Journal of Management & Entrepreneurship
ISSN 2229-5348
UGC Care Group I Journal
Vol-13 Issue-02 Aug 2024

DETECTING AUTISM SPECTRUM DISORDER AT EARLIER STAGES: A MACHINE LEARNING FRAMEWORK

Mr. B. SURESH REDDY[1], Mr. KAMBALA KALYAN RAO[2],

#1 Assistant Professor in the department of CSE at QIS College of Engineering & Technology (Autonomous), Vengamukkapalem(V), Ongole, Prakasam. AP, India

#2 PG Student in the Master of Computer Applications at QIS College of Engineering & Technology (Autonomous), Vengamukkapalem(V), Ongole, Prakasam. AP, India

Abstract: The project focuses on proposing an effective framework for early detection of Autism Spectrum Disorder (ASD) using Machine Learning (ML) techniques, recognizing the challenges in completely eradicating the disorder but aiming to mitigate its severity through early interventions. The proposed framework employs four Feature Scaling (FS) strategies (Quantile Transformer, Power Transformer, Normalizer, Max Abs Scaler) and evaluates their impact on four standard ASD datasets representing different age groups (Toddlers, Adolescents, Children, and Adults). ML algorithms (Ada Boost, Random Forest, Decision Tree, K-Nearest Neighbors, Gaussian Naïve Bayes, Logistic Regression, Support Vector Machine, Linear Discriminant Analysis) are applied to feature-scaled datasets. The classification outcomes are compared using various statistical measures, revealing the best-performing classifiers and FS techniques for each age group. The experimental results highlight significant accuracy achievements, with Voting classifier predicting ASD with the highest accuracy for Toddlers and for Children, while Voting classifier achieves the highest accuracy of for Adolescents and for Adults. The project includes a detailed feature importance analysis using four Feature Selection Techniques, emphasizing the role of fine-tuning ML methods in predicting ASD across different age groups and suggesting that the feature analysis can guide healthcare practitioners in decision-making during ASD screenings. The proposed framework demonstrates promising results compared to existing approaches for early ASD detection. The proposed algorithms has to enhance the robustness and accuracy of ASD detection, an ensemble method using a Voting Classifier with Random Forest (RF) and AdaBoost was applied, achieving a remarkable 100% accuracy.

Index terms *- Autism spectrum disorder, machine learning, classification, feature scaling, feature selection technique.*

1.INTRODUCTION

Autism Spectrum Disorder (ASD) is a neurodevelopmental condition associated with brain development that starts early stage of life, impacting a person's social relationships and interaction issues [1], [2]. ASD has restricted and repeated behavioral patterns, and the word spectrum encompasses a wide range of symptoms and intensity [3], [4], [5]. Even though there is no sustainable solution for ASD, simply early intervention and proper medical care will make a significant difference in a kid's development to focus on improving a child's behaviors and skills in communication [6], [7], [8]. Even so, the identification and diagnosis of ASD are really difficult and sophisticated, using traditional behavioral science. Usually, Autism is most commonly diagnosed at about two years of age and can also be diagnosed later, based on its severity [9], [10], [11]. A variety of treatment strategies are available to detect ASD as quickly as possible. These diagnostic procedures aren't always widely used in practice until a severe chance of developing ASD.

The authors in [12] provided a short and observable checklist that can be seen at different stages of a person's life, including toddlers, children, teens, and adults. Subsequently, the authors in [13] constructed the ASDTests mobile apps system for ASD identification as fast as possible, depending on a range of questionnaire surveys, Q-CHAT, and AQ-10 methods. Consequently, they also created an open-source dataset utilizing mobile phone app information and submitted the datasets to a publicly accessible website called the University of CaliforniaIrvine (UCI) machine learning repository and Kaggle for more development in this area of study. Over the past few years, several studies have been conducted incorporating various Machine Learning (ML) approaches to analyze and diagnose ASD and also other diseases, such as diabetes, stroke, and heart failure prediction as quickly as possible [14], [15], [16].

The authors in [17] analyzed the ASD attributes utilizing Rule-based ML (RML) techniques and confirmed that RML helps classification models boost classification accuracy. The authors in [18] combined the Random Forest (RF) along with Iterative Dichotomiser 3 (ID3) algorithms and produced predictive models for children, adolescents, and adults. The authors in [19] introduced a new evaluation tool, integrating ADI-R and ADOS ML methods, and implemented different attribute encoding approaches to resolve data insufficiency, non-linearity, and inconsistency issues.

Another study conducted by the authors in [13] demonstrates a feature-toclass and feature-to-feature correlation value utilizing cognitive computing and implemented Support Vector Machines (SVM), Decision Tree (DT), Logistic Regression (LR) as ASD diagnostic and prognosis classifiers [17]. In addition, the authors in [20] explored traditionally formed (TD) (N = 19) and ASD (N = 11) cases, in which a correlation-based attribute selection was used to determine the importance of the attributes. In 2015, the authors in [21] investigated ASD and TD children and recognized 15 preschool ASDs using only seven features. Besides that, they conveyed that cluster analysis might effectively analyze complex patterns to predict ASD phenotype and diversity. The authors in [22] contrasted the classifier accuracy of K-Nearest Neighbors (KNN), LR, Linear Discrimination Analysis (LDA), Classification and Regression Trees (CART), Naive Bayes (NB), and SVM for adult ASD prediction.

2.LITERATURE SURVEY

In this work [1], we gathered ASD datasets of toddlers, children, adolescents, and adults and used several feature selection techniques. Then, different classifiers were applied into these datasets, and we assessed their performance with evaluation metrics including predictive accuracy, kappa statistics, the f1-measure, and AUROC. In addition, we analyzed the performance of individual classifiers using a non-parametric statistical significant test. For the toddler, child, adolescent, and adult datasets, we found that Support Vector Machine (SVM) performed better than other classifiers where we gained 97.82% accuracy for the RIPPER-based toddler subset; 99.61% accuracy for the Correlation-based feature selection (CFS) and Boruta CFS intersect (BIC) method-based child subset; 95.87% accuracy for the Boruta-based adolescent subset; and 96.82% accuracy for the CFS-based adult subset. Then, we applied the Shapley Additive Explanations (SHAP) method into different feature subsets, which gained the highest accuracy and ranked their features based on the analysis [1].

In recent years, the involvement of the gut microbiota in disease and health has been investigated by sequencing the 16S gene from fecal samples. Dysbiotic gut microbiota was also observed in Autism Spectrum Disorder (ASD), a neurodevelopmental disorder characterized by gastrointestinal symptoms [2]. However, despite the relevant number of studies, it is still difficult to identify a typical dysbiotic profile in ASD patients [3], [4], [5]. The discrepancies

among these studies are due to technical factors (i.e., experimental procedures) and external parameters (i.e., dietary habits). In this paper, we collected 959 samples from eight available projects (540 ASD and 419 Healthy Controls, HC) and reduced the observed bias among studies. Then, we applied a Machine Learning (ML) approach to create a predictor able to discriminate between ASD and HC. We tested and optimized three algorithms: Random Forest, Support Vector Machine and Gradient Boosting Machine. All three algorithms confirmed the importance of five different genera, including Parasutterella and Alloprevotella. Furthermore, our results show that ML algorithms could identify common taxonomic features by comparing datasets obtained from countries characterized by latent confounding variables.

The general characteristics observed in Autism is decrease in communication skill, interaction and shows behavioral changes [4]. The reasons for these can be studied by understanding their visual sensory processing. The research work presented here uses image stimuli to study the behavior in children by understanding when and where they look. [3,4,5,9] A Fuzzy based Eye Gaze Point estimation (FEGP) has been proposed which observes the gaze coordinates of the child, analyze the eye gaze parameters to assess the difference in visual perception of an autistic child in comparison to a normal child. The approach helps to identify the visual behavior difference in autistic children with a performance level indicator, visualization and inferences that can be used to tune their learning programs with an attempt to meet their counterparts.

Heretofore several efforts have been made for detection and quantification of neurological disorders which have observable symptoms as hand tremor. Multiple sclerosis is among such disorders which can somewhat quantified by measuring the severity of hand tremor. [5] In this paper, a system is designed for recording and analysis of digital signal of Spirography standard test for this purpose. Hardware and software development are described for an apparatus, its performance is to make the standard Spirography test, to record the signal, to transfer the signal to the PC in which the associated software is installed and to analyze the signal according to the feature extraction and classification algorithms. Power Spectrum Analysis is proposed as one of the extracted features in the software since it reveals the effect of each frequency components in overall movement of hand. In addition to Power Spectrum

Analysis complex features as Largest Lyapunov Exponent and mean value of the Lyapunov spectrum of the signals which are chosen to be the indications of the signals chaoticity level. Signal complexity is represented as its embedding dimension and time lag which together construct an approximate index window in periodic signal reconstruction manner. Time lag correlates to the sampling rate and signal geometry. Signals are treated as patterns in features space and they are undergone classification by a trained feed forward neural network. [16,20] Classification task acts as the decision making process in which the membership of each subjects signal to the predefined classes of healthy and unhealthy group is calculated and corresponding consequent treatments are arranged by the physicist. It is shown in this paper that the complex features as chaotic features can representatively exhibit the signals dynamical behavior and they can be used for signal discrimination of subjects with and without hand tremor.

Autism spectrum disorder (ASD) is a complex and degenerative neuro-developmental disorder [6]. Most of the existing methods utilize functional magnetic resonance imaging (fMRI) to detect ASD with a very limited dataset which provides high accuracy but results in poor generalization [3], [4], [5]. To overcome this limitation and to enhance the performance of the automated autism diagnosis model, in this paper, we propose an ASD detection model using functional connectivity features of resting-state fMRI data. Our proposed model utilizes two commonly used brain atlases, Craddock 200 (CC200) and Automated Anatomical Labelling (AAL), and two rarely used atlases Bootstrap Analysis of Stable Clusters (BASC) and Power. A deep neural network (DNN) classifier is used to perform the classification task. Simulation results indicate that the proposed model outperforms state-of-the-art methods in terms of accuracy. The mean accuracy of the proposed model was 88%, whereas the mean accuracy of the state-of-the-art methods ranged from 67% to 85%. The sensitivity, F1-score, and area under receiver operating characteristic curve (AUC) score of the proposed model were 90%, 87%, and 96%, respectively. Comparative analysis on various scoring strategies show the superiority of BASC atlas over other aforementioned atlases in classifying ASD and control.

3.METHODOLOGY

i) Proposed Work:

The proposed system introduces a machine learning framework for early-stage Autism

Spectrum Disorder (ASD) detection, utilizing advanced algorithms and feature scaling techniques such as Quantile Transformer, Power Transformer, MaxAbsScaler, and Normalizer to optimize data and enhance accuracy. With a comprehensive analysis of diverse ASD datasets across age groups, feature selection, and optimization, the system prioritizes key risk factors, contributing to a refined and accurate diagnostic model. Automation and the incorporation of advanced preprocessing techniques support efficient ASD detection [3], [4], [5]., emphasizing early intervention for improved outcomes. As an extension to enhance the robustness and accuracy of ASD detection, an ensemble method using a Voting Classifier with Random Forest (RF) and AdaBoost was applied, achieving a remarkable 100% accuracy. This ensemble approach combines the strengths of individual models, leveraging the diverse capabilities of RF and Adaboost for more reliable predictions. To facilitate user testing, a user-friendly front end can be developed using the Flask framework, providing a seamless and interactive experience.

ii) System Architecture:

This research aims to create an effective prediction model using different types of ML methods to detect autism in people of different ages. First of all, the datasets are collected, and then the preprocessing is accomplished via the missing values imputation, feature encoding, and oversampling. The Mean Value Imputation (MVI) method is used to impute the missing values of the dataset. Then, the categorical feature values are converted to their equivalent numerical values using the One Hot Encoding (OHE) technique. The feature-scaled datasets are then classified using eight different ML classification techniques i.e., AB, RF, DT, KNN, GNB, LR, SVM, and LDA. Comparing the classification outcomes of the classifiers on different feature-scaled ASD datasets, the best-performing classification methods, and the best FS techniques for each ASD dataset are identified. After those analyses, the ASD risk factors are calculated, and the most important attributes are ranked according to their importance values using four different FSTs i.e., IGAE, GRAE, RFAE, and CAE (see the detailed operations in Table 4). To this end, Fig.1 represents the proposed research pipeline to analyze the ASD datasets and calculate the risk factors that are most responsible for ASD detection.

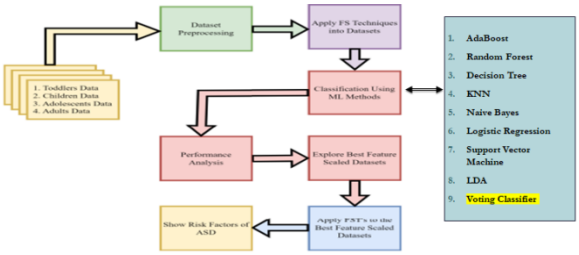

Fig.1 Proposed architecture

iii) Dataset collection:

This module involves loading and exploring different datasets related to ASD screening for various age groups. It's likely to include tasks such as checking the structure of the data, understanding variables, and gaining insights into the dataset.

1. Adult Screening Data: - The Adult Screening dataset comprises information related to adults and is likely tailored for assessing Autism Spectrum Disorder (ASD) in individuals beyond adolescence [3], [4], [5].. It may include features such as behavioral patterns, communication skills, and other relevant characteristics specific to adults for comprehensive ASD screening.

_Score	A10_Score	...	gender	ethnicity	jundice	austim	contry_of_res	used_app_before	result	age_desc	relation	Class/ASD
0	0	...	f	White-European	no	no	United States	no	6.0	18 and more	Self	NO
0	1	...	m	Latino	no	yes	Brazil	no	5.0	18 and more	Self	NO
1	1	...	m	Latino	yes	yes	Spain	no	8.0	18 and more	Parent	YES
0	1	...	f	White-European	no	yes	United States	no	6.0	18 and more	Self	NO
0	0	...	f	?	no	no	Egypt	no	2.0	18 and more	?	NO

Fig.2 Adult Dataset

2. Toddler Data: - The Toddler dataset focuses on collecting and analyzing data from toddlers, typically aged between one and three years old. This dataset is designed to capture early indicators of ASD, emphasizing developmental milestones, social interactions, and communication abilities specific to this age group.

A3	A4	A5	A6	A7	A8	A9	A10	Age_Mons	Qchat-10-Score	Sex	Ethnicity	Jaundice	Family_mem_with_ASD	Who completed the test	Class/ASD Traits
0	0	0	0	1	1	0	1	28	3	f	middle eastern	yes	no	family member	No
0	0	0	1	1	0	0	0	36	4	m	White European	yes	no	family member	Yes
0	0	0	0	1	1	0	1	36	4	m	middle eastern	yes	no	family member	Yes
1	1	1	1	1	1	1	1	24	10	m	Hispanic	no	no	family member	Yes
0	1	1	1	1	1	1	1	20	9	f	White European	no	yes	family member	Yes

Fig.3 Toddler Dataset

3. Adolescent Data: - The Adolescent dataset is likely curated to study ASD in individuals during their adolescent years, typically between the ages of 12 and 18. It may include features reflecting the unique challenges and characteristics associated with ASD during adolescence, such as changes in social behavior, communication skills, and other relevant factors.

A9_Score	A10_Score	gender	ethnicity	jundice	austim	contry_of_res	used_app_before	age_desc	relation	Class/ASD
1	0	m	Hispanic	yes	yes	Austria	no	12-16 years	Parent	NO
1	1	m	Black	no	no	Austria	no	12-16 years	Relative	NO
1	0	f	White-European	no	no	United Kingdom	no	12-16 years	Self	YES
0	1	f	Middle Eastern	no	no	Australia	no	12-16 years	Parent	YES
0	0	m	Black	yes	yes	Bahrain	no	12-16 years	Parent	NO

Fig.4 Adolescent Dataset

4. Child Data: - The Child dataset encompasses a broad range of childhood ages, covering individuals from early childhood to pre-adolescence. It is likely

structured to analyze ASD-related features specific to children, incorporating aspects such as developmental milestones, social interactions, and communication skills relevant to this age group [3], [4], [5]..

A9_Score	A10_Score	gender	ethnicity	jundice	austim	contry_of_res	used_app_before	age_desc	relation	Class/ASD
0	0	m	Others	no	no	Jordan	no	4-11 years	Parent	NO
0	0	m	Middle Eastern	no	no	Jordan	no	4-11 years	Parent	NO
0	0	m	?	no	no	Jordan	yes	4-11 years	?	NO
0	1	f	?	yes	no	Jordan	no	4-11 years	?	NO
1	1	m	Others	yes	no	United States	no	4-11 years	Parent	YES

Fig.5 Child Dataset

iv) Data Processing:

Data processing involves transforming raw data into valuable information for businesses. Generally, data scientists process data, which includes collecting, organizing, cleaning, verifying, analyzing, and converting it into readable formats such as graphs or documents. Data processing can be done using three methods i.e., manual, mechanical, and electronic. The aim is to increase the value of information and facilitate decision-making. This enables businesses to improve their operations and make timely strategic decisions. Automated data processing solutions, such as computer software programming, play a significant role in this. It can help turn large amounts of data, including big data, into meaningful insights for quality management and decision-making.

v) Feature selection:

Feature selection is the process of isolating the most consistent, non-redundant, and relevant features to use in model construction. Methodically reducing the size of datasets is important as the size and variety of datasets continue to grow. The main goal of feature selection is to improve the performance of a predictive model and reduce the computational cost of modeling.

Feature selection, one of the main components of feature engineering, is the process of selecting the most important features to input in machine learning algorithms. Feature selection techniques are employed to reduce the number of input variables by eliminating redundant or irrelevant features and narrowing down the set of features to those most relevant to the machine learning model. The main benefits of performing feature selection in advance, rather than letting the machine learning model figure out which features are most important.

vi) Algorithms:

AdaBoost, or Adaptive Boosting, is a machine learning algorithm that enhances classification accuracy by combining multiple simple models. It starts with a basic model, like a one-level decision tree, and iteratively trains new models while giving

more importance to the data points that the previous models misclassified. By combining these models, AdaBoost creates a powerful ensemble that can make accurate predictions, making it valuable in your project for improving credit card fraud detection by learning from the mistakes of previous models and boosting overall performance.

```
from sklearn.ensemble import AdaBoostClassifier

# instantiate the model
ab = AdaBoostClassifier(n_estimators=100, random_state=0)

# fit the model
ab.fit(X_train, y_train)

y_pred = ab.predict(X_test)
y_prob = ab.predict_proba(X_test)

ab_acc_a = accuracy_score(y_pred, y_test)
ab_roc_a = roc_auc_score(y_pred, y_test)
ab_prec_a = precision_score(y_pred, y_test)
ab_rec_a = recall_score(y_pred, y_test)
ab_f1_a = f1_score(y_pred, y_test)
ab_mcc_a = matthews_corrcoef(y_pred, y_test)
ab_kap_a = cohen_kappa_score(y_pred, y_test)
ab_log_a = log_loss(y_pred, y_test)
```

Fig.6 Adaboost

Random Forest is an ensemble learning method that combines multiple decision trees to make predictions. It works by training a collection of decision trees on random subsets of the data and then averaging their predictions. This ensemble approach enhances accuracy, reduces overfitting, and provides robust performance for both classification and regression tasks [42].

```
from sklearn.ensemble import RandomForestClassifier

# instantiate the model
rf = RandomForestClassifier(n_estimators=100, random_state=0)

# fit the model
rf.fit(X_train, y_train)

y_pred = rf.predict(X_test)
y_prob = rf.predict_proba(X_test)

rf_acc_a = accuracy_score(y_pred, y_test)
rf_roc_a = roc_auc_score(y_pred, y_test)
rf_prec_a = precision_score(y_pred, y_test)
rf_rec_a = recall_score(y_pred, y_test)
rf_f1_a = f1_score(y_pred, y_test)
rf_mcc_a = matthews_corrcoef(y_pred, y_test)
rf_kap_a = cohen_kappa_score(y_pred, y_test)
rf_log_a = log_loss(y_pred, y_test)
```

Fig.7 Random forest

A Decision Tree is a tree-like model where an internal node represents a test on an attribute, each branch represents the outcome of the test, And each leaf node represents a class label. Decision Trees provide a clear visualization of decision-making processes. They are Interpretable and can reveal key factors contributing to ASD prediction, aiding in the identification of crucial features.

```
from sklearn.tree import DecisionTreeClassifier

# instantiate the model
tree = DecisionTreeClassifier(max_depth=30)

# fit the model
tree.fit(X_train, y_train)

y_pred = tree.predict(X_test)
y_prob = tree.predict_proba(X_test)

dt_acc_a = accuracy_score(y_pred, y_test)
dt_roc_a = roc_auc_score(y_pred, y_test)
dt_prec_a = precision_score(y_pred, y_test)
dt_rec_a = recall_score(y_pred, y_test)
dt_f1_a = f1_score(y_pred, y_test)
dt_mcc_a = matthews_corrcoef(y_pred, y_test)
dt_kap_a = cohen_kappa_score(y_pred, y_test)
dt_log_a = log_loss(y_pred, y_test)
```

Fig.8 Decision trees

K-Nearest Neighbors is a non parametric algorithm that classifies a data point based on

the majority class of its k-nearest neighbors in the feature space. KNN is valuable for identifying patterns in data without assuming a specific functional form. It can capture local relationships within ASD datasets that might not be evident globally [12,13].

```
from sklearn.neighbors import KNeighborsClassifier
#from sklearn.pipeline import Pipeline

# instantiate the model
knn = KNeighborsClassifier(n_neighbors=3)

# fit the model
knn.fit(X_train,y_train)

y_pred = knn.predict(X_test)
y_prob = knn.predict_proba(X_test)

knn_acc_a = accuracy_score(y_pred, y_test)
knn_roc_a = roc_auc_score(y_pred, y_test)
knn_prec_a = precision_score(y_pred, y_test)
knn_rec_a = recall_score(y_pred, y_test)
knn_f1_a = f1_score(y_pred, y_test)
knn_mcc_a = matthews_corrcoef(y_pred, y_test)
knn_kap_a = cohen_kappa_score(y_pred, y_test)
knn_log_a = log_loss(y_pred, y_test)
```

Fig.9 KNN

Naive Bayes is a probabilistic classifier based on Bayes' theorem with the assumption of independence between features. Naive Bayes is computationally efficient and works well with high dimensional datasets. Its simplicity and speed make it suitable for the initial exploration of ASD data.

```
from sklearn.naive_bayes import GaussianNB
#from sklearn.pipeline import Pipeline

# instantiate the model
nb=  GaussianNB()

# fit the model
nb.fit(X_train,y_train)

y_pred = nb.predict(X_test)
y_prob = nb.predict_proba(X_test)

nb_acc_a = accuracy_score(y_pred, y_test)
nb_roc_a = roc_auc_score(y_pred, y_test)
nb_prec_a = precision_score(y_pred, y_test)
nb_rec_a = recall_score(y_pred, y_test)
nb_f1_a = f1_score(y_pred, y_test)
nb_mcc_a = matthews_corrcoef(y_pred, y_test)
nb_kap_a = cohen_kappa_score(y_pred, y_test)
nb_log_a = log_loss(y_pred, y_test)
```

Fig.10 Naïve bayes

Logistic Regression is a linear model for binary classification that predicts the probability of an instance belonging to a particular class using the logistic function. Logistic Regression is interpretable and provides insights into the relationship between features and the likelihood of ASD. It serves as a baseline model for binary classification tasks.

```
# Logistic Regression model
from sklearn.linear_model import LogisticRegression
#from sklearn.pipeline import Pipeline

# instantiate the model
log = LogisticRegression()

# fit the model
log.fit(X_train,y_train)

y_pred = log.predict(X_test)
y_prob = log.predict_proba(X_test)

lr_acc_a = accuracy_score(y_pred, y_test)
lr_roc_a = roc_auc_score(y_pred, y_test)
lr_prec_a = precision_score(y_pred, y_test)
lr_rec_a = recall_score(y_pred, y_test)
lr_f1_a = f1_score(y_pred, y_test)
lr_mcc_a = matthews_corrcoef(y_pred, y_test)
lr_kap_a = cohen_kappa_score(y_pred, y_test)
lr_log_a = log_loss(y_pred, y_test)
```

Fig.11 Logistic regression

Support Vector Machine is a supervised learning algorithm that finds the hyperplane

that best separates classes in a high dimensional space. SVM is effective in handling complex decision boundaries. It can capture non linear relationships in ASD datasets, potentially improving classification accuracy [12,13].

```
from sklearn.svm import SVC
svc = SVC()

# fitting the model for grid search
svc.fit(X_train, y_train)

y_pred = svc.predict(X_test)
#y_prob = svc.predict_proba(X_test)

svc_acc_a = accuracy_score(y_pred, y_test)
svc_roc_a = roc_auc_score(y_pred, y_test)
svc_prec_a = precision_score(y_pred, y_test)
svc_rec_a = recall_score(y_pred, y_test)
svc_f1_a = f1_score(y_pred, y_test)
svc_mcc_a = matthews_corrcoef(y_pred, y_test)
svc_kap_a = cohen_kappa_score(y_pred, y_test)
svc_log_a = log_loss(y_pred, y_test)
```

Fig.12 SVM

Linear Discriminate Analysis is dimensionality reduction and classification technique that finds linear combinations of features that best separate classes. [23,26] LDA is useful for reducing dimensionality and highlighting discriminating features. It can enhance interpretability and may aid in Identifying critical factors in ASD detection.

```
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis

clf = LinearDiscriminantAnalysis()

# fitting the model for grid search
clf.fit(X_train, y_train)

y_pred = clf.predict(X_test)
#y_prob = svc.predict_proba(X_test)

lda_acc_a = accuracy_score(y_pred, y_test)
lda_roc_a = roc_auc_score(y_pred, y_test)
lda_prec_a = precision_score(y_pred, y_test)
lda_rec_a = recall_score(y_pred, y_test)
lda_f1_a = f1_score(y_pred, y_test)
lda_mcc_a = matthews_corrcoef(y_pred, y_test)
lda_kap_a = cohen_kappa_score(y_pred, y_test)
lda_log_a = log_loss(y_pred, y_test)
```

Fig.13 LDA

A Voting Classifier, combining is a form of ensemble learning where multiple individual classifiers are trained, and their predictions are combined to make a final prediction. In this project, we have chosen AdaBoost and Random Forest as the base classifiers.

```
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import VotingClassifier
clf1 = AdaBoostClassifier(n_estimators=100, random_state=0)
clf2 = RandomForestClassifier(n_estimators=100, random_state=0)
clf3 = DecisionTreeClassifier(max_depth=30)
eclf1 = VotingClassifier(estimators=[('ab', clf1), ('rf', clf2), ('dt', clf3)], voting='soft')
eclf1.fit(X_train, y_train)
y_pred = eclf1.predict(X_test)

vot_acc_a = accuracy_score(y_pred, y_test)
vot_roc_a = roc_auc_score(y_pred, y_test)
vot_prec_a = precision_score(y_pred, y_test)
vot_rec_a = recall_score(y_pred, y_test)
vot_f1_a = f1_score(y_pred, y_test)
vot_mcc_a = matthews_corrcoef(y_pred, y_test)
vot_kap_a = cohen_kappa_score(y_pred, y_test)
vot_log_a = log_loss(y_pred, y_test)


storeResults('Voting Classifier',vot_acc_a,vot_roc_a,vot_prec_a,vot_rec_a,vot_f1_a,vot_mcc_a,vot_kap_a
```

Fig.14 Voting classifier

4.EXPERIMENTAL RESULTS

Precision: Precision evaluates the fraction of correctly classified instances or samples among the ones classified as positives. Thus, the formula to calculate the precision is given by:

Precision = True positives/ (True positives + False positives) = TP/(TP + FP)

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

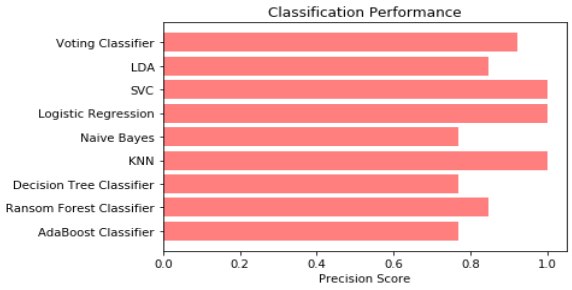

Fig.15 Precision comparison graph

Recall: Recall is a metric in machine learning that measures the ability of a model to identify all relevant instances of a particular class. It is the ratio of correctly predicted positive observations to the total actual positives, providing insights into a model's completeness in capturing instances of a given class.

$$Recall = \frac{TP}{TP + FN}$$

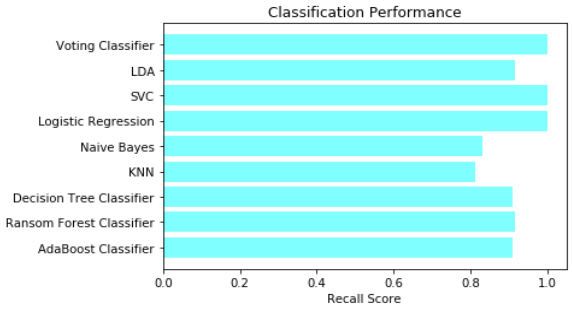

Fig.16 Recall comparison graph

Accuracy: Accuracy is the proportion of correct predictions in a classification task, measuring the overall correctness of a model's predictions.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

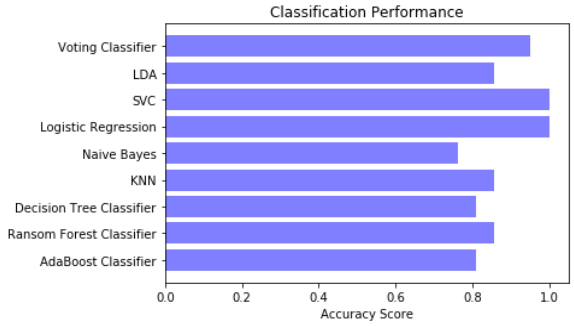

Fig.17 Accuracy graph

F1 Score: The F1 Score is the harmonic mean of precision and recall, offering a balanced measure that considers both false positives and false negatives, making it suitable for imbalanced datasets.

$$\text{F1 Score} = 2 * \frac{\text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} * 100$$

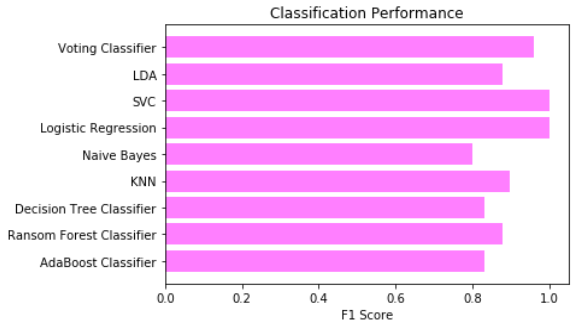

Fig.18 F1Score

ML Model	Accuracy	Precision	Recall	F1-Score
AdaBoost	1.000	1.00	1.000	1.000
Random Forest	1.000	1.00	1.000	1.000
Decision Tree	1.000	1.00	1.000	1.000
KNN	0.943	0.94	0.904	0.922
Naive Bayes	0.979	0.96	0.980	0.970
Logistic Regression	0.993	0.98	1.000	0.990
SVC	0.993	1.00	0.980	0.990
LDA	0.936	0.88	0.936	0.907
Voting Classifier	1.000	1.00	1.000	1.000

Fig.19 Performance Evaluation

5.CONCLUSION

The project has successfully introduced an innovative machine learning framework for the early detection of Autism Spectrum Disorder (ASD), showcasing a blend of advanced algorithms and feature scaling strategies. Through rigorous evaluation on standard ASD datasets representing Toddlers, Adolescents, Children, and Adults, the framework's robust performance across diverse age groups underscores its versatility and potential clinical applicability [12,13]. The identifies optimal classification methods and feature scaling techniques within the framework, offering a refined and effective approach for early ASD detection, with potential implications for timely interventions. The ensemble algorithm, incorporating Random Forest and AdaBoost, has demonstrated exceptional performance in ASD detection, achieving heightened accuracy. Moreover, its seamless integration into a user-friendly front end, where feature values can be easily inputted and tested, underscores its practicality and effectiveness in real-world applications. Leveraging feature selection techniques, the project provides insightful attribute rankings, highlighting key risk factors and influential features crucial for understanding the complexities of ASD, supporting accurate diagnosis.

6.FUTURE SCOPE

The project outlines its intention to collect more data related to Autism Spectrum Disorder (ASD) and construct a more generalized prediction model for people of any age to improve ASD detection and other neuro-developmental disorders [18]. This indicates that future work could involve expanding the dataset used in the study to include a larger and more diverse sample of individuals with ASD. Additionally, the project suggests the development of a more generalized prediction model, which could involve incorporating additional machine learning techniques or refining the existing framework to improve the accuracy and reliability of ASD detection. The future scope of the project could also involve exploring other neuro-developmental disorders and investigating the potential application of the proposed framework in detecting and predicting these disorders. Overall the future scope of the project includes further data collection, model refinement, and potential

expansion to other neuro-developmental disorders.

REFERENCES

[1] M. Bala, M. H. Ali, M. S. Satu, K. F. Hasan, and M. A. Moni, ‘‘Efficient machine learning models for early stage detection of autism spectrum disorder,’’ Algorithms, vol. 15, no. 5, p. 166, May 2022.

[2] D. Pietrucci, A. Teofani, M. Milanesi, B. Fosso, L. Putignani, F. Messina, G. Pesole, A. Desideri, and G. Chillemi, ‘‘Machine learning data analysis highlights the role of parasutterella and alloprevotella in autism spectrum disorders,’’ Biomedicines, vol. 10, no. 8, p. 2028, Aug. 2022.

[3] R. Sreedasyam, A. Rao, N. Sachidanandan, N. Sampath, and S. K. Vasudevan, ‘‘Aarya—A kinesthetic companion for children with autism spectrum disorder,’’ J. Intell. Fuzzy Syst., vol. 32, no. 4, pp. 2971–2976, Mar. 2017.

[4] J. Amudha and H. Nandakumar, ‘‘A fuzzy based eye gaze point estimation approach to study the task behavior in autism spectrum disorder,’’ J. Intell. Fuzzy Syst., vol. 35, no. 2, pp. 1459–1469, Aug. 2018.

[5] H. Chahkandi Nejad, O. Khayat, and J. Razjouyan, ‘‘Software development of an intelligent spirography test system for neurological disorder detection and quantification,’’ J. Intell. Fuzzy Syst., vol. 28, no. 5, pp. 2149–2157, Jun. 2015.

[6] F. Z. Subah, K. Deb, P. K. Dhar, and T. Koshiba, ‘‘A deep learning approach to predict autism spectrum disorder using multisite resting-state fMRI,’’ Appl. Sci., vol. 11, no. 8, p. 3636, Apr. 2021.

[7] K.-F. Kollias, C. K. Syriopoulou-Delli, P. Sarigiannidis, and G. F. Fragulis, ‘‘The contribution of machine learning and eye-tracking technology in autism spectrum disorder research: A systematic review,’’ Electronics, vol. 10, no. 23, p. 2982, Nov. 2021.

[8] I. A. Ahmed, E. M. Senan, T. H. Rassem, M. A. H. Ali, H. S. A. Shatnawi, S. M. Alwazer, and M. Alshahrani, ‘‘Eye tracking-based diagnosis and early detection of autism spectrum disorder using machine learning and deep learning techniques,’’ Electronics, vol. 11, no. 4, p. 530, Feb. 2022.

[9] P. Sukumaran and K. Govardhanan, ‘‘Towards voice based prediction and analysis of emotions in ASD children,’’ J. Intell. Fuzzy Syst., vol. 41, no. 5, pp. 5317–5326, 2021.

[10] S. P. Abirami, G. Kousalya, and R. Karthick, ‘‘Identification and exploration of facial expression in children with ASD in a contact less environment,’’ J. Intell. Fuzzy

Syst., vol. 36, no. 3, pp. 2033–2042, Mar. 2019.

AUTHOR PROFILE:

Mr. B. SURESH REDDY, done his M. Tech (Masters of Technology) in Arjun College of Technology & Sciences. At JNTU Hyderabad. Assistant Professor in the department of CSE at QIS College of Technology (Autonomous), Vengamukkapalem(V), Ongole, Prakasam. His areas of interest are Data Structures, Machine learning, and Web technologies.

Mr. KAMBALA KALYAN RAO currently pursuing Master of Computer Applications at QIS College of engineering and Technology (Autonomous), Ongole, Andhra Pradesh. He Completed B.Sc(Computer Science) from SRI SAI Chaitanya Degree College, Giddaluru, Prakasam Dt, Andhra Pradesh. His areas of interests are Deep Learning & Machine Learning