Predicting Autism Spectrum Disorder Using Machine Learning Classifiers

Technical Report · January 2024

DOI: 10.13140/RG.2.2.35833.44646

Nouhaila Nigrou
National school of applied sciences AL Hoceima

Predicting Autism Spectrum Disorder Using Machine Learning Classifiers

NIGROU Nouhaila

National school of applied sciences, AL HOCEIMA

nouhaila.nigrou@etu.uae.ac.ma

*Abstract*_Autism Spectrum Disorder (ASD) is a neurodevelopmental condition marked by a spectrum of challenges in social interaction, communication, and behavior. Typically diagnosed in early childhood, the manifestation of symptoms during the first two years of life underscores the importance of timely identification. This research paper focuses on the use of machine learning algorithms to predict ASD, recognizing the significance of early diagnosis for effective intervention. While ASD is commonly diagnosed in childhood, the complexity of detection increases in adolescence and adulthood, posing challenges for accurate identification. In this study, we analyze a comprehensive dataset incorporating behavioral features, applying Support Vector Machine, Logistic Regression, Random Forest, XGBoost and Multi-Layer Perceptron to develop predictive models. Through rigorous training and validation on the Dataset, the models are assessed using key performance metrics. The results reveal promising accuracy rates in ASD prediction, underscoring the potential of machine learning to contribute to early identification. **Keywords:** Autism Spectrum Disorder, Machine Learning,Predictive Models, Early Diagnosis.

I. INTRODUCTION

The diagnostic journey for Autism Spectrum Disorder (ASD) is a prolonged and intricate process, often spanning up to six months. Involving consultations with specialists like developmental pediatricians, neurologists, psychiatrists, or psychologists, this comprehensive approach aims to capture the complexity of ASD symptoms. However, this slow process not only adds to the emotional burden on families but also delays the important early intervention. Early Intervention has been shown to significantly improve outcomes for individuals on the autism spectrum. The extended diagnostic timeline gets in the way of starting these important interventions on time.

Machine Learning methods emerge as a promising solution to address this challenge in healthcare. By leveraging algorithms such as Support Vector Machine, Logistic Regression, Random Forest, XGBoost, and MLP. we aim to significantly expedite the ASD diagnostic process. This study emphasizes the transformative potential of Machine Learning in providing faster, more accessible, and accurate diagnoses for ASD, ultimately enhancing the prospects for Early Intervention and improved outcomes for autistic individuals.

II. PREVIOUS STUDIES

Some researchers have adopted a hybrid strategy, amalgamating Deep Learning and Explainable Artificial Intelligence (XAI) to forecast ASD in toddlers. The 4-layer neural network of the Deep Learning Model leverages behavioral traits from ASD screening datasets, demonstrating impressive accuracies ranging from 92% to 98% across diverse case studies. Additionally, the XAI component, employing SHAP values, interprets model decisions and highlights crucial features such as simple gestures and gaze following for precise predictions. This dual-model methodology not only elevates prediction accuracy but also furnishes clinicians with valuable, interpretable insights. The results emphasize the pivotal role of specific behavioral attributes in early ASD detection, offering actionable recommendations for clinicians and stakeholders in pediatric care.[1]

Several similar studies on diverse topics and datasets have been conducted. One noteworthy example involves the application of Support Vector Machine (SVM) for Magnetic Resonance Imaging (MRI) stroke classification. This study focused on classifying MRI images related to brain strokes. Gabor filters and Histograms were employed to extract features from the MRI images, and Support Vector Machine (SVM) with various kernels was utilized for classification.[2]

III. MATERIAL AND METHODOLOGY

From these kind of researches mentioned above here, we gather knowledge and related information that we have used for our work.

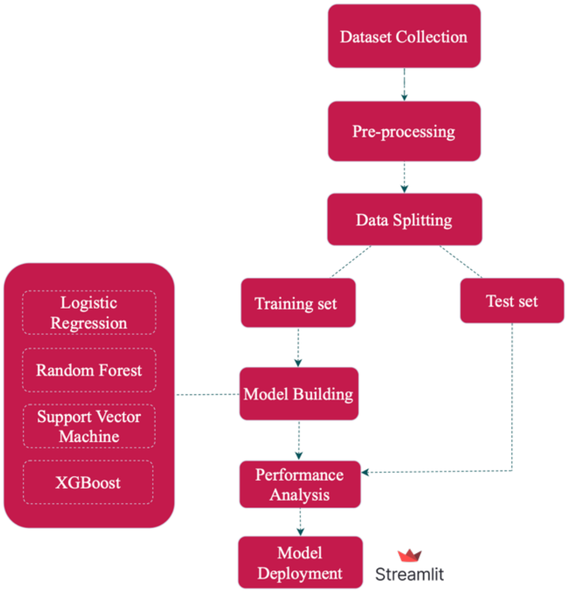


Figure1 : Project Workflow

Figure1 shows the steps in the proposed workflow which involves the data collection, pre-processing of data, training, and testing with specified models, and evaluation of the results and prediction of ASD.

1. Data Collection

For this project, we used an open-source dataset from the UCI Machine Learning repository, a well-established resource in computer science research. The dataset, curated by Tabtah, a respected researcher in machine learning and data mining, includes demographic information and the Autism Quotient Test (AQ) with 10 questions. Merging this dataset with another provided by Tabtah (after some modifications to ensure the compatibility) creates a comprehensive dataset, enriching our data for building a robust Autism Spectrum Disorder predictive model. We chose this dataset due to its widespread use in ASD studies, with the primary measure being the AQ questionnaire from the Autism Research Center at Cambridge University. This questionnaire evaluates attention switching, attention to details, communication, and imagination, providing a score that indicates 'Autistic-like' behavior based on ten specified questions filled out by parents, family members or professionals.

Here's a detailed description of data set features used in this study:

- **A1_Score :** I often notice small sounds when others do not.
- **A2_Score :** I usually concentrate more on the whole picture, rather than the small details.
- **A3_Score :** I find it easy to do more than one thing at once.
- **A4_Score :** If there is an interruption, I can switch back to what I was doing very quickly.
- **A5_Score :** I find it easy to read between the lines when someone is talking to me.
- **A6_Score :** I know how to tell if someone listening to me is getting bored.
- **A7_Score :** When I'm reading a story, I find it difficult to work out the character's intention.
- **A8_Score :** I like to collect information about categories of things (e.g. types of car, types of bird, types of train, types of plant etc).
- **A9_Score :** I find it easy to work out what someone is thinking or feeling just by looking at their face.
- **A10_Score :** I find it difficult to work out people's intentions.
- **Age :** Age of the patient in years.
- **Gender :** Gender of the patient.
- **Ethnicity :** Ethnicity of the patient.

- **Jaundice :** Whether the patient had jaundice at the time of birth.
- **Autism :** Whether an immediate family member has been diagnosed with autism.
- **Country of residence :** Country of residence of the patient.
- **Used app before :** Whether the patient had undergone a screening test before
- **Result:** Score for AQ1-10 screening test.
- **Age_desc:** Age of the patient.
- Relation: Relation of the patient who completed the test.
- **Class/ASD:** The target column, classified result as 0 or 1.
  0 represents No and 1 represents Yes.

Note:

- Score 1 for A1_Score, A7_Score, A8_Score, and A10_Score indicates agreement.
- Score 1 for A2_Score, A3_Score, A4_Score, A5_Score, A6_Score, A9_Score indicates disagreement.

This information is derived from the AQ questionnaire developed by the Autism Research Center at Cambridge University.

2. Data Understanding:

2.1. Features

The Dataset consists of 11 individual characteristics and 10 behavioral features. They are the following:

A1_Score	Binary(0,1)
A2_Score	Binary(0,1)
A3_Score	Binary(0,1)
A4_Score	Binary(0,1)
A5_Score	Binary(0,1)
A6_Score	Binary(0,1)
A7_Score	Binary(0,1)
A8_Score	Binary(0,1)
A9_Score	Binary (0,1)
A10_Score	Binary (0,1)
Age	String
Gender	Boolean('f', 'm')
Ethnicity	String
Jaundice	Boolean('YES', 'NO)
Autism	Boolean('YES', 'NO)
Country of Residence	String
Used appbefore	Boolean('YES', 'NO)
Relation	String
Class/ASD	Boolean(YES=1,NO=0)

3. Pre-Processing

3.1. Dropping irrelevent features

In this section, we address the exclusion of irrelevant features like country of residence, age_desc, Used_app_before, and result. We made this decision due to their limited impact on our analysis,

3.2. Missing Values Imputation

The dataset exhibited a small number of missing values in the age column which we addressed by filling them with the mean.

3.3. Handling Categorical Data

To manage non-numerical values, we applied the Label Encoder technique, a method that converts categorical data into numerical format. This process allows our machine learning models to understand and work with the data effectively.

3.4. Feature Scaling

Feature scaling normalizes the range of dataset features. Using the MinMaxScaler scales values between 0 and 1, preventing features with larger scales from dominating the model's training.

3.5. Data Splitting

This involves splitting the dataset into two partitions for training and testing, ensuring the model's effectiveness with new, unseen data. In our study, we divided the entire dataset into training and testing sets, maintaining an 80%-20% ratio.

3.6. SMOTE for balancing the Dataset

To address class imbalance, SMOTE (Synthetic Minority Over-sampling) was applied, it focuses on the minority class. and generates synthetic samples by interpolating between that instance and its nearest neighbors in feature space. This creates new synthetic instances that are similar to the existing minority class instances.

→ Dataset before oversampling :

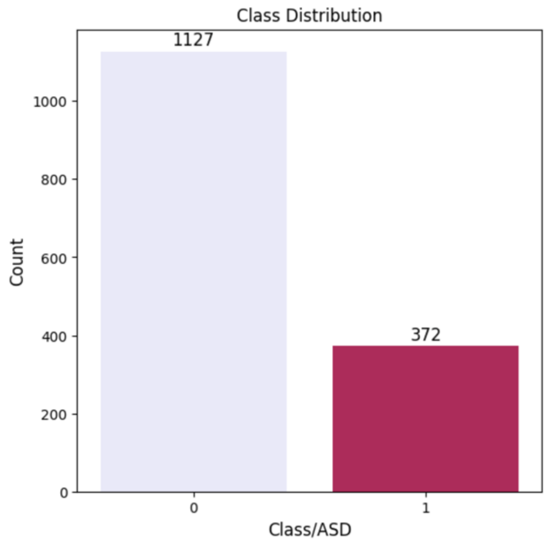


Figure2 : Dataset before balancing

→ Dataset after oversampling :

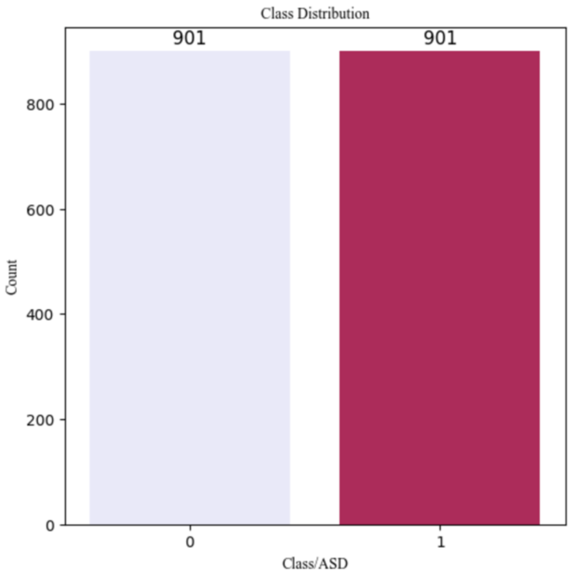


Figure3 : Dataset after balancing

4. Approaches

4.1. Models

In this section, we describe our machine learning approach to predict Autism Spectrum Disorder (ASD), emphasizing the chosen classifiers' roles in refining early diagnosis.

4.1.1. Logistic Regression (LR)

Logistic Regression (LR) is a statistical model tailored for binary classification tasks with outcomes restricted to 0 or 1. LR establishes the relationship between a dependent binary variable and a nominal or ordinal variable, characterized by the sigmoidal function. It computes the probability of an event based on input features, making it effective for binary predictions.

Sigmoid Function: $f(x) = \frac{1}{1+e^{-x}}$

To measure the model's accuracy, Logistic Regression employs a cost function. Specifically, the cross-entropy loss is used to evaluate the difference between predicted and actual outcomes. Expressed as following:

$$Logloss = \frac{1}{N}\sum_{i=1}^{N} -(yi \times \log(\hat{y}\,i) + (1 - yi) \times \log(1 - \hat{y}\,i))$$

4.1.2. Random Forest (RF)

Random Forest is an ensemble learning algorithm that builds a multitude of decision trees during training. The decision trees are trained on random subsets of the data, and the final prediction is a result of the aggregation (voting or averaging) of the predictions from individual trees. It is chosen for its ability to handle complex relationships in the data, resist overfitting and provide robustness against noise. In the realm of decision trees, The Gini Impurity, Information Gain, and Entropy are fundamental to the construction of decision trees, and they play a crucial role in the Random Forest Algorithm

Gini Impurity: is a measure of how often a randomly chosen element would be incorrectly labeled. It ranges from 0 (perfectly pure) to 1 (impure).

$$Gini(X) = 1 - \sum_{i=1}^{k} (pi)^2$$

Information Gain: is the difference between uncertainty of the starting node and weighted impurity of the two child nodes. It helps decide the most informative feature for splitting a node in a decision tree.

$$Information\ Gain = Entropy(parent) - \sum_{i=1}^{k} \frac{\mathrm{N}i}{\mathrm{N}} \times Entropy(\mathrm{child}i)$$

Entropy: is a measure of impurity(disorder) in a set.

$$Entropy(X) = -\sum_{i=1}^{k} pi\ \mathbf{log}2(pi)$$

4.1.3. XGBoost

eXtreme Gradient Boosting is an ensemble learning designed for speed and performance. It sequentially builds a series of decision trees, each correcting errors of the previous one. It employs a regularization term to control model complexity and prevent overfitting and uses a gradient-based optimization strategy for efficient training.

The primary objective of XGBoost is to minimize a loss function that measures the difference between the predicted values and the true labels. The overall objective function of XGBoost is a sum of two components:

$$Objective = -\sum_{i=1}^{n} Loss(yi, \widehat{yi}) + \sum_{k=1}^{K} \Omega(fk)$$

The first term represents the loss on individual data points, where yi is the true label, and $\widehat{yi}$ is the predicted value.

The second term is a regularization term that penalizes the complexity of the model to prevent overfitting, by summing over all the trees (K) in the ensemble.

4.1.4. Support Vector Machine (SVM)

SVM is a supervised learning algorithm used for classification tasks. It aims to find a hyperplane that best separates the data into different classes. The "support vectors" are the data points closest to the decision boundary, and the margin (the distance between the hyperplane and the observations closest to the hyperplane), is maximized to enhance robustness, that means maximum distances between the two classes.

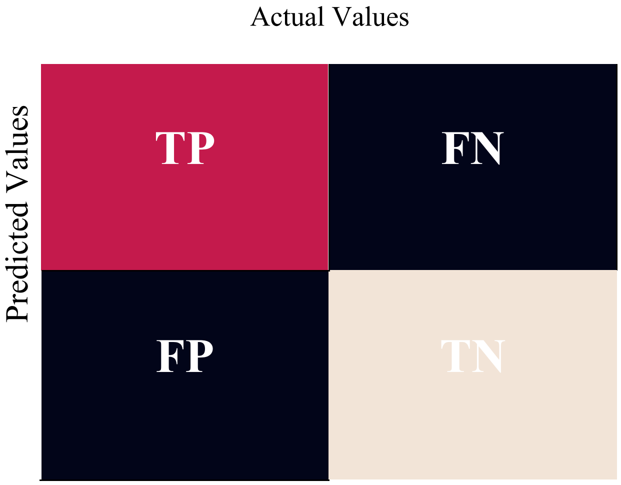
4.1.5. Multi-Layer Perceptron

MLP is a type of artificial neural network designed for learning and making predictions. It consists of layers of interconnected nodes, including input, hidden, and output layers. During training, the network adjusts the connection weights to minimize the difference between predicted and actual outputs. The nodes use activation functions to introduce non-linearity, enabling the MLP to capture complex patterns in data.

5. Model Evaluation

In the response dataset, Autism Spectrum Disorder (ASD) diagnosis is validated using four classes:

True Positive (TP): Correctly recorded cases with autism.
True Negative (TN): Cases correctly identified as not having autism.
False Positive (FP): Incorrectly recorded cases with autism.
False Negative (FN): Incorrectly predicted cases without autism



TP	FN
FP	TN

Figure4 : Confusion matrix

Figure 4 provides a detailed breakdown of the model's performance by comparing predicted and actual class labels. This information is useful for calculating various performance metrics including: Accuracy, Precision, Recall and F1-Score.

Accuracy : is a fundamental metric that measures the overall correctness of a model by calculating the ratio of correctly predicted instances to the total instances.

$$\boldsymbol{Accuracy} = \frac{Number\ of\ Corrected\ Predictions}{Total\ Number\ of\ Predictions}$$

ROC : Receiver Operating Characteristic
A ROC curve is a graph where the x-axis represents the false positive rate, and the y-axis represents the true positive rate. Each point on the curve corresponds to a different threshold for classifying positive instances.
A model with a higher ROC curve (closer to the top-left corner) indicates better discrimination between positive and negative instances.

AUC : Area under the curve
It quantifies the overall performance of a binary classification model by measuring the area under the ROC curve, the values range from 0.5 (indicating random chance) to 1.0 (perfect discrimination). A model with a higher AUC is better at distinguishing between positive and negative instances .

Precision: Correctly predicted autism cases out of all predicted positive cases.

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

Recall: Correctly identified autism cases out of all actual positive cases.

$$Recall \ = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

F1-Score: Harmonic mean of precision and recall, offering a balanced measure.

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

These measures collectively evaluate the accuracy and effectiveness of the classifiers in predicting ASD.

6. Model Tuning

Model tuning, or hyperparameter optimization, involves adjusting our model’s settings to find the most effective configuration for our task. Making it more effective in making accurate predictions on new, unseen data.

7. Results and Discussions

7.1. Comparison between different Classifiers

To determine which classifier performs best, we're comparing results from four case studies. The table shows the outcomes (After Tuning), helping us identify the most effective classifier using key metrics and statistical validation.

	LR	RF	XGB	SVM	MLP
Accuracy	**0.92000**	**0.91667**	**0.90667**	**0.9267**	**0.923**
Precision	**0.77778**	**0.845**	**0.81944**	**0.842**	**0.839**
Recall	**0.94596**	**0.811**	**0.79727**	**0.864**	**0.864**
F1-Score	**0.85369**	**0.827**	**0.80829**	**0.853**	**0.847**
ROC-AUC	**0.92875**	**0.881**	**0.86988**	**0.905**	**0.903**

TABLE 1 : Evaluation Metrics Results

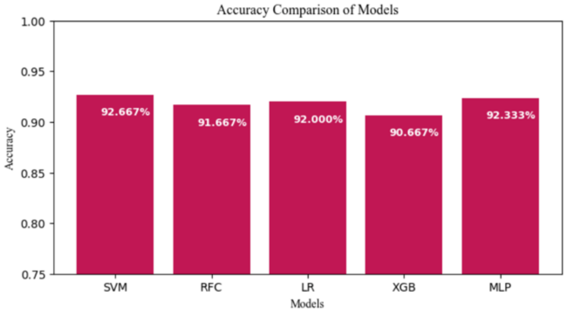


Figure 5 : Models Accuracy Comparison

7.2. Findings

In our project, we examined how well different classifiers—namely, SVM, Random Forest, Logistic Regression, XGBoost, and MLP—performed. According to Figure 5 and TABLE 1, SVM proved to be the most accurate at 92.67%, Logistic Regression stood out in terms of ROC-AUC with a score of 92.87%, highlighting its effectiveness in capturing positive rates. Random Forest demonstrated balanced performance with an accuracy of 91.7%. XGBoost performed competitively but slightly lagged with an accuracy of 90.67%, and MLP showed the third-highest accuracy among all models. These results emphasize the unique strengths of each classifier, with SVM and Logistic Regression appearing as promising choices for applications where accuracy is crucial.
Furthermore, the study uncovered that SVM's effectiveness is attributed to its ability to handle complex patterns, especially in non-linear data. On

the other hand, even though Logistic Regression achieved the highest ROC-AUC score, it might face challenges when dealing with complex relationships. As a result, SVM stand out as the optimal choice, delivering robust performance for our classification task.

7.3. Discussion

➔ ROC Curve

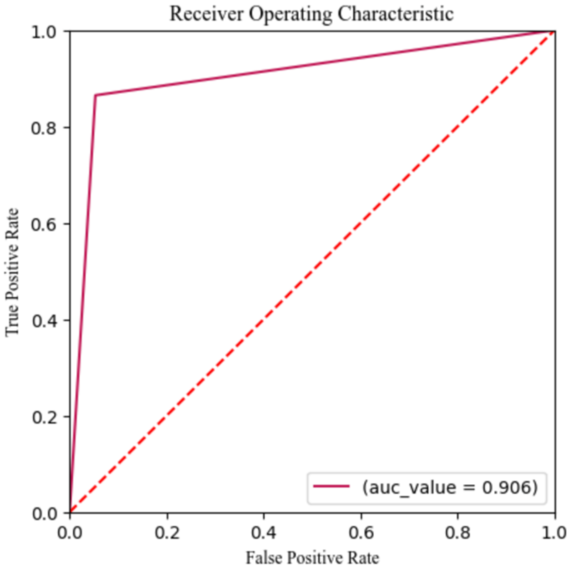


Figure 6 : ROC Curve

As illustrated in Figure 5. The Area Under the Curve (AUC) value is 0.906, indicating a high performance level for our classification model. Additionally, the ROC curve shows that the ideal point is closer to the top-left corner, reinforcing the indication of strong performance.

➔ Confusion Matrix

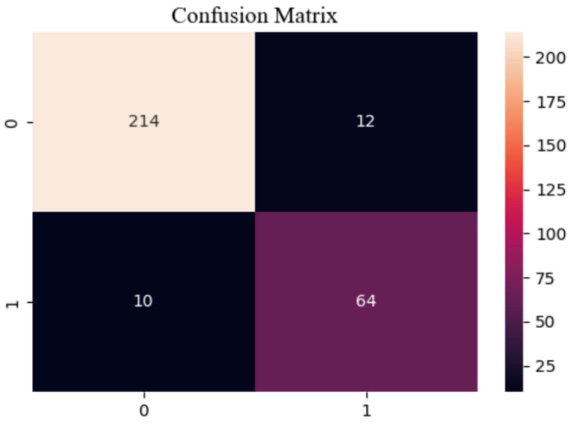


Figure 7 : Resulted Confusion matrix

Figure 6 reveals the performance of our autism prediction model. With 214 TP and 64 TN, the model accurately identified both autism and non-autism cases. However it also produced 10 FP, incorrectly predicting autism, and 12 FN, missing some autism cases.

7.4. Feature Importance

➔ SHAP values (XAI Methods)

We used an eXplainable AI (XAI) method to emphasize feature importance in our model, gaining insights into how specific features affect final predictions. This method aims to offer detailed and interpretable insights into the model's prediction process, including understanding individual feature contributions, providing explanations for specific predictions and the model's reliance on feature interactions.

Additionally, SHAP (Shapley Additive exPlanations) is a combined framework designed to interpret predictions by prioritizing the most crucial feature needed to make an accurate decision.

➔ Summary plot

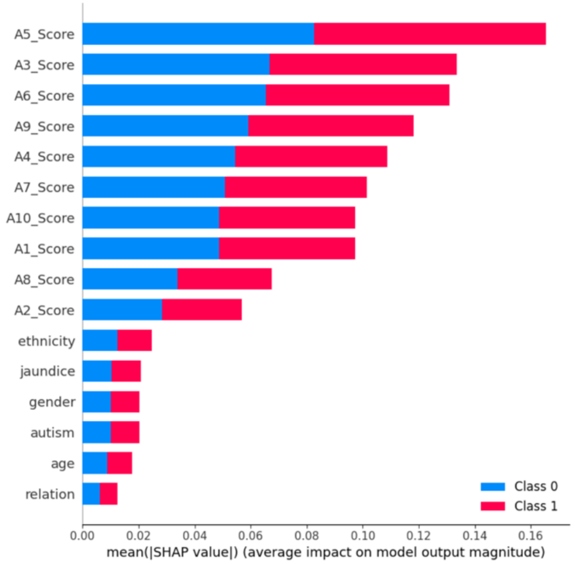


Figure 8 : Features contribution to the accuracy of the model based on XAI

Figure 8 indicates the contribution of different features towards the classification accuracy. From the results, it has been observed that features namely, A5_Score, A3_Score and A6_Score influence the prediction more in comparison to other remaining features. The XAI results markedly indicate and recommend the features or symptoms (presented in the Figure) which are highly co-related to prediction of the ASD traits.

➔ Summary plot of label 0

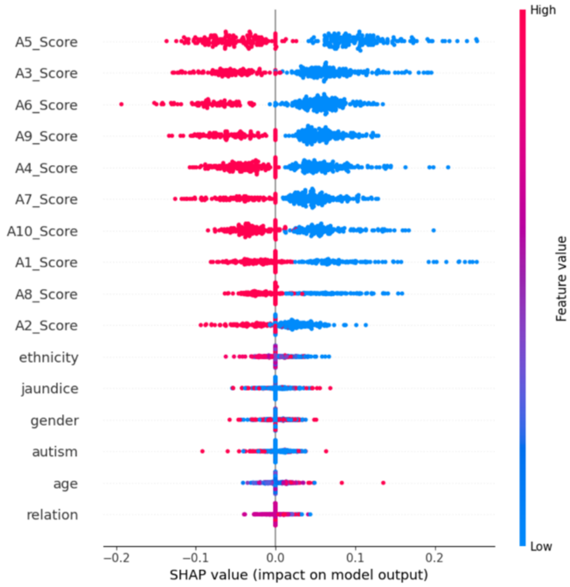


Figure 9 : Summary plot of the label '0', showing the impact of all features

In Figure 9, the color represents the value of the corresponding feature, with red indicating high values and blue indicating low values. If we look at the feature 'A5_Score', we will see that it is mostly high with a negative SHAP value. It means in class 0, a higher value of this feature is associated with a negative impact on the model's output. In other words, higher counts of 'A5_Score' contribute to a lower likelihood of the instance being classified as class 0.
For label '1' the visualization will be flipped.

➔ Force plot

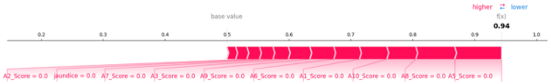


Figure 10: Features contributing to the '0' result

We can clearly see that zero value as a value in jaundice feature and all AQ test scores except A4_Score have contributed to negative ASD results.

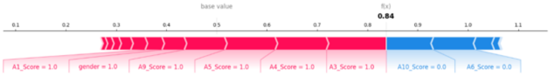


Figure 11: Features contributing to the '1' result

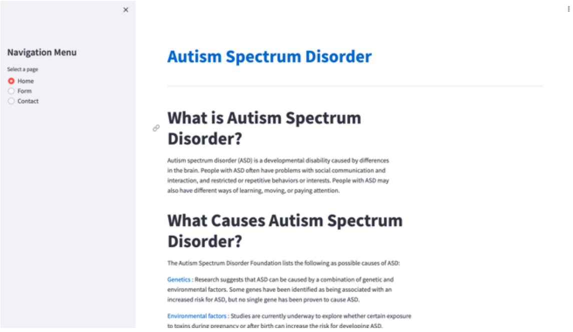
Figure 11 shows that 4 features "A9_Score", "A5_Score","A4_Score", "A3_Score" with 1 as a value contributes the most to positive ASD results.

7.5. Feature Insights

After considering all interpretations and conducting a feature importance analysis, we can conclude that Autism Spectrum Disorder is closely associated with **difficulties in multitasking**, **challenges in quickly resuming the current task after an interruption**, **limitations in inferring meaning beyond explicit communication**, and a **reduced ability to understand others' thoughts through facial expressions.**

8. Model Deployment

In this practical application of our research, we deploy an interactive Streamlit app. This user-friendly interface provides seamless access to our predictive model, enhancing engagement and accessibility.



Select a page
● Home
○ Form
○ Contact

Autism Spectrum Disorder

What is Autism Spectrum Disorder?

Autism spectrum disorder (ASD) is a developmental disability caused by differences in the brain. People with ASD often have problems with social communication and interaction, and restricted or repetitive behaviors or interests. People with ASD may also have different ways of learning, moving, or paying attention.

What Causes Autism Spectrum Disorder?

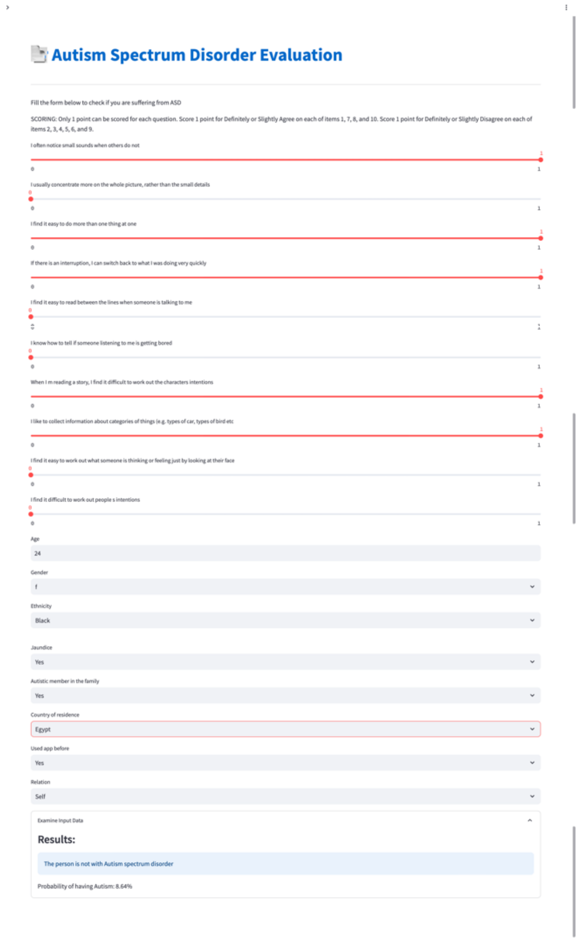
The Autism Spectrum Disorder Foundation lists the following as possible causes of ASD:

Genetics : Research suggests that ASD can be caused by a combination of genetic and environmental factors. Some genes have been identified as being associated with an increased risk for ASD, but no single gene has been proven to cause ASD.

Environmental factors : Studies are currently underway to explore whether certain exposure to toxins during pregnancy or after birth can increase the risk for developing ASD.

Figure 12: Interactive Streamlit Application

Our deployed Streamlit application goes beyond a mere visualization platform, incorporating a user-friendly form for inputting essential behavioral data. Using this input, the system then employs our predictive model to assess the likelihood of the user having Autism Spectrum Disorder (ASD).



Autism Spectrum Disorder Evaluation

Fill the form below to check if you are suffering from ASD

SCORING: Only 1 point can be scored for each question. Score 1 point for Definitely or Slightly Agree on each of items 1, 7, 8, and 10. Score 1 point for Definitely or Slightly Disagree on each of items 2, 3, 4, 5, 6, and 9.

I often notice small sounds when others do not: 1

I usually concentrate more on the whole picture, rather than the small details: 0

I find it easy to do more than one thing at one: 1

If there is an interruption, I can switch back to what I was doing very quickly: 1

I find it easy to read between the lines when someone is talking to me: 0

I know how to tell if someone listening to me is getting bored: 0

When I m reading a story, I find it difficult to work out the characters intentions: 1

I like to collect information about categories of things (e.g. types of car, types of bird etc: 1

I find it easy to work out what someone is thinking or feeling just by looking at their face: 0

I find it difficult to work out people s intentions: 0

Age: 24

Gender: f

Ethnicity: Black

Jaundice: Yes

Autistic member in the family: Yes

Country of residence: Egypt

Used app before: Yes

Relation: Self

Examine Input Data

Results:

The person is not with Autism spectrum disorder

Probability of having Autism: 8.64%

Figure 13 : Evaluation Form

The application not only delivers predictions but also provides transparency by displaying the associated probability of the predicted outcome as illustrated in Figure 13. This interactive approach empowers users to engage directly with the predictive capabilities of our model.

IV. FUTURE WORK

Looking forward, our focus on future work revolves around improving the accuracy of our ASD predictions models. Firstly, expanding our datasets to ensure a broader range of information aims to refine the precision of our predictions since our dataset lacked diversity or completeness. Secondly, transitioning from machine learning to advanced deep learning methodologies will empower us to handle larger and more intricate datasets, ultimately improving the reliability of our Autism Spectrum Disorder predictions.

V. CONCLUSION

Our study focused on predicting autism spectrum disorder through machine learning algorithms, specifically exploring models such as Logistic Regression, Random Forest, MLP and XGBoost to improve early diagnosis. After a thorough assessment based on key performance metrics—accuracy, precision, recall, and F1-score—Support Vector Machine (SVM) emerged as the most suitable model for our dataset, boasting the highest accuracy at 92%, precision at 0.845, recall at 0.865, and an F1-score of 0.853. Even after Model Tuning, where the radial basis function (RBF) kernel proved optimal, the scores remained unchanged. Employing eXplainable AI (XAI) via the SHAP model highlighted the crucial features contributing to accurate predictions, with the top 7 ranked features achieving an impressive contribution to the final predictions. Handling medical datasets posed unique challenges, making the identification of the most potent classifier a significant achievement in our study.

VI. REFERENCES

[1] Anupam Garg, Anshu Parashar, Dipto Barman, Sahil Jain, Divya Singhal, Mehedi Masud and Mohamed Abouhawwash. "Autism Spectrum Disorder Prediction by an Explainable Deep Learning Approach", Computers, Materials & Continua DOI:10.32604/cmc.2022.022170

[2] A.S.Shanthi and A.S.Shanthi. "Support Vector Machine for MRI Stroke Classification", International Journal on Computer Science and Engineering (IJCSE), ISSN: 0975-3397 in April, 2014.

[3] Soul, J. S. & Spence, S. J. Predicting autism spectrum disorder in very preterm infants. Paediatrics 146(4), e2020019448 (2020).