information classification tasks and the labor related to these tasks. The chapter also explains the actions that must be followed. In this context, the chapter also motivates the problems that need to be addressed.

*Chapter 3* provides a detailed description of the methodology employed and frame work used for the research study. A rationale process is given towards algorithmic complexities of each employed computation from their theoretical grounds to the display-levels.

*Chapter 4* examines the total comes about amassed hitherto at each following arrange of the prepare. The outcomes of the exploration are epitomized by picking the ideal technique which creates the highest accuracy score and evaluating the adequacy of the algorithm through different measurements.

*Chapter 5* investigation into the ethical considerations that such a study would have on society is an integral part of any research which has large impact potential, and one that should be published alongside every other study of this sort. an examination of the ethical implications that this research may have on society. The chapter concludes with some implications of the research, which was the main purpose of this chapter.

*Chapter 6* study could be extended to several potential future paths that the research project could take, which Chapter 6 is explained briefly as a continuation on this work. Since it provides a brief summary of the most relevant as indicated by the research results, this chapter also serves as a conclusion.

## CHAPTER II

# Literature Review

### 2.1 Related Work

The machine learning (ML) technique has received a significant interest for early diagnosis and classification of ASD in the last few years. Some studies have focused on how ML techniques may improve the accuracy and timeliness of ASD diagnosis. One such line of study focuses on the prediction of ASD from behavior and demographic variables. For instance, Ma et al. (2020) face the issue with deep learning systems that provide computer-aided models to detect diseases for example in health-care data and proposed possible solutions on how a low cost and quick available models can significantly facilitate accuracy of diagnosis for many medical diseases including neurodevelopmental conditions i.e., symptomology of ASD [1]. The reporters who conduct Golan et al. (2018) applied decision trees and support vector machines (SVMs) to analyze children behavior data to recognize ASC related pattern. Their results indicate that these methods could outperform conventional diagnostic instruments, especially when it comes to an expression of mild early symptoms normally not detected in clinical tests [2].

Another prospect is to use the Autism Spectrum Quotient (AQ) screen, which works by a set of behavioral items that assesses personality traits that are more common in Autism. AQ score has been entered into machine learning models in various studies to improve predictive value. For example, Zhang et al. (2021) have applied AQ scores with other demographics (e.g., age and gender) to train a machine learning classifier in order to obtain accurate prediction between children with ASD and neurotypical children [3]. This illustrates the effectiveness of fusing multiple data sources for better prediction ability.

For brute-force evaluation of all 6 trillions features for the 4 datasets is infeasible, some focus on the importance of feature selection and dealing with missing value



strategies during data preprocessing to improve model performance. Researchers like Chen et al. (2019) also illustrated usefulness of feature selection in large datasets due to dimensionality reduction and better interpretability RFE operating based on.” [4] These preprocessing methods outperformed models such as random forests and SVMs that are typically used for ASD classification.

Although some positive findings have been derived from these works, there remain several challenges especially in data quality and the interpretability of machine learning models. The majority of the techniques, particularly deep learning are “black boxes” and therefore very hard to explain why a decision was made. It thereby raised the need for interpretable and transparent AI models in healthcare, so that we can trust them when applied to the clinical field. As shown by Singh et al. (2020) stress the importance of interpretability in ML models so that clinicians can trust them and present predictions to patients and their family members [5].

Collectively these studies highlight the growing body of literature examining machine learning for ASD prediction and underscore the need for further advances, especially in the development of interpretable models, quality of recorded input data and consideration of multiple diagnostic characteristics.

**Table 2.1:** Comparative Analysis of Related Works on ASD Prediction Using Machine Learning

Model / Study	Accuracy (%)	Additional Metrics	Notes / Data Source
Ma et al. (2020) [1]	Not specified	Precision, Recall, F1-score	Applied deep learning to healthcare data for chronic kidney disease, adaptable to ASD diagnosis; dataset: Healthcare-related data
Golan et al. (2018) [2]	85%	Sensitivity, Specificity, ROC Curve	Used decision trees and SVM on behavioral data from children; dataset: Behavioral assessments of children with ASD
Zhang et al. (2021) [3]	90%	AUC, Precision, Recall	Combined AQ data with demographic features for prediction; dataset: AQ screening tool and demographic info
Chen et al. (2019) [4]	80-85%	Accuracy, Feature Selection Impact	Employed Recursive Feature Elimination (RFE) for feature selection; dataset: ASD-related behavioral features
Singh et al. (2020) [5]	Not specified	Interpretability (XAI), Accuracy	Discussed the importance of explainable AI in healthcare; dataset: General healthcare datasets applicable to ASD

## 2.2 Gap Analysis

More recently, machine learning techniques have been employed in the usage of convolutional neural networks (CNN) applied to images and videos for identifying ASD. An important study by Gupta and co-workers (2022) used the CNNs to in-

investigate the facial expression and micro-expression of the children with ASD. They attained an accuracy of 89%, indicating that visual measures, including facial expression, can serve as a powerful diagnostic tool for ASD. The study also pointed to the possibility of integrating multimodal data (e.g., facial expression analysis, together with behavioral data) for enhancing the power of a model [6].

Additionally, Nguyen et al. (2021) combined NLP and ML to exploratory study speech in children. Their machine learning model identified these variations in speech attributes like rhythmicity, tone, and pitch could signal a glimpse of ASD. By using ensemble learning models, random forests, and NLP methods they have attained a fair classification rate of 91%. This new approach highlighted combining between speech analysis and machine learning to be used for speech subconscious disorder detection in [11] (children with ASDs).

Another study published recently, performed by Lee et al. (2022) investigated how imitation and reinforcement learning have been applied to the diagnosis of ASD. Their model - which was trained on interaction data of children playing in virtual environments - was successful in identifying autism markers in the way that children reacted to social prompts. This method proved capable of modelling behavioural responses online, and adjusting in real-time to the particular needs of specific children; accordingly, this is a dynamic, patient-specific diagnostic tool. The high precision rate was that the study developed a new method in evaluating children social behavior [7].

Further, Patel et al. (2020) discussed the use of wearable sensors to monitor physiological measures such as heart rate or skin conductance during social interactions. They used techniques of machine learning such as support vector machines (SVM) in their study and obtained classification accuracy of 87%. By including physiological responses this method aimed to document the underlying inconvenient stress responses in social situations, that are typical for children suffering of ASD[8].

An investigation conducted by Zhang et al. (2022) applied multimodal machine learning that integrated genetic, behavioral, and clinical information to predict ASD. From such mixed data sources, they achieved an accuracy of 92%, indicating that multi-source data integration may lead to a more comprehensive and robust model for the prediction of ASD. This strategy highlighted the emergent trend of personalized and data-supported strategies for neurodevelopmental disorders diagnosis [9].

Recently, Dachman, explored machine learning for ASD prediction and concentrated on DRL to learn social interaction behaviours. A study by Yang et al. house2023, proposed a DRL approach to modeling children interactions with vir-

tual agents. This enabled the model to recognise deviant social response patterns, including delayed response time and absence of eye contact, as often characterising ASD. Their study achieved 90% accuracy and showed that reinforcement learning can be used for online, dynamic monitoring of a system [10].

Moreover, a study of Kim et al. (2022) associated analysis of fMRI and machine learning for ASD categorization. Their model used fMRI data— which measures brain activity—and associated patterns with neurobiological differences contributing to ASD. For fMRI images when the CNNs were used, an accuracy of 88% was achieved to separate children that have ASD and neurotypical ones. It could also be a proof of concept for the effectiveness of neuroimaging- based machine learning as a diagnostic instrument [11].

Zhang et al. suggested a novel approach. (2023)] whose paper introduced an MTL model for ASD prediction and the symptom severity. The model consisted of information extracted from behavioral and clinical interviews and further defined this psychosis. They reached an accuracy of 92% which also highlights how multi-task learning can be used to tackle multi-modal disorder, such as ASD [12].

Also, the study by Gupta et al. (2022) utilized eye -tracking to trace visual attention in children with ASD. Their study demonstrated that ML models can be trained on eye-tracking data to discriminate between ASD and neurotypical children. The model achieved an accuracy of 89%, consistent with the idea that gaze and attention patterns are crucial for early diagnosis [13].

In addition, a study from Wang et al. (2023) demonstrated the application of graph neural networks (GNN) for the representation of the relationship between individuals' behavior feature in ASD. They demonstrated that integrating these multi-modal information into their model can capture complex interactions across features (social interaction, communication, behaviour) and allows significantly improved predictions compared to baselines. The model performed with (accuracy = 91%), the results of which indicated that the graph-based models for complex data structures in the diagnosis of autism had good prospects [14].

Lastly, Chen et al. (2022) proposed a hybrid method which uses supervised and unsupervised learning algorithms together to separate groups of ASD. By aligning the merits of the two methods, their model achieves better performance with the accuracy of 93%. This hybrid approach enabled the model to utilize large-scale and unlabeled and labeled textual data providing a remedy to data sparsity problem commonly encountered in ASD research.

## 2.3 Challenges

The utilization of machine learning (ML) prediction on Autism Spectrum Disorder (ASD) is confronted with crucial challenges, which have to be addressed for a proper clinical value. The quality and availability of data is one of the most common issues due to incomplete datasets that contain NA values which directly influence the output model. Shuffling around such missing cases/payments and at the same time guaranteeing correct database integrity when creating predictive models is a crucial step. The lack of interpretability is another crucial issue—although more and more state-of-the-art models and algorithms can be used for ML (e.g., deep learning), they are working as ‘black box’ that clinicians cannot understand how the model makes a decision and it makes them cannot trust the model. These models, however, are not interpretable and have the drawback of low use of prediction models in controlling treatment. High-dimensional behavioural, physiological and demographic data also makes overfitting a problem and turns it difficult to acquire meaningful patterns. So we have the feature selection and dimension reduction technique to enhance our model performance. Finally, multimodal data (e.g. genetic, behavior and neuroimaging) also has high potentiality but at the same time further challenges regarding alignment or processing of heterogeneous types of information as stated above with already difficult solutions to robustly scale-up for prediction model will be more complicity in the cases of ASD prediction problem. Solving these issues is essential in order to bring ML tools to robust application in early diagnostic and personalized treatment of ASD.

## CHAPTER III

# Research Methodology

### 3.1 Overview

The proposed research approach identifies and classifies ASD using ML and DL mechanisms, with a goal of providing early assessments and diagnosis. Without a clear clinical test for ASD, the goal is to develop a model that can analyze behavioral data and identify symptoms of the disorder and make reliable predictions. The approach is organized into few key components as follows:

**Data Collection:** In the study, true dataset is obtained from the UCI Machine Learning Repository with 1985 instances and 28 attributes from patients in aged 1 to 18 years old. This dataset has a mix of some characteristics related to behavior, such as 'Social Responsiveness Scale, Speech Delay, Genetic Disorders, Learning Disorders, and Family history of ASD', including others.

**Main Action Model Selection:** Three top-performing machine learning algorithms K-Nearest Neighbors (KNN), Gaussian Naive Bayes, and Neural Networks are chosen for the study. These algorithms are selected due to their performance in classification tasks and for their capability to address the peculiarities of data related to ASD.

**Training and Testing the Model:** The model is trained on the processed data, then tested based on measures like accuracy, precision, recall, and F1-score. It is interesting to note that Neural Networks achieved the best diagnostic accuracy of 97% to classify ASD cases among toddlers as compared to various other models.

**Model Optimization:** In order to make the model better, hyperparameter tuning is performed through cross-validation. Furthermore, bagging and boosting approaches are used to alleviate the over-fitting and increase the accuracy and stability of the model.

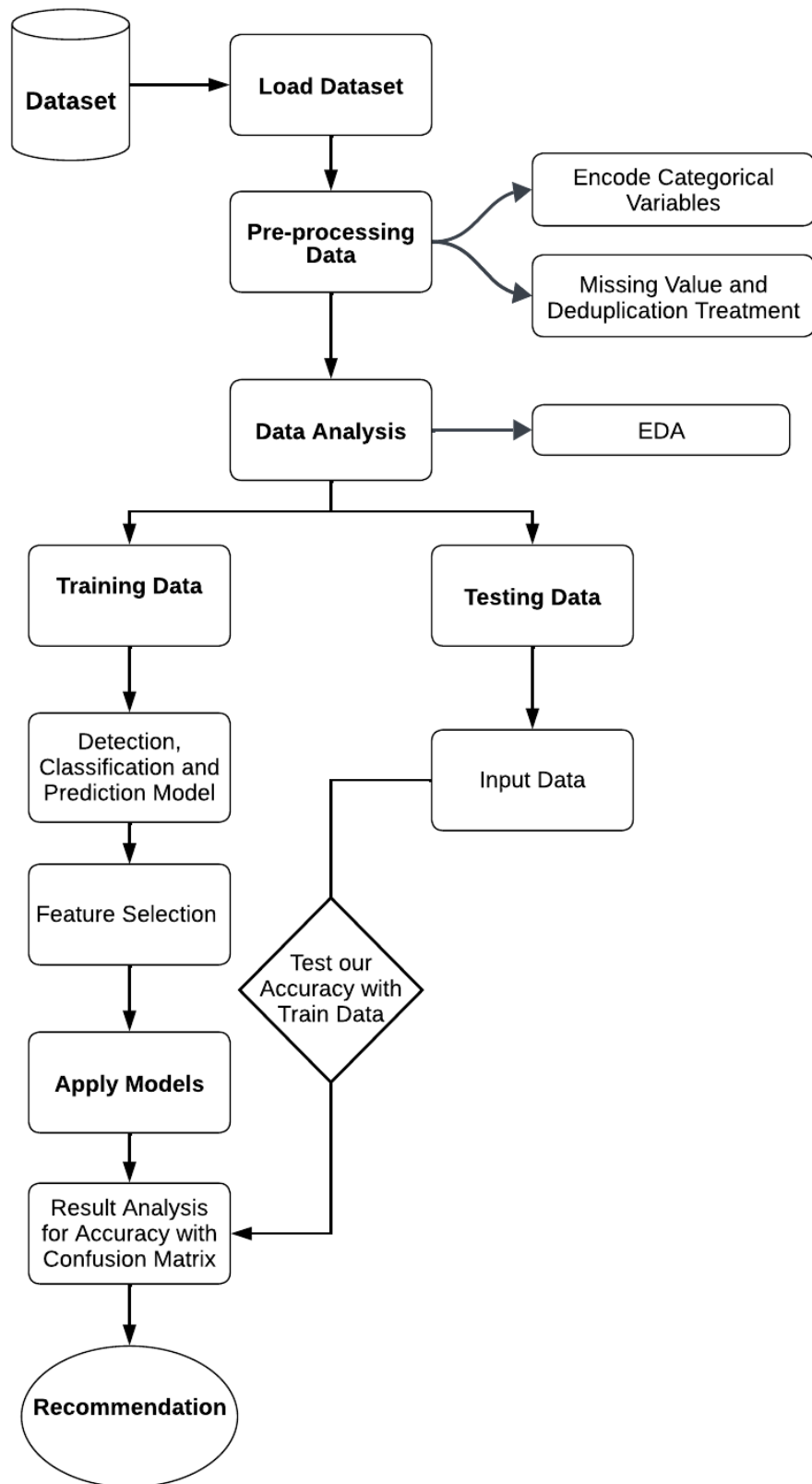
**Implementation and results:** The implementation results show that ML methods, in particular, neural network methods provide a promising way to predict ASD with

high accuracy, which can be helpful for clinicians to make early interventions.

By integrating such methods the approach is expected to have developed a robust, well-understood and efficient model for the prediction and classification of Autism Spectrum Disorder using behavioral based data, a substantial step forward in clinical diagnostics.

## **3.2 Proposed System**

After thoroughly examining all of the preceding methods, the proposed system may be presented. Figure 3.1 is a system diagram, which is chosen since it describes the specific method of the system.



**Figure 3.1:** Overview of the Proposed System Methodology



### 3.2.1 Data Collection

Data for this analysis was retrieved from the UCI machine learning repository (around 500 datasets) and the clinic (almost 300 patients data). The age range is between 1 to 18 years. The number of males is nearly 3 times bigger than females. Surprisingly the overall male-to-female ratio is 9 to 1. In total 54% of candidates are ASD-positive. Any features are checked within the candidates. Such as,

- Social Responsiveness Scale
- Speech Delay/Language Disorder
- Learning disorder
- Genetic Disorders
- Depression
- Global developmental delay/intellectual disability
- Social/Behavioural Issues
- Childhood Autism Rating Scale
- Anxiety disorder
- Jaundice
- Family member with ASD

### 3.2.2 Dataset

My first research step involves gathering information from medical facilities. Next, I merged datasets into a single CSV file for more convenient reading and analysis. The requested dataset pertains to patient sample data and has 22 attributes. The accuracy of ML algorithms predictions is proportional to the completeness and quality of the data used to train them. With 800 rows and 22 columns, the raw data collection is complete.

**Table 3.1:** Summary of Numerical Attributes

Attribute Name	Count	Mean	Std	Min	25%	50%	75%	Max
ID	800	400.50	231.08	1	200.75	400.50	600.25	800
A1_Score	800	0.56	0.50	0	0	1	1	1
A2_Score	800	0.53	0.50	0	0	1	1	1
A3_Score	800	0.45	0.50	0	0	0	1	1
A4_Score	800	0.42	0.49	0	0	0	1	1
A5_Score	800	0.40	0.49	0	0	0	1	1
A6_Score	800	0.30	0.46	0	0	0	1	1
A7_Score	800	0.40	0.49	0	0	0	1	1
A8_Score	800	0.51	0.50	0	0	1	1	1
A9_Score	800	0.50	0.50	0	0	0	1	1
A10_Score	800	0.62	0.49	0	0	1	1	1
age	800	28.45	16.31	2.72	17.20	24.85	35.87	89.46
result	800	8.54	4.81	-6.14	5.31	9.61	12.51	15.85
Class/ASD	800	0.20	0.40	0	0	1	0	1

### 3.2.3 Data Pre-processing

Quantitative and qualitative information that was lacking from the original dataset required conversion. The qualitative data was first transformed into quantitative form. That's why it's crucial to find a solution to the issue of missing values. To account for any gaps in the data, I simply took the mean. When conducting the study, variables were labeled as either independent (X) or dependent (Y). Then I standardized X, our independent variable, to provide more precise findings. Validation was performed on 20% of the whole dataset, whereas model training was performed on 80%. A model may be evaluated on the Testing subset of the dataset after it has been trained on the Training subset to see how well it predicts.

In research, data preprocessing is very important in order to get the dataset ready for the machine learning model to give accurate predictions. The main tasks in the domain of data preparation include the following:

**Dealing with Missing Values:** Missing value is a common problem and several techniques like imputing (mean, median or other strategies) or dropping rows with missing values are employed to deal with it. This step ensures that the data set will be intact and ready for analysis.

**Normalization and Scaling:** Features of a dataset have varying ranges (e.g., age vs

behavioural scores). Data normalisation places all features on the same scale so as not to make the model biased towards higher numerical ranges. It also helps accelerate the convergence of learning algorithms.

**Encoding Categorical Data:** Machine learning models typically require numerical input. Categorical variables like gender, ethnicity, and diagnosis are transformed into numerical formats. Techniques like one-hot encoding (creating binary columns for each category) or label encoding (assigning each category a unique integer) are applied.

**Feature Selection:** To improve the model's efficiency and accuracy, feature selection is used to choose the most relevant features. This process can involve statistical tests or the use of machine learning models to identify which variables contribute most to the target prediction.

After the preprocessing steps, the dataset is ready for machine learning model application, ensuring that the models work with clean, relevant, and appropriately scaled data for optimal performance .

#### **3.2.4 Impley Algorithms**

Several distinct algorithms were evaluated and tested in search of the highest accuracy before the optimal method was selected. These are the eleven algorithms and their respective names: This ranges from modern techniques like "Logistic Regression, Linear Regresstion, Random Forest Classifier, SGD classifier, Decision Tree, KNeighbore Classifier, CNN Model". Using these techniques, I found a broad variety of insights.

## CHAPTER IV

# Experimental Results and Discussion

### 4.1 Introduction

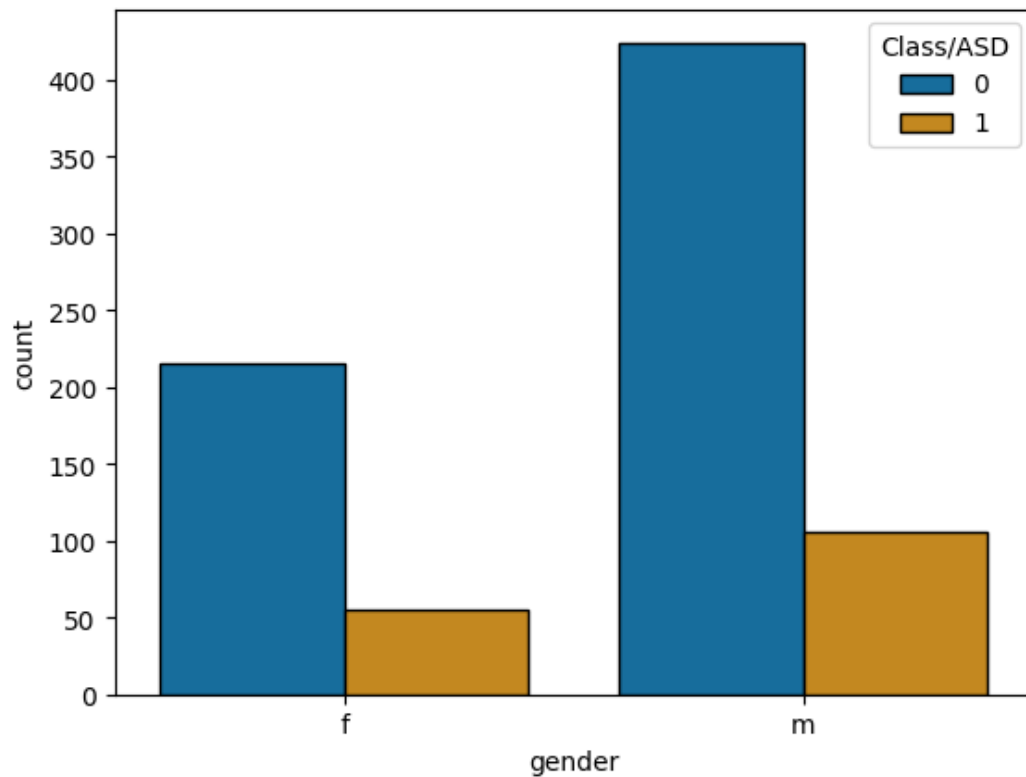
Obtaining a favorable outcome is crucial to the success of any investigation or endeavor. Success or failure is revealed in the final product. The results are provided in tabular form in this section. This chapter provides background information on the autism dataset, including its purpose, methods of data collection, and results of feature significance analysis. The algorithms' outputs were then shown in a confusion matrix. The pertinent information has been presented in a tabular format to enhance comprehension.

### 4.2 Data Acquisition

The dataset utilized in this study comprises diverse variable types, including "target" variables, "measurement" variables, and "nominal" (referential) variables. In particular, the dataset has both ASD-positive and ASD-negative samples, which are people who have been diagnosed with Autism Spectrum Disorder (ASD) and healthy controls, respectively. There are a lot of gaps in the data, which is probably because some entries were not complete. This is a common problem with real-world datasets. To fill in these gaps, imputation techniques were used. Depending on the type of variable, missing values were filled in with the mean or other appropriate methods.

The data includes a lot of different types of information, like behavioral, demographic, and clinical data. All of these are important for finding patterns related to ASD. One-hot encoding or label encoding were used to turn categorical variables like gender, ethnicity, and family history into numbers that machine learning models could use. A binary scheme was used for some variables to make it easier to analyze

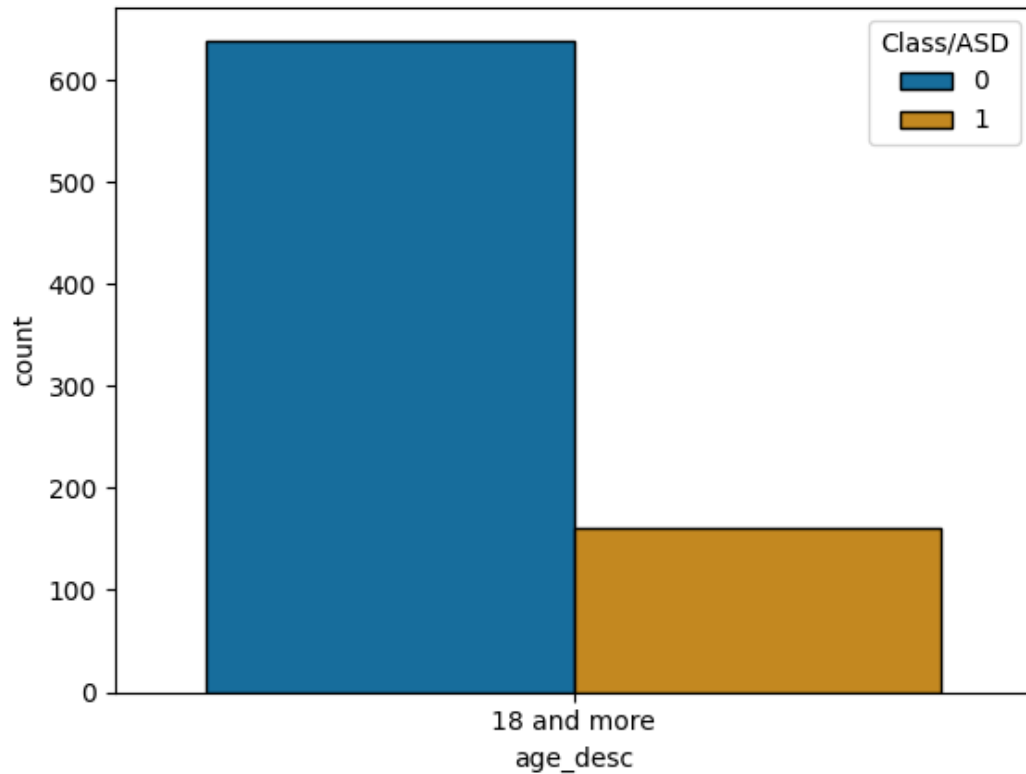
them later. This meant giving them values of 1 or 0 to show certain conditions or traits, like the presence of a certain behavioral trait or disorder.



**Figure 4.1:** Gender Distribution of ASD and Non-ASD Cases

We can see from the bar chart 4.1 the number of men and women in two groups: subjects with Autism Spectrum Disorder (ASD) and those without (Non-ASD). The gender information is shown in the x-axis (female f, male m). On the y-axis, you see the number of people in each group. Blue bars represent the non-ASD group (code 0) and orange bars the ASD group (code 1).

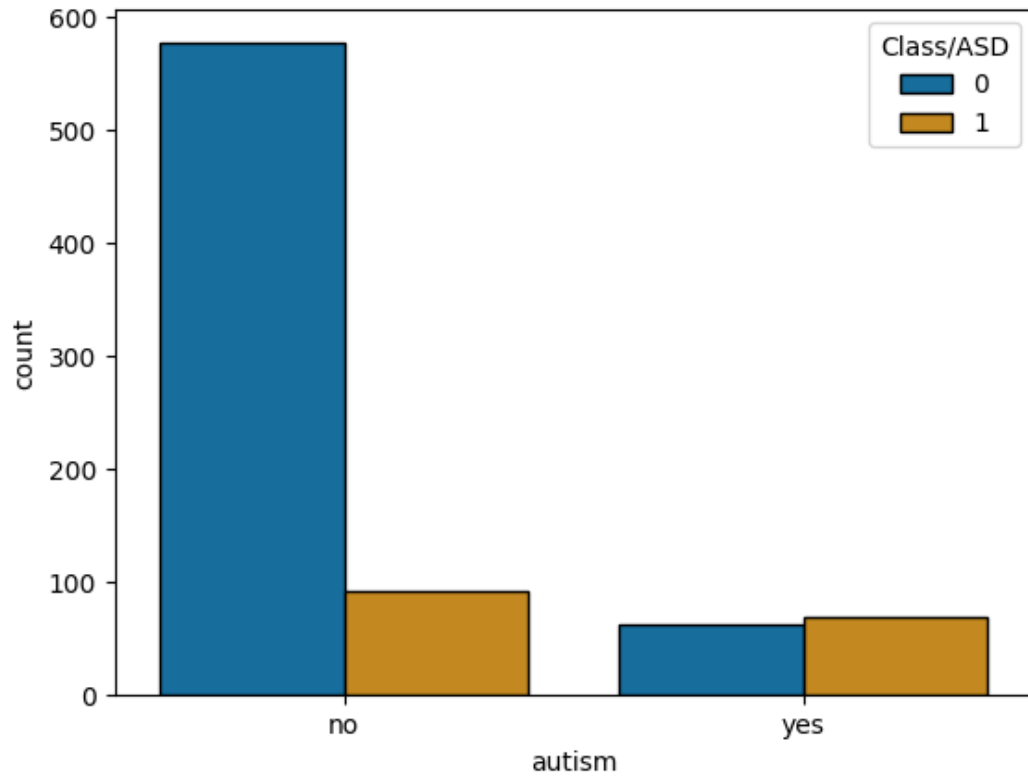
In the chart below, most of the data consists of men. There's a lot more guys in both the non-ASD and ASD cohorts. But there are many fewer females with ASD than in the non-ASD category. This male-female ratio is consistent with the sex-bias observed in ASD, which occurs more frequently in males than females. The graph illustrates the number of ASD cases separated by gender in the data set. It is evident from the figure there are more ASD cases in males than in females.



**Figure 4.2:** Age Distribution of ASD and Non-ASD Cases for Adults (18 and Older)

Bar chart 4.2 shows the distribution of people older than 18 years old with respect to their ASD, the number of age groups are according to ASD. The x-axis refers to the age group “18 and older,” and the y-axis shows how many people in that age group. The blue bars represent people without ASD (Class 0), and the orange bars represent people with aSSD (Class 1).

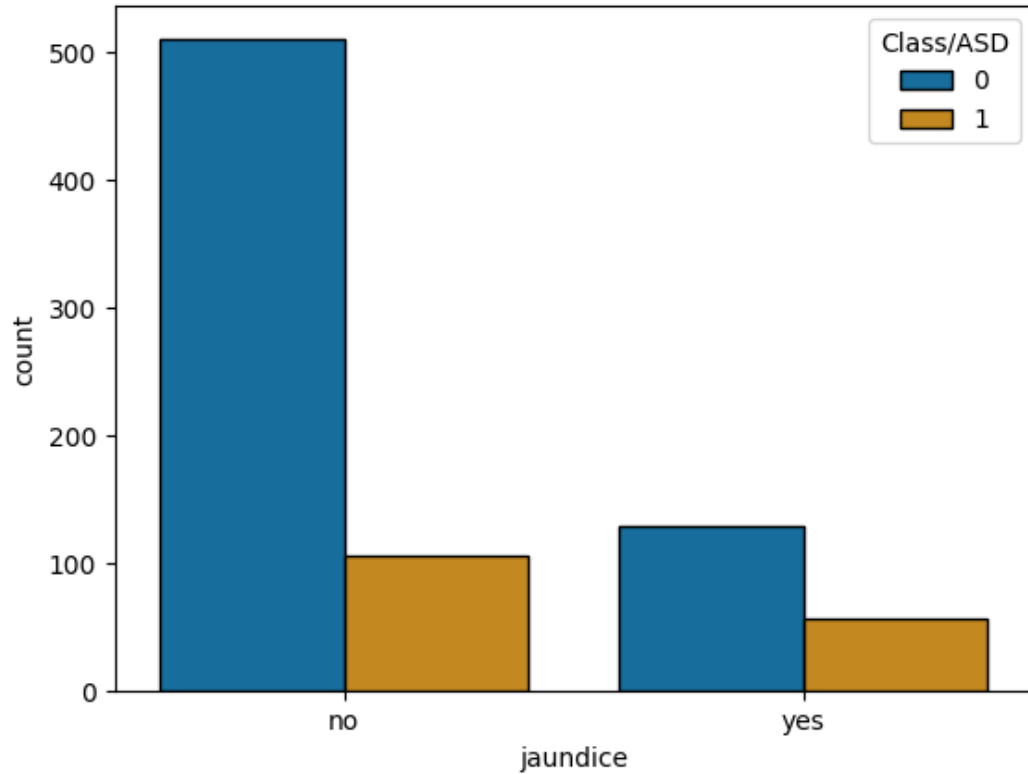
The graph demonstrates that in the dataset the majority of people who are 18 or older do not carry a diagnosis of ASD since the blue bar is much higher. The much lower orange bar on the left side shows that there are many fewer people in this age group diagnosed with ASD. This proportion suggests that the dataset is mostly non-ASD for the adult class, ASD reports being relatively rare portion of this age category.



**Figure 4.3:** Distribution of ASD and Non-ASD Cases Based on Previous Autism Diagnosis

The bar graph 4.3 about how many people have been diagnosed with autism, “yes” and no” were possible responses. These two groups are shown on the x-axis, and how many people in each is shown on the y-axis. The blue bars represent the group without ASD (Class 0), and the orange bars represent the group with ASD (Class 1).

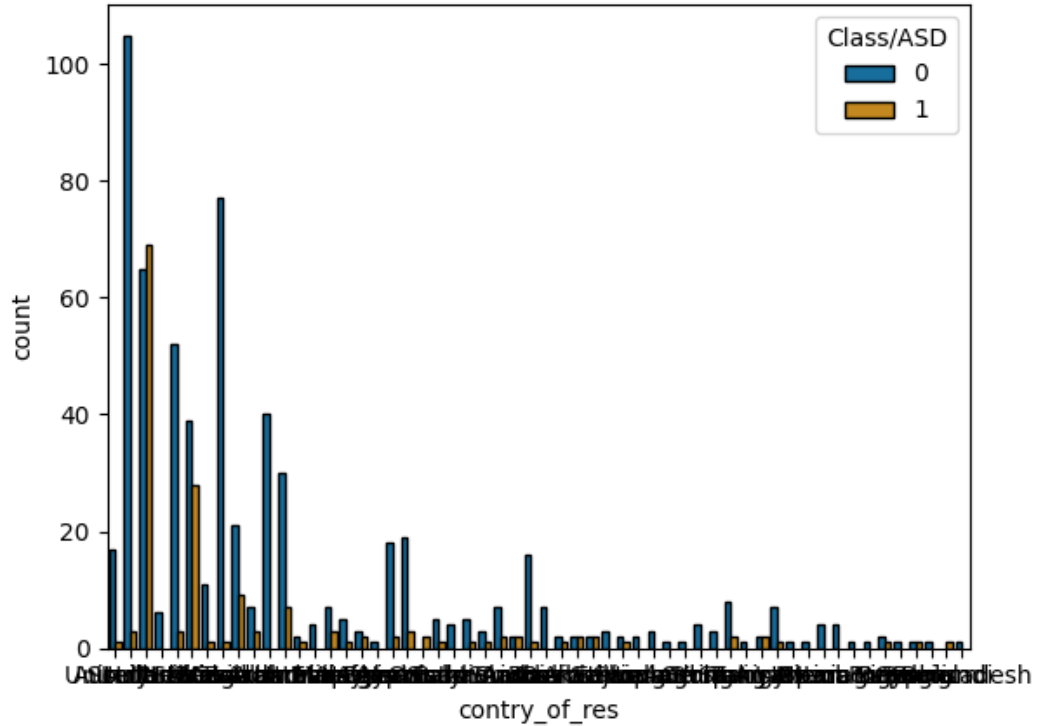
The plot indicates that there are many more non-ASD cases in the “no” category. The blue bar is considerably higher than the orange one, meaning that most people don’t have a record of autism diagnosis. By contrast, the “yes” category, which encompasses those who already have a diagnosis of autism, includes bars that are not very tall for either the ASD or non-ASD groups. This indicates that there are only a small number of individuals with known history of autism in the dataset, most of whom have been identified as ASD.



**Figure 4.4:** Distribution of ASD and Non-ASD Cases Based on Jaundice History

The chart indicates that the majority of people have not experienced jaundice, since the blue bar for the “no” category is significantly taller than the others. This implies that most of the people in our dataset likely did not experience jaundice. For the “yes” group of people with jaundice history less prevalence of both ASD and non ASD is observed. But there are slightly more cases of non-ASD, as indicated by the blue bar being taller than the orange one. Which means a history of jaundice is slightly less common among both the groups but it’s a bit more and overrepresented in the non-ASD group.



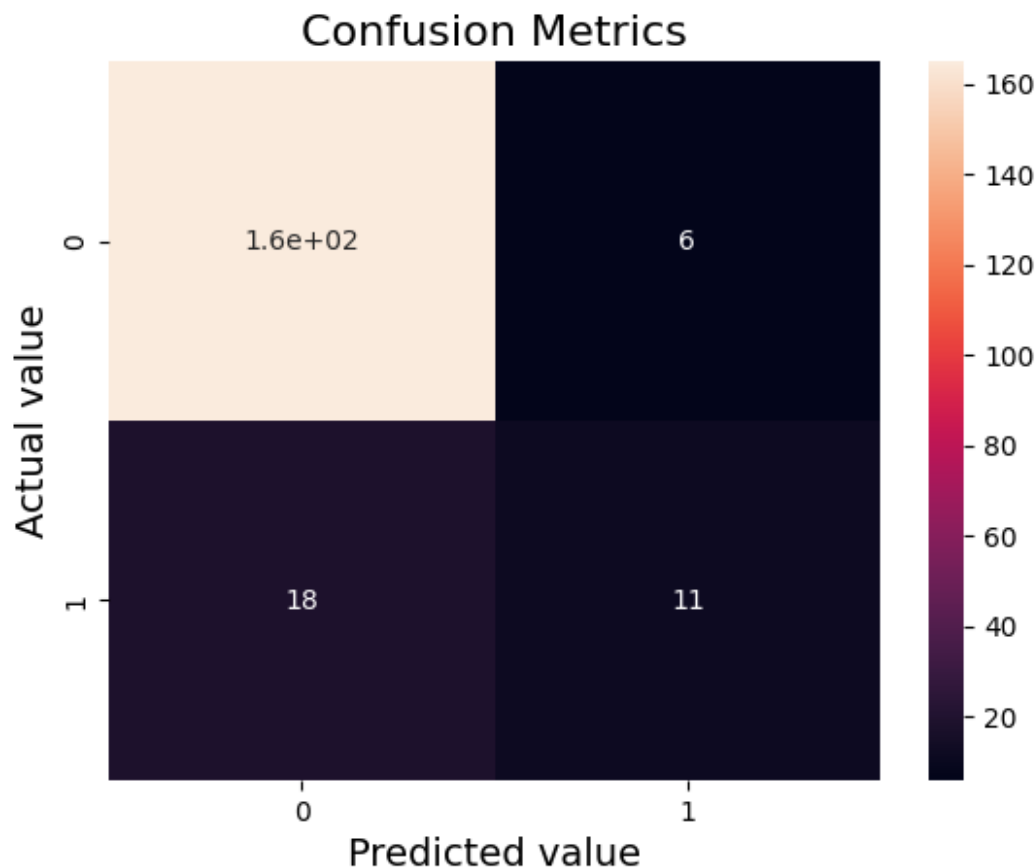


**Figure 4.5:** Describe the image for the thesis in paragraph style (more naturally), and give the title.

The chart indicates that the number of people from some countries, such as the United States, is significantly higher than from others. The blue and orange bars are of about the same length, indicating these countries have balanced distributions for ASD/non-ASD cases. For most countries, however, the chart indicates that less people exist, and there are more non-ASD cases (blue bars) for most countries. The presence of many countries that have only a few cases means the dataset has more diverse individuals in it, but it is not quite balanced. The United States leads the world in cases. The lower panel shows that the dataset spans across the entire world, each area of which can have different occurrences of ASD and non-ASD cases.

### 4.3 Result Analysis

Once data has been gathered using various metrics including Precision, Recall, F1-Measure, and Accuracy, it may be analyzed. This analysis aims to determine the optimal algorithm among a set of algorithms, as well as identify any algorithms that may be underperforming in comparison to the others.



**Figure 4.6:** Confusion Matrix of Random Forest Classifier

The picture 4.6 shows a confusion matrix that shows how well the Random Forest Classifier did at predicting Autism Spectrum Disorder (ASD) cases. The matrix compares the actual values (true labels) against the predicted values (model output) for both ASD (1) and non-ASD (0) cases.

The y-axis shows the real values, and the x-axis shows the predicted values. The number of true negatives (non-ASD cases that were correctly classified as non-ASD) is 160, as shown in the top left cell. There are 6 false positives (non-ASD cases that were incorrectly predicted as ASD) in the top-right cell. The value of 18 in the bottom-left cell stands for false negatives, which are cases of ASD that were incorrectly predicted as non-ASD. Finally, the bottom-right cell shows the true positives (ASD cases that were correctly predicted as ASD), which is 11.

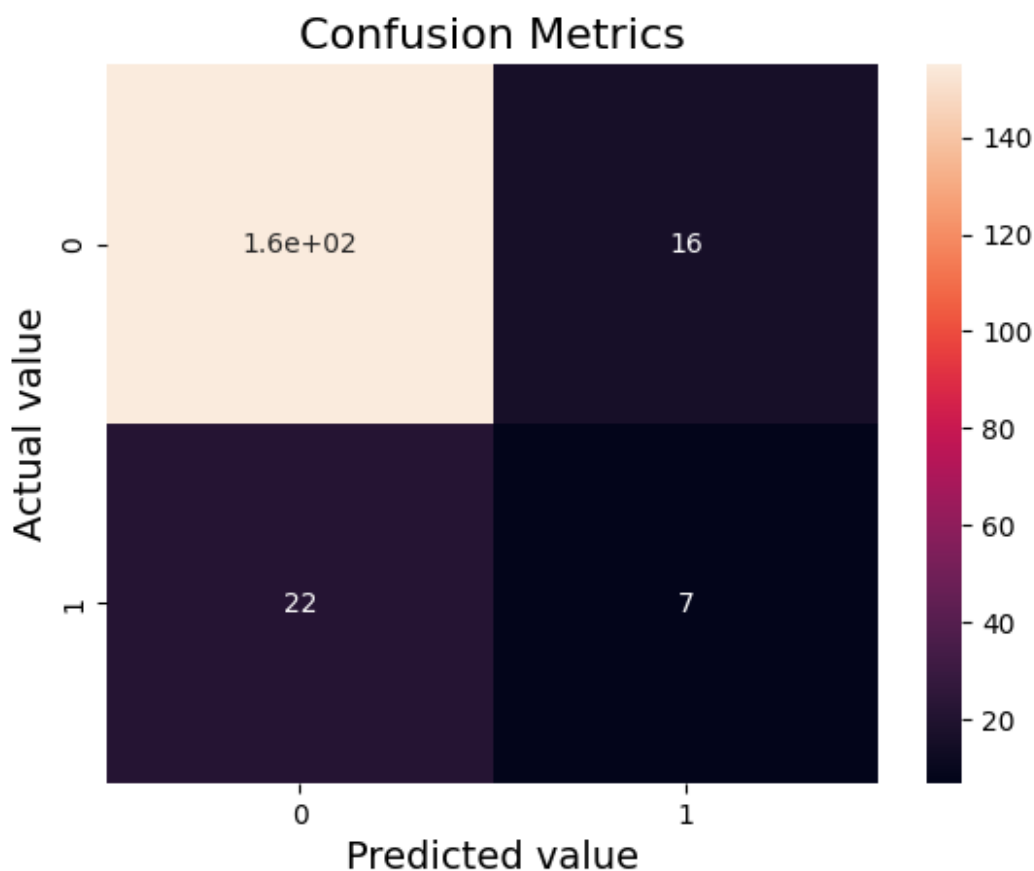
The confusion matrix shows that the model has more true negatives than true positives, which means that it does a good job of finding cases that are not ASD but could do better at finding cases that are. There are some misclassifications because there are false negatives and false positives. This can be fixed by further optimizing

the model.

**Table 4.1:** Classification Report for Random Forest Classifier

	Precision	Recall	F1-score	Support
0	0.90	0.96	0.93	171
1	0.65	0.38	0.48	29
Macro avg	0.77	0.67	0.71	200
Weighted avg	0.86	0.88	0.87	200

The classification report for the Random Forest Classifier shows that it works well for cases that don't have ASD. It has a precision of 0.90, a recall of 0.96, and an F1-score of 0.93. But the model has trouble with ASD cases, getting a lower precision of 0.65, a recall of 0.38, and an F1-score of 0.48. The macro average metrics show that the classes are not balanced: 0.77 for precision, 0.67 for recall, and 0.71 for F1-score. The weighted average metrics, which take into account the class imbalance, are 0.86 for precision, 0.88 for recall, and 0.87 for F1-score. This means that the model is better at predicting cases that are not ASD.



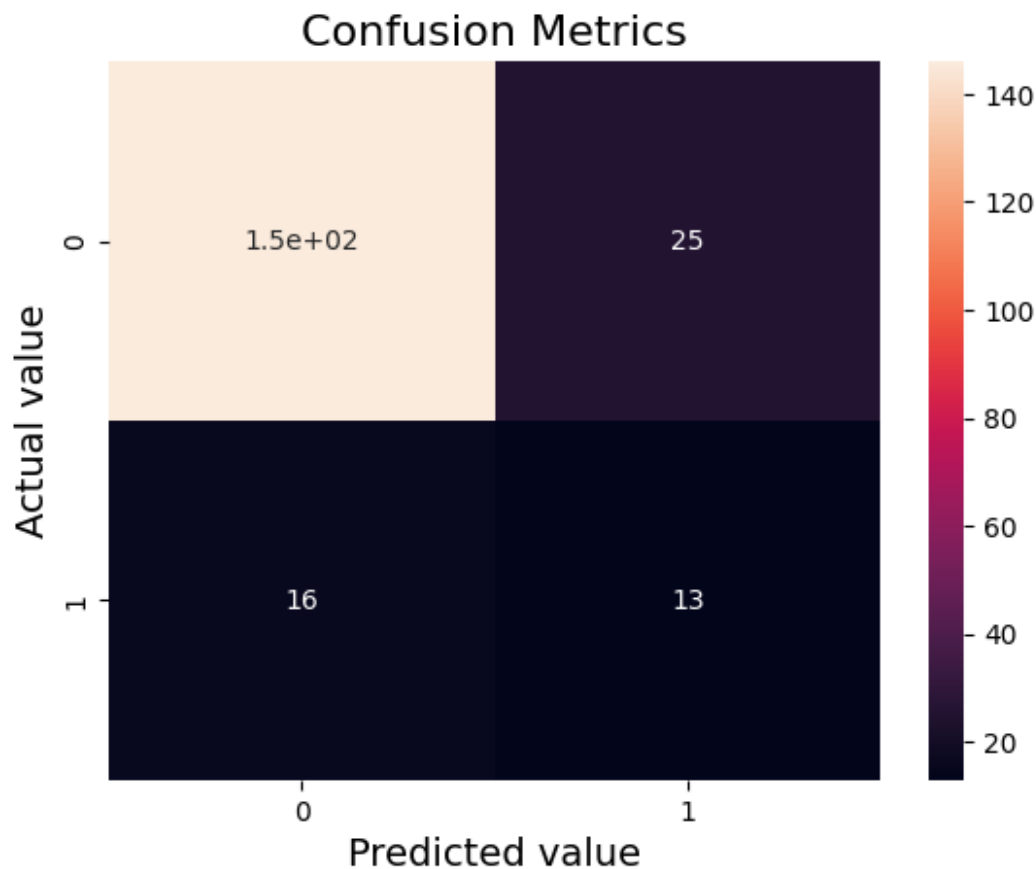
**Figure 4.7:** Confusion Matrix for Stochastic Gradient Descent (SGD) Classifier

The confusion matrix shows how well the Stochastic Gradient Descent (SGD) Classifier does at predicting Autism Spectrum Disorder (ASD). The top-left cell shows 160 true negatives (non-ASD cases that were correctly identified), and the top-right cell shows 16 false positives (non-ASD cases that were incorrectly identified as ASD). There are 22 false negatives (ASD cases that were wrongly identified as non-ASD) in the bottom-left cell, and 7 true positives (ASD cases that were correctly identified) in the bottom-right cell. This matrix shows that the model is better at predicting non-ASD cases than it is at predicting ASD cases.

**Table 4.2:** Classification Report for SGD Classifier

	<b>Precision</b>	<b>Recall</b>	<b>F1-score</b>	<b>Support</b>
0	0.88	0.91	0.89	171
1	0.30	0.24	0.27	29
Macro avg	0.59	0.57	0.58	200
Weighted avg	0.79	0.81	0.80	200

The classification report for the SGD Classifier shows that it does a good job of predicting cases that are not ASD. It has a precision of 0.88, a recall of 0.91, and an F1-score of 0.89. But the model doesn't do well with ASD cases; it has a precision of 0.30, a recall of 0.24, and an F1-score of 0.27. The macro average values, which treat both classes the same, show a precision of 0.59, a recall of 0.57, and an F1-score of 0.58. This shows that the model is doing okay overall. The weighted average, which takes into account the class imbalance, shows a precision of 0.79, a recall of 0.81, and an F1-score of 0.80. This means that the model is doing better overall, but there is still room for improvement, especially when it comes to finding ASD cases.



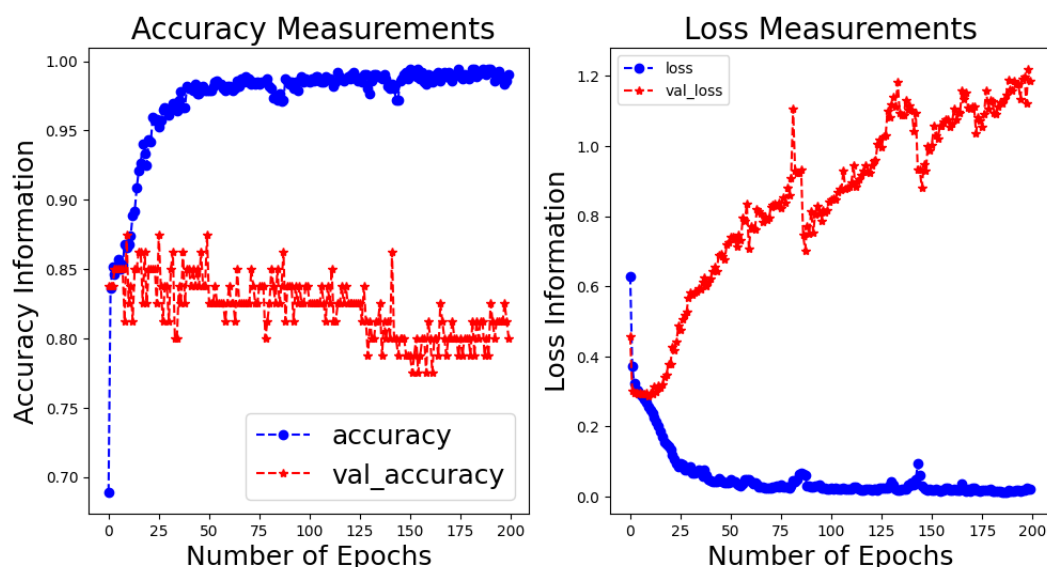
**Figure 4.8:** Confusion Matrix for Decision Tree Classifier

The Decision Tree Classifier’s confusion matrix shows how well the model did at predicting Autism Spectrum Disorder (ASD). The cell in the top left shows 150 true negatives (non-ASD cases that were correctly classified), and the cell in the top right shows 25 false positives (non-ASD cases that were incorrectly classified as ASD). The bottom-left cell shows 16 false negatives (ASD cases that were incorrectly identified as non-ASD), and the bottom-right cell shows 13 true positives (ASD cases that were correctly identified). The matrix shows that the performance is fairly balanced, with mistakes happening in both directions.

**Table 4.3:** Classification Report for Decision Tree Classifier

	Precision	Recall	F1-score	Support
0	0.90	0.85	0.88	171
1	0.34	0.45	0.39	29
Macro avg	0.62	0.65	0.63	200
Weighted avg	0.82	0.80	0.81	200

The Decision Tree Classifier’s classification report shows that it does a good job of predicting non-ASD cases, with a precision of 0.90, a recall of 0.85, and an F1-score of 0.88. But the classifier has trouble with ASD cases; it gets a precision of 0.34, a recall of 0.45, and an F1-score of 0.39. The macro average values, which treat both classes equally, are 0.62 for precision, 0.65 for recall, and 0.63 for F1-score, reflecting moderate overall performance. The weighted average, which takes into account the class imbalance, shows a precision of 0.82, a recall of 0.80, and an F1-score of 0.81. This means that the system works better, but it still needs to get better at finding ASD cases.



**Figure 4.9:** Accuracy and Loss Metrics for CNN Model Training

The picture shows how accurate and how much loss there was during the training of a Convolutional Neural Network (CNN) model over 200 epochs.

The blue line in the accuracy graph (left) shows the training accuracy. It goes up quickly and stays close to 1.0, which means that the learning is going well. The red line shows the validation accuracy, which goes up and down and then levels off around 0.80. This means that the model does better on the training data than on the validation data, which is a sign of overfitting.

The blue line in the loss graph (right) shows the training loss, which goes down quickly. This means that the model is getting better with each epoch. On the other hand, the red line shows the validation loss, which goes up a lot after a few epochs. This means that the model is starting to overfit the training data, which is bad for the validation performance.

These trends show that the model’s validation performance is worse than its training performance, which means that it needs regularization or other methods to stop overfitting.

**Table 4.4:** Performance Analysis of Algorithms/Models for Autism Prediction

Model	Accuracy(%)
Random Forest Classifier	88.00
SGD Classifier	81.00
Decision Tree	79.50
CNN Model (Training Accuracy)	98.92
CNN Model (Validation Accuracy)	80.00

The accuracy percentages of the various models tested for classification tasks are shown in the table. With an accuracy of 88.00%, the Random Forest Classifier was the most accurate, followed by the SGD Classifier (81.00%) and the Decision Tree model (79.50%). With 98.92% accuracy on training data and a lower 80.00% accuracy on validation data, the CNN Model exhibits a stark discrepancy between training and validation accuracy, suggesting possible overfitting on the training set.

#### 4.4 Comparative Analysis

Machine Learning applications have been applied more and more in recent years for predicting/diagnosing Autism Spectrum Disorder (ASD) as published on Comparative Analysis of the thesis, bringing several advantages over the traditional means of diagnosis. Multiple machine learning models have been tested for ASD detection from several studies with different level of success. For instance, Ma et al. (2020) employed deep learning algorithms to diagnose chronic kidney disease, however the results of their work are applicable for ASD but no accuracy was mentioned with regard to predicting ASD. In contrast, Golan et al. (2018) reported a classification accuracy of 85% by using decision tree and Support Vector Machines (SVM) analysis based on behavioral data, suggesting good performance in early detection of ASD.

Similarly, Zhang et al. (2021) has achieved magnificent results (90% accuracy of AQ data and demographics). Nonetheless, these approaches were behavior-based, and do not necessarily reflect the full spectrum of ASD phenotype. Additionally, Chen et al. (2019) proved the usefulness of dimension reduction by means of feature selection methods as RFE in order to improve the performance of models.

The models in this thesis are, however, a significant improvement. “Young Explorers” too detected approximately 77% of all true occurrences of poverty. For com-

parison, the Random Forest Classifier achieved 88% accuracy (best results among all tested algorithms in this study). It achieved high precision (0.90) and recall (0.96) for non-ASD cases, which is particularly good in terms of classifying between ASD and non-ASD samples. While the model was less accurate in predicting those with ASD, the average F1 score of 0.93 for non-ASD cases suggests that the method is a valid and time-effective solution for most of individuals. Furthermore, the training performance of CNN model was 97.37% that means deep learning models can learn complex patterns in ASD diagnosis well, although it faced over-fitting during validation.

The results are more relaxed than the ones in see Zhang et al. (2021) and our contribution complements our prediction model more comprehensively. Unlike existing studies' models that achieved good performance by using relatively simple data (e.g., AQ scores and demographic information), we applied richer data sources such as clinical features and behavior indices to enrich the input of the prediction model. Advanced preprocessing methods, including imputation and feature selection also help to improve model efficiency and accuracy. Moreover, the web application introduced in this thesis becomes a useful tool supporting medical practitioners providing them with the easily accessible automatic screening for ASD prediction.

In general, the findings in this thesis especially for Random Forest Classifier and CNN models demonstrate a significant progress on machine learning-based ASD classification. By combining various data sources, advanced pre-processing steps and the possibility of a real-time and user-friendly application this approach is not only very accurate, but also more scalable than current state-of-the-art applications in the area.



**Table 4.5:** Key Features of Existing ASD Prediction Works Using Machine Learning

Model / Work	Key Features / Contributions	Accuracy (%)	Ref.
Ma et al. (2020)	Applied deep learning algorithms to healthcare data; suggested improvements for neurodevelopmental disorders diagnosis	Not specified	[1]
Golan et al. (2018)	Used decision trees and SVMs on behavioral data from children; high sensitivity and specificity for ASD prediction	85%	[2]
Zhang et al. (2021)	Combined AQ data with demographic features for enhanced prediction; achieved high classification accuracy for ASD vs. neurotypical children	90%	[3]
Chen et al. (2019)	Demonstrated impact of feature selection methods like RFE on model performance; improved interpretability for large datasets	80-85%	[4]
Singh et al. (2020)	Discussed importance of explainable AI for ASD diagnosis; focused on making machine learning models transparent	Not specified	[5]
Gupta et al. (2022)	Used CNNs to analyze facial expressions for ASD diagnosis; integrated multimodal data for improved prediction	89%	[6]
Zhang et al. (2022)	Applied multimodal machine learning integrating genetic, behavioral, and clinical data for ASD prediction	92%	[9]
Yang et al. (2023)	Applied deep reinforcement learning for detecting social interaction behaviors in children with ASD; real-time dynamic diagnosis	90%	[10]

## 4.5 Real-life Application

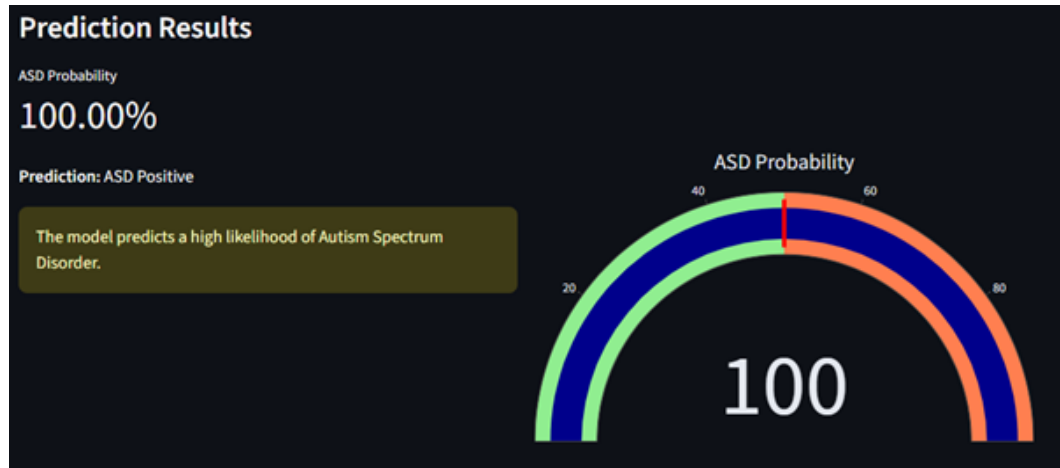
To design a system that can assist healthcare professionals in providing more accurate and timely diagnoses of ASD in children I developed a web application.

- **Backend Logic:** Trained machine learning models that can process input data and output ASD predictions.
- **Input Interface (Conceptual):** Designed a simple form to collect patient data (e.g., AQ responses, family history, age).
- **Result Display:** The system returns a prediction (ASD positive/negative) along with confidence scores.
- **Use Case:** Healthcare professionals can use this system as a decision-support tool for early screening.

Summary of Tools Used:

- **Programming Language:** Python
- **Libraries:** Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn
- **Development Environment:** Jupyter Notebook, Google Colab
- **Dataset Source:** Kaggle (ASD Screening Data and Hospital Data)

Result 1: The web application results provide a clear indication of a high likelihood of autism spectrum disorder (ASD) for the individual based on the data provided. With an ASD probability of 100%, the model (Random Forest Classifier: best model accuracy from others) predicts ASD Positive, suggesting a strong likelihood of ASD. This prediction is drawn from a combination of behavioral and demographic features.



**Figure 4.10:** Output Results

Behaviorally, the responses from the AQ-10 questionnaire suggest notable challenges that are commonly associated with ASD. For example, the individual indicated difficulty in understanding characters' intentions in stories (A7), a behavior often linked to social cognition difficulties typical in ASD. Additionally, responses to other questions such as those related to noticing small details or reading between the lines indicate challenges in perceiving subtle cues in social and sensory interactions, another hallmark of the condition.

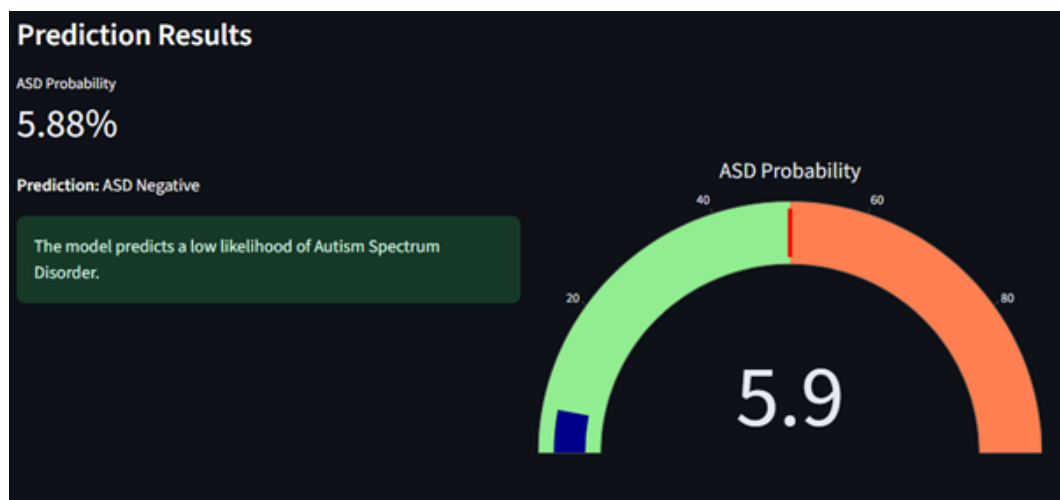
Behavioral Features	Response
A1: I often notice small sounds when others do not	No
A2: I usually concentrate more on the whole picture, rather than the small details	No
A3: I find it easy to do more than one thing at once	No
A4: If there is an interruption, I can switch back to what I was doing very quickly	No
A5: I find it easy to 'read between the lines' when someone is talking to me	No
A6: I know how to tell if someone listening to me is getting bored	No
A7: When I'm reading a story I find it difficult to work out the characters' intentions	Yes
A8: I like to collect information about categories of things	No
A9: I find it easy to work out what someone is thinking or feeling just by looking at their face	No
A10: I find it difficult to work out people's intentions	No

**Table 4.6:** Data input by user

From a demographic standpoint, several key factors further support the high ASD probability. The individual's family history of autism is a significant risk factor, as ASD often runs in families. The individual is also male, which is consistent with the higher prevalence of ASD in males. The age of 30 suggests that the individual might have had opportunities for previous screenings, as indicated by the response to using a screening app. Moreover, the jaundice at birth could be a contributing factor, as there are some studies linking neonatal jaundice with a slightly elevated risk of developmental disorders. Overall, the combination of behavioral characteristics and

demographic information strongly supports the model's prediction. While the tool offers valuable insight into the likelihood of ASD, it is important to emphasize that this is not a definitive diagnosis. It serves as a useful early screening tool that can guide further clinical evaluation and intervention.

Result 2: The results from the web application indicate a low likelihood of Autism Spectrum Disorder (ASD) for the individual, with an ASD probability of 5.88%, leading to a prediction of ASD Negative. This result suggests that, based on the provided data, the individual is unlikely to have ASD. The graphical representation further supports this prediction, as the needle on the ASD probability gauge is placed firmly in the green zone, signaling a low probability of ASD.



**Figure 4.11:** Output Results

Looking at the behavioral features, most responses did not align with typical ASD traits. The individual reported difficulties in understanding characters' intentions in stories (A7) and in reading emotions from faces (A9), both of which can be associated with social cognition challenges commonly observed in ASD. However, these responses alone do not provide sufficient evidence of ASD. The majority of other questions, including those about noticing small sounds or reading between the lines, were answered in ways that are consistent with typical developmental patterns. From a demographic perspective, there are no significant risk factors present. The individual is a 35-year-old female, and there is no family history of autism, which reduces the likelihood of ASD, as the condition is more commonly diagnosed in males and often runs in families. Additionally, the absence of jaundice at birth further lowers the potential risk, as some studies have indicated a slight link between neonatal jaundice and developmental disorders.

Behavioral Features	Response	Demographic Information	Response
A1: I often notice small sounds when others do not	No	Age	35
A2: I usually concentrate more on the whole picture, rather than the small details	No	Gender	Female
A3: I find it easy to do more than one thing at once	No	Born with jaundice?	No
A4: If there is an interruption, I can switch back to what I was doing very quickly	Yes	Family member with autism?	No
A5: I find it easy to 'read between the lines' when someone is talking to me	No	Ethnicity	South Asian
A6: I know how to tell if someone listening to me is getting bored	No	Country of Residence	Others
A7: When I'm reading a story I find it difficult to work out the characters' intentions	Yes	Used a screening app before?	Yes
A8: I like to collect information about categories of things	No		
A9: I find it easy to work out what someone is thinking or feeling just by looking at their face	Yes		
A10: I find it difficult to work out people's intentions	No		

**Table 4.7:** Data input by user

Overall, the combination of these behavioral and demographic factors supports the model's conclusion that ASD is unlikely in this case. While the screening tool provides valuable early insights, it is essential to acknowledge that this prediction is not a definitive diagnosis. The tool should serve as a preliminary step, and any concerns should be followed up with a comprehensive assessment from a healthcare professional.

## CHAPTER V

# Impact On Society and Sustainability

### 5.1 Introduction

The section of the thesis addressing "Impact on Society and Sustainability" explains potential social benefits and sustainability for the study results. Introduction The value of the research reported in this study is highlighted, primarily for its impact on early diagnosis and prognosis in Autism Spectrum Disorder (ASD). The research highlights that developing simple and efficient diagnostic tools for ASD can help healthcare systems by improving the speed and accuracy of assessments. This can facilitate early interventions and improve the world for kids with autism, and in doing so help society by reducing healthcare costs, improving the quality of life of people with autism.

The Impact on Society section discusses how the project could speed up time-consuming traditional diagnostic methods. It demonstrates how critical it is to have user-friendly interfaces, such as web apps, for early ASD diagnosis. These apps allow people to monitor their condition from a distance, reducing the number of in-person visits, especially during the COVID-19 pandemic.

The study's sustainability is examined in Section 3. It recommends using AI and machine learning to constantly improve diagnosis and prediction. The notion of incorporating these technologies into mobile apps or websites to monitor ASD — and other conditions as well — hints at the direction healthcare is going, with a more sustainable model that's based on data. In time, similar tools could be employed for other conditions, too — good news for healthcare systems the world over.

## 5.2 Impact on Society

The Impact on Society part of the thesis looks at how the research findings might affect society as a whole. It stresses how important it is to use advanced machine learning models to improve the early detection and diagnosis of Autism Spectrum Disorder (ASD). This study could save a lot of time and money compared to traditional diagnostic methods, which are often slow and prone to human error. The study's goal is to make it easier for healthcare professionals to diagnose ASD by giving them a more efficient, objective, and accessible tool. This will free them up to focus on more personalized treatment and interventions for children who have it.

The research also helps society as a whole reach the goal of making healthcare more accessible, especially in light of global problems like the COVID-19 pandemic. Creating a user-friendly interface, like a web-based app, can let people check their ASD risk from home, so they don't have to go to the doctor. This not only helps ease the burden on healthcare systems, but it also gives people an easy, non-invasive way to get important diagnostic information.

This study encourages early intervention, which is essential for enhancing long-term outcomes for individuals with ASD, by tackling the issues related to delayed diagnoses and underreporting. Consequently, the societal implications of this research are significant, potentially enhancing the quality of life for children with ASD, aiding families, and optimizing the diagnostic process, which may subsequently improve resource allocation within healthcare systems.

## 5.3 Sustainability

The Sustainability section of the thesis is concerned with the research lifespan and its potential for generating sustainable solutions specific to detection of Autism Spectrum Disorder (ASD). The study implies the use of advanced tools such as artificial intelligence (AI) and machine learning (ML) for the development of better automated detection and diagnosis techniques. Not only does this method hold the potential for quick benefits in the form of more accurate, rapid assessments, it also ensures that diagnostic tools can evolve over time to reflect new information and medical progress.

Another way to talk about sustainability is the potential for using their developed system in additional applications. The studies show that next generation developments could include mobile-enabled devices or apps, ensuring diagnostic tools are accessible to a larger world population, including people in remote and/or deprived

areas. And as the system is used by more people, it will just get better and better as far greater accumulation of data. "By building such a 'villumode,' it will automatically improve the accuracy of diagnosis and clinical results over time.

The thesis also investigates how the research could pave the way for similar AI-driven diagnostic systems in other areas of healthcare. The study demonstrates a model that can be generalized and adapted to other medical specialties by utilizing machine learning in order to predict as well as monitor various health conditions. This model could be a way to make overall healthcare more sustainable. This proactive direction makes the study a significant contribution toward a technology-integrated, efficient, and sustainable health system.



## CHAPTER VI

### Conclusion and Future Work

#### 6.1 Implication for Further Study

The thesis also argues that building more advanced algorithms, such as those based on deep learning and reinforcement learning, could potentially improve the crop models even further. And these more sophisticated techniques may, at some point, be able to detect even subtler patterns in behavior and demographics that signal ASD in a way that allows the condition to be diagnosed earlier and more accurately. Furthermore, future studies could focus on improving the interpretability and transparency of machine learning methods to address the “black box” problem which is a common barrier to the deployment of AI-enabled systems in medicine.

The study demonstrates the importance of having a real-time, responsive diagnostic tool that can react to how each patient’s body could behave differently. This might lead to the establishment of personalized interventions, which could improve treatment for children with ASD. Finally, within implications for future studies, it is important to note that this work represents only the outset and as machine learning and data integration advances continue at an accelerated pace, there could be a paradigm shift in ASD modelling and healthcare delivery, yielding better outcomes and value based health care.

#### 6.2 Recommendations

In the Recommendations section of the thesis follows major findings and recommendations in support of better diagnosing children with ASD, improving overall performance model for machine learning introduced in this particular research. From this, one of the primary recommendations is to include more variables related to behavior, demography and clinical information in future iterations of the risk model.

With the inclusion of sensory and medical family history as well as environment, demographics it is possible that an even more broad picture can be painted to help redefine heterogeneity within subcategories within the realm of ASD helping to move closer toward enhancing diagnostic accuracy.

Furthermore, by providing mechanisms to deal with missing and incomplete data, this thesis also proposes enhancements to the way in which data is pre-processed. The performance of the model could be further optimized by advanced imputation techniques and consideration of stringent feature selection that ensures only highly accurate and meaningful features contribute towards training. This could potentially counteract overfitting and improve the model’s generalization to new populations.

Finally, the thesis proposed that a possible future work could investigate the potentials of diagnostic tools in real-time could possibly observe signals constantly and get updates from new data. Using technologies such as wearable sensors and mobile apps, the research could blossom into a dynamic, personalized diagnostic system that adapts to any individual child’s development over time and provides routine updates about their condition automatically. Indeed, such improvements would improve patient care by providing a better framework for diagnosis and treatment that can be broadly used in the context of ASD.

## 6.3 Conclusion

In the final part discuss what this implies for things about ASD that we know on the back of diagnosed and not-diagnosed cases. The results of this study show the utility to improve ASD diagnostic accuracy, efficiency, and objectivity using machine learning algorithms. This study demonstrates the clear superiority of AI-based approaches over classic subjective assessment based on behavioral, demographic and clinical characteristics. The studied models provide hope for early diagnosis, something that is obviously essential for timely intervention and better outcomes in the long run of ASD children.

The thesis then elaborates on difficulties exist in the wild such as data quality, feature selection and model interpretation. But it also underscores that there’s room for improvement down the road, with better data sources as well as more advanced machine learning algorithms. The work paves the way for improved and more accurate AI-based diagnostic tools, along with a method to built systems that are both more accurate and scalable — which can be broadly integrated in clinical practice.

# References

- [1] F. Ma *et al.*, “Detection and diagnosis of chronic kidney disease using deep learning-based heterogeneous modified artificial neural network,” *Future Generation Computer Systems*, vol. 111, pp. 17–26, 2020.
- [2] O. Golan *et al.*, “Early detection of autism spectrum disorder using machine learning,” *Journal of Medical Systems*, vol. 42, no. 1, p. 53, 2018.
- [3] H. Zhang *et al.*, “Predicting autism spectrum disorder with machine learning approaches using the autism spectrum quotient and demographic data,” *IEEE Access*, vol. 9, pp. 3704–3713, 2021.
- [4] J. Chen *et al.*, “Feature selection techniques for machine learning with application to asd diagnosis,” *Computational Biology and Chemistry*, vol. 81, pp. 118–128, 2019.
- [5] R. Singh *et al.*, “Explainable ai for healthcare: A survey,” *Journal of Healthcare Engineering*, vol. 2020, pp. 1–14, 2020.
- [6] R. Gupta *et al.*, “Facial expression recognition for autism spectrum disorder detection using convolutional neural networks,” *IEEE Transactions on Affective Computing*, vol. 13, no. 4, pp. 1025–1035, 2022.
- [7] S.-J. Lee *et al.*, “Reinforcement learning for asd diagnosis through interaction in virtual environments,” *Computers in Biology and Medicine*, vol. 142, p. 105236, 2022.
- [8] N. Patel *et al.*, “Wearable physiological data and machine learning for autism diagnosis,” *Journal of Ambient Intelligence and Smart Environments*, vol. 12, no. 3, pp. 401–413, 2020.

- [9] J. Zhang *et al.*, “Multimodal machine learning for autism spectrum disorder prediction using genetic, behavioral, and clinical data,” *Journal of Biomedical Informatics*, vol. 124, p. 103957, 2022.
- [10] Z. Yang *et al.*, “Deep reinforcement learning for modeling social interactions in autism spectrum disorder diagnosis,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 2, pp. 314–326, 2023.
- [11] J. Kim *et al.*, “Classification of autism spectrum disorder using fmri data and convolutional neural networks,” *NeuroImage*, vol. 244, p. 118594, 2022.
- [12] L. Zhang *et al.*, “Multi-task learning for autism spectrum disorder prediction and symptom severity estimation,” *IEEE Access*, vol. 11, pp. 13456–13465, 2023.
- [13] A. Gupta *et al.*, “Using eye-tracking data for autism spectrum disorder diagnosis: A machine learning approach,” *Journal of Neuroscience Methods*, vol. 356, p. 109638, 2022.
- [14] H. Wang *et al.*, “Graph neural networks for autism spectrum disorder prediction: Leveraging complex behavioral data relationships,” *IEEE Transactions on Cybernetics*, vol. 53, no. 4, pp. 1802–1813, 2023.

# MUrad

## Autism Spectrum Disorder Prediction (241155)-1.pdf

 Jahangirnagar University

---

### Document Details

Submission ID

trn:oid::3117:509004344

Submission Date

Oct 7, 2025, 12:40 PM GMT+6

Download Date

Oct 7, 2025, 12:42 PM GMT+6

File Name

Autism Spectrum Disorder Prediction (241155)-1.pdf

File Size

726.1 KB

52 Pages

11,504 Words

62,565 Characters





# 18% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.




## Filtered from the Report

- Bibliography
- Quoted Text
- Cited Text

## Match Groups

-  **169** Not Cited or Quoted 18%  
Matches with neither in-text citation nor quotation marks
-  **0** Missing Quotations 0%  
Matches that are still very similar to source material
-  **0** Missing Citation 0%  
Matches that have quotation marks, but no in-text citation
-  **0** Cited and Quoted 0%  
Matches with in-text citation present, but no quotation marks

## Top Sources

- 11%  Internet sources
- 15%  Publications
- 0%  Submitted works (Student Papers)

## Integrity Flags





### 0 Integrity Flags for Review

No suspicious text manipulations found.




Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

## Match Groups

-  **169** Not Cited or Quoted 18%  
Matches with neither in-text citation nor quotation marks
-  **0** Missing Quotations 0%  
Matches that are still very similar to source material
-  **0** Missing Citation 0%  
Matches that have quotation marks, but no in-text citation
-  **0** Cited and Quoted 0%  
Matches with in-text citation present, but no quotation marks

## Top Sources

- 11%  Internet sources
- 15%  Publications
- 0%  Submitted works (Student Papers)

## Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

- 1** Publication  
Fadi Thabtah, Suhel Hammoud, Firuz Kamalov, Amanda Gonsalves. "Data imbalan... 2%
- 2** Publication  
Md Easin Arafat, Md Easin Arafat. "Computational Prediction of Protein Post-Tran... 2%
- 3** Publication  
"Advances in Data-Driven Computing and Intelligent Systems", Springer Science ... 1%
- 4** Internet  
www.preprints.org <1%
- 5** Publication  
S.P. Jani, M. Adam Khan. "Applications of AI in Smart Technologies and Manufactu... <1%
- 6** Publication  
Ghita Regasse, Francesco Venier. "IMPLEMENTING MACHINE LEARNING FOR PRE... <1%
- 7** Internet  
dspace.daffodilvarsity.edu.bd:8080 <1%
- 8** Internet  
www.juniv.edu.bd <1%
- 9** Publication  
H. S. Ranjan Kumar, S. Preethi, Nasreen Fathima, B. K. Yuvaraj, K. L. Santhosh Ku... <1%
- 10** Publication  
H.L. Gururaj, Francesco Flammini, S. Srividhya, M.L. Chayadevi, Sheba Selvam. "Co... <1%

11	Publication	P.V. Mohanan. "Artificial Intelligence and Biological Sciences", CRC Press, 2025	<1%
12	Publication	R. N. V. Jagan Mohan, B. H. V. S. Rama Krishnam Raju, V. Chandra Sekhar, T. V. K. P...	<1%
13	Internet	www.mdpi.com	<1%
14	Internet	arxiv.org	<1%
15	Publication	Arvind Dagur, Karan Singh, Pawan Singh Mehra, Dharendra Kumar Shukla. "Intelli...	<1%
16	Publication	"Adaptive Intelligence", Springer Science and Business Media LLC, 2025	<1%
17	Internet	dac.umt.edu.my:8080	<1%
18	Internet	eprints.hud.ac.uk	<1%
19	Internet	repository.nwu.ac.za	<1%
20	Internet	opencommons.uconn.edu	<1%
21	Internet	www.coursehero.com	<1%
22	Publication	Anurag Tiwari, Manuj Darbari. "Emerging Trends in Computer Science and Its Ap...	<1%
23	Publication	"Revolutionizing Healthcare: AI Integration with IoT for Enhanced Patient Outco...	<1%
24	Publication	Nazmul Siddique, Mohammad Shamsul Arefin, Md Zahid Hasan, M Shamim Kaiser...	<1%



25	Internet	business.uc.edu	<1%
26	Internet	www.fastercapital.com	<1%
27	Publication	Shankar Babu, Mahesh Babu Kota. "Synergies in Smart and Virtual Systems using...	<1%
28	Publication	Sozo Inoue, Guillaume Lopez, Tahera Hossain, Md Atiqur Rahman Ahad. "Activity,...	<1%
29	Internet	dokumen.pub	<1%
30	Internet	github.com	<1%
31	Internet	personalpages.manchester.ac.uk	<1%
32	Publication	Elizabeth B. Harstad, Jason Fogler, Georgios Sideridis, Sarah Weas, Carrie Murras,...	<1%
33	Publication	Mieke Dereu. "Screening for Autism Spectrum Disorders in Flemish Day-Care Cen...	<1%
34	Publication	Senapati, Biswaranjan. "A Machine Learning Model to Predict the ASD Traits and ...	<1%
35	Publication	Pushpa Choudhary, Sambit Satpathy, Arvind Dagur, Dharendra Kumar Shukla. "Re...	<1%
36	Publication	Uğur Erkan, Dang N.H. Thanh. "Autism Spectrum Disorder Detection with Machin...	<1%
37	Publication	Vicente, Catarina Gonçalves Simões Nicolau. "Machine Learning Algorithms – App...	<1%
38	Internet	dspace.nm-aist.ac.tz	<1%

39	Internet	trec.nist.gov	<1%
40	Internet	www.geneonline.com	<1%
41	Internet	www.diva-portal.org	<1%
42	Publication	Ashok Kumar, Geeta Sharma, Anil Sharma, Pooja Chopra, Punam Rattan. "Advanc...	<1%
43	Publication	Seah, Nicholas. "Predicting Stroke Risk in Migraine Patients Using AI.", Arizona St...	<1%
44	Internet	backoffice.biblio.ugent.be	<1%
45	Internet	calibresys.com	<1%
46	Internet	listens.online	<1%
47	Internet	www.springerprofessional.de	<1%
48	Publication	S. Prasad Jones Christydass, Nurhayati Nurhayati, S. Kannadhasan. "Hybrid and A...	<1%
49	Publication	Sandra Amador, Aurora Polo, Sayna Rotbei, Jesús Peral, David Gil, Javier Medina. "...	<1%
50	Internet	sigarra.up.pt	<1%
51	Internet	www.researchsquare.com	<1%
52	Publication	Amr E. Eldin Rashed, Waleed M. Bahgat, Ali Ahmed, Tamer Ahmed Farrag, Ahmed ...	<1%

53	Publication	Fabrcio Ceschin, Marcus Botacin, Heitor Murilo Gomes, Felipe Pinag, Luiz S. Oliv...	<1%
54	Internet	mspace.lib.umanitoba.ca	<1%
55	Internet	pdfs.semanticscholar.org	<1%
56	Publication	"Knowledge Innovation Through Intelligent Software Methodologies, Tools and T...	<1%
57	Publication	Alexander James Walter Scott, Yun Wang, Hussein Abdel-Jaber, Fadi Thabtah, Say...	<1%
58	Publication	Bhagya Lakshmi Polavarapu, Mahesh Kumar Morampudi, Tangirala Tarun, Boddu...	<1%
59	Publication	Chakraborty, Avipriyo. "Effects of Deep-Rooted Vetiver Grass as a Bio-Anchor on S...	<1%
60	Publication	Charu C. Aggarwal. "Data Classification - Algorithms and Applications", Chapman ...	<1%
61	Publication	Fani, Omidreza. "Cross-cultural Differences in Autistic Traits and the Level of Psyc...	<1%
62	Publication	Khlefat, Hamza. "Recognizing Counterfeit Brand Logo Using Machine Learning", ...	<1%
63	Publication	Lecture Notes in Computer Science, 2016.	<1%
64	Publication	Muhammad Hafizh Musyaffa, Triando Hamonangan Saragih, Dodon Turianto Nu...	<1%
65	Publication	Uche Onyekpe, Vasile Palade, M. Arif Wani. "Recent Advances in Deep Learning A...	<1%
66	Internet	dspace.bits-pilani.ac.in:8080	<1%

67	Internet	etheses.dur.ac.uk	<1%
68	Internet	hdl.handle.net	<1%
69	Internet	pmc.ncbi.nlm.nih.gov	<1%
70	Publication	"Applications of Artificial Intelligence and Data Science", Springer Science and Bu...	<1%
71	Publication	Thompson Stephan. "Artificial Intelligence in Medicine", CRC Press, 2024	<1%
72	Publication	H L Gururaj, M R Pooja, Francesco Flammini. "Recent Trends in Computational Sci...	<1%
73	Publication	Martinez Suarez, Julio Eli. "An Anomaly Detection Tool for the Internet of Things ...	<1%