Project Milestone

Early Detection of Autism Spectrum Disorder across Different Age Groups using Machine Learning

Senthilkumar, Malini
School of Applied Computational Science
Meharry Medical College
Nashville, TN, USA
msenthilkuma24@mmc.edu

Nixon, Mikaela
School of Applied Computational Science
Meharry Medical College
Nashville, TN, USA
mnixon21@mmc.edu

Graves, Sade
School of Applied Computational Science
Meharry Medical College
Nashville, TN, USA
sgraves24@mmc.edu

Abstract - This study addresses the critical need for early detection of Autism Spectrum Disorder (ASD) across various age groups—including toddlers, children, adolescents, and adults. Recognizing the profound impact of ASD on early development, particularly in toddlers (2-4yrs) and children (5-18 yrs), the study employs a diverse set of machine learning models such as Logistic Regression, K-Nearest Neighbors (KNN), Decision Trees (DT), Support Vector Machines (SVM), and Neural Networks (NN). A key aspect of the methodology is meticulous feature selection to identify the top four features most indicative of ASD across these age groups. By leveraging these algorithms and applying a stacking technique that combines the strengths of different classifiers, the study aims to develop highly accurate predictive models for early ASD detection.

Keywords —Risk Factor, Classification, environment interactions, autism spectrum disorder, screening.

I. INTRODUCTION

Autism Spectrum Disorder (ASD), commonly known as autism, is a complex condition that affects behavior and communication. It often involves repetitive actions and challenges in social interactions, including difficulties in understanding others' thoughts and feelings. Symptoms typically manifest within the first three years of life and include communication difficulties, social contact challenges, narrowed interests, and repetitive behaviors.

Early detection of autism is crucial because intervention and therapy can significantly improve communication skills and overall development. Symptoms usually begin to appear between 12 to 18 months of age, making timely diagnosis essential for effective intervention.

Diagnosing ASD is challenging due to the lack of conventional medical tests; instead, physicians rely on observational and psychological assessments, examining various aspects of an individual's daily life.

Recent research has enhanced the identification of autism by using advanced machine-learning algorithms for feature selection to pinpoint the most important characteristics indicative of ASD. Many of the current ML approaches leverages the strengths of multiple classifiers to provide more accurate and reliable results for early autism detection.

To address these inefficiencies, a precise and scalable computational model is required—one that can accurately predict ASD risk by integrating ASD data with advanced deep learning techniques. This would streamline the identification process, reducing the need for repetitive, labor-intensive analyses and enhancing the ability to detect novel risk genes with greater speed and accuracy.

MOTOR DIFFICULTIES
COMMUNICATION DIFFICULTIES
...
AUTISM
SPECTRUM DISORDER
REPETITIVE BEHAVIOR
SOCIAL IMPEDIMENTS
?
SENSITIVITY TO CHANGE AND STIMULATION

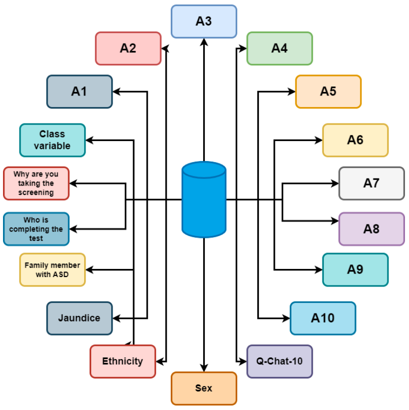


Fig 1. Features noted with ASD diagnosis

II. DATA DESCRIPTION

A. Autistic Spectrum Disorder Screening Data: UCI Machine Learning Repository

The Autistic Spectrum Disorder Screening Data is a publicly available dataset hosted by the UCI Machine Learning Repository. This dataset is designed to support research in the early detection of autism spectrum disorder ASD) in toddlers, adults, children and adolescents by providing screening data that can be used to develop and evaluate machine learning models.

Dataset Characteristics:

The dataset used is multivariate, consisting of 21 categorical and integer attributes, and includes toddlers, children, adolescents, and adults.

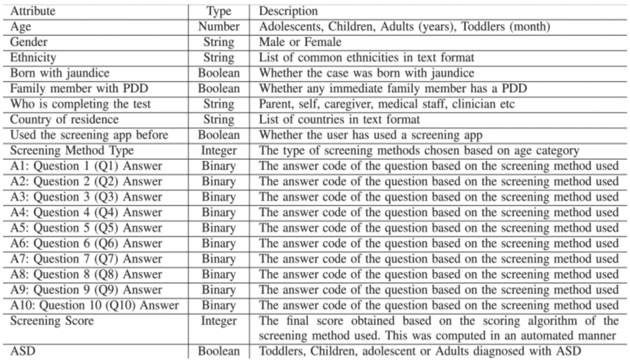
- Type: Multivariate
- Number of Instances:
  - Toddlers: 1054
  - Children: 292
  - Adolescent: 104
  - Adults: 704

- Number of Attributes: 21
- Attribute Types: Categorical, Integer
- Associated Tasks: Classification

The study utilizes four ASD datasets covering toddlers, adolescents, children, and adults, sourced from public repositories on Kaggle and UCI ML. These datasets, originally developed through the ASDTests smartphone app, use QCHAT-10 and AQ-10 screening tools to score individuals on a scale of 0 to 10, where a score of 6 or higher suggests a positive ASD diagnosis. Data from the ASDTests app, alongside open-source databases, supports ongoing research in ASD detection.

Detailed descriptions of the datasets are provided in Table 1 and Table 2.

Table 1. Datasets Description



Attribute	Type	Description
Age	Number	Adolescents, Children, Adults (years), Toddlers (month)
Gender	String	Male or Female
Ethnicity	String	List of common ethnicities in text format
Born with jaundice	Boolean	Whether the case was born with jaundice
Family member with PDD	Boolean	Whether any immediate family member has a PDD
Who is completing the test	String	Parent, self, caregiver, medical staff, clinician etc
Country of residence	String	List of countries in text format
Used the screening app before	Boolean	Whether the user has used a screening app
Screening Method Type	Integer	The type of screening methods chosen based on age category
A1: Question 1 (Q1) Answer	Binary	The answer code of the question based on the screening method used
A2: Question 2 (Q2) Answer	Binary	The answer code of the question based on the screening method used
A3: Question 3 (Q3) Answer	Binary	The answer code of the question based on the screening method used
A4: Question 4 (Q4) Answer	Binary	The answer code of the question based on the screening method used
A5: Question 5 (Q5) Answer	Binary	The answer code of the question based on the screening method used
A6: Question 6 (Q6) Answer	Binary	The answer code of the question based on the screening method used
A7: Question 7 (Q7) Answer	Binary	The answer code of the question based on the screening method used
A8: Question 8 (Q8) Answer	Binary	The answer code of the question based on the screening method used
A9: Question 9 (Q9) Answer	Binary	The answer code of the question based on the screening method used
A10: Question 10 (Q10) Answer	Binary	The answer code of the question based on the screening method used
Screening Score	Integer	The final score obtained based on the scoring algorithm of the screening method used. This was computed in an automated manner
ASD	Boolean	Toddlers, Children, adolescent or Adults diagnosed with ASD

Table 2. Feature Description of the ASD Datasets

Description of Child's Behavior

Item	Description
A1	When you call their name, does your youngster make eye contact in return?
A2	How comfortable is it for your youngster to look you in the eye?
A3	When anything, like a toy that's out of reach, is desired by your kid, does he or she point?
A4	Is your youngster indicate that they both find something interesting, like a fascinating sight?
A5	Does the kid engage in pretend play, like caring for playthings or talking on a dummy phone?

A6	Does the youngster follow your gaze to see where you are looking?
A7	Does your child exhibit any outward signals of wanting to console someone who appears distressed, such as petting their hair or offering them a hug?
A8	What would you say about your child's initial verbal exchanges?
A9	Is your youngster wave you off or make other basic gestures?
A10	Does your youngster sometimes look at nothing for no apparent reason?

C. Purpose and Use Cases:

- Early Detection of ASD: The dataset aids in developing tools for early screening of ASD in toddlers, adults, children and adolescents, which is critical for timely intervention.
- Machine Learning Model Development: Researchers can utilize the dataset to train and test classification algorithms (e.g., CNN-LSTM models) to predict ASD risk based on behavioral data.
- Behavioral Analysis: Analysis of responses can help identify key behavioral indicators that are strong predictors of ASD. While the dataset focuses on screening results rather than biological mechanisms, it provides insights into behavioral and demographic factors associated with ASD.

Understanding these factors can contribute to:

- Identifying Behavioral Risk Factors: Highlighting which behaviors are most indicative of ASD risk.
- Exploring Demographic Influences: Examining how age, gender, ethnicity, and family history correlate with ASD screening outcomes.
- Supporting Multimodal Research: When combined with biological data (e.g., environment and genetic data), it can help build comprehensive models` that consider both behavioral and biological risk factors.

I. RELATED WORK

In recent years, machine learning has significantly advanced the detection of Autism Spectrum Disorder (ASD) across different age groups, from toddlers to adults. Research in this field has leveraged various approaches, including traditional classifiers, deep learning, and ensemble methods, aiming to improve the accuracy and reliability of ASD screening. One prominent direction is the application of feature selection and stacking techniques, combining multiple classifiers to improve prediction accuracy for early ASD detection, particularly in children. Studies have also explored the use of federated learning (FL), which allows training on distributed data sources while preserving privacy.

Overall, related work in ASD detection increasingly focuses on integrating innovative machine learning techniques, such as federated learning and stacked ensemble models, to address challenges in early diagnosis, data privacy, and diverse data representation across age groups. These advancements hold promise for more accurate and accessible ASD screening tools, potentially transforming early intervention strategies.

A. Detection of autism spectrum disorder (ASD) in children and adults

- Farooq, M.S., Tehseen, R., Sabir, M. *et al.*, 2023, Detection of autism spectrum disorder (ASD) in children and adults using machine learning.

This study focuses on detecting Autism Spectrum Disorder (ASD) in both children and adults by employing machine learning techniques with a federated learning (FL) framework. ASD impacts social and cognitive functions, leading to repetitive behaviors, restricted interests, and communication challenges. Early detection is essential for reducing the severity and long-term effects of ASD.

Key contributions of this paper include:

- Application of Federated Learning (FL) for ASD Detection: The study uniquely applies FL, a decentralized approach, to ASD detection. This method enables local training of machine learning models, specifically logistic regression and support vector machine (SVM), on different datasets without sharing raw data, thereby maintaining data privacy and reducing communication costs.

- Meta-Classifier for Enhanced Accuracy: Results from locally trained classifiers are aggregated on a central server, where a meta-classifier is trained to identify the most effective model for ASD detection across children and adults. This approach allows for higher accuracy in predicting ASD by leveraging the strengths of multiple classifiers.

- Dataset Diversity and Performance: The researchers utilized four distinct ASD datasets, each containing over 600 records from affected children and adults to train and evaluate the model. The model achieved a high prediction accuracy, with 98% accuracy in children and 81% accuracy in adults, demonstrating the effectiveness of FL and machine learning for age-specific ASD detection.

This paper highlights the potential of federated learning and machine learning models in improving diagnostic accuracy for ASD while respecting data privacy, presenting a significant advancement in the early and precise detection of ASD in varied age groups.

B. Newly proposed technique for autism spectrum disorder-based machine learning

- Dr. Sherif Kamel and Rehab Al-harbi

This study addresses the urgent need for effective and easily deployable screening methods to identify Autism Spectrum Disorder (ASD) among toddlers, given the rapid rise in ASD diagnoses. The key contribution is the development of a Logistic Regression-based machine learning model tailored to predict ASD in young children based on behavioral data from healthcare datasets. The authors emphasize the benefits of machine learning in reducing the time required for ASD diagnosis, allowing for quicker intervention and improved outcomes for affected children.

The paper highlights challenges in implementing machine learning solutions in healthcare, such as data accessibility, tool integration, and resource allocation, which limit the broader adoption of AI-driven diagnostic tools. By demonstrating how Logistic Regression can be applied to behavior classification for ASD detection, the study contributes to the growing body of research that seeks to streamline ASD screening processes and support early intervention efforts in healthcare.

C. Machine learning for autism spectrum disorder diagnosis–challenges

- X. Cao and J. Cao

Over the past decade, there has been significant progress in computer-assisted diagnosis (CAD) using machine learning, particularly with advancements in deep learning and transfer learning. These techniques leverage large public datasets to develop robust representations, fine-tuning them for specific clinical fields. This approach has proven beneficial in mental health, as demonstrated by Kirtley et al. (2022), who applied machine learning algorithms to predict suicidal ideation, showcasing the potential of CAD systems in mental illness prevention. However, the application of machine learning to Autism Spectrum Disorder (ASD) remains challenging due to ASD's heterogeneity and its overlapping cognitive features with other conditions, making data collection and CAD design difficult. Key contributions of this work include highlighting the effectiveness of machine learning in mental health diagnostics and identifying challenges specific to ASD applications.

D. Machine Learning for Diagnostic Assessment of ASD with Comorbidities

- Schulte-Ruther et al.

Schulte-Ruther et al. focused on enhancing ASD diagnosis amidst overlapping disorders, such as anxiety, ADHD, and conduct disorder, using machine learning. Their study involved training random forest models on Autism Diagnostic Observation Schedule (ADOS) item scores from 1,251 samples, including individuals with ASD, anxiety disorders, ADHD, and conduct disorder. A key contribution of this work is its focus on predicting ASD in the presence of other diagnoses and identifying item importance profiles for ASD and comorbid conditions. This study emphasizes the potential of machine learning to refine diagnostic accuracy and effectively address ASD's diagnostic complexity when comorbid conditions are present.

E. Use of Machine Learning to Improve Autism Screening and Diagnostic Instruments: Effectiveness, Efficiency, and Multi-Instrument Fusion

- T. Shrivastava, V. Singh, and A. Agrawal

This paper explores the application of machine learning (ML) to enhance the accuracy, efficiency, and integration of autism screening and diagnostic tools.

Key contributions include:

1. Improvement in Screening Efficiency: The study utilizes ML algorithms to streamline autism screening processes, significantly improving diagnostic speed while maintaining accuracy. This efficiency addresses the need for scalable and timely autism assessments, particularly in resource-limited settings.
2. Multi-Instrument Fusion: By integrating data from multiple diagnostic tools, the study demonstrates that ML can create a cohesive model that synthesizes insights across instruments, leading to more robust and reliable screening outcomes.
3. Enhanced Diagnostic Accuracy: The research highlights that ML models can improve the predictive accuracy of traditional autism diagnostic instruments, providing clinicians with higher confidence in their assessments and minimizing false positives and negatives.
4. Data-Driven Decision Support: This approach enables ML-driven decision support, assisting healthcare professionals in interpreting complex datasets to make more informed decisions about autism diagnosis and interventions.

Overall, this study contributes to the field by showcasing how ML can optimize autism screening and diagnostics through improved accuracy, efficiency, and the effective integration of multiple assessment tools.

II. APPROACH

Our approach integrates the application of machine learning (ML) to enhance the accuracy, efficiency, and integration of autism screening and diagnostic tools.

1. Improvement in Screening Efficiency: The study utilizes ML algorithms to streamline autism screening processes, significantly improving diagnostic speed while maintaining accuracy. This efficiency addresses the need for scalable and timely autism assessments, particularly in resource-limited settings.

2. Multi-Instrument Fusion: By integrating data from multiple diagnostic tools, the study demonstrates that ML can create a cohesive model that synthesizes insights across instruments, leading to more robust and reliable screening outcomes.

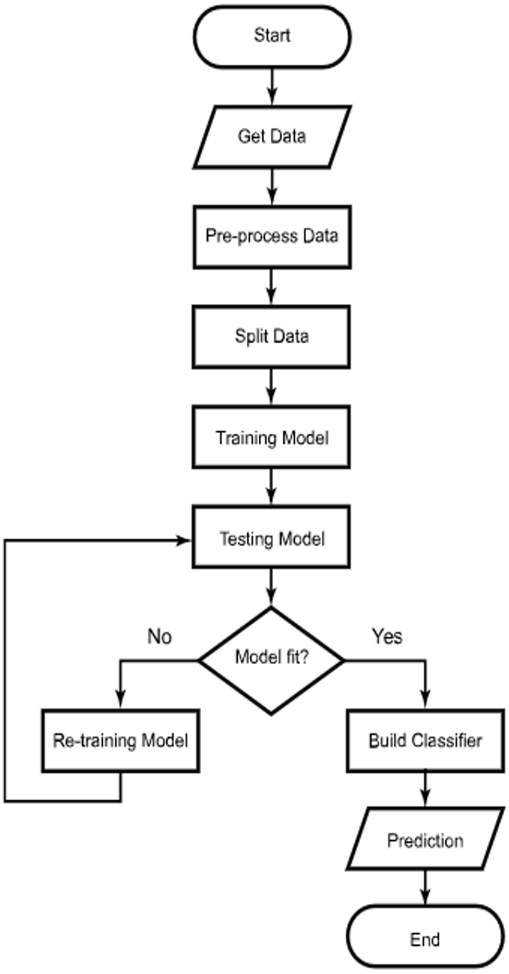
3. Enhanced Diagnostic Accuracy: The research highlights that ML models can improve the predictive accuracy of traditional autism diagnostic instruments, providing clinicians with higher confidence in their assessments and minimizing false positives and negatives.

4. Data-Driven Decision Support: This approach enables ML-driven decision support, assisting healthcare professionals in interpreting complex datasets to make more informed decisions about autism diagnosis and interventions.

In this research, we present a comprehensive machine learning (ML) framework aimed at improving early-stage detection of Autism Spectrum Disorder (ASD) across various age groups. Our approach begins by addressing the issue of imbalanced class distribution with Random Over Sampling to ensure that our models do not develop bias toward majority class samples, thus promoting balanced predictions. To further enhance model accuracy, we can select the most suitable Feature Scaling (FS) method for each ASD dataset, allowing us to normalize feature values effectively and improve overall prediction performance.

We apply and analyze the performance of multiple effective ML algorithms on each feature-scaled dataset, systematically identifying the best FS technique for each age-specific dataset. Additionally, we conduct an in-depth feature importance analysis on the most accurately scaled datasets, utilizing four feature selection techniques to pinpoint key risk factors predictive of ASD.
Finally, we will validate our approach by performing extensive experiments and comparisons across four standardized ASD datasets—spanning toddlers, children, adolescents, and adults—demonstrating the robustness and adaptability of our framework in early ASD prediction across age groups.

ML Process Flowchart:



The machine learning models to be used in the study:

1. AdaBoost (AB): AdaBoost is an ensemble classifier that combines multiple weak classifiers to reduce errors. It assigns weights to instances and retrains classifiers iteratively to improve accuracy. The final prediction is a weighted combination of the classifiers, making it effective for boosting overall performance.

2. Random Forest (RF): Random Forest is an ensemble of decision trees built from random samples of the dataset. Each tree votes on the classification, and the final prediction is based on majority voting. This method enhances stability and accuracy by reducing overfitting.

3. Decision Tree (DT): Decision Trees use a top-down approach to split data based on information gain, choosing attributes that best separate classes. This process creates rules for classification, making DTs interpretable and suitable for capturing complex decision boundaries.

4. Gaussian Naïve Bayes (GNB): GNB assumes normal distribution for each feature and calculates the probability of class membership based on mean and standard deviation. It is computationally efficient and effective for cases where the independence assumption roughly holds.

5. Logistic Regression (LR): Logistic Regression models the probability of an outcome based on input variables, transforming odds using a logistic function. It updates coefficients through gradient descent, making it useful for binary classification.

III. DATA EXPLORATION

The initial data analysis is primarily focused on features associated with Autism Spectrum Disorder (ASD) in children, adolescent, adults.

Initial Data Preparation

1. Data Import and Decoding:
   - The dataset is loaded using scipy.io.arff and converted into a panda DataFrame.
   - Text encoding is adjusted by decoding UTF-8 encoded text fields.
2. Feature Renaming and Selection:
   - Column names are simplified and renamed to make them more interpretable (e.g., born_with_jaundice, Q1_Score through Q10_Score for screening questions, etc.).
   - The dataset is filtered to retain relevant features, including demographic details, screening scores, and a binary label for ASD diagnosis.
3. Handling Missing Values:
   - Missing values in age are replaced with 0, and a model-based imputation is applied later.
   - Records containing ambiguous values for ethnicity based on country are handled by substituting plausible values through a dictionary mapping.
4. Binary Encoding:
   - Various categorical fields, such as gender, born_with_jaundice, family_member_with_PDD, and used_screening_app_before, are binary encoded to simplify the model input.
5. Data Transformation:
   - Continuous fields are standardized, while screening scores and target labels (ASD_Label)

are converted to integer types for easy numerical analysis.

6. Creating and Filling Additional Columns:
    - Fields like ethnicity are fixed based on mappings from other columns.
    - Null values in age are predicted using a Random Forest model trained on the available age data.

Exploratory Data Analysis (EDA)

Distribution Analysis:The data is visualized by age, gender, country, and ethnicity to understand the distribution of ASD cases. Various bar plots and scatter plots depict ASD prevalence across these categories.

1. Age wise

Children: The chart (3.1) illustrates the age-wise distribution of children diagnosed with ASD (orange) and non-ASD (blue), highlighting that younger age groups, particularly 4 and 5 years old, exhibit higher ASD counts compared to older age groups.

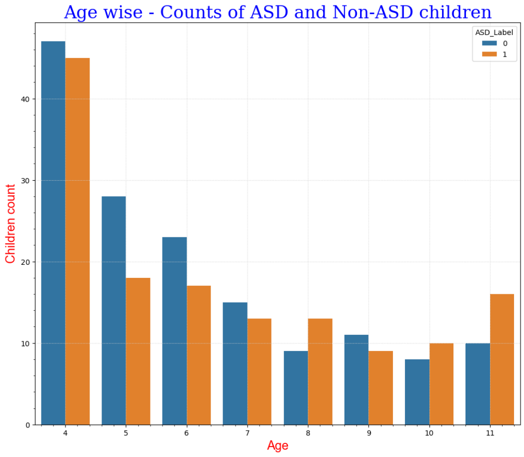


Fig 3.1 – Age-wise | Children

Adult: The plot (3.2) displays the age-wise distribution of adults diagnosed with ASD (orange) and non-ASD (blue), showing that younger adults, particularly those in their late teens and early twenties, have higher ASD counts compared to older age groups.

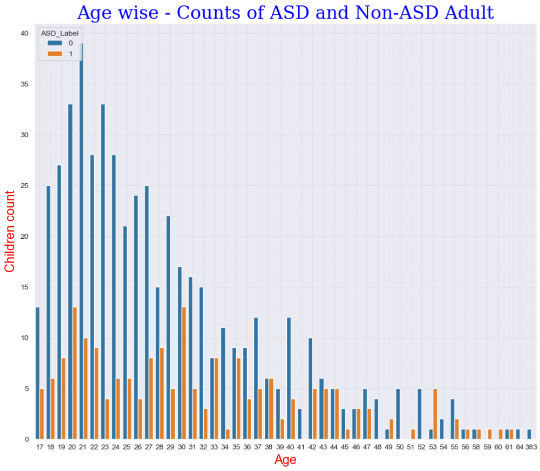


Fig 3.2 – Age-wise | Adult

2. Gender

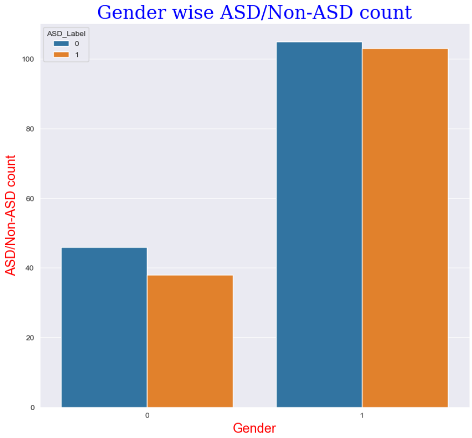


Fig 3.3 – Gender | Children

Children: The plot (3.3) illustrates the gender-wise distribution of ASD (orange) and non-ASD (blue) individuals, showing a higher representation of males (gender = 1) compared to females (gender = 0) in both categories, with ASD counts nearly equal among males.

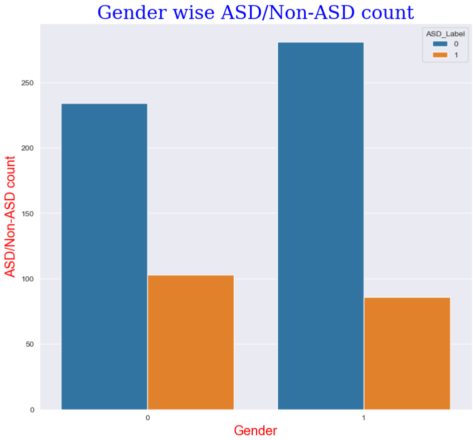


Fig 3.4 – Gender | Adult

Adult: The plot (3.4) depicts the gender-wise distribution of ASD (orange) and non-ASD (blue) individuals. Males (gender = 1) exhibit higher non-ASD counts compared to females (gender = 0), while ASD counts are similar across both genders, highlighting the gender disparity in the non-ASD population.

3. Country

Children: The plot (3.5) showcases the country-wise distribution of ASD (orange) and non-ASD (blue) children, highlighting a notable variation in ASD prevalence across countries, with the United States and "Others" showing relatively higher counts of ASD cases compared to other regions.

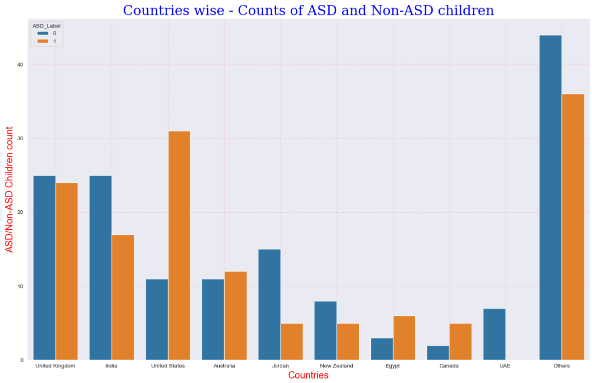


Fig 3.5 – Country | Children

Adult: The plot (3.6) displays the country-wise distribution of ASD (orange) and non-ASD (blue) adults, indicating higher counts of non-ASD individuals across most countries, with notable ASD cases in "Others," India, and the United States.

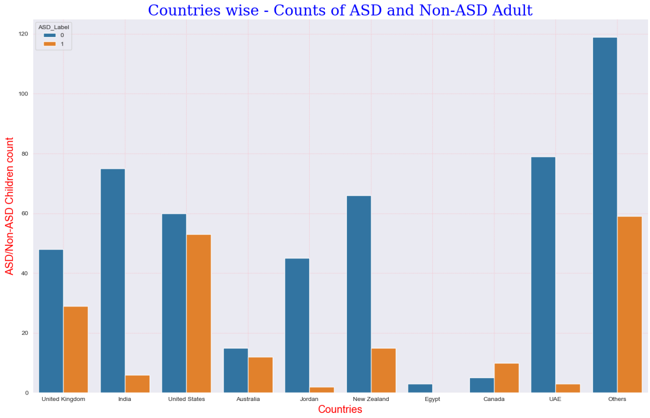


Fig 3.6 – Country | Adult

4. Ethnicity

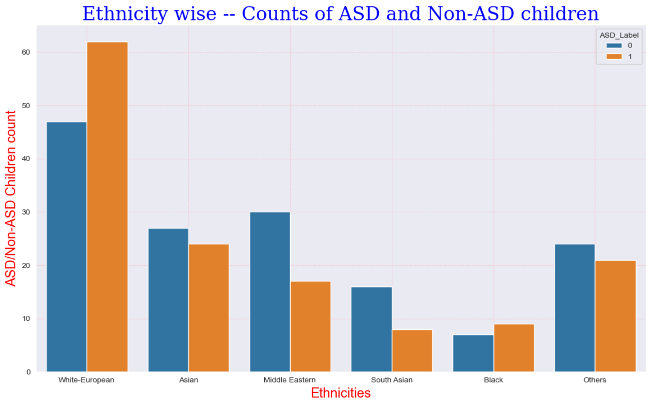


Fig 3.7 – Ethnicity | Children

Children: The plot (3.7) highlights that children of White-European ethnicity have the highest counts of ASD compared to other ethnic groups, while other ethnicities, such as South Asian and Black, show relatively lower ASD counts.

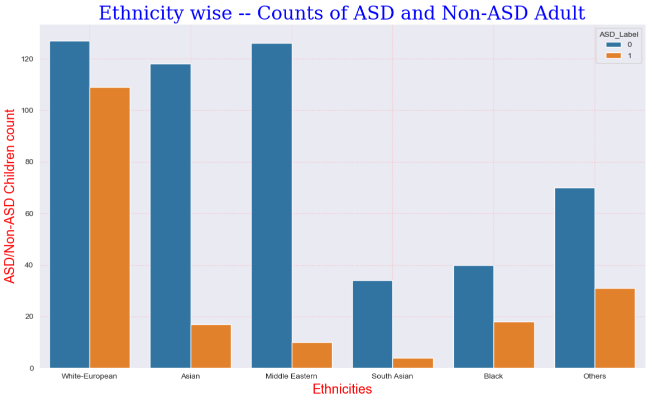


Fig 3.8 – Ethnicity | Adult

The plot shows that White-European adults have the highest counts of ASD compared to other ethnic groups, while South Asian and Black ethnicities show relatively lower counts for both ASD and Non-ASD categories.

ASD and Demographic Correlation:

- The analysis includes demographic-based correlations, showing insights such as how the presence of jaundice, family history of Pervasive Development Disorder (PDD), and prior usage of screening applications relate to ASD risk.

Feature Correlation Heatmap:

- A heatmap visualizes the correlation between features, highlighting screening_score as highly correlated with the ASD label, leading to its exclusion to avoid redundancy.

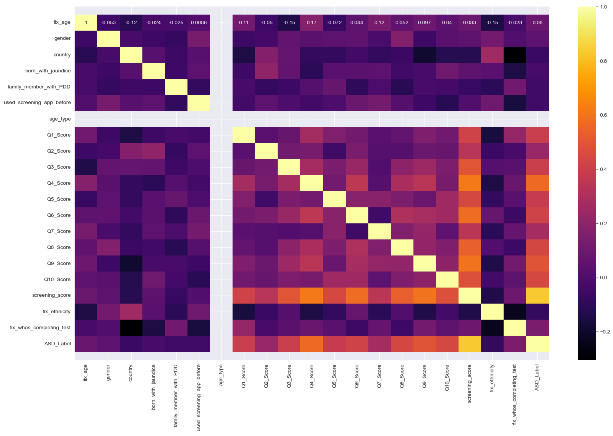


Fig 4.1 Correlation - Children

Key Observations (Fig 4.1)

1. Q1_Score to Q10_Score and ASD_Label:

The scores (Q1_Score to Q10_Score) show moderate to high positive correlations with the ASD_Label. This indicates that responses to the screening questions are strong predictors of whether a child has ASD or not. The screening_score feature, which is derived as the sum of the Q1_Score to Q10_Score, exhibits a strong positive correlation with ASD_Label. Dropping SCREENING_SCORE feature from the dataset as it is highly co-related with the TARGET Label. Other features like fix_age, gender, and fix_ethnicity show much weaker correlations with ASD_Label. This indicates that demographic variables alone are not as significant in predicting ASD compared to the screening scores.

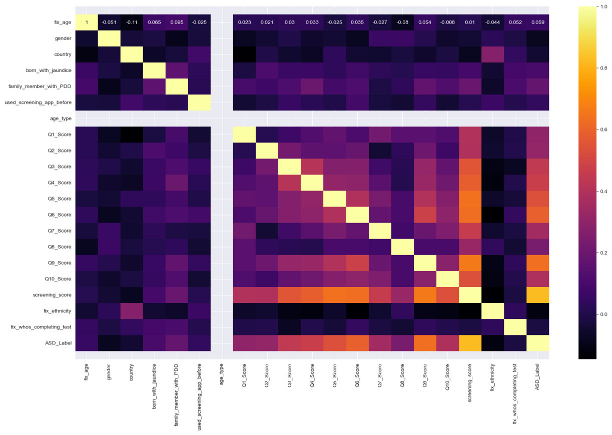


Fig 4.2 Correlation - Children

Key Observations (Fig 4.2)

Screening Scores (Q1_Score to Q10_Score) and ASD_Label:The screening scores exhibit moderate to high positive correlation with ASD_Label. Responses to the individual screening questions are strong predictors of whether an adult is ASD-positive or ASD-negative. It that directly reflects the likelihood of ASD and may overlap with individual screening scores, potentially introducing redundancy. Features like fix_age, gender, fix_ethnicity, family_member_with_PDD, and used_screening_app_before show weak correlations with ASD_Label. These demographic features are not as strong predictors of ASD compared to the screening scores.

Machine Learning Models

1. Data Split:
   - The dataset is divided into training and test sets.
   - Split Ratio – 70:30
2. Logistic Regression Classifier:
   - Logistic Regression is added as a baseline model.
   - The model is trained on the training set and evaluated on the test set for metrics.
3. Random Forest Classifier:
   - Random Forest Classifier is trained and evaluated for same metrics - precision, recall, accuracy, and AUC-ROC metrics. Cross-validation confirms the model's robustness.
4. Gradient Boosting and Naive Bayes Classifiers:
   - Gradient Boosting and Gaussian Naive Bayes classifiers are also trained, evaluated, and compared with the Random Forest model.

At the project midpoint, multiple classification models have been applied to predict Autism Spectrum Disorder (ASD) in children, including Random Forest, Gradient Boosting, Naive Bayes, Logistic Regression, and XGBoost. In further analysis, we plan to incorporate neural network and deep learning algorithms to enhance model performance. Each model has been evaluated using key metrics, including precision, recall, accuracy, and AUC-ROC.

IV. MODEL

We have implemented the following models to train and validate our data:

Logistic Regression: Logistic regression served as a baseline model due to its simplicity and interpretability. It models the probability of readmission as a linear combination of the features. The model's coefficients are straightforward to interpret, providing insights into the direction and strength of each feature's impact. However, it is limited in its ability to capture non-linear relationships in the data.

Random Forest: Random forest is an ensemble method that combines multiple decision trees, leveraging different subsets of data to enhance robustness and reduce overfitting. It performs well with non-linear data and provides feature importance metrics, making it useful for identifying key predictors. However, it can be computationally intensive and prone to overfitting if not carefully tuned.

Decision Tree: A decision tree is a simple, interpretable model that splits the data recursively based on feature thresholds. Each branch represents a decision rule, making it intuitive for users to understand. However, decision trees are prone to overfitting, especially in noisy datasets, and often require pruning or parameter tuning to improve generalizability.

XGBoost: XGBoost is an optimized version of gradient boosting designed for speed and efficiency. It incorporates regularization terms, making it less prone to overfitting and suitable for large datasets. While it offers high predictive power, it can be computationally intensive and requires careful parameter tuning.

LightGBM: LightGBM is a gradient-boosting framework optimized for efficiency and scalability. Its leaf-wise tree growth strategy reduces training time without sacrificing accuracy, making it ideal for large, high-dimensional datasets. However, it may be overfit on smaller datasets and requires careful handling of categorical features.

Deep Neural Network (DNN): Finally, we used a Deep Neural Network (DNN) to leverage its ability to model complex, non-linear relationships in the data. DNNs excel in predictive accuracy when trained with sufficient data and optimized effectively, providing a robust solution for our analysis. In this analysis, due to limited data, we did not use any DNN method for the prediction.

A. **Comparative Models Assessment – Children**

1. Random Forest Classifier

The performance metrics of the classification model indicate its effectiveness in predicting Autism Spectrum

- Accuracy: 95.89% — The model correctly predicted 95.89% of the cases.
- Precision: 92.59% — Among the positive predictions, 92.59% were true positives, showing low false positives.
- Recall: 96.15% — The model identified 96.15% of all actual ASD cases, indicating high sensitivity.
- F1 Score: 94.34% — A balanced measure of precision and recall, highlighting strong overall performance.

Confusion Matrix:

- True Negatives (TN): 45 — Non-ASD cases correctly classified.
- False Positives (FP): 2 — Non-ASD cases misclassified as ASD.
- False Negatives (FN): 1 — ASD cases misclassified as non-ASD.
- True Positives (TP): 25 — ASD cases correctly classified.

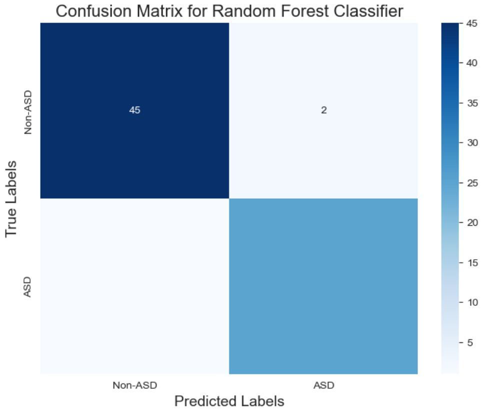


Fig 5.1 Confusion Matrix

- AUC: 0.96 — Demonstrates excellent discrimination ability between ASD and non-ASD samples, with a 96% chance of correctly distinguishing them.

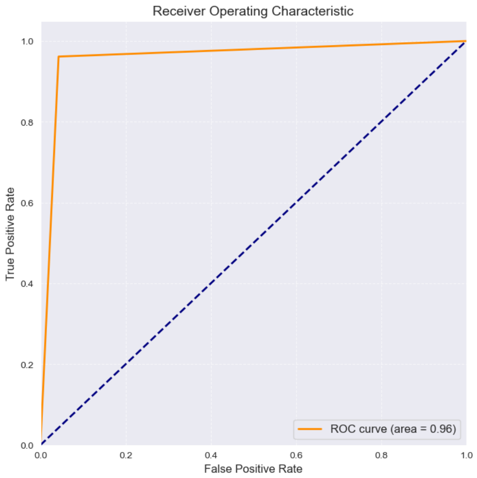


Fig 5.2 ROC

Conclusion:

The model exhibits strong performance, with excellent accuracy, precision, recall, and an AUC score. The low numbers of false positives and false negatives further emphasize its reliability, making it a robust classifier for distinguishing ASD from non-ASD cases in children.

2. Gradient Boosting Classifier

Disorder (ASD) cases:

- Accuracy: 98.63% — The model correctly predicted nearly all cases.
- Precision: 96.30% — Among the positive predictions, 96.30% were true positives, showing very few false positives.
- Recall: 100% — The model identified all actual ASD cases, indicating perfect sensitivity.
- F1 Score: 98.11% — A balanced measure of precision and recall, highlighting outstanding performance.
- AUC: 0.99 — Indicates excellent discrimination ability, with a 99% chance of correctly distinguishing between ASD and non-ASD samples.

Confusion Matrix:

- True Negatives (TN): 46 — Non-ASD cases correctly classified.
- False Positives (FP): 1 — Non-ASD cases misclassified as ASD.
- False Negatives (FN): 0 — No ASD cases were misclassified as non-ASD.
- True Positives (TP): 26 — All ASD cases were correctly classified.

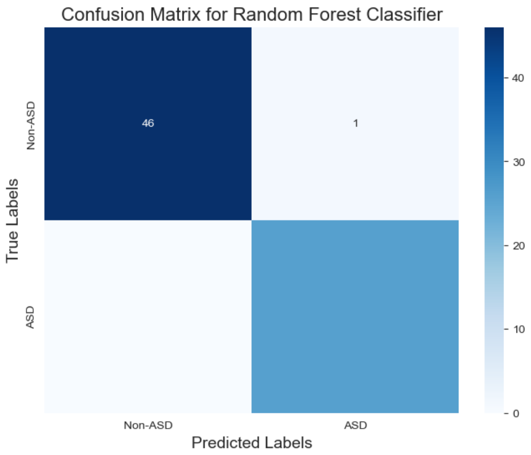


Fig 5.3 Confusion Matrix

Conclusion:

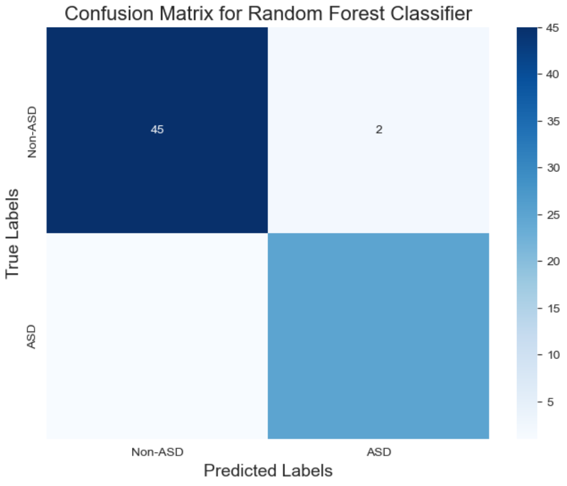
The model achieves near-perfect results, with extremely high accuracy, precision, recall, and AUC. Its ability to identify all true positives (ASD cases) while minimizing false positives and false negatives makes it an exceptionally reliable classifier for ASD detection.

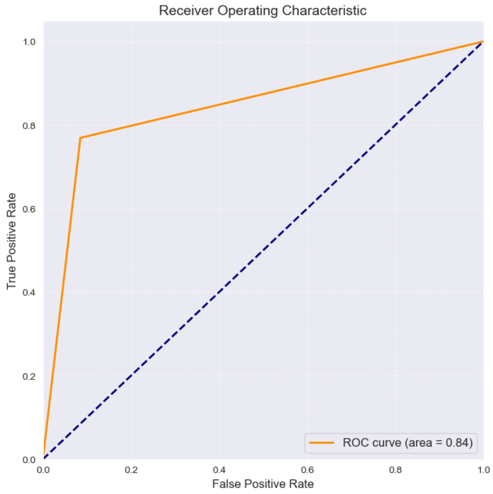
3. Naive Bayes Classifier

- Accuracy: 86.30% — The model correctly predicted 86.30% of the cases.
- Precision: 83.33% — Among the positive predictions, 83.33% were true positives, indicating a moderate level of false positives.
- Recall: 76.92% — The model identified 76.92% of all actual ASD cases, suggesting some missed cases (false negatives).
- F1 Score: 80.00% — Balances precision and recall, showing good overall performance.
- AUC: 0.84 — Indicates a strong ability to distinguish between ASD and non-ASD samples, with an 84% chance of correct classification.

Confusion Matrix:

- True Negatives (TN): 43 — Non-ASD cases correctly classified.
- False Positives (FP): 4 — Non-ASD cases misclassified as ASD.
- False Negatives (FN): 6 — ASD cases misclassified as non-ASD.
- True Positives (TP): 20 — ASD cases correctly classified.





Conclusion:

While the model shows strong accuracy and precision, the recall indicates that some ASD cases are being missed (false negatives). With an AUC of 0.84, the model performs well in distinguishing between ASD and non-ASD cases, but further optimization might improve sensitivity (recall) to ensure fewer missed ASD predictions.

B. **Comparative Models Assessment – Adult**

1. Random Forest Classifier

The performance metrics of the classification model indicate its effectiveness in predicting Autism Spectrum.

- Accuracy: 93.75% — The model correctly predicted 93.75% of the cases.
- Precision: 93.02% — Among the positive predictions, 93.02% were true positives, indicating very few false positives.
- Recall: 83.33% — The model identified 83.33% of all actual ASD cases, showing good sensitivity but with some missed cases (false negatives).
- F1 Score: 87.91% — Balances precision and recall, indicating robust overall performance.

Confusion Matrix:

- True Negatives (TN): 125 — Non-ASD cases correctly classified.
- False Positives (FP): 3 — Non-ASD cases misclassified as ASD.
- False Negatives (FN): 8 — ASD cases misclassified as non-ASD.
- True Positives (TP): 40 — ASD cases correctly classified.

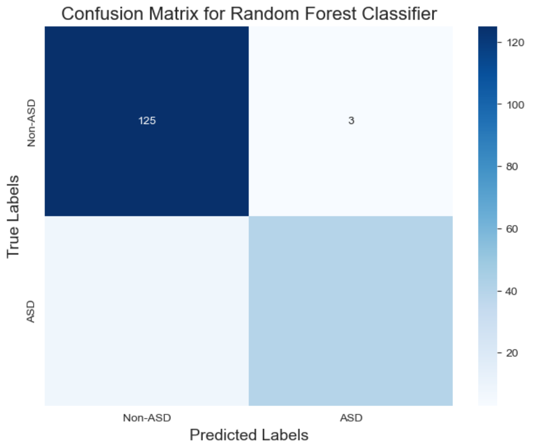


Fig 6.1 Confusion Matrix

AUC: 0.90 — Indicates excellent discrimination ability, with a 90% chance of correctly distinguishing between ASD and non-ASD samples.

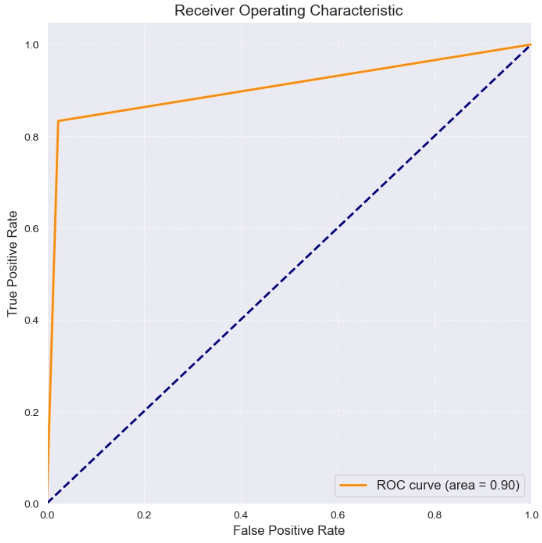


Fig 6.2 ROC

Conclusion:

The model exhibits strong performance, with high precision and accuracy. However, the recall suggests that a small percentage of ASD cases are being missed. With an AUC of 0.90, the model is excellent at distinguishing between ASD and non-ASD cases. Further improvements could focus on enhancing sensitivity to reduce the number of false negatives.

2. Gradient Boosting Classifier

- Accuracy: 95.45% — The model correctly predicted 95.45% of the cases.

- Precision: 93.48% — Among the positive predictions, 93.48% were true positives, showing very few false positives.
- Recall: 89.58% — The model identified 89.58% of all actual ASD cases, indicating strong sensitivity but with a small number of missed cases (false negatives).
- F1 Score: 91.49% — Balances precision and recall, highlighting robust overall performance.

Confusion Matrix:

- True Negatives (TN): 125 — Non-ASD cases correctly classified.
- False Positives (FP): 3 — Non-ASD cases misclassified as ASD.
- False Negatives (FN): 5 — ASD cases misclassified as non-ASD.
- True Positives (TP): 43 — ASD cases correctly classified.

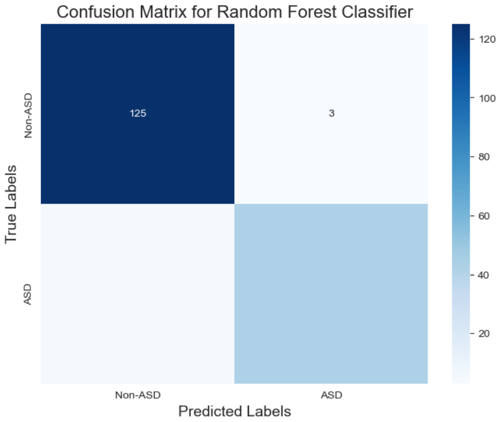


Fig 6.3 Confusion Matrix

AUC: 0.95 — Indicates excellent discrimination ability, with a 95% chance of correctly distinguishing between ASD and non-ASD samples.

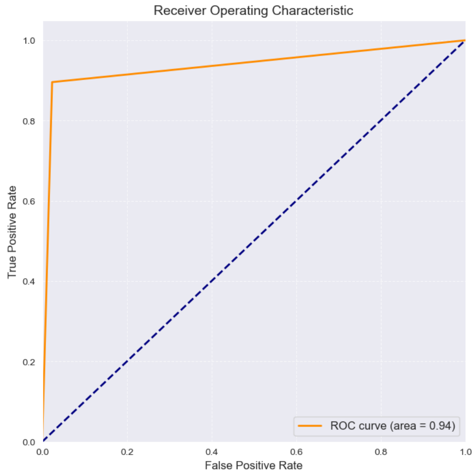


Fig 6.4 ROC

Conclusion:

The model achieves outstanding results, with high accuracy, precision, recall, and an AUC of 0.95. While a few ASD cases were missed (false negatives), the overall performance is highly reliable for distinguishing between ASD and non-ASD cases. Minor optimizations could focus on improving sensitivity to reduce the number of false negatives further.

3. Naive Bayes Classifier

- Accuracy: 96.02% — The model correctly classified 96.02% of the cases.
- Precision: 91.84% — Among the positive predictions, 91.84% were true positives, showing a low number of false positives.
- Recall: 93.75% — The model successfully identified 93.75% of all actual ASD cases, demonstrating high sensitivity.
- F1 Score: 92.78% — A balanced measure of precision and recall, reflecting strong overall performance.

Confusion Matrix:

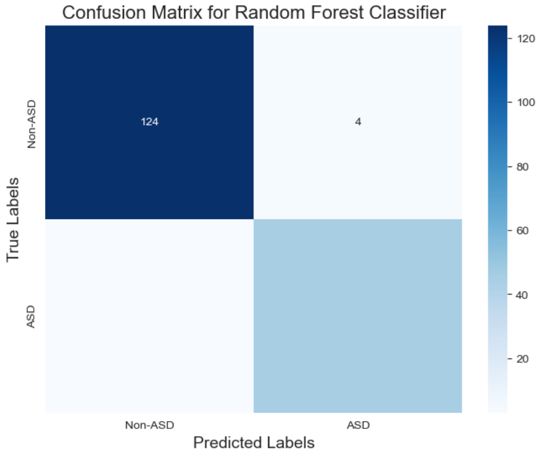


Fig 6.5 Confusion Matrix

- True Negatives (TN): 124 — Non-ASD cases correctly classified.
- False Positives (FP): 4 — Non-ASD cases misclassified as ASD.
- False Negatives (FN): 3 — ASD cases misclassified as non-ASD.
- True Positives (TP): 45 — ASD cases correctly classified.

AUC: 0.84 — Indicates excellent discrimination ability, with an 84% chance of correctly distinguishing between

ASD and non-ASD samples.

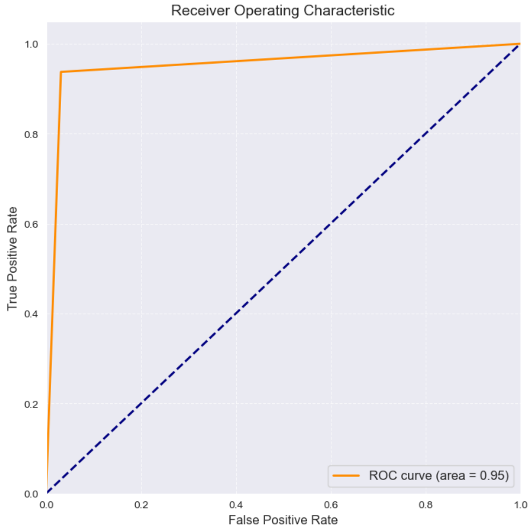


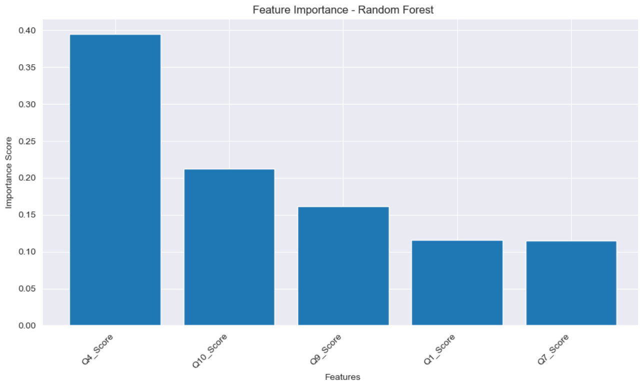
Fig 6.6 ROC

Conclusion:

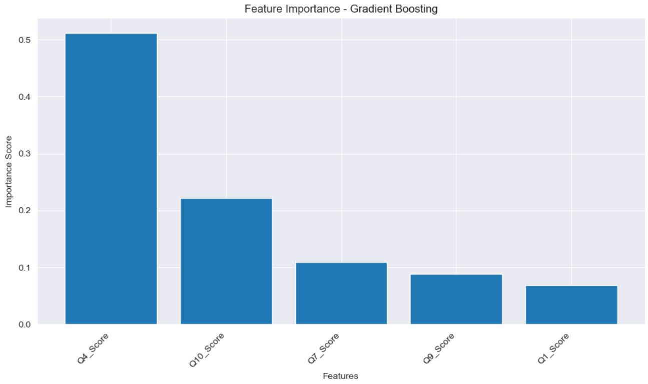
The model demonstrates excellent performance, with high accuracy, precision, recall, and F1 score. It has a low number of misclassifications, indicating strong reliability. While the AUC of 0.84 suggests good discriminative ability, further fine-tuning could enhance this metric for even better classification performance.

Feature Importance

Random Forest: Q4_Score is the most influential feature in the Random Forest model, with an importance of approximately 0.4. Q10_Score and Q3_Score also play significant roles, while Q1_Score and Q7_Score have lower but still meaningful contributions to the predictions.



Gradient Boosting: Q4_Score is the dominant feature in the Gradient Boosting model, with a slightly higher importance (~0.5) compared to Random Forest. Q10_Score is the second most important, while Q7_Score, Q3_Score, and Q1_Score have smaller yet notable impacts on the model's decisions.



CONCLUSION

Both Random Forest and Gradient Boosting models demonstrate strong predictive performance, with Q4_Score emerging as the most critical feature across both models. Its dominant importance highlights its decisive role in distinguishing cases effectively. Q10_Score consistently ranks as the second most influential feature, while Q3_Score, Q7_Score, and Q1_Score contribute to a lesser extent, but still provide valuable information. While Random Forest offers a balanced feature importance distribution, Gradient Boosting emphasizes Q4_Score even more, suggesting its sensitivity to key features.

These findings underline the consistency of feature relevance across both models, reinforcing the significance of Q4_Score and Q10_Score in prediction tasks. Overall, both models are robust and complementary in their feature utilization, making them reliable classifiers for the given task.

ACKNOWLEDGMENT

We thank Dr. Gupta for the opportunity and support in performing this assignment.

REFERENCES

[1] Das, Anupam & Pattanaik, Prasant & Bandopadhyay, Anjan & Mukherjee, Suchetan & Turjya, Sapthak Mohajon, 2024, Early Autism Spectrum Disorder Screening in Toddlers: A Comprehensive Stacked Machine Learning Approach

[2] Bryers A, Hawkes CA, Parkin E, Dawson N., 2022, Progress towards understanding risk factor mechanisms in the development of autism spectrum disorders.

[3] C. Lord, M. Elsabbagh, G. Baird, and J. Veenstra-Vanderweele, 2018, autism spectrum disorder.

[4] M. S. Farooq, R. Tehseen, M. Sabir, and Z. Atal, 2023, Detection of autism spectrum disorder (ASD) in children and adults using machine learning.

[5] S. Kamel and R. Al-harbi, 2021, Newly proposed technique for autism spectrum disorder-based machine learning.

[6] K. Vakadkar, D. Purkayastha, and D. Krishnan, 2021, Detection of autism spectrum disorder in children using machine learning techniques.

[7] Keil-Stietz K, Lein PJ., 2022, Gene × environment interactions in autism spectrum disorders.

[8] M. F. Rabbi, S. M. M. Hasan, A. I. Champa and M. A. Zaman, 2021, A Convolutional Neural Network Model for Early-Stage Detection of Autism Spectrum Disorder.

[9] Saleh, A. Y., & Chern, L. H., 2021, Autism Spectrum Disorder Classification Using Deep Learning. International Journal of Online and Biomedical Engineering (iJOE).

[10] Mouncef El ouardi, Ahmed Saad Squalli Houssaini, Mohammed Oukabli et al., 2024, autism spectrum disorder gene prediction using Machine learning model and Human brain Spatiotemporal gene expression Data.

[11] Dhuha Dheyaa Khudhur, Saja Dheyaa Khudhur, 2023, The classification of autism spectrum disorder by machine learning methods on multiple datasets for four age groups.

[12] Ashima Sindhu Mohanty, Priyadarsan Parida and K C Patra, 2021, Identification of Autism Spectrum Disorder using Deep Neural Network.

[13] L. Goel, S. Gupta, A. Gupta, S. N. Rajan, V. K. Gupta, A. Singh, and P. Gupta, 2024, Advancing asd detection: novel approach integrating attention graph neural networks and crossover boosted meerkat optimization.

[14] W. Nie, B. Zhou, Z. Wang, B. Chen, X. Wang, C. Hu, H. Li, Q. Xu, X. Xu, and H. Liu, 2024, Computational interpersonal communication model for screening autistic toddlers: A case study of response-toname.

[15] https://www.kaggle.com/datasets/fabdelja/autism-screening-for- toddlers

[16] https://github.com/mm909/Kaggle-Autism/tree/master/reference

[17] https://drive.google.com/drive/folders/1XQU0pluL0m3TIlXqntano12d68peMb8A

[18] https://archive.ics.uci.edu/datasets?search=Autistic%20Spectrum%20Disorder%20Screening%20Data%20for%20Children

[19] https://autismsciencefoundation.org/autism-research-in-2022/

[20] https://gene.sfari.org/database/human-gene/

[21] https://www.cell.com/action/showFullTableHTML?isHtml=true&tableId=tbl1&pii=S2666-979X%2823%2900037-X

[22] https://github.com/getzlab/rnaseqc/blob/master/README.md.

[23] https://archive.ics.uci.edu/dataset/419/autistic+spectrum+disorder+screening+data+for+children

[24] https://archive.ics.uci.edu/dataset/420/autistic+spectrum+disorder+screening+data+for+adolescent

[25] https://archive.ics.uci.edu/dataset/426/autism+screening+adult