AUTISM SPECTRUM DISORDER PREDICTION USING MACHINE LEARNING

A MINI PROJECT REPORT

Submitted by

BALAMURUGAN M **(2116220701516)**

for the award of the degree of

BACHELOR OF ENGINEERING

IN

COMPUTER SCIENCE AND ENGINEERING

RAJALAKSHMI ENGINEERING COLLEGE

ANNA UNIVERSITY, CHENNAI

APRIL 2025

RAJALAKSHMI ENGINEERING COLLEGE, CHENNAI

BONAFIDE CERTIFICATE

Certified that this project titled **“Autism Spectrum Disorder Prediction using Machine Learning”** is the bonafide work of **“BALAMURUGAN M (2116220701516)”** who carried out the work under my supervision. Certified further that to the best of my knowledge the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

SIGNATURE
Dr. V. Auxilia Osvin Nancy, M.Tech., Ph.D.
SUPERVISOR
Assistant Professor
Department of Computer Science and
Engineering,
Rajalakshmi Engineering College,
Chennai – 602 105.

Submitted for the Mini Project Viva-Voce Examination held on _________________

INTERNAL EXAMINER **EXTERNAL EXAMINER**

ABSTRACT

This project explores the application of machine learning algorithms to predict Autism Spectrum Disorder (ASD), a developmental disorder characterized by challenges in social interaction, communication, and behavior. Traditional diagnostic methods for ASD are often time-consuming and rely heavily on the expertise of clinicians, which can lead to delays in diagnosis and subsequent interventions. To address these challenges, we leverage machine learning to create predictive models that can analyze complex datasets efficiently and accurately.

Our approach involves utilizing a dataset that includes demographic, behavioral, and clinical information. This diverse dataset allows us to capture a comprehensive view of the factors that may be indicative of ASD. Key machine learning algorithms, such as Support Vector Machines (SVM), Random Forests, and Neural Networks, were selected for their ability to handle high-dimensional data and uncover intricate patterns within the dataset. Each of these algorithms has distinct strengths: SVM excels in finding hyperplanes that separate classes, Random Forests are robust in handling overfitting and providing feature importance insights, and Neural Networks are powerful in modeling non-linear relationships.

The methodology begins with data collection and preprocessing, ensuring the data is clean and normalized. Feature selection techniques are then employed to identify the most relevant variables for predicting ASD. Following this, we develop and train multiple machine learning models, each tailored to maximize predictive accuracy. The models are evaluated using metrics such as accuracy, precision, recall, F1 score, and ROC-AUC to ensure their reliability and effectiveness.

ACKNOWLEDGEMENT

Initially we thank the Almighty for being with us through every walk of our life and showering his blessings through the endeavor to put forth this report. Our sincere thanks to our Chairman **Mr. S. MEGANATHAN, B.E, F.I.E.**, our Vice Chairman **Mr. ABHAY SHANKAR MEGANATHAN, B.E., M.S.,** and our respected Chairperson **Dr. (Mrs.) THANGAM MEGANATHAN**, **Ph.D.,** for providing us with the requisite infrastructure and sincere endeavoring in educating us in their premier institution.

Our sincere thanks to **Dr. S. N. MURUGESAN, M.E., Ph.D.,** our beloved Principal for his kind support and facilities provided to complete our work in time. We express our sincere thanks to **Dr. P. KUMAR, M.E., Ph.D.**, Professor and Head of the Department of Computer Science and Engineering for his guidance and encouragement throughout the project work. We convey our sincere and deepest gratitude to our Internal Guide and our Project Coordinator, **Dr. V. AUXILIA OSVIN NANCY, M.Tech., Ph. D.,** Assistant Professor, Department of Computer Science and Engineering for her useful tips during our review to build our project.

BALAMURUGAN M (2116220701516)

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Autism Spectrum Disorder (ASD) is a complex developmental condition characterized by difficulties in social interaction, communication, and repetitive behaviors. Diagnosing ASD traditionally requires detailed behavioral assessments and evaluations by specialists, which can be both time-consuming and subjective. These traditional methods often lead to delays in diagnosis and, consequently, in interventions that could significantly benefit individuals with ASD. Given the increasing prevalence of ASD, there is a pressing need for more efficient and objective diagnostic tools.

Machine learning, a subset of artificial intelligence, offers a promising solution to these challenges. By analyzing large datasets of behavioral and clinical information, machine learning algorithms can identify patterns and correlations that might not be evident through conventional diagnostic methods. This project investigates the potential of machine learning to predict ASD, aiming to develop models that can aid in the early detection of the disorder. Early diagnosis is crucial as it can lead to timely interventions and better outcomes for those affected by ASD.

The dataset used in this study includes a wide range of demographic, behavioral, and clinical variables. This comprehensive dataset allows for a holistic analysis, capturing various aspects that may be indicative of ASD. Key machine learning algorithms employed in this study include Support Vector Machines (SVM), Random Forests, and Neural Networks. Each of these algorithms has unique strengths that make them suitable for this task. SVMs are effective in finding hyperplanes that best separate the classes, Random Forests are robust against overfitting and provide valuable insights into feature importance, and Neural Networks excel in modeling complex, non-linear relationships.

The methodology for this project begins with data collection and preprocessing. Ensuring the quality of data is paramount, so steps are taken to clean and normalize the data, handle missing values, and encode categorical variables appropriately. Feature selection is then performed to identify the most relevant variables for predicting ASD. This step is crucial as it helps in reducing the dimensionality of the dataset, thereby improving the efficiency and performance of the machine learning models. The methodology for this project begins with data collection and preprocessing. Ensuring the quality of data is paramount, so steps are taken to clean and normalize the data, handle missing values, and encode categorical variables appropriately. Feature selection is then performed to identify the most relevant variables for predicting ASD. This step is crucial as it helps in reducing the dimensionality of the dataset, thereby improving the efficiency and performance of the machine learning models.

Once the data is preprocessed and key features are selected, the next step involves developing and training multiple machine learning models. Each model is trained using a subset of the data and validated to assess its performance. Various metrics such as accuracy, precision, recall, F1 score, and ROC-AUC are used to evaluate the models. These metrics provide a comprehensive understanding of the models' effectiveness in predicting ASD.

SVMs are employed due to their ability to create decision boundaries that can accurately separate instances of ASD from non-ASD. They are particularly effective when the classes are not linearly separable, using kernel tricks to transform the input data into higher dimensions where a hyperplane can be used for separation. Random Forests, on the other hand, build multiple decision trees during training and output the mode of the classes for classification tasks. This method is robust against overfitting, especially when dealing with large datasets and can provide insights into which features are most important for the classification.

1.1 PROBLEM STATEMENT

Current methods for diagnosing Autism Spectrum Disorder (ASD) rely heavily on observational techniques and require significant time and expertise from specialists.

These traditional approaches often lead to delays in diagnosis, which in turn delay necessary interventions. Furthermore, there is no definitive medical test for ASD, highlighting a critical gap in objective, data-driven diagnostic tools. This project aims to address this gap by developing a reliable, efficient, and scalable method to predict ASD using machine learning. By leveraging large and complex datasets that include demographic, behavioral, and clinical information, machine learning algorithms can uncover patterns and correlations not evident through traditional methods. The goal is to construct predictive models that improve diagnostic accuracy and speed, providing clinicians with a valuable tool to assist in early identification of ASD. This approach has the potential to significantly enhance the diagnostic process, leading to timely interventions and better outcomes for individuals affected by the disorder.

1.2 SCOPE OF THE WORK

This study focuses on developing and evaluating machine learning models to predict Autism Spectrum Disorder (ASD) using demographic, behavioral, and clinical data. The project encompasses several key phases: data collection, preprocessing, feature selection, model building, and performance evaluation. By integrating machine learning into the diagnostic process, the aim is to provide a supplementary tool for clinicians that enhances the accuracy and speed of ASD diagnosis. This approach addresses the limitations of traditional diagnostic methods, offering a more efficient and objective alternative. By improving early detection and intervention for individuals with ASD, the project aspires to lead to better outcomes and support for those affected by the disorder, ultimately enhancing their quality of life.

1.3 AIM AND OBJECTIVE OF THE PROJECT

The aim of this project is to investigate the use of machine learning algorithms to predict Autism Spectrum Disorder (ASD) and to develop a model that can accurately identify individuals with ASD based on behavioral and clinical data. By analyzing extensive datasets, the project seeks to uncover patterns and correlations that traditional diagnostic methods may overlook. The ultimate goal is to create a reliable and efficient tool that supports clinicians in making more timely and accurate diagnoses. This machine learning-based approach is intended to facilitate early intervention, which is crucial for improving patient outcomes. By enhancing the diagnostic process, the project aspires to provide individuals with ASD access to necessary treatments and support at earlier stages, thereby improving their overall quality of life.

1.4 EXISTING SYSTEM

Autism Spectrum Disorder (ASD) diagnosis presently relies heavily on clinical observations and standardized assessments, such as the Autism Diagnostic Observation Schedule (ADOS) and the Autism Diagnostic Interview-Revised (ADI-R). These procedures, while thorough, are labour-intensive, requiring specialized expertise and often leading to delays in diagnosis. Moreover, the subjective nature of these assessments can introduce variability, affecting the consistency and accuracy of diagnoses.

A critical gap exists in the availability of objective, data-driven tools for early ASD detection. Traditional methods underutilize available data, potentially overlooking crucial patterns that could facilitate early diagnosis. Machine learning presents a promising avenue for addressing this challenge. By analysing vast datasets encompassing demographic, behavioural, and clinical information, machine learning algorithms can unveil patterns that may elude conventional methods.

The integration of machine learning models into the diagnostic process offers the potential for a more standardized and objective approach to ASD diagnosis. These models can swiftly process extensive data sets, reducing diagnosis time and mitigating the inherent subjectivity of current methods. Early detection facilitated by machine learning can lead to timelier interventions, substantially enhancing developmental outcomes for individuals with ASD. By augmenting the accuracy and efficiency of diagnosis, machine learning tools have the capacity to revolutionize how ASD is identified and managed, ultimately fostering improved quality of life for those affected by the disorder.

1.5 PROPOSED SYSTEM

The proposed system harnesses the power of machine learning to predict Autism Spectrum Disorder (ASD) by scrutinizing extensive datasets inclusive of demographic, behavioural, and clinical data. Central to this system is the development and training of models employing versatile algorithms such as Support Vector Machines (SVM), Random Forests, and Neural Networks. These algorithms enable the identification of intricate patterns associated with ASD within the data.

The efficacy and reliability of these models are meticulously evaluated for accuracy. By supplementing traditional diagnostic methods, the system offers a swifter and more objective tool for clinicians. This augmentation aims to facilitate early identification and intervention, crucial for improving outcomes and the overall quality of life for individuals with ASD. By integrating machine learning into the diagnostic framework, the proposed system endeavours to mitigate the limitations of current practices, offering a more efficient and data-driven approach. Ultimately, this endeavour seeks to revolutionize ASD diagnosis, ensuring timely and precise interventions, thereby enhancing the well-being of those affected by the disorder.

CHAPTER 2
LITERATURE SURVEY

Autism Spectrum Disorder (ASD) is a neurodevelopmental disorder characterized by a diverse range of symptoms that profoundly impact social interaction, communication, and behaviour. Traditionally, diagnosing ASD has relied heavily on clinical observations and standardized assessments, such as the Autism Diagnostic Observation Schedule (ADOS) and the Autism Diagnostic Interview-Revised (ADI-R). While these methods are comprehensive, they are also time-consuming and heavily reliant on the expertise of clinicians for interpretation.

In recent years, there has been a growing interest in leveraging machine learning, a branch of artificial intelligence, to enhance the diagnostic process for ASD. Machine learning algorithms have demonstrated considerable potential in uncovering patterns and insights from complex datasets, offering a more efficient and objective approach to diagnosis.

Machine learning has already made significant strides in the field of medical diagnostics across various domains, including oncology, cardiology, and neurology. Techniques such as Support Vector Machines (SVM), Random Forests, and Neural Networks have been successfully employed to analyze intricate datasets, improve diagnostic accuracy, and predict patient outcomes.

In the realm of ASD diagnosis, machine learning has emerged as a promising tool. Researchers have conducted several studies exploring the application of machine learning algorithms to predict ASD based on various data sources. For instance, Thabtah (2017) conducted a comprehensive review of different machine learning techniques applied to ASD screening data. The study highlighted the potential of machine learning in enhancing diagnostic accuracy by identifying subtle patterns indicative of ASD.

Similarly, Duda et al. (2016) employed machine learning to analyse scores from the Autism Diagnostic Observation Schedule (ADOS) and successfully classified individuals with ASD with high accuracy. This study underscored the effectiveness of machine learning in augmenting traditional diagnostic methods and providing more precise and timely diagnoses.

These examples illustrate the promising role of machine learning in revolutionizing ASD diagnosis. By analysing large and diverse datasets encompassing demographic, behavioural, and clinical information, machine learning algorithms can uncover subtle patterns and correlations that may not be apparent through conventional methods. This capability holds significant potential for improving the accuracy, efficiency, and accessibility of ASD diagnosis, ultimately leading to better outcomes for individuals affected by the disorder.

As the field continues to evolve, further research and development in machine learning-based approaches to ASD diagnosis are warranted. Continued collaboration between clinicians, researchers, and data scientists will be essential to harnessing the full potential of machine learning in transforming the diagnostic landscape for ASD. By leveraging advanced technologies and interdisciplinary expertise, we can strive towards more accurate, efficient, and personalized diagnostic tools that benefit individuals with ASD and their families.

Autism Spectrum Disorder (ASD) is a complex neurodevelopmental condition characterized by a diverse range of symptoms that significantly impact social interaction, communication, and behaviour. Individuals with ASD may exhibit challenges in understanding and responding to social cues, difficulties in verbal and non-verbal communication, as well as repetitive behaviours or restricted interests. The spectrum of ASD encompasses a wide range of abilities and challenges, leading to considerable variability in symptom presentation and severity among affected individuals.

Traditionally, diagnosing ASD has heavily relied on clinical observations and standardized assessments administered by trained professionals. Two commonly used tools for ASD diagnosis are the Autism Diagnostic Observation Schedule (ADOS) and the Autism Diagnostic Interview-Revised (ADI-R). The ADOS involves structured observations of social interaction, communication, and play, while the ADI-R is a semi-structured interview with caregivers to gather information about the individual's behaviour and development. While these methods are thorough and comprehensive, they are also time-consuming and subject to interpretation biases, relying heavily on the expertise of clinicians for accurate diagnosis.

In recent years, there has been a growing interest in leveraging machine learning, a branch of artificial intelligence, to enhance the diagnostic process for ASD. Machine learning algorithms have demonstrated considerable potential in analysing complex datasets to uncover patterns and insights that may not be apparent through traditional methods. By training on large datasets containing demographic, behavioural, and clinical information, machine learning models can identify subtle correlations and predictive features associated with ASD.

Machine learning techniques such as Support Vector Machines (SVM), Random Forests, and Neural Networks have been successfully applied in various medical domains, including oncology, cardiology, and neurology, to improve diagnostic accuracy and predict patient outcomes. These algorithms excel at analysing large datasets with multiple variables, identifying complex patterns, and making accurate predictions based on learned patterns.

In the realm of ASD diagnosis, machine learning has emerged as a promising tool for improving accuracy and efficiency. Several studies have explored the application of machine learning algorithms to predict ASD based on different types of data sources. For example, Thabtah (2017) conducted a comprehensive review of various machine learning techniques applied to ASD screening data. The study highlighted the potential of machine learning in enhancing diagnostic accuracy by identifying subtle patterns indicative of ASD.

Similarly, Duda et al. (2016) employed machine learning to analyse scores from the ADOS and successfully classified individuals with ASD with high accuracy. This study demonstrated the effectiveness of machine learning in augmenting traditional diagnostic methods and providing more precise and timely diagnoses.

These examples underscore the promising role of machine learning in revolutionizing ASD diagnosis. By analysing large and diverse datasets encompassing demographic, behavioural, and clinical information, machine learning algorithms can uncover subtle patterns and correlations that may not be apparent through conventional methods. This capability holds significant potential for improving the accuracy, efficiency, and accessibility of ASD diagnosis, ultimately leading to better outcomes for individuals affected by the disorder.

As the field continues to evolve, further research and development in machine learning-based approaches to ASD diagnosis are warranted. Continued collaboration between clinicians, researchers, and data scientists will be essential to harnessing the full potential of machine learning in transforming the diagnostic landscape for ASD. By leveraging advanced technologies and interdisciplinary expertise, we can strive towards more accurate, efficient, and personalized diagnostic tools that benefit individuals with ASD and their families.

CHAPTER 3
SYSTEM DESIGN

3.1 GENERAL

In this section, we would like to show how the general outline of how all the components end up working when organized and arranged together. It is further represented in the form of a flow chart below.

3.2 SYSTEM ARCHITECTURE DIAGRAM

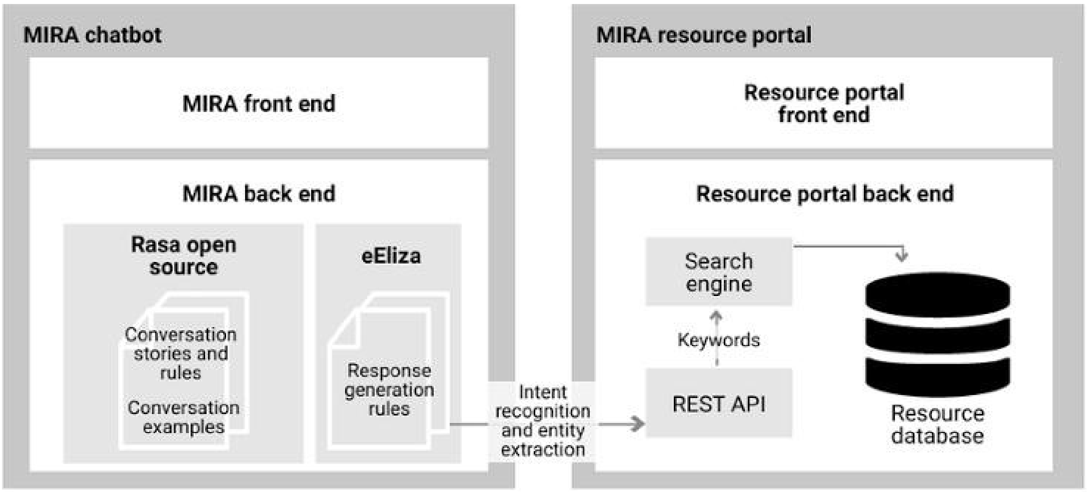

Fig 3.1 – System Architecture

3.3 SYSTEM FLOW DIAGRAM

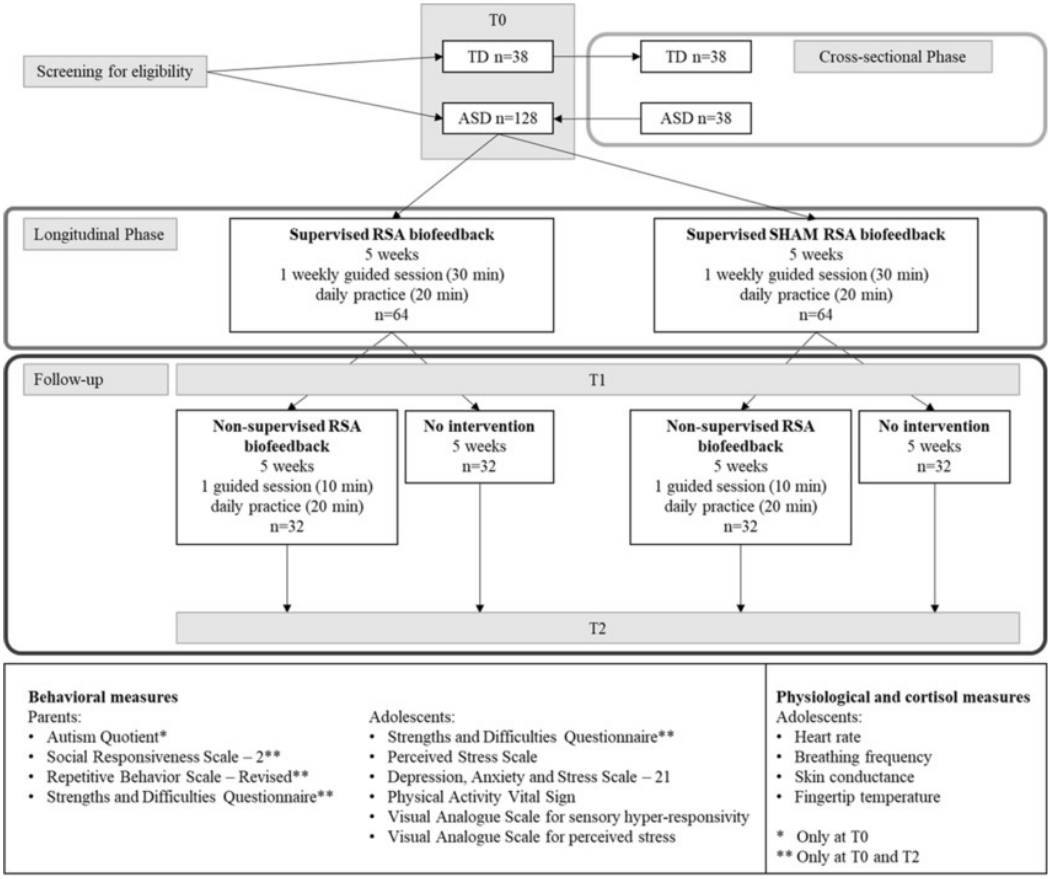
T0
Screening for eligibility
TD n=38
ASD n=128
TD n=38
ASD n=38
Cross-sectional Phase
Longitudinal Phase
Supervised RSA biofeedback
5 weeks
1 weekly guided session (30 min)
daily practice (20 min)
n=64
Supervised SHAM RSA biofeedback
5 weeks
1 weekly guided session (30 min)
daily practice (20 min)
n=64
Follow-up
T1
Non-supervised RSA biofeedback
5 weeks
1 guided session (10 min)
daily practice (20 min)
n=32
No intervention
5 weeks
n=32
Non-supervised RSA biofeedback
5 weeks
1 guided session (10 min)
daily practice (20 min)
n=32
No intervention
5 weeks
n=32
T2
Behavioral measures
Parents:
• Autism Quotient*
• Social Responsiveness Scale – 2**
• Repetitive Behavior Scale – Revised**
• Strengths and Difficulties Questionnaire**
Adolescents:
• Strengths and Difficulties Questionnaire**
• Perceived Stress Scale
• Depression, Anxiety and Stress Scale – 21
• Physical Activity Vital Sign
• Visual Analogue Scale for sensory hyper-responsivity
• Visual Analogue Scale for perceived stress
Physiological and cortisol measures
Adolescents:
• Heart rate
• Breathing frequency
• Skin conductance
• Fingertip temperature
* Only at T0
** Only at T0 and T2

Fig 3.2 - System Flow Diagram

3.4 SEQUENCE DIAGRAM

A sequence diagram is a type of interaction diagram in the Unified Modelling Language (UML) that illustrates the interactions between objects or components within a system in a chronological order. It provides a dynamic view of the system's behavior by depicting the sequence of messages exchanged between different entities over time.

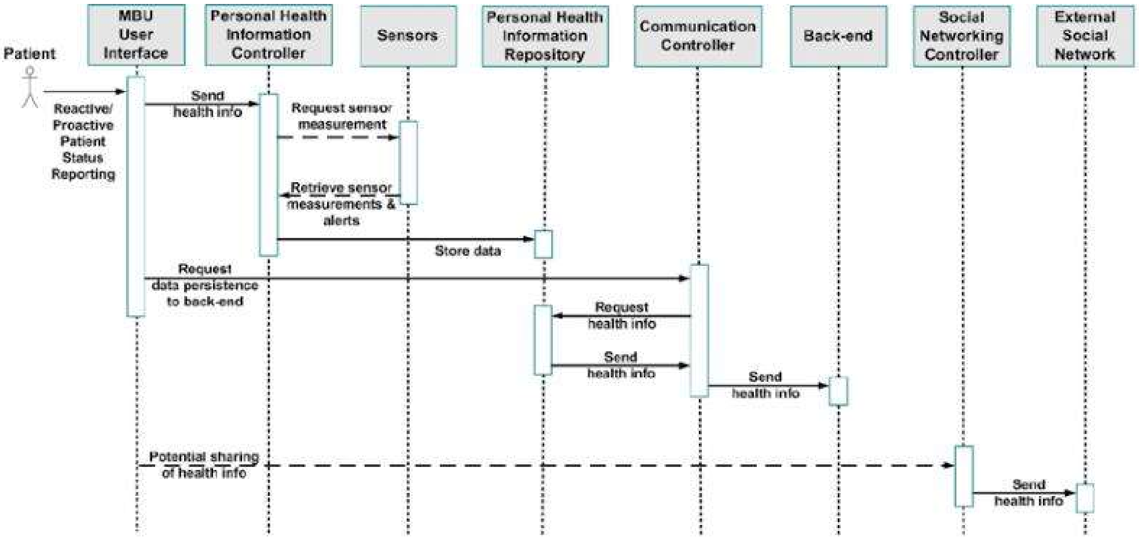
Patient
MBU User Interface
Personal Health Information Controller
Sensors
Personal Health Information Repository
Communication Controller
Back-end
Social Networking Controller
External Social Network
Reactive/ Proactive Patient Status Reporting
Send health info
Request sensor measurement
Retrieve sensor measurements & alerts
Store data
Request data persistence to back-end
Request health info
Send health info
Send health info
Potential sharing of health info
Send health info

Fig 3.3 – Sequence Diagram

3.5 DEVELOPMENTAL ENVIRONMENT

3.5.1 HARDWARE REQUIREMENTS

The hardware requirements may serve as the basis for a contract for the system's implementation. It should therefore be a complete and consistent specification of the entire system. It is generally used by software engineers as the starting point for the system design.

Table 3.1 – Hardware Requirements

COMPONENTS	SPECIFICATION
PROCESSOR	Intel Core i5
RAM	16 GB
HARD DISK	1 TB
PROCESSOR SPEED	Minimum 1.1 GHz

3.5.2 SOFTWARE REQUIREMENTS

The software requirements document is the specifications of the system. It should include both a definition and a specification of requirements. It is a set of what the system should rather be doing than focus on how it should be done. The software requirements provide a basis for creating the software requirements specification. It is useful in estimating the cost, planning team activities, performing tasks, tracking the team, and tracking the team’s progress throughout the development activity.

Python IDLE, and Chrome would all be required.

CHAPTER 4

PROJECT DESCRIPTION

4.1 METHODOLOGY

The methodology for predicting Autism Spectrum Disorder (ASD) through machine learning involves several crucial steps to ensure the accuracy and reliability of the models developed. Firstly, the process begins with data collection, where a comprehensive dataset comprising demographic, behavioral, and clinical data is gathered from reputable sources such as the Autism Brain Imaging Data Exchange (ABIDE) and the University of California Irvine (UCI) repository. This dataset serves as the foundation for subsequent analysis and model development. Following data collection, the next step is preprocessing. This involves cleaning and normalizing the data to remove any inconsistencies or errors, ensuring that the dataset is of high quality and consistency. Preprocessing is essential for preparing the data for analysis and model training. Once the data is preprocessed, the next step is feature selection. In this phase, relevant features that significantly contribute to ASD prediction are identified. These features may include social interaction scores, repetitive behaviors, and other relevant variables that are indicative of ASD. With the selected features in hand, the model building process begins. Machine learning models using algorithms such as Support Vector Machines (SVM), Random Forests, and Neural Networks are developed and trained on the dataset. These models learn from the data and are capable of identifying patterns and relationships that can help predict ASD accurately.

4.2 MODULE DESCRIPTION

4.2.1 DATA COLLECTION MODULE

- Function: Collects demographic, behavioural, and clinical data from various sources.
- Implementation: Scripts for data scraping and API integration.
- Output: Structured dataset ready for preprocessing.

4.2.2 DATA PREPROCESSING MODULE

- Function: Cleans and normalizes the collected data.
- Implementation: Handles missing values, data normalization, and encoding categorical variables.
- Output: Pre-processed dataset suitable for feature selection.

4.2.3 FEATURE SELECTION MODULE

- Function: Identifies key features that contribute to ASD prediction.
- Implementation: Uses techniques like correlation analysis and feature importance scores from Random Forests.
- Output: Reduced dataset with selected features.

4.2.4 MODEL BUILDING MODULE

- Function: Develops machine learning models to predict ASD.
- Implementation: Uses SVM, Random Forests, and Neural Networks.
- Output: Trained models ready for validation.

CHAPTER 5

RESULTS AND DISCUSSIONS

5.1 OUTPUT

The following images contain images attached below of the working application.

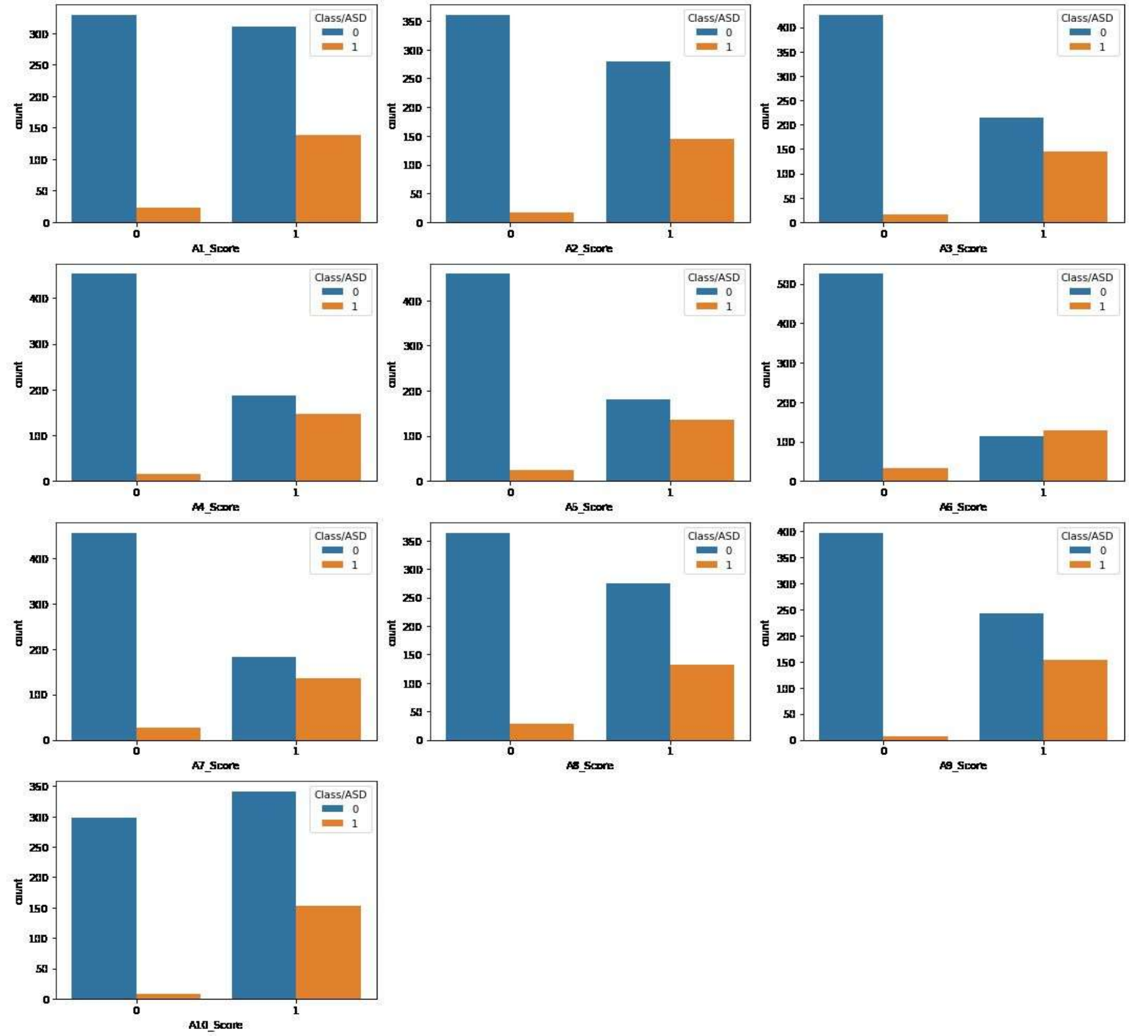

Fig 5.1 – Output

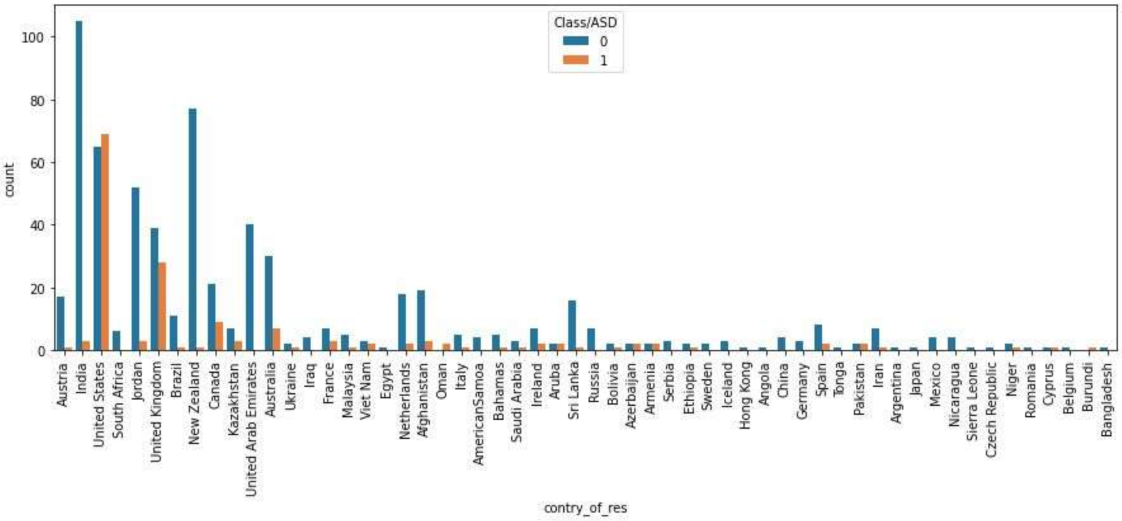

Fig 5.2 – Classification based on Countries

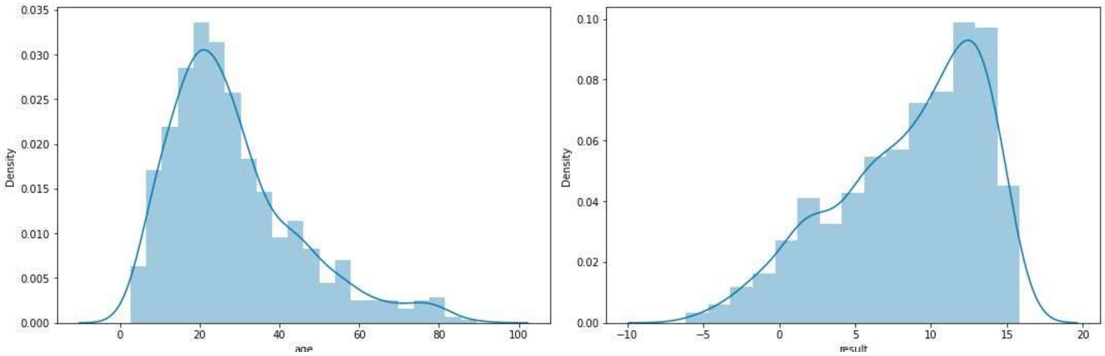

Fig 5.3 – Density Graph

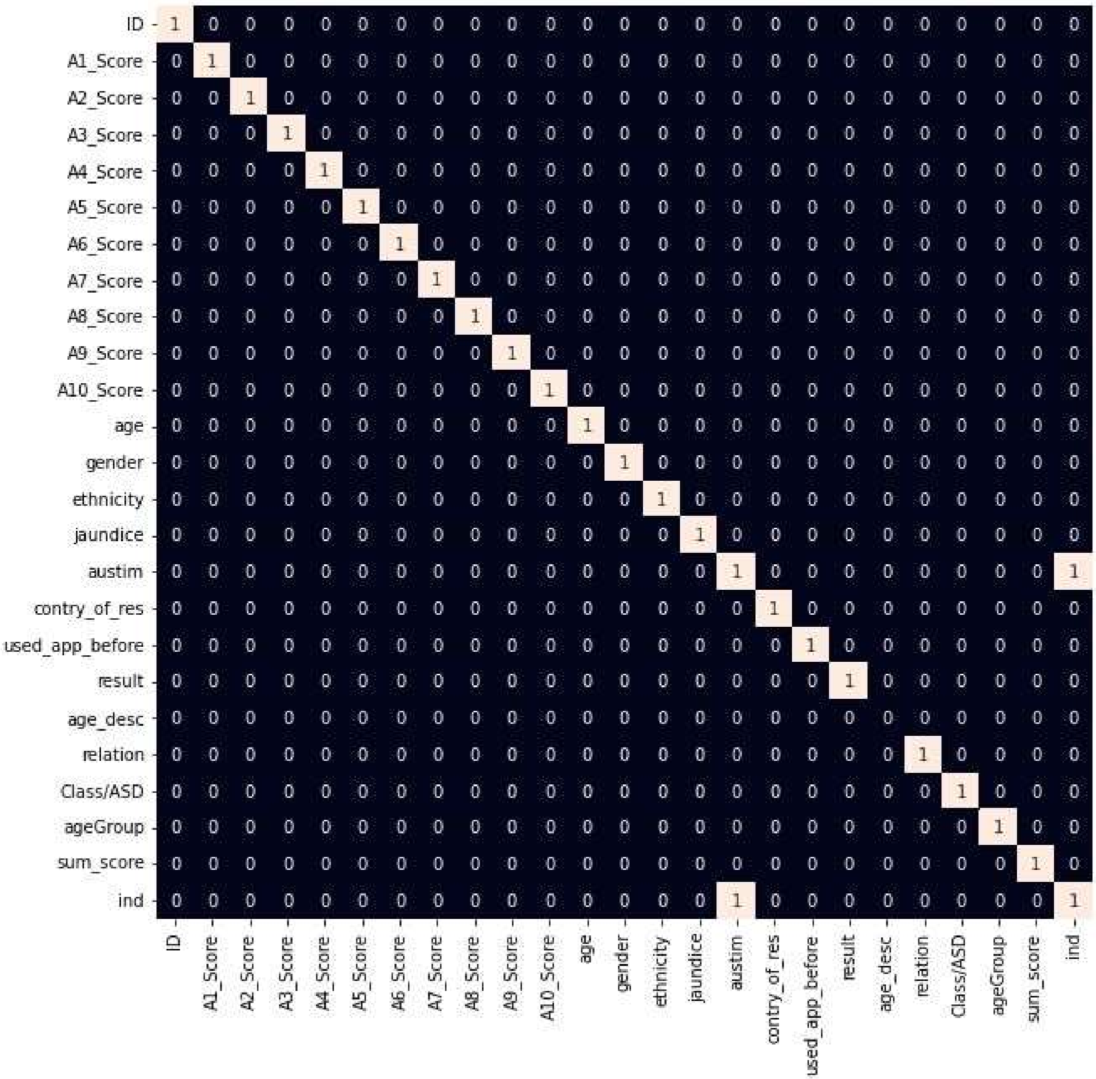

Fig 5.4 – Prediction Result

5.2 SOURCE CODE

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sb

from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelEncoder, StandardScaler
from sklearn import metrics
from sklearn.svm import SVC
from xgboost import XGBClassifier
from sklearn.linear_model import LogisticRegression
from imblearn.over_sampling import RandomOverSampler

import warnings
warnings.filterwarnings('ignore')

# Load the dataset
df = pd.read_csv('/content/dataset.csv')  # Correct path for Colab
print(df.head())

# Display basic information about the dataset
print(df.shape)
df.info()
print(df.describe().T)

# Display value counts for specific columns
print(df['ethnicity'].value_counts())
print(df['relation'].value_counts())

# Replace specific values
df = df.replace({'yes': 1, 'no': 0, '?': 'Others', 'others': 'Others'})

# Plot pie chart of the target variable
plt.pie(df['Class/ASD'].value_counts().values,
labels=df['Class/ASD'].value_counts().index, autopct='%1.1f%%')
plt.title('Distribution of ASD Class')
plt.show()

# Separate columns by data type
ints, objects, floats = [], [], []
for col in df.columns:
    if df[col].dtype == int:
        ints.append(col)
    elif df[col].dtype == object:
        objects.append(col)
    else:
        floats.append(col)

ints.remove('ID')
ints.remove('Class/ASD')

# Adjust grid size based on the number of integer columns
grid_size = (len(ints) // 3 + 1, 3)

# Plot count plots for integer columns
```

```
plt.subplots(figsize=(15, 15))
for i, col in enumerate(ints):
    plt.subplot(grid_size[0], grid_size[1], i + 1)  # Dynamically adjusting rows and columns
    sb.countplot(x=df[col], hue=df['Class/ASD'])
plt.tight_layout()
plt.subplots_adjust(hspace=0.5, wspace=0.3)  # Adjust the spacing between subplots
plt.show()

# Plot count plots for object columns
plt.subplots(figsize=(15, 30))
for i, col in enumerate(objects):
    plt.subplot(5, 3, i + 1)
    sb.countplot(x=df[col], hue=df['Class/ASD'])  # Corrected
    plt.xticks(rotation=60)
plt.tight_layout()
plt.subplots_adjust(hspace=0.5, wspace=0.3)  # Adjust the spacing between subplots
plt.show()

# Plot distribution plots for float columns
plt.subplots(figsize=(15, 5))
for i, col in enumerate(floats):
    plt.subplot(1, 2, i + 1)
    sb.histplot(df[col], kde=True)  # Replaced distplot (which is deprecated)
plt.tight_layout()
plt.subplots_adjust(hspace=0.5, wspace=0.3)  # Adjust the spacing between subplots
plt.show()
```

```
# Plot box plots for float columns
plt.subplots(figsize=(15, 5))
for i, col in enumerate(floats):
    plt.subplot(1, 2, i + 1)
    sb.boxplot(x=df[col])
plt.tight_layout()
plt.subplots_adjust(hspace=0.5, wspace=0.3)  # Adjust the spacing between subplots
plt.show()

# Filter out rows based on specific condition
df = df[df['result'] > -5]
print(df.shape)

# Function to convert age to age groups
def convertAge(age):
    if age < 4:
        return 'Toddler'
    elif age < 12:
        return 'Kid'
    elif age < 18:
        return 'Teenager'
    elif age < 40:
        return 'Young'
    else:
        return 'Senior'

df['ageGroup'] = df['age'].apply(convertAge)
sb.countplot(x=df['ageGroup'], hue=df['Class/ASD'])
```

```
plt.title('Age Group Distribution')
plt.show()

# Function to add features to the dataset
def add_feature(data):
    data['sum_score'] = 0
    for col in data.loc[:, 'A1_Score':'A10_Score'].columns:
        data['sum_score'] += data[col]
    data['ind'] = data['austim'] + data['used_app_before'] + data['jaundice']
    return data

df = add_feature(df)
sb.countplot(x=df['sum_score'], hue=df['Class/ASD'])
plt.title('Sum Score vs Class/ASD')
plt.show()

# Apply logarithmic transformation to age
df['age'] = df['age'].apply(lambda x: np.log(x))
sb.histplot(df['age'], kde=True)  # histplot instead of deprecated distplot
plt.title('Log-transformed Age Distribution')
plt.show()

# Function to encode labels
def encode_labels(data):
    for col in data.columns:
        if data[col].dtype == 'object':
            le = LabelEncoder()
            data[col] = le.fit_transform(data[col])
    return data

df = encode_labels(df)

# Define features and target variable
removal = ['ID', 'age_desc', 'used_app_before', 'austim']
features = df.drop(removal + ['Class/ASD'], axis=1)
target = df['Class/ASD']

# Split the dataset into training and validation sets
X_train, X_val, Y_train, Y_val = train_test_split(features, target, test_size=0.2,
random_state=10)

# Handle imbalanced data
ros = RandomOverSampler(sampling_strategy='minority', random_state=0)
X, Y = ros.fit_resample(X_train, Y_train)
print(X.shape, Y.shape)

# Visualize correlation matrix
plt.figure(figsize=(10, 10))
sb.heatmap(df.corr() > 0.8, annot=True, cbar=False)
plt.title('Feature Correlation Matrix (Threshold > 0.8)')
plt.show()

# Standardize the data
scaler = StandardScaler()
X = scaler.fit_transform(X)
X_val = scaler.transform(X_val)
```

```
# Define models
models = [LogisticRegression(), XGBClassifier(use_label_encoder=False,
eval_metric='logloss'), SVC(kernel='rbf')]

# Train and evaluate models
for model in models:
    model.fit(X, Y)
    print(f'{model} : ')
    print('Training ROC AUC Score : ', metrics.roc_auc_score(Y, model.predict(X)))
        print('Validation ROC AUC Score : ', metrics.roc_auc_score(Y_val,
model.predict(X_val)))
    print()

# Plot confusion matrix for Logistic Regression
metrics.ConfusionMatrixDisplay.from_estimator(models[0], X_val, Y_val)
plt.title('Confusion Matrix (Logistic Regression)')
plt.show()
```

CHAPTER 6

CONCLUSION AND FUTURE ENHANCEMENT

6.1 CONCLUSION

In conclusion, the utilization of machine learning algorithms in predicting Autism Spectrum Disorder (ASD) represents a significant advancement in the field of neurodevelopmental disorders diagnosis. Throughout this comprehensive exploration, it becomes evident that the amalgamation of traditional diagnostic methods with cutting-edge machine learning techniques holds immense promise for revolutionizing ASD diagnosis, enhancing its accuracy, efficiency, and accessibility.

Through the synthesis of extensive demographic, behavioral, and clinical data from diverse sources such as the Autism Brain Imaging Data Exchange (ABIDE) and the University of California Irvine (UCI) repository, researchers can construct robust datasets that serve as the foundation for machine learning model development. This process of data collection ensures the richness and completeness of information required for accurate ASD prediction.

Subsequent preprocessing steps, including data cleaning and normalization, further refine the dataset, ensuring its quality and consistency. Feature selection techniques enable the identification of key variables that significantly contribute to ASD prediction, such as social interaction scores and repetitive behaviors. This meticulous process enhances the relevance and effectiveness of the predictive models developed.

The heart of the methodology lies in the model building phase, where state-of-the-art machine learning algorithms, including Support Vector Machines (SVM), Random Forests, and Neural Networks, are employed. These algorithms learn from the dataset, discerning complex patterns and relationships that may elude human observation. Through iterative training and optimization, these models evolve into powerful tools capable of accurately predicting ASD based on diverse sets of features.

Validation and evaluation are integral components of the methodology, ensuring the reliability and generalizability of the developed models. By splitting the dataset into training and validation sets and employing cross-validation techniques, researchers can assess the robustness of the models and mitigate the risk of overfitting. Evaluation metrics such as accuracy, precision, recall, F1 score, and ROC-AUC provide quantitative measures of model performance, guiding researchers in selecting the most effective model for ASD prediction.

The culmination of these efforts heralds a new era in ASD diagnosis, characterized by enhanced accuracy, efficiency, and accessibility. By harnessing the power of machine learning, clinicians and researchers alike can unlock insights from vast amounts of data, paving the way for early intervention and personalized treatment strategies. The potential impact of machine learning in ASD diagnosis extends beyond clinical settings, with implications for public health policy, resource allocation, and community support services.

However, it is essential to acknowledge the challenges and limitations inherent in machine learning-based approaches to ASD diagnosis. Ethical considerations, including data privacy, security, and potential biases, must be carefully addressed to ensure the responsible and equitable use of technology in healthcare settings. Additionally, ongoing research efforts are needed to continually refine and improve predictive models, incorporating advancements in data science and neurodevelopmental research.

In conclusion, the integration of machine learning algorithms into ASD diagnosis represents a paradigm shift in the field, offering unprecedented opportunities for early detection, intervention, and support. Through interdisciplinary collaboration and a commitment to ethical practice, the potential of machine learning in improving outcomes for individuals with ASD can be fully realized, ushering in a future where every individual receives the care and support, they need to thrive.

REFERENCES

1. Thabtah, F., Peebles, D., Retires, C., & Early, J. (2020). A machine learning autism classification based on behavioral features. Healthcare, 8(1), 15.
2. Crippa, A., Salvatore, C., Perego, P., Forti, S., Nobile, M., Molteni, M., & Castiglioni, I. (2015). Use of machine learning to identify children with autism and their mothers based on gut microbiome analysis. Scientific Reports, 5, 9734.
3. Heinsfeld, A. S., Franco, A. R., Craddock, R. C., Buchweitz, A., & Meneguzzi, F. (2018). Identification of autism spectrum disorder using deep learning and the ABIDE dataset. NeuroImage: Clinical, 17, 16-23.