**COMPARATIVE ANALYSIS OF MACHINE LEARNING MODELS IN EARLY DETECTION AND DIAGNOSIS OF AUTISM SPECTRUM DISORDER AMONG TODDLERS**

**AUSTINE OMONDI SCM224-0602/2019**

**CHARLES MAINA SCM224-0585/2019**

**CHRICENCIA WINNY SCM224-0601/2019**

**KEVIN KIBET SCM224-0579/2019**

**A RESEARCH PROPOSAL SUBMITTED TO THE DEPARTMENT OF STATISTICS AND ACTUARIAL SCIENCES IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE AWARD OF BSC. BIOSTATISTICS IN JOMO KENYATTA UNIVERSITY OF AGRICULTURE AND TECHNOLOGY**

**JULY 2023**

**DECLARATION**

This project is an original work and has not been presented as an award of any degree in nay University.

Austine Omondi Date: 26th July 2023

SCM224-0602/2019 ------------------------

Charles Maina Date: 26th July 2023

SCM224-0585/2019 ------------------------

Chricencia Winny Date: 26th July 2023

SCM224-0601/2019 ------------------------

Kevin Kibet Date: 26th July 2023

SCM224-0579/2019 ------------------------

This project has been submitted to the University for examination with my approval as the University supervisor.

Dr. Susan Mwelu Date: 27th July 2023

Supervisor ------------------------

# **DEDICATION**

This project has been dedicated to God Almighty for his great love and wisdom that He granted us during the research, and during generation of ideas for this project. We also dedicate this research to our lovely parents to reciprocate their constant encouragement and relentless prayers during the entire period of our educational journey.

# **ACKNOWLEDGEMENT**

First of all, we would like to express our profound gratitude to Dr Samuel Mutua, and Dr. Susan Mwelu who has been there to assist us and provide invaluable advice, and mentorship during the entire period of this project. We sincerely acknowledge her support, guidance, and for always holding us to higher standards.

Secondly, we are grateful to all our lecturers and the department of Statistics and Actuarial Sciences for giving us invaluable knowledge that has assisted us in throughout this project.

Finally, we would like to thank our families for their moral support, inspiration, and all form of help accorded to us during the entire period of our study.

Table of Contents

[**DEDICATION** iii](#_Toc150440156)

[**ACKNOWLEDGEMENT** iv](#_Toc150440157)

[**LIST OF TABLES** vii](#_Toc150440158)

[**LIST OF FIGURES** viii](#_Toc150440159)

[**LIST OF ABBREVIATIONS** ix](#_Toc150440160)

[**CHAPTER ONE** 1](#_Toc150440161)

[**1.1** **Background of the Study** 1](#_Toc150440162)

[**1.1.1 Introduction** 1](#_Toc150440163)

[**1.1.2 Historical Background of Autism Spectrum Disorder** 2](#_Toc150440164)

[**1.1.3 Utilization of Machine Learning Models in Diagnosis of Autism Spectrum Disorder** 4](#_Toc150440165)

[**1.2 Statement of the Problem** 6](#_Toc150440166)

[**1.3** **Research Objectives** 6](#_Toc150440167)

[**1.3.1** **General Objective** 6](#_Toc150440168)

[**1.3.2 Specific Objectives** 6](#_Toc150440169)

[**1.4** **Justification of the Research** 6](#_Toc150440170)

[**1.5 Scope of the Research** 7](#_Toc150440171)

[**CHAPTER TWO** 8](#_Toc150440172)

[**2.1 Introduction** 8](#_Toc150440173)

[**2.2 Theoretical Review** 8](#_Toc150440174)

[**2.3 Overview of the Literature Review** 19](#_Toc150440175)

[**CHAPTER THREE** 20](#_Toc150440176)

[**3.1 Introduction** 20](#_Toc150440177)

[**3.2 Data Source and Type** 20](#_Toc150440178)

[**3.3 Data Preprocessing** 20](#_Toc150440179)

[**3.3.1 Data Cleaning** 20](#_Toc150440180)

[**3.3.2 Feature Selection** 21](#_Toc150440181)

[**3.3.3 Encoding Categorical Variables** 21](#_Toc150440182)

[**3.3.4 Normalization** 22](#_Toc150440183)

[**3.3.5 Handling Imbalance** 22](#_Toc150440184)

[**3.3.6 Data Splitting** 22](#_Toc150440185)

[**3.4 Modelling** 23](#_Toc150440186)

[**3.4.1 Support Vector Machine (SVM)** 23](#_Toc150440187)

[**3.4.2 Random Forest (RF)** 25](#_Toc150440188)

[**3.4.3 Artificial Neural Network** 26](#_Toc150440189)

[**3.5 Evaluation Metrics** 28](#_Toc150440190)

[**3.6 Model Deployment** 29](#_Toc150440191)

[**CHAPTER FOUR** 30](#_Toc150440192)

[**4.1 Introduction** 30](#_Toc150440193)

[**4.2 Descriptive Statistics** 30](#_Toc150440194)

[**4.3 Data Preprocessing Results** 30](#_Toc150440195)

[**4.4 Model Performance Results** 31](#_Toc150440196)

[**4.5 Feature Importance Analysis Results** 31](#_Toc150440197)

[**4.6 Discussion of Results** 32](#_Toc150440198)

[**4.7 Model Comparison** 32](#_Toc150440199)

[**4.8 Model Deployment** 32](#_Toc150440200)

[**CHAPTER FIVE** 34](#_Toc150440201)

[**Budget** 41](#_Toc150440202)

[**Work Plan** 42](#_Toc150440203)

# **LIST OF TABLES**

[Table 1 Budget 36](#_Toc141203067)

[Table 2 Work plan 37](#_Toc141203068)

# **LIST OF FIGURES**

[Figure 1 Distinct classes partitioned using hyperplane 24](#_Toc141190412)

[Figure 2 Random Forest Tree presentation 26](#_Toc141190413)

[Figure 3 A typical neural network 27](#_Toc141190414)

# **LIST OF ABBREVIATIONS**

fMRI Functional MRI

sMRI Structural MRI

ANN Artificial Neural Network

ASD Autism Spectrum Disorder

AUC Area Under Curve

DSM-5 Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition

DT Decision Tree

DL Deep Learning

KNN K-Nearest Neighbor

LR Logistic Regression

NB Naïve Bayes

PCA Principal Component Analysis

RF Random Forest

ROC Receiver Operator Characteristic

SMOTE Synthetic Minority Oversampling Technique

SVM Support Vector Machine

WEKA Waikato Environment for Knowledge Analysis

DSM Diagnostic and Statistical Manual of Mental Disorders

ML Machine Learning

AI Artificial Intelligence

# **CHAPTER ONE**

**INTRODUCTION**

## **Background of the Study**

### **1.1.1 Introduction**

Autism Spectrum Disorder (ASD) is a neurological disorder which might have a lifelong impact on the language learning, speech, cognitive and social skills of the individual. Currently, ASD diagnosis is done using clinical standardized tests as the only methods which requires prolonged diagnostic time and faces a steep increase in medical costs (Halligan, 2020).

Autism Spectrum Disorder is a term that defines a group of neurological disorders attributed to difficulties in social skills, communication and behavior by an individual (Van’tHof et al., 2021). Autistic individuals may have problems with pleasant diversions, boredom and social interaction. Others may show difficulty in paying attention in class or following directions. Individuals with ASD may have difficulties in social and interaction skills for example reluctance in making eye contact, and showing no signs of sensitivity to the feelings of others, etc. Also, they may exhibit restricted interests or repetitive behavior. Other symptoms may include; delayed language development, delayed motor development, delayed cognitive or learning development, abnormal eating and sleeping patterns, gastrointestinal problems, anxiety, stress, lack of fear or dread in excess (Talukdar, 2023).

According to Kim et al. (2019) there are varying causal and risk factors due to the spectrum nature of autism. The causal factors include; Genetics, and environmental factors. While the risk factors include; sex, family history, extremely preterm babies and parents age during birth. The clinical diagnosis of autism in children is currently done in two stages. The first stage is the screening for general development during well-child checkups where the child is screened at 9, 18, 24 and 30 months of age. The second stage is the diagnostic evaluation which include neurological and medical tests, evaluation of the kid’s cognitive skills, evaluation of the child’s language skills, Observing the child’s actions etc., (Talukdar, 2023).

Autism Spectrum Disorder has varying interventions. Some therapies aim at minimizing troubling behaviors and building communication and social skills while others focus on sensory integration problems, motor skills, emotional issues and food sensitivities. Some of the common autism treatments are behavior therapy, physical therapy, occupational therapy and nutritional therapy (Smith, 2023).

### **1.1.2 Historical Background of Autism Spectrum Disorder**

To date, identifying the initial point when disorders and conditions were first described is an effortless task. However, this has not been the case for autism. The diagnostic criteria of autism have been unclear. In fact, in the past five decades, autism has been described in distinct ways that were less direct, with several branching out. Despite the deviations throughout time, clinicians have come up with a basic definition of autism published in the Diagnostic and Statistical Manual of Mental Disorders (DSM).

The psychiatrist Eugen Bleuler introduced the term "autism" in 1908, initially employing it to describe a schizophrenic individual who had retreated into their internal world. In using the term "autism," Bleuler aimed to convey the notion of excessive self-absorption and the tendency to withdraw into oneself (Mandal, 2019). Over time, the DSM, which serves as a guide for physicians in diagnosing various conditions, including autism, has undergone revisions that have impacted the diagnostic criteria and the portrayal of autism. These changes in the DSM editions offer valuable insights into the evolution of autism diagnosis throughout the years, particularly in the United States. By examining the modifications made to the manual, we can gain a comprehensive understanding of how the diagnosis of autism has transformed over time (Ostimo, 2023).

Leo Kanner, an Austrian-American psychiatrist, and physician, provided the initial description of autism in 1943. Kanner's observations portrayed autism as an emotional disorder rather than a cognitive or developmental one. Consequently, the second edition of the DSM-II, published in 1952, characterized autism as a psychiatric condition (Ostimo, 2023). According to the manual, autism was regarded as a variant of childhood schizophrenia. In the 1950s, Bruno Bettelheim described that the main cause of autism in children was the coldness of their mothers, but this theory was disregarded in the 1970s following the discovery that autism developed as a result of some environmental factors and genetic etiology (Mandal, 2019).

In 1980, the third edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM-III) differentiated autism from schizophrenia and defined it as a distinct "pervasive developmental disorder." This revision established autism as a separate and independent diagnosis, recognizing its unique characteristics and distinguishing it from other mental health conditions. The evolution of autism diagnosis did not conclude with the previous revision. In 1987, the DSM-III underwent further modifications, substantially altering the criteria for autism in various ways (Ostimo, 2023). As a result of this revision, autism encompassed a more comprehensive scope, incorporating a new category known as "pervasive developmental disorder-not otherwise specified" (PDD-NOS) to include cases of "mild autism." Moreover, the previously required 30-month developmental delay was no longer a mandatory criterion.

In the DSM-IV, which was published in 1994 and revised in 2000, autism was formally recognized as a spectrum disorder. In this particular edition, the manual listed five conditions, each with distinct features. Alongside autism and PDD-NOS, other conditions were incorporated within the autism spectrum, including Asperger's syndrome (Mandal, 2019). The current version of the DSM – DSM-V – consolidated all the condition subcategories under the comprehensive diagnosis known as an Autism Spectrum Disorder, resulting in the integration of Asperger Syndrome into the broader spectrum. As a consequence of this consolidation, ASD is currently characterized by two main categories: impaired social communication and restricted and repetitive behaviors (Ostimo, 2023).

### **1.1.3 Utilization of Machine Learning Models in Diagnosis of Autism Spectrum Disorder**

The use of machine learning (ML) models in detecting and diagnosing Autism Spectrum Disorder (ASD) has gained significant attention in recent years (Bohr & Memarzadeh, 2020). By combining ML techniques with neuroimaging modalities and other diagnostic tools, researchers have made progress in accurately classifying individuals with ASD (Prova et al., 2021). This section provides an overview of the evolution of ML in ASD diagnosis and highlights key peer-reviewed literature that demonstrates the effectiveness of ML models in detecting and diagnosing ASD.

Neuroimaging-based single subject prediction of brain disorders, including ASD, has been the focus of numerous studies. These studies employ various neuroimaging modalities such as structural, functional, and diffusion MRI, along with ML techniques, to accurately classify patients with ASD. A comprehensive review by Arbabshirani (2017) summarizes over 200 reports in this field, providing valuable insights into the advancements and challenges in neuroimaging-based single subject prediction of ASD.

Children diagnosed with ASD often exhibit atypical motor patterns, which can be identified using ML techniques with high accuracy. Nagesh et al. (2022) demonstrated the effectiveness of ML in identifying these motor patterns in children with ASD. By leveraging ML algorithms, researchers can accurately classify individuals with ASD based on their motor patterns.

Early diagnosis of ASD is crucial for effective intervention and support. ML methods have been employed to develop intelligent medical diagnostic systems for early detection of ASD. Safara (2021) proposed a multilayer perceptron neural network for ASD detection, achieving an accuracy of 99.6%. This study highlights the potential of ML in identifying and diagnosing ASD in the early stages, leading to improved outcomes for individuals with ASD.

Resting-state functional MRI (fMRI) has been used in conjunction with ML techniques to diagnose ASD and identify the brain regions contributing to the diagnosis decision. ElNakieb et al. (2023) developed a pipelined framework using fMRI, which achieved a global balanced accuracy of 98.8% in diagnosing ASD. This study demonstrates the potential of ML models in accurately diagnosing ASD based on resting-state fMRI data.

Machine learning has also been utilized to improve screening and diagnostic instruments for ASD. Bone et al. (2016) utilized ML algorithms to derive ASD instrument algorithms, aiming to enhance widely-used screening and diagnostic tools. By leveraging ML techniques, researchers can develop more accurate and efficient instruments for ASD screening and diagnosis.

The use of ML models in the diagnosis of ASD has made significant progress in recent years. Neuroimaging-based single subject prediction, motor pattern analysis, intelligent medical diagnostic systems, resting-state fMRI analysis, and improvement of screening and diagnostic instruments are some of the key areas where ML has been successfully applied. These studies demonstrate the effectiveness of ML models in accurately detecting and diagnosing ASD. Further research and advancements in ML techniques hold great potential for improving the diagnosis and intervention for individuals with ASD. In this project, we will utilize behavioral features and cultural features to determine the best-performing machine learning model that can be used by healthcare practitioners in the early detection and diagnosis of Autism Spectrum Disorder.

## **1.2 Statement of the Problem**

Currently, neuroimaging is being employed in machine learning models for the diagnosis of Autism Spectrum Disorder (ASD), but this approach is both expensive and time-consuming. It is then crucial to identify the most effective supervised machine learning model, taking into account various classification metrics, to address the issue of costly and time-consuming ASD diagnosis.

## **Research Objectives**

### **General Objective**

The general objective of this study is to carry out a comparative analysis of machine learning models in early detection and diagnosis of autism spectrum disorder using behavioral and cultural features.

### **1.3.2 Specific Objectives**

1. To utilize SVM, RF, and ANN in diagnosing Autism Spectrum Disorder.
2. To c­­­ompare the performance of the chosen machine learning models in diagnosing Autism Spectrum Disorder.
3. To deploy the best performing machine learning model using a web-based application.

## **Justification of the Research**

The early detection and diagnosis of Autism Spectrum Disorder in toddlers are crucial for timely interventions and better outcomes. Through evaluating and comparing various machine learning algorithms, this research aims to identify the most effective models for accurately diagnosing ASD in toddlers. This can lead to the development of a standardized and objective screening tool which is time and cost effective. By optimizing the selected models and considering different classification metrics, diagnostic accuracy can be improved, providing reliable results. The study's findings have the potential to revolutionize ASD diagnosis, providing clinicians with a validated and efficient tool for early detection, leading to improved outcomes for toddlers with ASD.

## **1.5 Scope of the Research**

The scope of this research is to optimize and compare machine learning models, namely SVM, RF, and ANN, for early detection and diagnosis of Autism Spectrum Disorder (ASD) in toddlers. The research will utilize behavioral and cultural features, evaluating various classification metrics like accuracy, sensitivity, specificity, and AUC-ROC to assess model effectiveness. Cultural variations in interpreting ASD symptoms will be explored for developing a standardized screening tool. The study also aims to optimize and fine-tune selected models to enhance accuracy and efficiency. Ultimately, the best performing model will be deployed through a web-based application.

# **CHAPTER TWO**

**LITERATURE REVIEW**

## **2.1 Introduction**

This chapter entails the depiction of literature relevant to the study. The chapter will comprise theoretical literature, and overview of literature.

## **2.2 Theoretical Review**

One study by Mellema et al. (2021) optimized and compared the performance of 12 popular and powerful machine learning models for the diagnosis of ASD using neuroimaging features from functional MRI (fMRI) and structural MRI (sMRI). The study aimed to develop reproducible neuroimaging features that can accurately diagnose ASD. The researchers collected neuroimaging data from a large cohort of individuals with ASD and typically developing individuals. They then used machine learning algorithms, such as support vector machines, random forests, and deep learning models, to classify the participants into ASD and control groups based on the neuroimaging features. The results of the study showed that several machine learning models achieved high accuracy in diagnosing ASD using neuroimaging features. The random forest model performed the best, with an accuracy of 87.5%. The study also identified specific neuroimaging features that were most informative for distinguishing between individuals with ASD and typically developing individuals. These features included alterations in functional connectivity patterns, gray matter volume, and cortical thickness in specific brain regions implicated in social communication and sensory processing.

Another study by Wingfield et al. (2020) focused on developing a predictive model for pediatric autism screening. The study highlighted the cultural variations in the interpretation of ASD behavioral symptoms and the need for more accurate screening tools. The researchers collected behavioral data from a diverse sample of children from different cultural backgrounds and used machine learning algorithms to develop a predictive model for ASD. The model incorporated a wide range of behavioral features, such as social interaction, communication skills, and repetitive behaviors. The results of the study showed that the machine learning model achieved high accuracy in predicting ASD in the diverse sample of children. The model was able to identify specific behavioral patterns that were indicative of ASD across different cultural contexts. The study emphasized the importance of considering cultural factors in the development of screening tools and the potential of machine learning models to improve the accuracy of ASD diagnosis in diverse populations.

Sonia et al. (2020) proposed an intelligent framework to predict autism in infants using machine learning. The study emphasized the cost and time efficiency of using machine learning models for early detection of autism. Early identification of ASD in infants is crucial for early intervention and improved outcomes. The proposed framework utilized machine learning algorithms to analyze various data sources, including behavioral observations, parental reports, and genetic data, to predict the likelihood of ASD in infants. The results of the study showed that the machine learning model achieved high accuracy in predicting ASD in infants. The model was able to identify specific patterns in the data that were indicative of ASD, allowing for early intervention and support. The study highlighted the potential of machine learning models to assist healthcare professionals in the early detection of ASD, leading to timely interventions and improved outcomes for infants at risk of developing ASD.

Similarly, Kanchanamala and Sagar (2019) conducted a review of machine learning models for predicting ASD and highlighted the lack of standard diagnosis and treatment for ASD. The review examined various machine learning algorithms and their application in predicting ASD based on different types of data, including neuroimaging, genetic, and behavioral data. The authors emphasized the need for standardized diagnostic criteria and the integration of multiple data sources to improve the accuracy of ASD prediction models. The review identified several machine learning models that achieved high accuracy in predicting ASD. These models utilized different types of data and algorithms, highlighting the importance of a multidimensional approach to ASD prediction. The authors also discussed the challenges and limitations of using machine learning models for ASD prediction, such as the need for large and diverse datasets, interpretability of the models, and ethical considerations.

While the majority of the articles provided are from 2020 or earlier, they still provide valuable insights into the use of machine learning models for the diagnosis and prediction of ASD. For example, Zhou et al. (2014) used multiparametric MRI characterization and machine learning techniques to identify neuroimaging features associated with ASD. The study found alterations in gray matter volume, cortical thickness, and white matter volume in children with ASD compared to typically developing children. The findings highlighted the potential of neuroimaging and machine learning models in understanding the underlying neurobiological mechanisms of ASD.

Based on the urgent need to incorporate machine learning models in healthcare, Modzi et al. (2022) assessed six machine learning algorithms from existing studies to find the best classifier for analyzing the ASD screening training dataset. This study was the continuation of a similar research conducted by Thabtah (2019) who concluded that machine learning allows expedition of ASD screening and diagnosis. They aimed to identify the best classifier for ASD screening. Modzi et al. (2022) utilized classification algorithms that presented optimal accuracy in classifying other diseases. The classification algorithms were as follows: NB, LR, KNN, J48, RF, SVM, and DL. Moreover, they applied the data imputation method to the dataset to cater for missing values and compared the classification algorithms before and after the process was employed.

Modzi et al. (2022) utilized the WEKA modeler environment to measure the performance metrics of each classification algorithm on the ASD dataset with and without missing values to study how both conditions altered the classifier’s predictive power. They did this by comparing the results of ROC plots to classify each algorithm’s specificity and sensitivity. They also used the WEKA experimenter to verify each classifier’s performance based on a corrected paired T-Test mode. Lastly, they produced a confusion matrix to portray the sensitivity, specificity, and accuracy of the best classifier.

J48 emerged as the best algorithm when classifying the ASD screening dataset with or without missing values registering 100% performance in both cases. It also required a short time to classify the dataset (Modzi et al., 2022). Additionally, the J48 classifier registered a 95% of confidence interval when tested with the incomplete dataset using the WEKA experimenter implying that J48 is a powerful classifier when dealing with datasets containing missing values. J48’s confusion matrix illustrated that in both cases, with and without missing values, J48 did not produce any false positive or false negative classification. Also, J48 correctly classified 151 and 141 instances of true positives and true negatives respectively. Modzi et al. (2022) concluded that J48 could be utilized by clinicians to make better decisions while analyzing datasets for efficient and effective results. They also suggested future studies to study the kappa values of each classifier evaluated in their study, utilize other imputation methods to address missing values and compare other classification algorithms (Modzi et al., 2022).

Hassan and Taher (2022) analyzed and studied ASD data among children in the Dohuk governorate, Iraq, using three classification algorithms; DT, ANN, and KNN. Their primary objective was to uncover risk factors that make children susceptible to ASD and to protect other children from developing the infection. Before employing any classification algorithm, the data was preprocessed using three methods. Data normalization was employed to ensure all variables in the dataset contributed equally to model fitting. Moreover, to mitigate dimensionality by only extracting the relevant features, PCA was carried out. Lastly, the data collected was imbalanced with the number of samples in one class being greater than the other class. The SMOTE was applied in the study to eliminate the skewness observed in the distribution of classes (Hassan & Taher, 2022).

After implementing the three classification algorithms, Hassan and Taher (2022) utilized the confusion matrix to assess the performance of each algorithm. In all the algorithms, the data was split into training and testing sets. The evaluation metrics for comparison employed comprised the F-score, precision, specificity, sensitivity, and accuracy. Remarkably, the comparison was carried out on four data types: the original, normalized, (normalized + PCA), and (normalized + PCA + SMOTE) (Hassan & Taher, 2022). Generally, the results obtained indicated a variation in the performance of the three classification algorithms based on the data preprocessing methods employed (normalization, PCA, and SMOTE). In this study, the ANN classifier registered the best classification results for the original data with an F-score of 87.96% and a 92.96% accuracy. Moreover, Hassan and Taher (2022) found that the ANN classifier was the best model overall in terms of accuracy and F1 score. They concluded that the three classification algorithms could be employed in ASD diagnosis.

Dewi and Imah (2020) utilized four classification algorithms for ASD classification in children to find the one with the best performance, based on the results of DSM-5. To assess the performance, each algorithm was assigned various parameters. Dewi and Imah (2020) employed the KNN, SVM, Random Forest, Deep Learning, and Backpropagation algorithms for the classification of two datasets. Performance evaluation was based on sensitivity, specificity, ROC, recall, precision, and accuracy. The best classification algorithm in this study in terms of classification time, kappa statistic, and the value of accuracy was the Random Forest algorithm (Dewi & Imah, 2020). Remarkably, the Random Forest algorithm had a kappa statistic of 1 implying that classifying ASD using the algorithm provides a high level of true confidence. However, Dewi and Imah (2020) also noted that the DL algorithm was not suitable for this study since both datasets used had few features.

According to Maenner et al. (2019), in comparison to Linear Discriminant Analysis, Latent Semantic Analysis, multinomial Naïve Bayes, Support Vector Machine, random forest, NB-SVM, neural networks, and basic hyperparameter optimization models, the random forest or neural network are suitable for conducting a study in a fully automated workflow since they give more prevalence estimates with consideration of individual-level predictive quality. This is because Random Forest and Neural Networks do not sacrifice much in the individual-level predictive quality. The random forest technique stands out as a good model for doing surveillance in all the analyzed models due to its ability to interpret the importance of its features. Moreover, the random forest and neural network automatically produce predicted class probabilities which can aid in supporting the recent surveillance workflow by helping health experts in concentrating on evaluations which may be difficult to evaluate. Additionally, the NB-SVM may be used as an additional tool in doing a review of written evaluations especially in situations where cross validation has been used to obtain the non-thresholded probability estimates that are well calibrated to the true distribution of class levels (Maenner et al., 2019). Although none of the models under their study was able to meet the interrater agreement levels, both random forest and NB-SVM were close whereby they achieved 89% on many train-test cycles. This implies when the health experts are reviewing evaluations, they base their decision of whether a child is autistic or not on more than just the text they contain.

Alkahtani et al. (2023) states that the use of facial landmarks is promising as a tool used for detection of Autism Spectrum Disorder. According to the authors, the use of algorithms that analyze facial landmarks is useful in reducing the gap between autism categorization and facial analysis, thereby making automated autism categorization another method for detection of autism spectrum disorder. This is better in terms of cost and time spent in diagnosis and reduces the chances of a false positive or false negative due to expert prejudice as in the case of traditional ways of hospitalization of individuals with ASD. Also, it is crucial in creating a definite and informative paradigm for the diagnosis of autism spectrum which helps in reducing the challenges that comes with dealing with toddlers. Alkahtani et al. (2023) found that creating this systems for diagnosis can be quite challenging due to the behavioral nature of persons with autism spectrum disorder. They used deep learning models (MobileNet and VGG-16) with the aid of various machine learning models to achieve their objectives where they found that their model of choice achieved a 92% accuracy. In the use of facial landmarks, a single picture may be all that it takes to diagnose whether a child is autistic or otherwise. According to the authors, coming up with a mobile application that can do a facial scan of the child and diagnose autism spectrum disorder will enable even both parents and clinicians to conduct the tests.

Chaitra et al. (2020) investigated the efficacy of FC (Functional Connectivity) and complex network measures in the diagnosis of ASD compared to matched controls in a machine learning framework. Their hypothesis was that a feature set containing both connectivity and graph measures would give a better diagnosis of ASD in terms of accuracy than with the individual sets alone since connectivity and graph measures are partially different. They tested and proved this hypothesis as the combined feature set gave a statistically notable high accuracy than with the contrary. In the study they also noted that complex network features do not produce a better classification accuracy on their own than traditional FC. They also derived from their research that the predictivity of the combined feature set is 3% higher than in diagnosis using conventional FC feature sets. The main issue with their study was the data they used was obtained from different scanners which was not harmonized and also used different imaging protocols which will add undesirable variance in the data obtained.

According to Vakadkar et al. (2020), in current times, clinical standardized tests are the only methods that are being used to diagnose ASD. This not only requires prolonged diagnostic time but also faces a steep increase in medical costs. To improve the precision and time required for diagnosis, the researchers used machine learning techniques to complement the conventional methods. They compared models such as Support Vector Machines (SVM), Random Forest Classifier (RFC), Naïve Bayes (NB), Logistic Regression (LR), and KNN based on various features, such as age, sex, ethnicity, etc., and evaluated each classifier to determine the best-performing model to their dataset and constructed predictive models based on the outcome. Their key objective was to determine if the child is susceptible to neurological disorders such as ASD in its nascent stages, which would help streamline the diagnosis process. From this research, results show that Logistic Regression gives the highest accuracy for the selected dataset.

In another research by Chen et al. (2022), early identification is vital for children with ASD to ensure their access to timely intervention and to optimize long-term outcomes The researchers demonstrated the feasibility of predicting ASD diagnosis at early ages using health claims data and machine learning models. In doing this, they were able to compare the predictive power of the Random Forest model and Linear Regression models. The results proved that LR and RF models achieved an overall AUROC (area under the receiver operating characteristic) above 0.75 when predicting ASD diagnosis at the age of 24 months. The results also showed that prediction performance increased with age at the time of prediction. This is reasonable because more clinical information accumulated over a longer follow-up period since birth may contain more distinctive patterns to effectively differentiate children with ASD. The study highlighted two limitations; diagnosis of ASD established only based on existing diagnosis codes from claims data could be inaccurate and unreliable sometimes in practice. Also, the absence of ASD diagnosis codes in one’s health record may not necessarily indicate an individual not having ASD, especially for children born in later years, due to limited follow-up time prior to the cut-off date in the database. The model provided a limited value for individuals who do not present comorbid conditions from past healthcare encounter data.

A common study by Hasan et al. (2022) proposed a machine-learning framework for ASD detection in people of different ages (Toddlers, Children, Adolescents, and Adults). The main aim of the research was to create an effective prediction model using different types of ML methods in early detection of autism. After completing the initial data processing, those ASD datasets were scaled using four different types of feature scaling (Quantile Transformer, Power Transformer, Max Abs Scaler) techniques, classified using eight different ML classifiers Ada Boost (AB), Random Forest (RF), Decision Tree (DT), K-Nearest Neighbors (KNN), Gaussian Naïve Bayes (GNB), Logistic Regression (LR), Support Vector Machine (SVM) and Linear Discriminant Analysis (LDA). The researchers then analyzed each feature-scaled dataset’s classification performance and identified the best-performing FS and classification approaches. They considered different statistical evaluation measures such as accuracy, ROC, F1-Score, precision, recall, Mathew’s correlation coefficient (MCC), kappa score, and Log loss to justify the experimental findings. After analyzing the experimental outcomes of different classifiers on feature-scaled ASD datasets, results showed that AB predicted ASD with the highest accuracy of 99.25%, and 97.95% for Toddlers and Children, respectively and LDA predicted ASD with the highest accuracy of 97.12% and 99.03% for Adolescents and Adults datasets, respectively.

Another study by Wang and Avillach, (2021) generated a genetic diagnostic classifier (Deep Autism) based on a deep learning architecture using 100 significant common variants and accurately distinguished ASD from controls within the SSC data set. The diagnostic classifier was able to correctly classify individuals with ASD with an accuracy of 88.6% and an AUC (area under the receiver operating characteristic curve) of 0.955. Their findings showed that the sensitivity and specificity of the classifier when applied to identify ASD were 88% and 89%, respectively. It is notable that the sensitivity for identifying cases is highly desirable for screening purposes. They also investigated the classification performance of different approaches and the corresponding proportion of subjects who did not have ASD who could be reliably classified as controls. From the results, Deep Autism can be suggested as an alternative to conventional shallow machine-learning approaches. In the comparisons among the classifiers, Deep Autism performed the best, followed by random forest. Both these classifiers are nonlinear models. Therefore, the causes of ASD are not a simple linear combination of common variants.

Further on this study, Wang and Avillach (2021) employed Naive Bayes, logistic regression, support vector machine, random forest, and deep neural network classifiers to compare the prediction of ASD diagnosis. They applied five-fold cross-validation to evaluate the selected significant common variants. The classifier performed better than the conventional machine learning techniques in terms of AUC, accuracy, specificity, sensitivity, and F1 score. Accuracy was 0.886 in the case of DeepAutism, followed by 0.808 for the random forest in the same test data set for ASD diagnosis prediction. DeepAutism also yielded the best sensitivity of 0.881 for the prediction of ASD and the best specificity of 0.893 for non-ASD prediction. The false positive (discriminatory) rate is minimum for DeepAutism at 7% compared with other machine learning techniques.

In another research, Masum et al. (2021) applied traditional ML classifiers (e.g., decision tree classifier, random forest classifier, naïve Bayes classifier, and logistic regression classifier) and neural network-based architecture to classify ASD. In the post-processing, they applied several feature importance techniques such as Random Forest-based importance, Permutation-based importance, and Shap value-based importance to investigate the important features in ASD prediction. Results show that Decision Tree and Random Forest classifier outperform other models by achieving the highest accuracy, F-beta, Recall, and precision scores for all the datasets. On the other hand, the Naïve Bayes classifier provides poor performance in terms of all performance metrics compared to other methods. Both DT and RF provide 100% accuracy, F-beta, recall, and precision score along with 0.00 standard deviation.

Lee et al. (2019) applied 8 supervised learning algorithms to predict whether children meet the case definition for ASD based solely on the words in their evaluations. They compared the algorithms’ performance across 10 random train-test splits of the data, using classification accuracy, F1 score, and a number of positive calls to evaluate their potential use for surveillance. Across the 10 train-test cycles, the random forest and support vector machine with Naive Bayes features (NB-SVM) each achieved slightly more than 87% mean accuracy. The NB-SVM produced significantly more false negatives than false positives (*P* = 0.027), but the random forest did not, making its prevalence estimates very close to the true prevalence in the data. The best-performing neural network performed similarly to the random forest on both measures. The random forest performed as well as more recently available models like the NB-SVM and the neural network, and it also produced good prevalence estimates. NB-SVM may not be a good candidate for use in a fully-automated surveillance workflow due to increased false negatives. More sophisticated algorithms, like hierarchical convolutional neural networks, may not be feasible to train due to the characteristics of the data. Current algorithms might perform better if the data are abstracted and processed differently and if they take into account information about the children in addition to their evaluations.

## **2.3 Overview of the Literature Review**

In summary, the use of machine learning models for the diagnosis and prediction of Autism Spectrum Disorders is a rapidly evolving field of research. These models have the potential to improve the accuracy and efficiency of screening and diagnosis, leading to earlier intervention and improved outcomes for individuals with ASD. The studies discussed in this literature review demonstrate the effectiveness of machine learning models in integrating neuroimaging and behavioral data to accurately diagnose and predict ASD. However, further research is needed to validate and optimize these models for clinical use. Additionally, the development of standardized diagnostic criteria and the consideration of cultural factors are important for the successful implementation of machine learning models in ASD diagnosis and prediction.

# **CHAPTER THREE**

**RESEARCH METHODOLOGY**

## **3.1 Introduction**

This chapter outlines the methodology employed in the study, which aims to determine the best-performing machine learning model for the detection and diagnosis of Autism Spectrum Disorder using behavioral and cultural features. It entails a description of the research design, data source, data selection criteria, data preprocessing, feature extraction, optimization and model selection.

## **3.2 Data Source and Type**

The data source for this study is a publicly available dataset named "Autistic Spectrum Disorder Screening Data for Toddlers." This dataset was created by Dr. Fadi Thabtah from the Department of Digital Technology at Manukau Institute of Technology in Auckland, New Zealand. The dataset was developed to address the urgent need for easily implemented and effective screening methods for Autism Spectrum Disorder among toddlers.

## **3.3 Data Preprocessing**

Data preprocessing is a crucial step in preparing the dataset for analysis. It involves various techniques to ensure data quality, handle missing values, and standardize the data before applying machine learning algorithms. The following sub-topics outline the key steps involved in data preprocessing for this study

### **3.3.1 Data Cleaning**

Data cleaning aims to identify and address any inconsistencies, errors, or outliers present in the dataset. In this study, the dataset will be thoroughly examined for missing values in each variable. If missing values are identified, various strategies will be implemented to handle them. For numerical variables such as age or the score by Q-Chat-10, missing values will be imputed using techniques such as mean imputation, median imputation, or regression imputation, depending on the distribution and nature of the data. Categorical variables, such as “sex”, “ethnicity”, or “who is completing the test”, will be handled by creating an additional category for missing values or using specific imputation techniques tailored for categorical data.

### **3.3.2 Feature Selection**

Feature selection is the process of identifying the most relevant features from the dataset that contribute significantly to the prediction task. In this study, the feature selection process will be iterative and data-driven, combining statistical tests, model-based rankings, regularization techniques, and expert knowledge. The performance of the selected feature set will be evaluated using appropriate validation techniques, such as cross-validation, to ensure that the chosen features generalize well to unseen data. The final selected features will serve as the input for the subsequent model selection, optimization, and evaluation phases, allowing for the identification of the best-performing machine learning model for ASD detection and diagnosis using behavioral and cultural features.

### **3.3.3 Encoding Categorical Variables**

Categorical variables often require encoding to convert them into a numerical representation suitable for machine learning algorithms. This step ensures that categorical features can be effectively utilized by the machine learning models. Binary variables in the dataset are already represented as "yes" or "no" so they will be retained in their original binary representation. Nominal categorical variables in the dataset include "sex" (Male or Female) and "ethnicity" (a list of common ethnicities). These variables will be encoded using one-hot encoding. Each category within the variables will be transformed into a separate binary variable, where a value of 1 will indicate the presence of that category, and 0 will indicate its absence.

### **3.3.4 Normalization**

Normalization or scaling is performed to standardize the numerical features in the dataset. This step aims to bring all the features to a similar scale, preventing any bias that may arise from differences in magnitude. In this study, min-max scaling will be applied to achieve standardized feature values. The dataset includes continuous variables such as "age" and "score by Q-Chat-10. Min-max scaling will transform the values to a specific range, typically between 0 and 1, preserving the relative relationships among the values.

### **3.3.5 Handling Imbalance**

Imbalanced data refers to a situation where the classes in the dataset are not represented equally, leading to biased predictions. We will first conduct Exploratory Data Analysis to visualize the distribution of the target variable (ASD traits or No ASD traits) to assess the extent of the imbalance. In case of imbalance, we will employ techniques such as oversampling the minority class (e.g., Synthetic Minority Over-sampling Technique - SMOTE) or undersampling the majority class to balance the class distribution. This ensures that the machine learning models can learn from both positive and negative instances effectively.

### **3.3.6 Data Splitting**

To evaluate the performance of the machine learning models properly, the dataset is typically split into training and testing sets. The dataset will be split into three subsets: a training set (70%), a validation set (10%), and a testing set (20%). The training set will be used to train the models, the validation set will be used for model selection and hyperparameter tuning, and the testing set will be used to assess the final performance of the models.

## **3.4 Modelling**

The modelling phase is a critical step in the research process. Here, machine learning models will be trained using the preprocessed and cleaned dataset from the steps 3.4. The chosen machine learning models that will be trained are SVM, RF, and ANN. The following approach will be taken for modelling:

### **3.4.1 Support Vector Machine (SVM)**

SVM is a powerful classification algorithm that aims to find the best hyperplane to separate the data points into different classes. However, it is primarily employed for classification tasks (SVM Algorithm, n.d.). The SVM algorithm creates the best decision boundary or line – called a hyperplane – that segregates n-dimensional space into classes where we can categorize new data points in future. The aim of the SVM algorithm is to choose the extreme vectors or points – also called support vectors – that assist in developing the hyperplane; hence the name Support Vector Machine (Saini, 2023). The diagram below illustrates two distinct classes partitioned using a hyperplane.



Figure 1 Distinct classes partitioned using hyperplane

However, there could be several hyperplanes giving 100% accuracy which implies that we have to find the hyperplane with the maximum margin, known as the optimal hyperplane. This is the main aim of the SVM algorithm (Saini, 2023). Margin is the distance between the hyperplane and support vectors. Significantly, support vectors are data points closest to the hyperplane from both classes that alter the hyperplane’s position (Saini, 2023). Consider *n* training examples where each example *x* has two classes , D dimension, and hyperplane , the decision rule will be:

-------------------------------------------------------------------------- (3.1)

We then compute the minimum value of:

-------------------------------------------------------------------------- (3.2)

And obtain the optimal hyperplane with these conditions:

------------------------------------------------------------------------- (3.3)

Next, we calculate the decision function to determine the data category:

----------------------------------------------------------------- (3.4)

Where is a kernel function, is each data’s weight, and *m* represents many support vectors (Dewi & Imah, 2020). In this study, we will employ the scikit-learn library which contains the necessary modules for SVM classification.

While modelling an SVM algorithm, we consider some kernel functions that allow us to manipulate the data. Major kernel functions for classification tasks include the linear kernel employed when data is linear separable, Gaussian Kernel Radial Basis Function, that performs transformation when there is no prior knowledge about data, and polynomial kernel, that represents the similarity of vectors in the training set in a feature space over polynomials of the original variables used in the kernel. In this study, experimentation will be used to determine the most suitable kernel.

### **3.4.2 Random Forest (RF)**

Random Forest is an ensemble learning algorithm that combines multiple decision trees to make predictions. Random forest leverages the power of multiple models to provide robust and reliable results. The algorithm builds multiple decision trees by randomly selecting subsets of the original dataset, a process known as bootstrapping.

Mathematically, for Random Forest model's classification, the prediction can be represented as:

ȳ = mode(ȳ₁, ȳ₂, ..., ȳₙ) ------------------------------------------------------------------------- (3.5)

where ȳ represents the final predicted class, ȳⱼ is the predicted class of the j-th decision tree, and mode() is the function that returns the most common (majority) class among the predictions.

For regression tasks, the prediction can be represented as:

ȳ = (ȳ₁ + ȳ₂ + ... + ȳₙ) / n ------------------------------------------------------------------------- (3.6)

where ȳ represents the final predicted value, ȳⱼ is the predicted value of the j-th decision tree, and n is the total number of decision trees in the Random Forest.

In this study, we will train each decision tree on different subsets of the data, which introduces randomness and diversity into the forest. For each decision tree in the forest, the algorithm will select a random subset of features at each node to make a split based on the selected features and their corresponding thresholds. This random feature selection will aid in reducing correlation among the trees. Training of each decision tree will be done using a recursive process called recursive partitioning. This process continues until a stopping criterion is met. Once all the decision trees are constructed, they will be used to make predictions. For classification tasks, the class that receives the majority of votes from the trees is chosen as the final prediction. Below is a diagram showing the working of a random Forest algorithm.

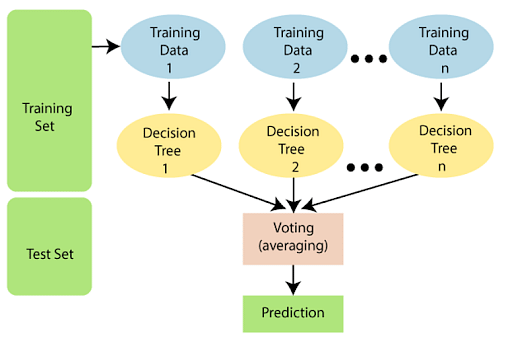


Figure 2 Random Forest Tree presentation

### **3.4.3 Artificial Neural Network**

An artificial neural network (ANN) is a computational model that attempts to account for the parallel nature of the human brain. An (ANN) is a network of highly interconnecting processing elements (neurons) operating in parallel. These elements are inspired by biological nervous systems. As in nature, the connections between elements largely determine the network function. A subgroup of processing element is called a layer in the network. The first layer is the input layer and the last layer is the output layer. Between the input and output layer, there may be additional layer(s) of units, called hidden layer(s). Figure 3 represents a typical neural network. A neural network can be trained to perform a particular function by adjusting the values of the connections (weights) between element.

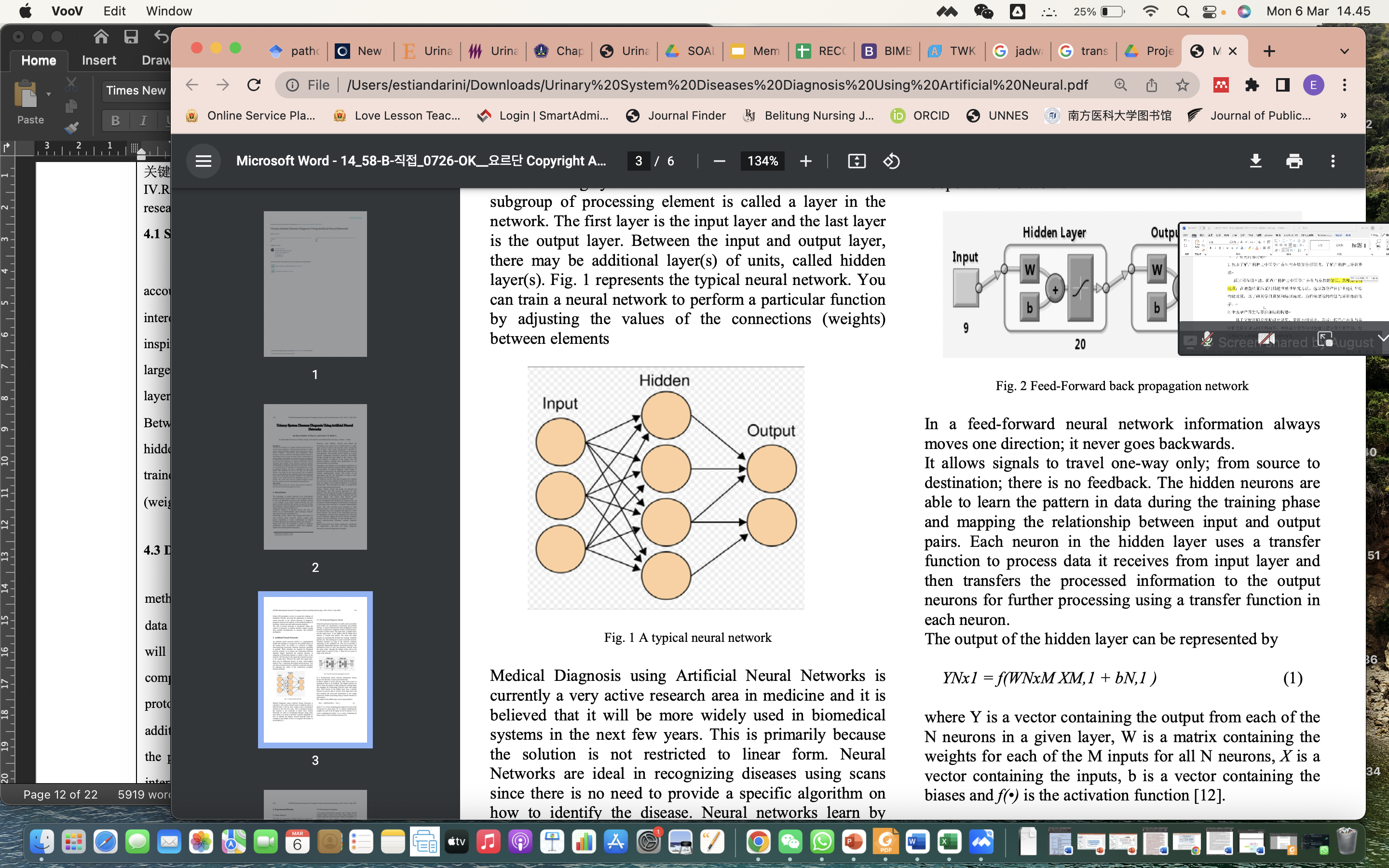


Figure 3 A typical neural network

In the case of ASD detection and diagnosis, ANN can be trained on the dataset using the behavioral and cultural features. The mathematical representation of ANN involves the computations performed in each neuron and the forward and backward propagation of data through the network for training. Let X and Y be defined as before. The mathematical representation of a single neuron in an ANN can be expressed as follows:

z = w^T \* X + b ------------------------------------------------------------------------- (3.7)

a = g(z) ------------------------------------------------------------------------- (3.8)

Here, w represents the weight vector, b is the bias term, X is the input feature vector, z is the weighted sum of inputs and biases, a is the activation function applied to z, and g() is the activation function.

Feed-forward neural networks are widely and successfully used models for classification, forecasting and problem solving. A typical feed-forward back propagation neural network is proposed to detection and diagnosis of ASD. It consists of three layers: the input layer, a hidden layer, and the output layer. The input and target samples are automatically divided into training, validation and test sets as explained in part 3.3.6. The training set is used to teach the network. Training continues as long as the network continues improving on the validation set. The test set provides a completely independent measure of network accuracy. The information moves in only one direction, forward, from the input nodes, through the hidden nodes and to the output nodes. In a feed-forward neural network information always moves in one direction; it never goes backwards. It allows signals to travel one-way only; from source to destination; there is no feedback. The hidden neurons are able to learn the pattern in data during the training phase and mapping the relationship between input and output pairs. Each neuron in the hidden layer uses a transfer function to process data it receives from the input layer and then transfers the processed information to the output neurons for further processing using a transfer function in each neuron.

## **3.5 Evaluation Metrics**

The performance of the optimized machine learning models will be evaluated using various classification metrics, including Accuracy, Precision, Recall (Sensitivity), F-1 Score, Specificity, and the area under the receiver operating characteristic curve (AUC-ROC). These metrics will assess the models' ability to effectively detect and diagnose ASD using the behavioral and cultural features.

Accuracy is the proportion of correctly classified instances to the total number of instances.

Precision is the proportion of true positive predictions to the total number of positive predictions.

Recall (Sensitivity or True Positive Rate) is the proportion of true positive predictions to the total number of actual positive instances.

The F1 score is the harmonic mean of precision and recall, providing a balance between the two metrics.

Specificity (True Negative Rate) is the proportion of true negative predictions to the total number of actual negative instances.

## **3.6 Model Deployment**

The best performing machine learning model from the previous section will be deployed into a web-based application. The application will provide a user interface which prompts the user to enter behavioral and cultural factors. The predicted result is then displayed to the user.

# **CHAPTER FIVE**

**CONCLUSION AND RECOMMENDATIONS**

## **5.1 Summary of Findings**

Summarize the key results from Chapter Four.

## **5.2 Implications**

Discuss the implications of your findings.

Relate your results to the broader field of study.

## **5.3 Recommendations**

Provide recommendations based on your results.

Suggest practical applications or policy implications.

## **5.4 Future Work**

Identify avenues for future research.

Discuss questions your study raises that could be explored further.

## **5.5 Conclusion**

Summarize the main points.

Conclude your research journey.

**References**

Alkahtani, H., Aldhyani, T. H., & Alzahrani, M. Y. (2023). Deep Learning Algorithms to Identify Autism Spectrum Disorder in Children-Based Facial Landmarks. *Applied Sciences*, *13*(8), 4855.

Arbabshirani, M., Plis, S., Sui, J., Calhoun, V. (2017). Single Subject Prediction Of Brain Disorders In Neuroimaging: Promises and Pitfalls. NeuroImage, (145), 137-165. https://doi.org/10.1016/j.neuroimage.2016.02.079

Bohr, A., & Memarzadeh, K. (2020). The rise of artificial intelligence in healthcare applications. In *Artificial Intelligence in healthcare* (pp. 25-60). Academic Press.

Bone, D., Bishop, S., Black, M., Goodwin, M., Lord, C., Narayanan, S. (2016). Use Of Machine Learning To Improve Autism Screening and Diagnostic Instruments: Effectiveness, Efficiency, And Multi‐instrument Fusion. J Child Psychol Psychiatr, 8(57), 927-937. <https://doi.org/10.1111/jcpp.12559>.

Chaitra, N., Vijaya, P. A., & Deshpande, G. (2020). Diagnostic prediction of autism spectrum disorder using complex network measures in a machine learning framework. *Biomedical Signal Processing and Control*, *62*, 102099.

Chen, Y. H., Chen, Q., Kong, L., & Liu, G. (2022). Early detection of autism spectrum disorder in young children with machine learning using medical claims data. *BMJ Health & Care Informatics*, *29*(1), e100544.

Dewi, E. S., & Imah, E. M. (2020, November). Comparison of machine learning algorithms for autism spectrum disorder classification. In International joint conference on science and engineering (IJCSE 2020) (pp. 152-159). Atlantis Press.

ElNakieb, Y., Ali, M., Elnakib, A., Shalaby, A., Mahmoud, A., Soliman, A., … & El-Baz, A. (2023). Understanding the Role Of Connectivity Dynamics Of Resting-state Functional Mri In The Diagnosis Of Autism Spectrum Disorder: A Comprehensive Study. Bioengineering, 1(10), 56. <https://doi.org/10.3390/bioengineering10010056>.

Gong, M. (2021). A novel performance measure for machine learning classification. *International Journal of Managing Information Technology (IJMIT) Vol*, *13*. <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3807764>.

Hasan, S. M., Uddin, M. P., Al Mamun, M., Sharif, M. I., Ulhaq, A., & Krishnamoorthy, G. (2022). A Machine Learning Framework for Early-Stage Detection of Autism Spectrum Disorders. *IEEE Access*.

Hassan, M. M., & Taher, S. A. (2022). Analysis and Classification of Autism Data Using Machine Learning Algorithms. Science Journal of University of Zakho, 10(4), 206-212.

Kim, J. Y., Son, M. J., Son, C. Y., Radua, J., Eisenhut, M., Gressier, F., ... & Fusar-Poli, P. (2019). Environmental risk factors and biomarkers for autism spectrum disorder: an umbrella review of the evidence. The Lancet Psychiatry, 6(7), 590-600.

Lee, S. H., Maenner, M. J., & Heilig, C. M. (2019). A comparison of machine learning algorithms for the surveillance of autism spectrum disorder. *PloS one*, *14*(9), e0222907.

Lee, S. H., Maenner, M. J., & Heilig, C. M. (2019). A comparison of machine learning algorithms for the surveillance of autism spectrum disorder. *PloS one*, *14*(9), e0222907.

Mandal, A. (2019, February 26). Autism History. *News-Medical.Net*. Retrieved July 3, 2023, from <https://www.news-medical.net/health/Autism-History.aspx>

Masum, M., Nur, I., Faruk, M. H., Adnan, M., & Shahriar, H. (2022). A comparative study of machine learning-based autism spectrum disorder detection with feature importance analysis. In *COMPSAC 2022: Computer Software and Applications Conference* (Vol. 3).

Mellema, C., Nguyen, K., Treacher, A., Montillo, A. (2021). Reproducible Neuroimaging Features For Diagnosis Of Autism Spectrum Disorder With Machine Learning. <https://doi.org/10.1101/2021.10.21.21265162>.

Mohd Radzi, S. F., Hassan, M. S., & Mohd Radzi, M. A. H. (2022). Comparison of classification algorithms for predicting autistic spectrum disorder using WEKA modeler. BMC medical informatics and decision making, 22(1), 1-15.

Nagesh, N., Patil, P., Patil, S. (2022). An Architectural Framework For Automatic Detection Of Autism Using Deep Convolution Networks and Genetic Algorithm. IJECE, 2(12), 1768. <https://doi.org/10.11591/ijece.v12i2.pp1768-1775>.

Otsimo. (2023, March 27). History of Autism: When Was Autism First Diagnosed? *Special needs*. Retrieved July 3, 2023, from <https://otsimo.com/en/history-of-autism-when-was-autism-first-diagnosed/>

Pandian, S. (2022). A Comprehensive Guide on Hyperparameter Tuning and its Techniques. *Analytics Vidhya*. <https://www.analyticsvidhya.com/blog/2022/02/a-comprehensive-guide-on-hyperparameter-tuning-and-its-techniques/>

P, K., Sagar, G. (2019). A Review Of Machine Learning Models For Predicting Autism Spectrum Disorder. helix, 1(9), 4797-4801. <https://doi.org/10.29042/2019-4797-4801>.

Prova, A., Ali, M., Ali, M. (2021). Prediction Of Autism Spectrum Disorder Using Feature Selection and Machine Learning Algorithms., 133-148. <https://doi.org/10.1007/978-981-16-4103-9_12>.

Rahman, M. M., Usman, O. L., Muniyandi, R. C., Sahran, S., Mohamed, S., & Razak, R. A. (2020). A review of machine learning methods of feature selection and classification for autism spectrum disorder. Brain sciences, 10(12), 949.

Safara, F. (2021). Autism Spectrum Diagnosis Using Adaptive Learning Algorithm For Multiple Mlp Classifier. JISIoT, 33-44. <https://doi.org/10.54216/jisiot.020201>.

Saini, A. (2023). Guide on Support Vector Machine (SVM) Algorithm. *Analytics Vidhya*. <https://www.analyticsvidhya.com/blog/2021/10/support-vector-machinessvm-a-complete-guide-for-beginners/#:~:text=optimize%20in%20SVM.-,Margin%20in%20Support%20Vector%20Machine,and%20b%20is%20an%20offset.&text=If%20the%20value%20of%20w,it%20is%20a%20negative%20point>.

Senapati, B. (2023). A Machine Learning Model to Predict the ASD Traits and Challenging Behaviors in Children (Doctoral dissertation, University of Arkansas at Little Rock).

Smith, M., MA. (2023). Autism Treatments, Therapies, and Interventions. HelpGuide.org. <https://www.helpguide.org/articles/autism-learning-disabilities/autism-treatments-therapiesinterventions.htm#:~:text=Common%20autism%20treatments%20include%20behavior,way%20that%20can%20be%20sustained>.

Sonia\*, S. (2020). Intelligent Frame Work To Predict Autism In Infants Using Machine Learning. IJRTE, 1(9), 675-679. <https://doi.org/10.35940/ijrte.a1997.059120>.

Support Vector Machine (SVM) Algorithm - Javatpoint. (n.d.). www.javatpoint.com. <https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm>

Talukdar, J., Gogoi, D. K., & Singh, T. P. (2023). A comparative assessment of most widely used machine learning classifiers for analysing and classifying autism spectrum disorder in toddlers and adolescents. Healthcare Analytics, 3, 100178.

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

Thabtah, F. (2019). Machine learning in autistic spectrum disorder behavioral research: A review and ways forward. *Informatics for Health and Social Care*, *44*(3), 278-297.

Vakadkar, K., Purkayastha, D., & Krishnan, D. (2021). Detection of autism spectrum disorder in children using machine learning techniques. *SN Computer Science*, *2*, 1-9.

Vujović, Ž. (2021). Classification model evaluation metrics. International Journal of Advanced Computer Science and Applications, 12(6), 599-606. <https://www.researchgate.net/profile/ZeljkoVujovic/publication/352902406_Classification_Model_Evaluation_Metrics/links/60dede1592851ca9449f17bb/Classification-Model-Evaluation-Metrics.pdf>.

Wang, H., & Avillach, P. (2021). Diagnostic classification and prognostic prediction using common genetic variants in autism spectrum disorder: Genotype-based deep learning. *JMIR medical informatics*, *9*(4), e24754.

Wingfield, B., Miller, S., Yogarajah, P., Kerr, D., Gardiner, B., Seneviratne, S., … & Coleman, S. (2020). A Predictive Model For Paediatric Autism Screening. Health Informatics J, 4(26), 2538-2553. <https://doi.org/10.1177/1460458219887823>.

Zhou, Y., Yu, F., Duong, T. (2014). Multiparametric MRI Characterization and Prediction In Autism Spectrum Disorder Using Graph Theory And Machine Learning. PLoS ONE, 6(9), e90405. <https://doi.org/10.1371/journal.pone.0090405>.

Breiman, L. (2001). Random forests. Machine learning, 45, 5-32.

Biau, G., & Scornet, E. (2016). A random forest guided tour. Test, 25, 197-227.

Jeon, Y., & Lin, Y. (2006). An effective method for high-dimensional log-density ANOVA estimation, with application to nonparametric graphical model building. Statistica Sinica, 353-374.

Biau, G., & Scornet, E. (2016). A random forest guided tour. Test, 25, 197-227.

# **CHAPTER FOUR**

**RESULTS AND DISCUSSION**

## **4.1 Introduction**

This section unfolds the findings of our in-depth exploration into machine learning models for early Autism Spectrum Disorder (ASD) detection among toddlers. In this chapter, we present the empirical results derived from our model training and evaluation. The subsequent discussions delve into the nuances of each model's performance, shedding light on their respective strengths and limitations. Through an evidence-based comparison, we aim to pinpoint the most effective model in the context of ASD detection, thereby contributing valuable insights to early detection of Autism Spectrum Disorder.

## **4.2 Descriptive Statistics**

The analyzed data provided insights about several variables. Descriptive statistics shows the spread of the numerical variables in the dataset.

4.2.1 Summary Statistics

In this section, we present summary statistics for the key variables involved in the assessment of machine learning models for early detection and diagnosis of Autism Spectrum Disorder (ASD).

**4.2.1.1Demographic Characteristics**

* **Age at Assessment:**

The age distribution of participants in the study ranged from 12 to 36 months, with an average age of approximately 27.87 months. The majority of participants fall within the 23 to 36 months age range, as indicated by the interquartile range (IQR).A standard deviation of 7.98 months indicated moderate variability.

**Participant Characteristics**

* **Gender Distribution:**

The study comprised 735 male participants and 319 female participants, reflecting a higher representation of males.

* **History of Jaundice**

A majority of participants (766) reported no history of jaundice, while 288 participants indicated a history of jaundice.

* **Family Members with ASD:**

Among the participants, 884 reported no family history of ASD, while 170 participants had family members with ASD.

* **Test Completion by:**

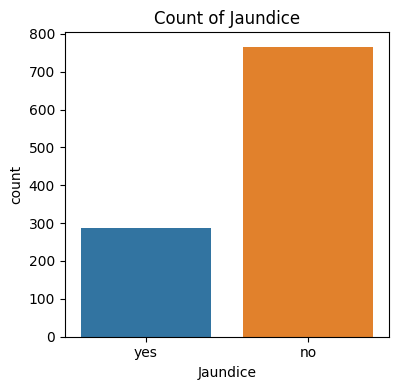
The majority of assessments (1018) were completed by family members, with a smaller number completed by health care professionals(29), individuals themselves(4), or others(3)

* **ASD Traits Identified:**

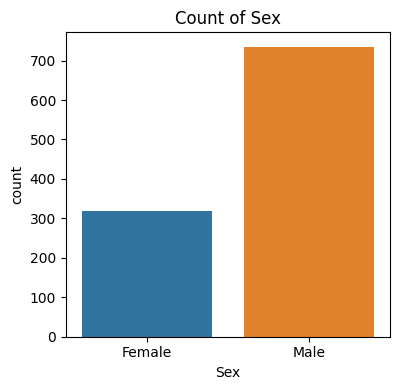
Of the participants, 728 exhibited traits associated with ASD, while 326 did not show such traits.

4.2.2 Data Distribution

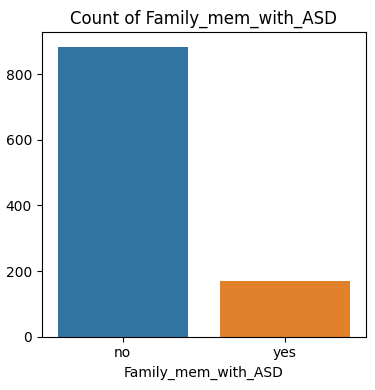
**Bar Chart of Jaundice History distribution**



The bar chart displays the distribution of participants with and without a history of jaundice. The majority (73%) report no jaundice history (766), while 27% indicate a previous history (288). Understanding the distribution of this variable is important for exploring potential correlations between jaundice and the manifestation of Autism Spectrum Disorder traits.



The Bar plot visually represents the gender distribution, indicating that 70% of the participants are male (735) and 30% are female (319). This distribution is essential to consider when evaluating gender-related patterns in the context of Autism Spectrum Disorder, potentially influencing the interpretation of diagnostic outcomes.



This bar chart visualizes the distribution of participants with and without family members diagnosed with Autism Spectrum Disorder. A significant proportion (84%) report no family history of ASD (884), while 16% have family members with the condition (170). This distribution aids in understanding the familial context, which may influence the heritability and presentation of ASD traits.

**Discussion**

In this section, the dataset's descriptive statistics shed light on key demographic and contextual attributes of the study participants. The average age at assessment was found to be approximately 27.87 months, with a notable standard deviation of 7.98 months, indicating significant variability. The interquartile range (IQR) spanning from the 25th to the 75th percentiles reveals a concentration of participants between 23 and 36 months. Gender distribution highlights a higher representation of males, constituting 735 participants compared to 319 females. Moreover, a substantial portion of participants reports no history of jaundice (766), and the majority does not have family members diagnosed with Autism Spectrum Disorder (ASD) (884). Notably, family members predominantly completed the assessments (1018), underscoring the role of familial involvement in the diagnostic process. The prevalence of ASD traits is evident, with 728 participants exhibiting such traits. These findings underscore the importance of considering demographic and familial factors in the early detection and diagnosis of ASD, emphasizing the need for tailored interventions and support strategies based on a comprehensive understanding of participant characteristics

## **4.3 Data Preprocessing Results**

[Content here…]

4.3.1 Handling Missing Values

* Discuss the efficacy of employed strategies for addressing missing values in numerical and categorical variables.

4.3.2 Feature Selection Outcomes

* Expound on the results of the feature selection process, emphasizing the significance of selected features.

## **4.4 Model Performance Results**

[Content here…]

4.4.1 Support Vector Machine (SVM)

* Provide a detailed account of SVM model performance, backed by relevant statistical metrics and visual representations.

4.4.2 Random Forest (RF)

* Elaborate on RF model outcomes, emphasizing classification metrics, and elucidating any observed nuances.

4.4.3 Artificial Neural Network (ANN)

* Present comprehensive insights into the performance of the ANN model, substantiating findings with statistical rigor.

## **4.5 Feature Importance Analysis Results**

* Articulate the outcomes of feature importance analysis, delineating the pivotal role of specific features in predictive accuracy.
* Use visualizations or tables to accentuate the prominence of identified features.

## **4.6 Discussion of Results**

4.6.1 Interpretation

* Systematically interpret the results vis-à-vis the stipulated research questions, addressing both anticipated and unanticipated findings.

4.6.2 Implications

* Discuss the broader implications of the results for the field of Autism Spectrum Disorder detection and diagnosis.

4.6.3 Comparison with Existing Literature

* Establish connections between obtained results and extant literature, corroborating or challenging prevailing insights.

## **4.7 Model Comparison**

* Undertake a robust comparative analysis of SVM, RF, and ANN models, employing statistical tests to discern disparities in performance.
* Discern the unique strengths and limitations of each model.

## **4.8 Model Deployment**

Content here

**4.9 Limitations**

* Conscientiously delineate the limitations inherent in the study, acknowledging potential biases and constraints.
* Provide insights into how these limitations may have influenced the interpretative landscape.

**Top of Form**

# **Budget**

Table 1 Budget

|  |  |  |  |
| --- | --- | --- | --- |
| **Item name** | **Unit** | **Cost (Ksh.)** | **Total (Ksh.)** |
| Printing | 5 | 500 | 2500 |
| Internet |  | 2500 | 2500 |
| Transport | 4 | 500 | 2000 |
| Laptop & Stationery |  | 60000 | 60000 |
| Miscellaneous | 4 | 300 | 1200 |
| **Total** |  |  | **68200** |

# **Work Plan**

Table 2 Work plan

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Date**  **Activity** | **May** | **June** | **June**  **–**  **July** | **27th July** | **July**  **–**  **Sept** | **Sept**  **Nov** | **Nov** | **1st Dec** |
| Preliminary work |  |  |  |  |  |  |  |  |
| Proposal writing |  |  |  |  |  |  |  |  |
| Proposal presentation |  |  |  |  |  |  |  |  |
| Data Analysis |  |  |  |  |  |  |  |  |
| ML modelling |  |  |  |  |  |  |  |  |
| Web development |  |  |  |  |  |  |  |  |
| Documentation |  |  |  |  |  |  |  |  |
| Project presentation |  |  |  |  |  |  |  |  |