Early Childhood Autism Identification

Satwik Reddy Sripathi
sripathi.sa@northeastern.edu

Naga Kushal Ageeru
ageeru.n@northeastern.edu

Akula Dhanush
akula.d@northeastern.edu

I. Objectives and significance

A neuro-developmental disorder affecting a broad spectrum of abilities and functioning is known as autism spectrum disorder (ASD). Improving the quality of life for children with ASD requires early diagnosis and intervention. The early diagnosis of autism spectrum disorder (ASD) in toddlers, which enables prompt intervention and support, is one of the major problems in the field of autism research and healthcare.

A child's or individual's diagnosis of autism is based on their struggles with repetitive tasks, verbal and non-verbal communication, and social interaction. In today's world, autism spectrum disorder (ASD) is growing more common. The WHO estimates that 1 in 68 children suffer from ASD. As a result, approximately 68 million individuals worldwide including over 2 million in the US alone have ASD. Thus being identified as its much better to treat autism as early as it is recognized.

A promising approach to improving and automating the screening and diagnosis of ASD in early children is through machine learning techniques. Toddlers may exhibit early indications of autism, and it is critical to identify these signals as soon as possible. Early signs could be speech and language development delays or deficiencies, difficulties interacting with others, repetitive or restricted activities, and strong interests in particular subjects or objects. Early intervention services can help toddlers learn critical skills and reduce the disorder's long-term effects if these signs and symptoms are recognized in a timely manner.

Early autism screening presents a number of difficulties. We must not only take into consideration the various and complex ways that autism manifests itself in different cultural and ethnic contexts, but also deal with the stigma associated with the diagnosis.

Our project's focuses on creating an extensive and effective system for diagnosing autism in children by utilizing a range of machine learning approaches like Naïve bayes, SVM, Linear regression / logistic regression, Random forest, XGBoost and then using evaluation metrics like Stratified cross validation, MCC and Leave One Out Cross Validation. This research recognizes the need for new easily available, reasonably priced, and non-invasive techniques to spot possible ASD cases in their early developmental stages. We facilitate the early identification and intervention of ASD by utilizing sophisticated computational techniques and data-driven methodologies.

II. Background

Machine learning have been applied in a number of research to enhance and expedite the diagnosis of ASD. Using a 65-item Social Responsiveness Scale, Duda et al. [5] used forward feature selection in conjunction with under-sampling to distinguish between autism and ADHD. Brain activity measures were employed by Deshpande et al. [4] as a predictor of ASD. Additionally, soft computing methods including classifier combination, ANNs, and probabilistic reasoning have been applied [2]. Multiple studies have discussed automated machine learning models that just rely on characteristics as input features. Several research also utilized brain neuroimaging data.

In the exploration of Autism Spectrum Disorder (ASD) detection, a pivotal contribution is found in the research paper titled "Analysis and Detection of

Paper ID	Findings	Limitations.
[1]	This paper uses the SVM approach and has an accuracy of 97 percent	The major limitation is that the dataset is highly imbalanced
[2]	They created a unique algorithm that blends the functional and structural aspects Many illustrations of the brain's functional connections were drawn out in chalk.	When compared to earlier research, the ML models employed for Autism show a paltry 4.2% gain in prediction accuracy.
[3]	They use the sensor data and an AI system to analyze the patient's status by analyzing their emotions and facial expressions. The system comprises of a smart wristband that is linked to a mobile application and has an interactive monitor and camera. With an accuracy of 78.56%, the Inception-ResNetV2 architecture outperformed all other models.	The accuracy is comparatively less for the previous works using similar approaches
[4]	Metrics based on brain activity used for prediction of ASD Used SVM to obtain an accuracy of 95.9% with 2 clusters and 19 features	Has a very constrained sample size.
[5]	Here six machine learning models were trained and tested using score sheets from 2925 people with ASD or ADHD on the 65 Social Responsiveness Scale. It was discovered that, with a 96.4% accuracy rate, 5 of the 65 behaviors were adequate to differentiate between ASD and ADHD.	Due to the dataset's primary compilation from autism-based collections, there was a noticeable imbalance favoring the ASD class.
[7]	The paper establishes the efficacy of machine learning techniques, with Convolutional Neural Network (CNN) models demonstrating superior accuracy rates of 99.53%, 98.30%, and 96.88% for ASD screening in adults.	the study is constrained by reliance on relatively small publicly available datasets and lacks in-depth exploration of age-specific nuances in ASD detection.

Table 1 : Literature Survey Key Findings and Limitations

Autism Spectrum Disorder Using Machine Learning Techniques" [7]. This paper addresses the generic nature of ASD, a lifelong neuro-disorder impacting an individual's communication and interaction abilities. ASD can manifest at any point in one's life, typically showing symptoms in the first two years, thereby being categorized as a "behavioral disease." The study recognizes the increasing prevalence of machine learning techniques in medical diagnosis research and endeavors to leverage Naïve Bayes, Support Vector Machine, Logistic Regression, KNN, Neural Network, and Neural Network for predicting and analyzing ASD across different age groups—children, adolescents, and adults [10].

The research evaluates the proposed techniques using publicly available dataset specifically tailored for ASD screening. The dataset utilized in the study are representative of different age groups: one for ASD screening in children (292 instances, 21 attributes), another for ASD screening in adults (704 instances, 21 attributes), and a third for ASD screening in adolescents (104 instances, 21 attributes) [8]. Through the application of various machine learning techniques and meticulous

handling of missing values, the results reveal that Neural Network based prediction models exhibit superior performance, achieving accuracy rates of 99.53%, 98.30%, and 96.88% for ASD screening in adult, child, and adolescent dataset, respectively [9]. This insightful paper not only highlights the efficacy of machine learning in ASD detection but also provides empirical evidence supporting the superiority of CNN-based models in this context [7]

Li et al. [**?**] took 6 personal characteristics out of 851 people in the ABIDE database and used a cross-validation approach to train and test the machine learning models. This used as a classification tool for patients with and without ASD. In addition to identifying ASD symptoms, Thabtah et al. [**?**] introduced a novel machine learning technique called Rules-Machine Learning (RML), which provides users with a knowledge base of rules for comprehending the underlying causes of the classification. In order to help ASD patients deal with the COVID-19 epidemic, where Al Banna et al. [**?**] used a customized AI-based system for monitoring and support.

Table 1 depicts the key findings in a paper and their limitations

III. Methods

A. Data Collection and Processing

We acquired data sets from Kaggle, focused on autism detection. One dataset exclusively included children, while the other encompassed both children and adults. Both data sets followed a uniform recording process, utilizing the same survey questionnaire. The Figures Fig. 2 and Fig 3 below illustrate the initial structure of these dataset.

The dataset exhibited variations in certain features, including differences in data types and capitalization used for class representation. Data Frame A presented age in months, while Data Frame B used years. Addressing these disparities, we harmonized the dataset, ensuring uniform formatting. Subsequently, we combined them to create the final dataset, depicted below in Fig. 4.

The dataset exhibited neither missing values nor outliers, as it primarily comprised survey-recorded rows. The final dataset comprises 17 features and a target class, totaling 2178 records. Within this, there are 1335 positive (Yes) samples and 843 negative (No) samples.

Feature	Unique Values
A1	0, 1
A2	0, 1
A3	0, 1
A4	0, 1
A5	0, 1
A6	0, 1
A7	1, 0
A8	1, 0
A9	0, 1
A10	1, 0
Age_Years	2, 3, 1, 4, 5, 6, 7, 8, 9
Qchat-10-Score	3, 4, 10, 9, 8, 5, 6, 2, 0, 7, 1
Sex	f, m
Ethnicity	middle eastern, White European, Hispanic, black, asian, south asian, Native Indian, Others, Latino, mixed, Pacifica, Mixed, PaciFica
Jaundice	yes, no
Family mem with ASD	no, yes
Who completed the test	family member, Health Care Professional, Health care professional, Self, Others, family member
Class	No, Yes

Fig. 1: Features

B. Features

Q-chart : A questionnaire, encompasses questions from A1 to A10. Each question correlates with specific autism-related symptoms. Data collectors use a binary response system, marking 1 for the presence of a trait and 0 for its absence. This approach aids in assessing autism-related traits based on observed symptoms and collected responses. refer to Fig. 5 for the Q chart

From the Fig. 1 we are able to get that

Age-Years: Represents the age of the child in years, ranging from 1 to 10.

Qchat-10-Score: Indicates the sum of scores obtained from the Q-chat questionnaire, ranging from 0 to 10.

	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	Age_Years	Qchat-10-Score	Sex	Ethnicity	Jaundice	Family_mem_with_ASD	Who completed the test	Class
0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	1.0	2	3.0	f	middle eastern	yes	no	family member	No
1	1.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	3	4.0	m	White European	yes	no	family member	Yes
2	1.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	1.0	3	4.0	m	middle eastern	yes	no	family member	Yes
3	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	2	10.0	m	Hispanic	no	no	family member	Yes
4	1.0	1.0	0.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	2	9.0	f	White European	no	yes	family member	Yes

Fig. 2: **DataFrame A**

	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	Age_Years	Qchat-10-Score	Sex	Ethnicity	Jaundice	Who completed the test	Who_completed_the_test	Class
0	0	0	0	0	0	0	0	0	0	0	1	0	m	middle eastern	yes	no	family member	No
1	0	0	0	0	0	0	0	0	0	0	3	0	m	asian	yes	yes	family member	Yes
2	0	0	0	0	0	0	0	0	0	0	4	0	m	black	yes	no	family member	No
3	0	0	0	0	0	0	0	0	0	0	4	0	f	asian	no	no	Health Care Professional	No
4	0	0	0	0	0	0	0	0	0	0	5	0	m	White European	yes	no	Health Care Professional	No

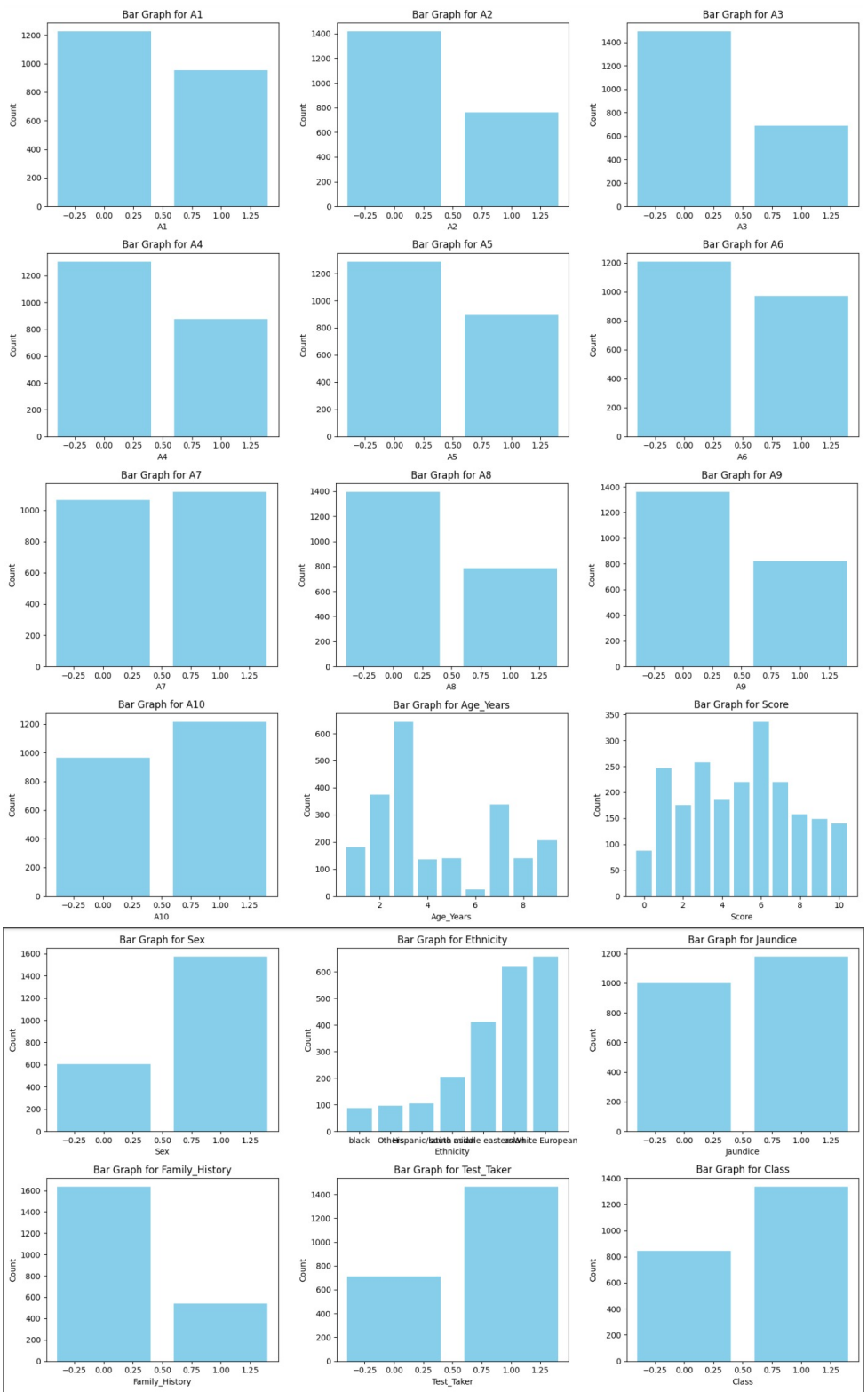
Fig. 3: **DataFrame B**

	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	Age_Years	Qchat-10-Score	Sex	Ethnicity	Jaundice	Family_mem_with_ASD	Who completed the test	Class
0	0	0	0	0	0	0	1	1	0	1	2	3	f	middle eastern	yes	no	family member	No
1	1	1	0	0	0	1	1	0	0	0	3	4	m	White European	yes	no	family member	Yes
2	1	0	0	0	0	0	1	1	0	1	3	4	m	middle eastern	yes	no	family member	Yes
3	1	1	1	1	1	1	1	1	1	1	2	10	m	Hispanic	no	no	family member	Yes
4	1	1	0	1	1	1	1	1	1	1	2	9	f	White European	no	yes	family member	Yes

Fig. 4: **Total DataFrame**

		A	B	C	D	E
1	Does your child look at you when you call his/her name?	Always	Usually	Sometimes	Rarely	Never
2	How easy is it for you to get eye contact with your child?	Very easy	Quite easy	Quite difficult	Very difficult	Impossible
3	Does your child point to indicate that s/he wants something? (e.g. a toy that is out of reach)	Many times a day	A few times a day	A few times a week	Less than once a week	Never
4	Does your child point to share interest with you? (e.g. pointing at an interesting sight)	Many times a day	A few times a day	A few times a week	Less than once a week	Never
5	Does your child pretend? (e.g. care for dolls, talk on a toy phone)	Many times a day	A few times a day	A few times a week	Less than once a week	Never
6	Does your child follow where you're looking?	Many times a day	A few times a day	A few times a week	Less than once a week	Never
7	If you or someone else in the family is visibly upset, does your child show signs of wanting to comfort them? (e.g. stroking hair, hugging them)	Always	Usually	Sometimes	Rarely	Never
8	Would you describe your child's first words as:	Very typical	Quite typical	Slightly unusual	Very unusual	My child doesn't speak
9	Does your child use simple gestures? (e.g. wave goodbye)	Many times a day	A few times a day	A few times a week	Less than once a week	Never
10	Does your child stare at nothing with no apparent purpose?	Many times a day	A few times a day	A few times a week	Less than once a week	Never

Fig. 5: **Q Chart**



Sex: Denotes the gender of the child, where 'm' stands for male, and 'f' stands for female.

Ethnicity: Specifies the ethnic background of the child.

Jaundice: Represents whether the child had a history

of jaundice, with responses 'yes' or 'no'.

Family-mem-with-ASD: Indicates the presence of a family member with Autism Spectrum Disorder (ASD), with responses 'yes' or 'no'.

Who-completed-the-test: Specifies the person who completed the questionnaire, such as a family member or health care professional.

Class: Denotes the presence (Yes) or absence (No) of autism in the child.

C. Data Analysis

Maintaining balanced class priors is crucial to prevent model bias and ensure fair representation of each record type across features. This proportional distribution prevents the model from leaning towards trivial predictions, fostering a more accurate and unbiased learning process. The Fig 6 represents the class prior data mapping.

The visual examination of the provided images reveals a balanced distribution of records across all features, indicating roughly proportional representation for each category. This equilibrium is crucial for machine learning models, preventing bias and ensuring fair learning from diverse data. Balanced class priors, where the ratio of records for each type is maintained, guard against the model becoming overly influenced by any single category, thereby averting the risk of trivial or biased predictions. This thoughtful duration of data fosters a more comprehensive and unbiased learning experience, ultimately enhancing the model's ability to generalize effectively across a spectrum of real-world scenarios. Fig 7 depicts the Score Vs Target plot

Children with autism are more likely to have higher Q-CHAT-10 scores, indicating a greater likelihood of having ASD. The graph also shows that there is some overlap between the two distributions. This means that some children with autism may have Q-CHAT-10 scores that are in the same range as children without autism. This overlap is due to the fact that ASD is a complex spectrum disorder with a wide range of presentations. Fig 8 depicts the Jaundice Vs Target plot

The chart shows that there are more children with autism who have jaundice (around 60%) than children without autism (around 40%). This suggests that there may be a link between jaundice and autism. The variance of the jaundice distribution for children with autism is higher than the variance of the jaundice distribution for children without autism. This indicates that the jaundice distribution is more spread out for children with autism. This could be due to a number of factors, such as the underlying cause of the autism or the severity of the autism. Fig 9 depicts the Age Vs Target plot

The stacked bar chart shows the distribution of age (in years) for children with and without autism (class 0 and class 1, respectively). The y-axis shows the count of children, and the orange stack of bars represents autism yes, while the blue stack of bars represents autism no. The chart shows that there are more children with autism in the younger age groups. This is consistent with the fact

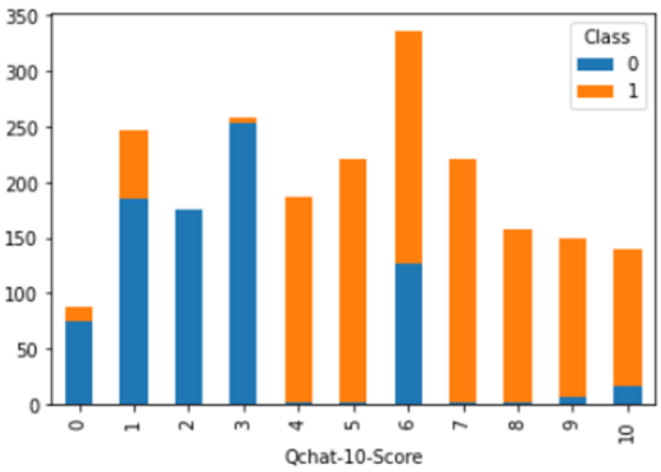


Fig. 6: **Score vs Target**

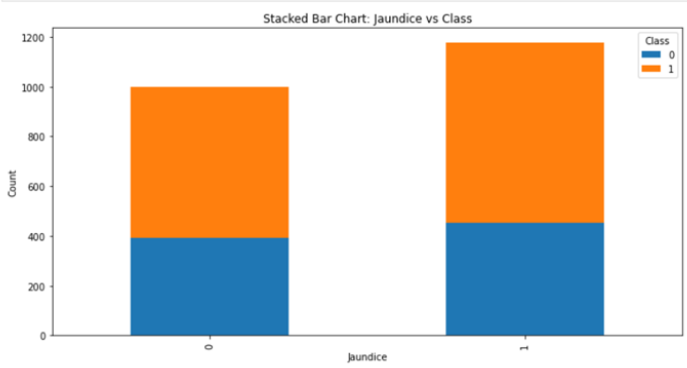


Fig. 7: **Jaundice vs Target**

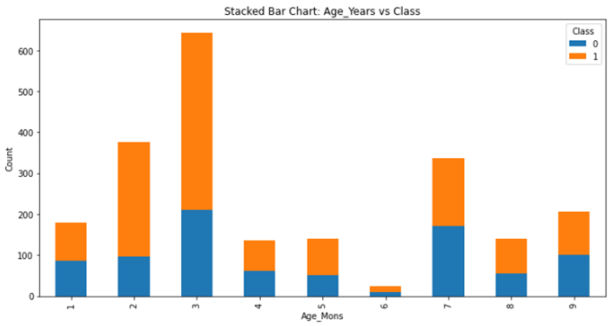


Fig. 8: **Age vs Target**

that autism is typically diagnosed in early childhood. The chart also shows that there is a small but steady increase in the number of children with autism across the age groups.

D. Feature Engineering

1) ***Label Mapping****:* The provided label mapping is a technique used to convert categorical variables with binary outcomes into a numeric format suitable for machine learning models. In this specific case:

- For the 'Sex' column, 'm' (male) is encoded as 1, and 'f' (female) is encoded as 0.
- In the 'Jaundice' column, 'yes' is encoded as 1, and 'no' is encoded as 0.
- The 'Family-mem-with-ASD' column is mapped so that 'yes' is represented as 1, and 'no' is represented as 0.
- Finally, the 'Class' column is encoded with 'Yes' as 1 and 'No' as 0.

This mapping simplifies the categorical data, enabling machine learning algorithms to process and analyze the information effectively.

2) ***Changing column names*** *:* The dataset has been updated with clearer column names for better understanding. For instance, 'Age-Mons' is now 'Age', 'Qchat-10-Score' is 'Score', 'Family-mem-with-ASD' is 'Family-History', and 'Who completed the test' is 'Test-Taker'. Fig 11 depicts on how the new dataframe is changed.

3) ***Feature Selection****:* Feature selection is a critical step in building robust machine learning models, ensuring that only the most relevant attributes contribute to predictive accuracy. By eliminating irrelevant or redundant features, we streamline the model, prevent overfitting, and enhance its interpretability. This process not only improves model performance but also reduces computational complexity.

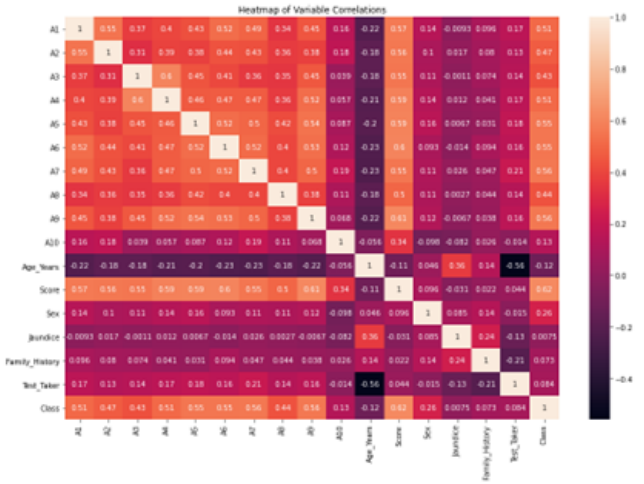


Fig. 9: **Feature Selection**

In our feature selection process, we leveraged Pearson correlation analysis to identify the degree and direction of linear relationships between variables. The resulting correlation heat map provided a visual guide to assess the strength and nature of connections. Bright regions highlighted strong correlations, aiding in the identification of influential features. This data-driven approach empowers us to make informed decisions about which attributes significantly impact our model's performance, fostering a more effective and efficient machine learning pipeline. Fig 10 shows the heat map for feature selection

Looking at the heat map, we found that some features like Jaundice and Test-Taker have very little connection with other features. So, we might think about removing them. But, Family-History stays because, you know, certain things, like ASD, can run in families. So, it's good to keep an eye on that. By doing this, we're making sure our model considers the family link, which is pretty important for understanding and predicting ASD.

4) ***Feature Encoding****:* We performed one-hot encoding on categorical features like Ethnicity to convert

	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	Age_Years	Score	Sex	Ethnicity	Jaundice	Family_History	Test_Taker	Class
0	0	0	0	0	0	0	1	1	0	1	2	3	0	middle eastern	1	0	1	0
1	1	1	0	0	0	1	1	0	0	0	3	4	1	White European	1	0	1	1
2	1	0	0	0	0	0	1	1	0	1	3	4	1	middle eastern	1	0	1	1
3	1	1	1	1	1	1	1	1	1	1	2	10	1	Hispanic/latino	0	0	1	1
4	1	1	0	1	1	1	1	1	1	1	2	9	0	White European	0	1	1	1

Fig. 10: **Data Frame after Column Names Changed**

them into numerical format for our model. After this transformation, each unique category in these feature got its own column, marked as True or False i.e,Ethnicity was split into several columns like Ethnicity-Others, Ethnicity-White European, Ethnicity-Asian, and so on.

A1	int64
A2	int64
A3	int64
A4	int64
A5	int64
A6	int64
A7	int64
A8	int64
A9	int64
A10	int64
Age Years	int64
Score	int64
Sex	int64
Family History	int64
Class	int64
Ethnicity Others	bool
Ethnicity White European	bool
Ethnicity asian	bool
Ethnicity black	bool
Ethnicity middle eastern	bool
Ethnicity south asian	bool

Fig. 11: **Final Features**

This method ensures that the model understands and interprets categorical data correctly, assigning numerical values to each category. With these transformed features, our data is now ready for machine learning algorithms that require numerical inputs, offering better compatibility and accuracy in predicting the Autism Spectrum Disorder (ASD) class. The Final features are shown below along with their data type in the above image i.e. Fig 12

5) ***Data upsampling****:* We initially applied the Synthetic Minority Over-sampling Technique (SMOTE) to the original dataset, ensuring equal representation of classes. Subsequently, we expanded this balanced dataset to a total of 2500 records by employing the Random Sampling with Replacement for upsampling whole dataset. This approach enhances the model's ability to learn from both classes, mitigating biases arising from class imbalances and fostering a more robust classification model. The upsampling is depicted in Fig 13.

6) ***Data Splitting****:* The Final-dataset has been split further training and testing purposes .The training set, denoted as (X-train1, y-train1), comprises 80% of the original data. This subset is used to train and build the machine learning model. The remaining 20% forms the testing set, labeled as (X-test, y-test), and is reserved to assess the model's performance on unseen data, ensuring an objective evaluation of its predictive capabilities.

E. Methodology

Our aim of the project is to pre-process the data collected, in order to perform a comparative analysis of the performance between a wide range of machine learning models. The performance is evaluated with various metric. Following it the proposed flow of work for this project:

1) ***Exploratory analysis****:* In this section, we are going to perform a basic analysis based on the features, to get to know the key features which play a major role in deciding whether the toddler has traits of autism. Feauture Engineering is carried out, aiding in the selection of crucial features for enhanced model interpretation and optimization. This section involves various data visualization techniques and plotting using libraries like matplotlib, seaborn and plotly.

2) ***Pre-Processing****:* In this section, Data Cleaning is performed to handle the missing values, i.e., to Decide

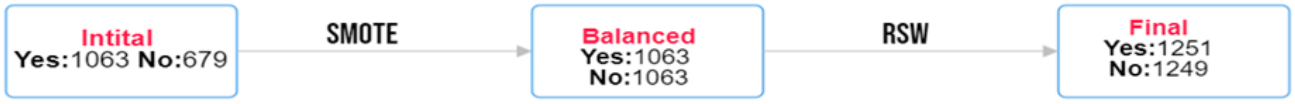


Fig. 12: **Upsampling**

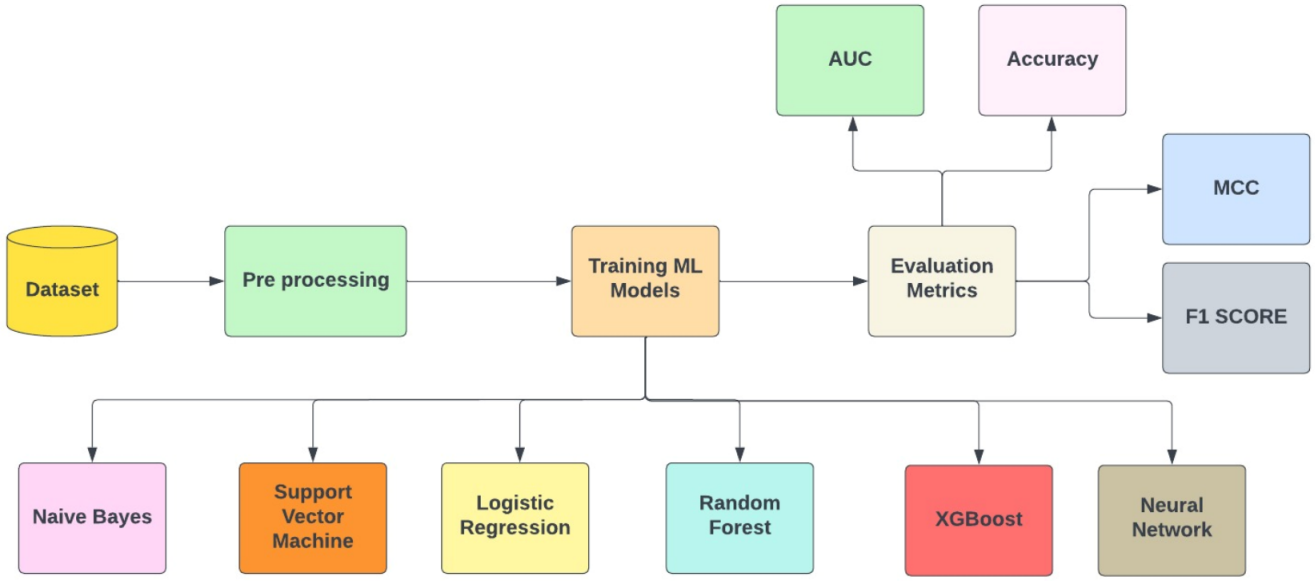


Fig. 13: **Pipeline**

whether to remove, fill, or impute missing data and outlier detection and treatment to identify and handle outliers that can skew your model. A series of data transformation techniques are also used like Feature scaling to normalize or standardize numerical features to ensure they have similar scales and encoding categorical variables to Convert categorical variables into numerical format, such as one-hot encoding or label encoding. And after this data is split into training and test sets to evaluate the model's performance and prevent over-fitting.

*3) **Model Building**:* In this section proposed machine learning models are trained, which are Naïve Bayes, Logistic Regression, Support Vector Machine, Random Forests and XGBoost. The following are the descriptions for the models and the evaluation metrics which we used.

Naïve bayes classification: Naive Bayes is based on Bayes' theorem, which computes the chance of a class label given features vector. It estimates the probability of a class label given a feature vector in classification.

Logistic Regression: In logistic regression, you use a logistic function (also known as a sigmoid function) to model the probability that a given input belongs to one of two classes. Any real-valued number can be mapped to the range [0, 1] using the logistic function

$$F(x) : F(x) = 1/(1 + e^{-x})$$

. The logistic function has an S-shaped curve, making it appropriate for modeling probabilities.

Support Vector Machine: SVM is a powerful binary classification algorithm that is used in machine learning. SVM determines a hyperplane that best separates data

points from two classes while maximizing the margin between them. Mathematically, the hyperplane is represented by the equation:

$$w^T * x + b = 0,$$

where: w is the normal vector to the hyperplane. b is the bias or intercept term. The classification is determined by the sign of the equation: if

$$w^T * x + b > 0,$$

then it belongs to one class, and if

$$w^T * x + b < 0,$$

it belongs to the other class.

Random Forest: Random Forest is based on decision trees, and it constructs multiple decision trees to make predictions before combining their outputs to make a final result. To classify, the algorithm combines ensemble learning techniques with the decision tree framework, the bootstrapping Random Forest algorithm generates multiple randomly selected decision trees from the data. The results are then averaged to produce an output that frequently produces accurate predictions or classifications.

XGBoost: It is based on gradient boosting and is known for its ability to provide excellent predictive performance across a wide range of tasks. When training a model, XGBoost measures the model's performance using a particular objective function. The "logistic" objective function for binary classification uses logistic regression to model the likelihood of the positive class. The log-likelihood of the observed labels is what the model seeks to minimize.

The evaluation strategy employed in this study holds paramount importance in assessing the efficacy of the machine learning model. It serves as a critical means to compare and contrast multiple models, enabling informed decision-making. To comprehensively evaluate the performance of the model, we will utilize a variety of metrics and techniques. The Receiver Operating Characteristic-Area Under the Curve (ROC-AUC) curve will be analyzed to gauge the model's ability to discriminate between positive and negative cases. Furthermore, we will assess key metric scores, including accuracy, precision, recall, and the F1 score, which collectively provide a comprehensive overview of the model's classification performance. Additionally, the Matthews Correlation Coefficient (MCC) will be considered, as it offers a balanced measure of performance, particularly valuable in scenarios with imbalanced dataset. This rigorous evaluation strategy is designed to ensure the robustness and reliability of the machine learning model, ultimately contributing to the validity and applicability of our research findings.

Matthews Correlation Coefficient (MCC) : It stands as a pivotal metric for the assessment of classification models, particularly in scenarios characterized by imbalanced dataset. MCC provides a measure of a model's efficacy in effectively distinguishing between different classes within the dataset. Its significance is most pronounced when the distribution of instances across classes is unequal, as it accurately gauges a model's performance while accounting for such class imbalance. In essence, MCC offers a balanced evaluation that considers the true positive, true negative, false positive, and false negative classifications, making it a valuable tool for the robust assessment of classification models in real-world applications where class distributions may be uneven.

Stratified cross validation: It is a technique commonly used in machine learning to assess the performance of a predictive model. It is an extension of k-fold cross-validation where the dataset is divided into k subsets or folds. In standard k-fold cross-validation, the data is randomly split into k folds, and each fold is used as a testing set exactly once. Stratified cross-validation, on the other hand, takes into account the distribution of the target variable when creating folds. It ensures that each fold maintains the same distribution of the target variable as the entire dataset. This is particularly useful when dealing with imbalanced dataset where the distribution of the classes is uneven.

The process of stratified cross-validation can be out-

lined as follows:

- Data Splitting: The dataset is divided into k folds, maintaining the distribution of the target variable in each fold.
- Training and Testing: The model is trained on k-1 folds and tested on the remaining fold. This process is repeated k times, with a different fold used as the test set in each iteration.
- Performance Evaluation: The performance metrics (e.g., accuracy, precision, recall) are calculated for each iteration, providing a more robust estimate of the model's performance.

Leave-One-Out Cross-Validation (LOOCV) : LOOCV is a special case of cross-validation where only one data point is used as the test set while the rest of the data is used for training. This process is repeated for each data point in the dataset. LOOCV is an extreme case of k-fold cross-validation where k is equal to the number of data points Leave-One-Out Cross-Validation (LOOCV) can be a useful technique in Autism Spectrum Disorder (ASD) classification for several reasons:

- Limited Data: In medical and healthcare-related domains, dataset can often be limited, especially when dealing with rare conditions like ASD. LOOCV allows you to make the most of the available data by leaving out one sample at a time for testing, ensuring that each data point contributes to both training and testing.
- Individualized Assessment: ASD is a spectrum disorder with significant individual variability. LOOCV provides a personalized evaluation for each subject, considering their unique characteristics. This is crucial in a disorder where symptoms and manifestations can vary widely between individuals.
- High Sensitivity to Individual Differences: LOOCV is particularly sensitive to individual differences because it trains the model on all but one data point. This can be beneficial when trying to capture subtle patterns or specific features that are indicative of ASD, as these may vary significantly between individuals.
- Reducing Bias: LOOCV tends to provide an unbiased estimate of the model's performance, as each instance serves as a test set exactly once. This can be important when evaluating the generalization capability of the model, especially in cases where imbalances or specific characteristics in the data need to be carefully considered.

F. Evaluation strategy

Selecting an appropriate evaluation strategy is paramount when addressing medical conditions such as autism. In this context, minimizing type 2 errors is crucial to avoid situations where the model incorrectly predicts the absence of autism in a child who is, in fact, affected. These misclassifications can have significant real-world consequences, making the choice of evaluation metrics critical.

Accuracy : Accuracy stands as a foundational metric, offering a holistic measure of correctness. It serves as a fundamental indicator of how well the model classifies instances, providing an overall snapshot of performance. However, in medical scenarios, where false negatives can have significant consequences, additional metrics are essential.

F1 Score : It is a valuable metric that considers both precision and recall. This is particularly relevant when aiming to strike a balance between identifying positive cases and avoiding false positives.

Matthews Correlation Coefficient (MCC) : It takes into account true and false positives and negatives, offering a balanced performance measure. AUC-ROC Curve : The AUC-ROC Curve takes a different approach by offering a graphical representation of the model's discrimination ability. This becomes crucial when dealing with models that output probabilities rather than discrete predictions. The curve illustrates how well the model distinguishes between the classes at different probability thresholds, providing insights into its overall performance.

IV. RESULTS

In this , we will discuss about the results obtained from training models on the dataset and discuss on the values obtained on the test data set. The training accuracies of various machine learning models on the autism dataset provide valuable insights into their performance. The cross-validation accuracy (CVA) and leave-one-out cross-validation accuracy (LOOCVA) metrics offer a comprehensive view of how well each model generalizes to unseen data. Table I depicts the results for training accuracies on autism dataset.

XGBoost stands out with the highest training accuracy, achieving a CVA and LOOCVA of 0.9920. This indicates that XGBoost effectively captures intricate patterns in the data and exhibits strong generalization capabilities. The consistency between CVA and LOOCVA suggests minimal overfitting, reinforcing XGBoost's reliability.

Random Forest, while demonstrating a commendable CVA of 0.9584, shows a slightly lower LOOCVA of 0.9520. This discrepancy suggests potential overfitting, implying that the model may be fitting too closely to the training data. However, Random Forest still performs well and showcases its ability to learn from the dataset.

SVM and logistic regression yield respectable training accuracies, with CVAs of 0.9564 and 0.9084, respectively. Notably, their LOOVCAs are marginally lower, hinting at some level of overfitting. Both models demonstrate an understanding of the data's underlying patterns, although caution is warranted when considering potential over-fitting effects.

In contrast, Naive Bayes presents the lowest training accuracy, scoring a CVA of 0.8768 and a LOOCVA of 0.8796. This indicates that Naive Bayes might struggle to capture the intricate relationships within the dataset compared to the other models.

The variations in training accuracies emphasize the importance of model selection and the need to assess performance on unseen data. While the models exhibit strong learning capabilities during training, it is crucial to validate their effectiveness on a separate test set to ensure their generalization to real-world scenarios.

Now lets talk about the performance of these models on the test set.

Table II depicts the performance of these models on the test set

Logistic regression achieved a commendable test accuracy of 0.8991, showcasing a minimal variance from the training accuracy. This alignment suggests the model's parameters are well-tuned, achieving a desirable generalized fit, a favorable outcome for any model. The MCC score nearing 0.8 indicates a performance superior to random chance, signifying that the model's predictions align well with the true outcomes.

Naive Bayes, with a test accuracy of 0.8716, maintains a relatively consistent performance, resembling its training accuracy. The MCC of 0.7270 suggests the model performs better than random chance.

SVM exhibits a strong test accuracy of 0.9427, closely mirroring its training performance. With an MCC of 0.8807, the model excels in aligning predictions with actual outcomes.

Random Forest, with a test accuracy of 0.9450, sustains its robust training performance. An MCC of 0.8918 indicates a high level of correlation between predictions and true outcomes.

XGBoost stands out with an impressive test accuracy of 0.9794, maintaining the excellence observed in its training phase. The MCC of 0.9563 reflects a superior predictive capability, well beyond random chance.

The final parameters obtained for models are mentioned in Table III .

In the aforementioned analysis, XGBoost demonstrated superior performance across various metrics, showcasing its exceptional generalization and accuracy compared to other models. Below shows (Fig 14 & Fig 15) the outputs obtained by a sample from x-test and tested by xgboost to give y-pred and then compared to true-values of y.

It gives overall idea and a snapshot of working of the best model (out of all 5) on the test dataset.

To further validate its effectiveness, we conducted an experiment to assess its performance against a neural

TABLE I: Training Accuracies of Models on Autism Dataset

Model	Cross Validation Accuracy	Leave One Out CV Accuracy
Logistic Regression	0.9084	0.9076
Naive Bayes	0.8768	0.8796
SVM	0.9564	0.9560
Random Forest	0.9584	0.9520
XGBoost	0.9920	0.9920

TABLE II: Performance Metrics of Models on Autism Dataset

Model	Test Accuracy	Precision	Recall	F1-Score	MCC
Logistic Regression	0.8991	0.9254	0.9118	0.9185	0.7862
Naive Bayes	0.8716	0.9000	0.8934	0.8967	0.7270
SVM	0.9427	0.9732	0.9338	0.9531	0.8807
Random Forest	0.9450	1.0000	0.9118	0.9538	0.8918
XGBoost	0.9794	0.9888	0.9779	0.9834	0.9563

TABLE III: Model Parameters

Model	Parameters
Logistic Regression	[C: 1.0 , maxiter: 100 , solver: 'lbfgs']
Naïve Bayes	[No special setting]
SVM	[decision_function_shape: 'ovr', degree: 3, gamma: 'scale', kernel: 'rbf']
Random Forest	[criterion: 'gini' , min_samples_leaf: 1, min_samples_split: 2, n_estimators: 100]
XGBoost	[objective: 'binary:logistic']

	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	Age_Years	Score	Sex	Family_His	Ethnicity_(	Ethnicity_\	Ethnicity_(	Ethnicity_I	Ethnicity_r	Ethnicity_south asian
2082	1	1	1	1	1	1	1	0	1	1	5	9	1	0	0	1	0	0	0	0
1687	0	0	0	0	0	0	0	0	0	1	3	6	0	0	0	0	0	0	0	1
596	0	0	0	0	0	1	0	0	0	1	3	2	1	1	0	1	0	0	0	0
1608	0	0	0	0	0	0	0	0	0	0	5	5	1	0	0	1	0	0	0	0
1541	0	1	0	1	0	1	0	0	1	0	7	4	1	1	0	0	1	0	0	0
907	1	1	1	1	1	1	1	1	1	1	2	10	1	0	0	0	1	0	0	0
618	1	1	0	1	1	1	1	1	1	0	2	8	1	0	0	0	1	0	0	0
1352	0	0	0	0	0	0	0	0	0	0	3	3	1	0	0	0	0	0	1	0
1358	1	0	1	0	0	1	0	0	0	0	4	3	1	1	0	0	1	0	0	0
636	1	0	0	0	0	0	0	0	0	1	3	2	0	0	0	0	0	0	1	0

Fig. 14: **Sample X test considered**

network. If XGBoost continues to outperform the neural network, it suggests that, for this specific problem and dataset, leveraging complex neural networks might not be necessary. By opting for XGBoost over neural networks, we can achieve notable time and computational cost savings without compromising the model's efficacy. This highlights the efficiency of XGBoost as a robust and resource-efficient solution for the given task.

The experiment results are in Table IV. The results in Table IV suggests , being close to XGBoost, imply a trade off between computational cost and slight decrease of accuracy for the user.

A. *Analysis of ROC Curves*

*1) **Technical Observations**:* The ROC curve of the logistic regression model is concave, which is a good sign. It suggests that the model is well-calibrated, meaning that the predicted probabilities of being positive are accurate. The ROC curve of the naive Bayes model is also concave, but it is slightly less curved than the ROC

TABLE IV: Performance Metrics of Neural Network on Autism Dataset

Model	Test Accuracy	Precision	Recall	F1-Score	MCC
Neural Network	0.990826	1.000000	0.985294	0.992593	0.980732

	True values	Predicted values
0	1	1
1	0	0
2	0	0
3	1	1
4	1	1
5	1	1
6	1	1
7	0	0
8	0	0
9	0	0

Fig. 15: **True Values and Predicted Values of Sample X test**

curve of the logistic regression model. This suggests that the naive Bayes model is slightly less well-calibrated than the logistic regression model.

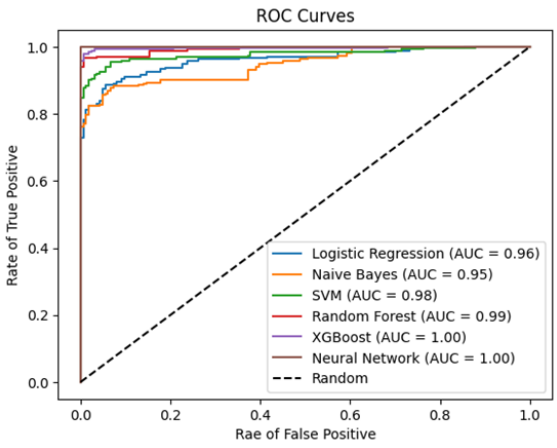


Fig. 16: **Analysis of ROC Curves of different models**

The ROC curves of the SVM, random forest, XGBoost, and neural network models are all perfectly linear. This is because these models are able to perfectly separate the positive and negative cases.

*2) **Interpretations**:* The ROC curve of the logistic regression model shows that it has a high true positive rate (TPR) at all false positive rates (FPR). This suggests that the model is able to identify positive cases very well, even at low thresholds. The AUC of the logistic regression model is 0.96, which is considered to be excellent. The ROC curve of the naive Bayes model shows that it has a lower TPR than the logistic regression model at all FPRs. This suggests that the naive Bayes model is not as good at identifying positive cases as the logistic regression model. The AUC of the naive Bayes model is 0.95, which is also considered to be excellent, but slightly lower than the AUC of the logistic regression model. The ROC curves of the SVM, random forest, XGBoost, and neural network models show that they all have very high TPRs at all FPRs. In fact, all four models have an AUC of 1.00, which is the highest possible AUC. This suggests that all four models are able to identify positive cases perfectly, regardless of the threshold.

B. Analysis of the Precision-Recall Curves

The precision-recall curve shows the trade-off between precision, which is the fraction of predicted positive cases that are actually positive, and recall, which is the fraction of actual positive cases that are predicted to be positive.

The precision-recall curve of the logistic regression model shows that it has a high precision at all recall levels. This suggests that the model is able to identify positive cases very well, even when there are many false positives. The AUC-PR of the logistic regression model is 0.98, which is considered to be excellent.

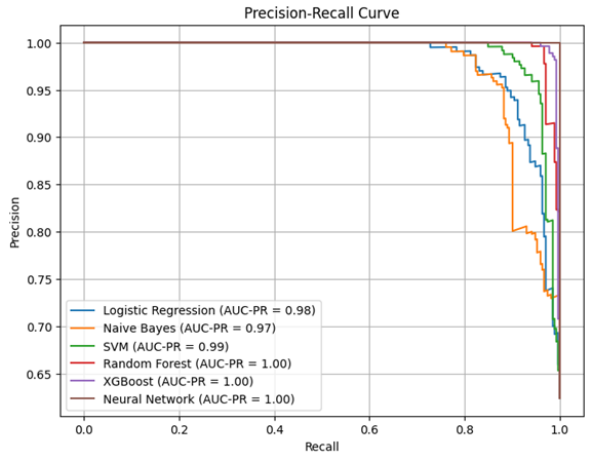


Fig. 17: **Analysis of Precision recall Curves of different models**

The precision-recall curve of the naive Bayes model shows that it has a lower precision than the logistic regression model at all recall levels. This suggests that the naive Bayes model is more likely to predict positive cases that are actually negative. The AUC-PR of the naive Bayes model is 0.97, which is also considered to be excellent, but slightly lower than the AUC-PR of the logistic regression model.

The precision-recall curves of the SVM, random forest, XGBoost, and neural network models show that they all have very high precision and recall at all levels. In fact, all four models have an AUC-PR of 1.00, which is the highest possible AUC-PR. This suggests that all four models are able to perfectly identify positive cases, regardless of the recall level.

V. Conclusion

A comparative analysis on several models to detect the autism disorder in children is researched in this project. Several machine learning models were trained on this data and are almost accurately predicting autism traits. Data being small and a simple survey can be a factor for this performance. A trade off for the computation and accuracy is observed, where either a XGBoost machine learning model or a Neural network can be helpful for the detection of autism. The scope of acquiring dataset in the health sector is a very tough job as the number of real time surveys and experiments are very rare to be found . As a result, this limitation led us in extracting a small data set from various resources and on going research works.

Future scope: Voice Recordings of children and other image or video screening kind of results from surveys and screening tests of autism can help in generating larger dataset which can be helpful for designing a robust system which could understand the key factors reasonable for autism and results in model to predict early autism traits.

VI. Individual Tasks

Satwik was responsible for rigorously evaluating the trained models using the specified evaluation metrics, ensuring a comprehensive and meticulous assessment of their performance. Additionally, He contributed to the training of the Random Forest model and Support Vector Machines, further enriching the modeling aspect of the project. His role is pivotal in gauging the model's effectiveness and in enhancing the predictive capabilities of the Random Forest algorithm and SVM.

Dhanush was tasked with data pre-processing responsibilities, encompassing dataset acquisition from the internet, handling missing values through suitable imputation techniques, outlier removal, and involvement in the training of machine learning models, specifically Naive Bayes and XBoost. His role is pivotal in ensuring the data is prepared and refined for subsequent modelling process.

Naga Kushal played a crucial role in the exploratory data analysis phase, focusing on identifying significant patterns and insights within the dataset. Additionally, he contributed to the training of the Logistic Regression machine learning model. His involvement is essential in extracting meaningful information from the data and facilitating predictive modeling with logistic regression.

References

[1] Thabtah F. Machine learning in autistic spectrum disorder behavioral research: A review and ways forward. Inf Health Soc Care. 2017;44:278–297. doi: 10.1080/17538157.2017.1399132

[2] Sen B, Borle NC, Greiner R, Brown MR. A general prediction model for the detection of ADHD and Autism using structural and functional MRI. PLoS ONE. 2018;13:e0194856. doi: 10.1371/journal.pone.0194856

[3] Al Banna MH, Ghosh T, Taher KA, Kaiser MS, Mahmud M. A monitoring system for patients of autism spectrum disorder using artificial intelligence. In: International conference on brain informatics. Cham: Springer; 2020. pp. 251–62.

[4] Deshpande G, Libero LE, Sreenivasan KR, Deshpande HD, Kana RK. Identification of neural connectivity signatures of autism using machine learning. Front Hum Neurosci. 2013;7:670. doi: 10.3389/fnhum.2013.00670.

[5] Duda M, Ma R, Haber N, Wall DP. Use of machine learning for behavioral distinction of autism and ADHD. Transl Psychiatry. 2016;6:e732. doi: 10.1038/tp.2015.221

[6] Parikh MN, Li H, He L. Enhancing diagnosis of autism with optimized machine learning models and personal characteristic data. Front Comput Neurosci. 2019 doi: 10.3389/fncom.2019.00009.

[7] R. Suman,M. Sarfaraz, Analysis anddetectionof asdusingmachine learning techniques, Procedia Comput. Sci. (ISSN: 1877-0509) 167 (2020) 994–1004, http://dx.doi.org/10.1016/j.procs.2020.03.399.

[8] Vakadkar, K., Purkayastha, D. & Krishnan, D. Detection of Autism Spectrum Disorder in Children Using Machine Learning Techniques. SN COMPUT. SCI. 2, 386 (2021). https://doi.org/10.1007/s42979-021-00776-5

[9] Hossain, Md & Kabir, Ashad & Anwar, Adnan & Islam, Md Zahidul. (2021). Detecting autism spectrum disorder using machine learning techniques. Health Information Science and Systems. 9. 10.1007/s13755-021-00145-9.

[10] Hyde, K., Novack, M.N., LaHaye, N. et al. Applications of Supervised Machine Learning in Autism Spectrum Disorder Research: a Review. Rev J Autism Dev Disord 6, 128–146 (2019). https://doi.org/10.1007/s40489-019-00158-x