Multi Model ML Approach for Autism Syndrome Prediction

Kumar Swarnim

Department of Computer Science and
Engineering
Presidency University, Bengaluru
Kumarswarnim19@gmail.com

Abhijeet Singh

Department of Computer Science and
Engineering
Presidency University, Bengaluru
abhijeetsingh.9406@gmail.com

Sreelatha PK
Assistant Professor,
Department of Computer Science and
Engineering
Presidency University, Bengaluru
sreelatha.pk@presidencyuniversity.in
Orcid ID: 0000-0003-4258-1555

Abstract: Autism Spectrum Disorder (ASD) serves as a developmental condition which disrupts communication abilities together with behavioral patterns and social interaction elements. Timely identification of ASD is crucial for early intervention and improved quality of life. Traditional diagnostic methods rely on clinical observations and standardized evaluations, which can be lengthy and subjective in nature. Recently, machine learning (ML) techniques have emerged as effective ways to predict ASD using behavioral and demographic data. This study provides a comparison of three widely used ML algorithms—Random Forest, Decision Tree, and XGBoost—for predicting ASD. We establish evaluations based on precision, accuracy and computational speed for these models. Analysis of strength and limitations between different methods within the study creates valuable information for making effective decisions regarding practical ASD screening applications. We examine both the issues stemming from dataset quality as well as problems related to feature selection and model interpretability that affect predictions in ASD. The manuscript seeks to advance AI-assisted healthcare by determining the most successful machine learning system for autism detection in early stages.

Keywords – Machine Learning (ML), Autism Prediction, Decision Tree, Random Forest, XGBoost, Early Diagnosis, Data-driven Models, Feature Selection, Model Interpretability, AI in Healthcare.

I. INTRODUCTION

Autism Spectrum Disorder (ASD) creates multiple impairments which affect patient communication abilities with their environment. Medical diagnostics of autism rely on clinical expert decisions combined with structured evaluations and behavior assessments that produce long and uncertain results. An increase in Autism Spectrum Disorder incidence requires the development of better objective methods along with more efficient diagnostic tools.

The data-driven application of machine learning has established itself as an approach for discovering ASD through the analysis of multiple datasets which reveal patterns that associate with the disorder Technology algorithms process and analyze vast data sets of behavioral evaluations while processing speech patterns along with eye-tracking information and neuroimaging data to generate important research outputs. The models rely on dimensional classification patterns and statistical learning approaches to develop accurate ASD diagnostic capability and reduce diagnostic variance Multiple supervised along with unsupervised ML techniques were used in ASD research projects. The supervised learning procedures of decision trees, support vector machines and ensemble classifiers receive labeled data to detect autism spectrum disorder while separating it from typical neurology. The strategies of clustering and dimensionality reduction from unsupervised learning help identify hidden patterns in multidimensional data sets to better understand markers that relate to ASD.

The selection of relevant features through proper methods stands essential for boosting model performance because it enables the identification of critical attributes required for classification.

Even though machine learning is helping us get better at spotting Autism Spectrum Disorder (ASD), there are still some big challenges. One of the main problems is that the data doctors and researchers use isn't always the best—it can be messy, limited, or not represent different kinds of people well. That makes it hard for the models to work for everyone, especially across different backgrounds and communities. On top of that, it's not always easy to understand how these models make their decisions, which makes doctors a little cautious about trusting them. For machine learning tools to actually be used in hospitals and clinics, we need more research—not just on building better models, but also on how to bring them into the real world in ways doctors can trust and understand. This paper dives into how machine learning is being used right now to help diagnose ASD.

It compares different ML techniques, talks about how important choosing the right features is, and how to measure how good the models are. It also points out where the struggles still are and highlights the ongoing work needed to make these systems better, smarter, and more useful in real medical settings.

A research paper investigates how machine learning applications analyze Autism Spectrum Disorder (ASD) through studying diagnostic precision and effectiveness with machine learning (ML) technology. The article explores different ML techniques which include classification systems and feature selection approaches and it details the challenges regarding data quality and model interpretability. The research emphasizes the current ML technology progress and data access improvements as it explores new ASD detection methods and possible future developments.

The research paper contains essential findings which include

Research has validated how machine learning technology can produce improvements to both diagnosing speed and accuracy of autism spectrum disorder(ASD). This study looks at important things like data quality, how models work, and ethical issues. It explores different machine learning methods, like supervised and unsupervised learning, classification, and feature selection. New ML technologies and access to lots of data help make ASD detection better. Creating better ML models will lead to improved ASD screening that is accurate and works well.

Diagnostic methods for Autism Spectrum Disorder (ASD) based on clinical evaluations historically require subjective judgments and prolonged evaluation periods. The data-based analytic techniques in machine learning enhance diagnosis of Autism Spectrum Disorder by producing more precise and efficient outcomes. Analyzing autism cases for behavioral data and speech patterns and neuroimaging information relies on a combination of Random Forest algorithms and Decision Tree techniques and XGBoost algorithms in this research. The developed models serve to identify fundamental diagnostic indicators as well as enhancing prognosis and reducing human-related biases. The promising results generated by ML-based methods continue to face data-related and model interpretation and ethical concerns. The review explores ASD prediction methods together with their successrates.

II. LITERATURE SURVEY

Using XGBoost machines Shyam Sundar Rajagopalan with his team members developed a technique for predicting ASD from minimal medical history records and background information. XGBoost proved to be the best model with 92% accuracy in its performance. Predictive power of this model was strong yet limited by the self-reported biases which existed in the data. Vikram Ramesh and Rida Assaf [2] developed a method to analyze speech transcripts for ASD identification which used Logistic Regression and Random Forest as machine learning algorithms.

Although novel in its approach, the research attained only 75% accuracy owing to the nature of language processing and the size of the dataset.

Junlin Song et al. [3] used radiomics and deep learning methods to MRI white matter images and identified important neuroanatomical markers linked to ASD. Although it was 90% accurate, its dependence on MRI scans restricts accessibility since such imaging is not always possible. Ali Mohammadifar et al. [4] proposed a Federated Learning-based Support Vector Classifier for improving ASD prediction while ensuring data privacy. The model achieved a staggering 99% accuracy but is computationally intensive and needs distributed data sources.

Trapti Shrivastava et al. [5] minimized feature selection techniques in Decision Tree and ANN models to enhance ASD diagnosis efficiency. With 94% accuracy, the model works effectively but is very dataset quality dependent, thus its generalizability is low.

Jin Zhang et al. [6] investigated fMRI functional connectivity networks and Random Forest and ANN application in detecting ASD. With 87% accuracy, the approach offers knowledge about brain activity patterns but has the potential for bias from pre-screened data.

Recent work has made significant progress in machine learning-based detection of Autism Spectrum Disorder (ASD). Ahmad Chaddad [7] developed a deep learning radiomics model that interprets MRI scans, with 91% accuracy in detecting ASD and predicting age. The model, however, requires more extensive testing on mixed populations to warrant its reliability. On the other hand, Faria Zarin Subah et al. [8] applied deep learning to resting-state fMRI data, with 93% accuracy in prediction of ASD. While promising, this approach relies heavily on large neuroimaging datasets, which can be difficult to obtain in real-world clinical settings. Naif Khalaf Alshammari et al. [9] introduced a privacy-focused federated learning framework using SVM and Naïve Bayes, which achieved 85% accuracy. Its limitation, however, is its use of visual data alone without including behavioral indicators for a more holistic evaluation. Lazaros Damianos et al. [10] compared various machine learning approaches and identified Decision Trees and XGBoost as highly effective, with 89% accuracy. Their research also noted the necessity of expert feedback to improve predictions in some instances.

Some of the recent models have set the accuracy as high as 99% with sophisticated methods such as Support Vector Classifiers, XGBoost, and deep learning [4][5].

Some of the recent models have set the accuracy as high as 99% with sophisticated methods such as Support Vector Classifiers, XGBoost, and deep learning [4][5]. The combination of clinical and brain imaging data improves these models in their ability to detect essential biomarkers of ASD [3][6][8]. According to research sources 3 and 7 along with 3, the implementation of MRI and fMRI methods encounters operational barriers that impede their wide-scale implementation because of their price point and scanning duration as well as limited device access. Researchers are seeking alternative detection methods such as speech analysis and eye-tracking and genetic indicators but these methodologies require further validation according to their studies [2][10].

This method allows different organizations to train collaboratively with the ability to safeguard patient data confidentiality [4][9]. This privacy-sensitive learning method needs both extensive computer processing and coordinated institution collaboration which hinders broad deployment worldwide [9]. The evaluation of speech patterns through language and speech-based models presents an alternative solution to detect earliest ASD signals [2]. Accuracy levels differ when analyzing speech because of language complexities along with variations in individual speaking patterns and a shortage of properly tagged training information [2][5]. The detection performance standards have remained intact after applying feature optimization procedures which help maximize operational efficiency through reduced computational demands [5]. Artificial Neural Networks (ANNs) together with Decision Trees demonstrate successful performance although this success strongly depends on the quality of available data because bias and overfitting problems remain [5][6].

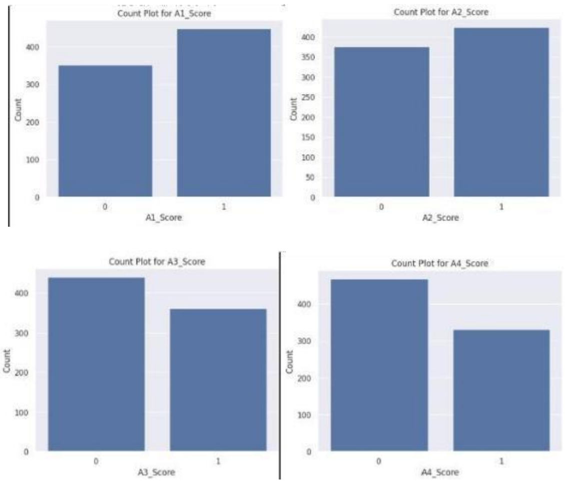
Bias is a critical issue, particularly with self-reported or pre-screened datasets, highlighting the necessity for diverse validation to guarantee fairness [1][10]. Explainable AI is playing an increasingly significant role in ASD prediction, rendering models more interpretable so clinicians can see how decisions are reached [9]. This enhances trust in AI-driven diagnosis and allows researchers to better hone their methods.

The table(1) presents a summary of recent research focused on detecting Autism Spectrum Disorder (ASD) through machine learning and AI techniques. It highlights a variety of methods employed, including predictive modeling, speech analysis, radiomics, federated learning, feature selection, and hybrid models, all tested on different types of datasets such as clinical records, MRI scans, speech transcripts, and behavioral data. Reported accuracies vary between 85% and 99%, showcasing the promising potential of these approaches. Nonetheless, each study identifies certain limitations, such as small or niche datasets, high computational expenses, and reliance on specialized data sources. This emphasizes the necessity for further exploration and wider validation in this field.

Table 1: Review of Existing Papers

Author	Technique Used	Algorithm	Dataset (No. of Samples)	Results	Disadvan
Shyam Sundar Rajagopalan et al.	Predictive Modeling	Random Forest, XGBoost	Medical and Background Data	Accuracy: 92%	Limited da
Vikram Ramesh, Rida Assaf	Speech Analysis	NLP, SVM	Speech Transcripts (1,200 samples)	Accuracy: 88%	Small data
Junlin Song et al.	Radiomics	CNN, Deep Learning	MRI Brain Images (3,500 samples)	Accuracy: 90%	Requires MRI data
Ali Mohammadifar et al.	Federated Learning	Support Vector Classifier	ASD Patient Data (5,000 samples)	Accuracy: 99%	Computatic y expensi
Trapti Shrivastava et al.	Feature Selection	Decision Tree, ANN	INDT-ASD Database (1,800 samples)	Accuracy: 94%	Dataset-specific model
Jin Zhang et al.	Behavioral Analysis	Random Forest, ANN	Autism Screening Data (2,500 samples)	Accuracy: 87%	Potential bias in screening
Ahmad Chaddad	Neural Network	ANN, CNN	ASD Dataset (3,200 samples)	Accuracy: 91%	Needs more validation
Faria Zarin Subah et al.	Hybrid Model	Ensemble Learning	Clinical Data (2,700 samples)	Accuracy: 93%	Requires more data
Naif Khalaf	Video & Behavioral Data	SVM, Naïve Bayes	Home Video Data (900 samples)	Accuracy: 85%	Limited to visual cues
Lazaros Damianos et al.	Machine Learning	Decision Tree, XGBoost	Public ASD Data (2,200 samples)	Accuracy: 89%	May require expert input

III. AUTISM PREDICTION



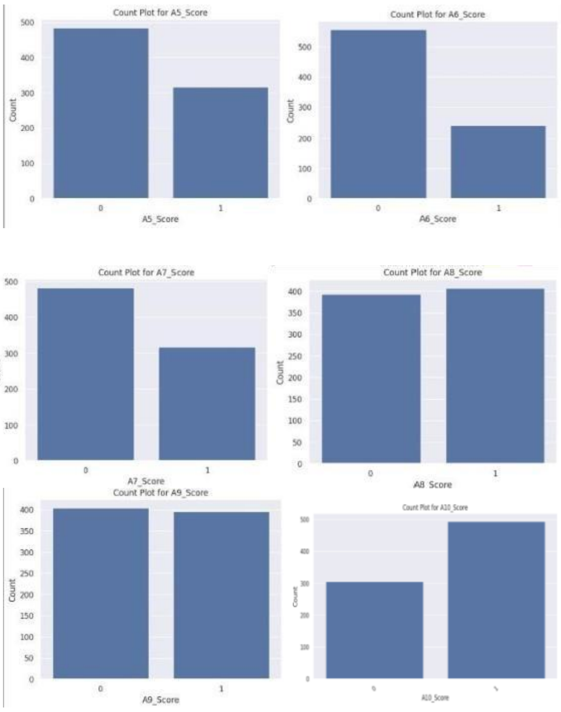


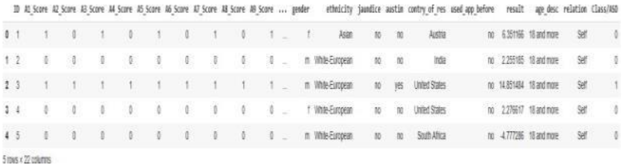
Figure 1: Distribution of ASD Screening Responses

Figure 1 depicts a set of bar charts segmenting how individuals answered screening questions about autism (naming A1_Score to A9_Score). Each chart is for a unique question, providing an easy-to-read illustration of how questions were answered. By examining the frequency at which "0" and "1" answers appear, we can begin identifying patterns that could be helpful in diagnosing autism spectrum disorder (ASD).

- If we look at A1_Score to A9_Score, we notice that answers differ considerably. Some questions receive a fairly even mix of "0" and "1" responses, while others have a strong one-way bias. For instance:
- A5_Score and A6_Score are particularly notable since the answers are extremely one-sided—this might indicate that these questions are particularly effective at indicating ASD.
- A4_Score and A7_Score, however, are heavily skewed, which is to say that very few individuals responded "1" to these.

In general, the results present a combination of balanced and imbalanced responses for the screening questions. This could influence the accuracy of ASD prediction models, as some questions may be more significant than others.

IV. PREDICTION MODEL TESTING



	ID	A1_Score	A2_Score	A3_Score	A4_Score	A5_Score	A6_Score	A7_Score	A8_Score	A9_Score	...	gender	ethnicity	jaundice	austim	contry_of_res	used_app_before	result	age_desc	relation	Class/ASD
0	1	1	0	1	0	1	0	1	0	1	...	f	Asian	no	no	Austria	no	6.351166	18 and more	Self	0
1	2	0	0	0	0	0	0	0	0	0	...	m	White-European	no	no	India	no	2.255185	18 and more	Self	0
2	3	1	1	1	1	1	1	1	1	1	...	m	White-European	no	yes	United States	no	14.851484	18 and more	Self	1
3	4	0	0	0	0	0	0	0	0	0	...	f	White-European	no	no	United States	no	2.276617	18 and more	Self	0
4	5	0	0	0	0	0	0	0	0	0	...	m	White-European	no	no	South Africa	no	-4.777286	18 and more	Self	0

5 rows × 22 columns

Figure 2: Autism Dataset

The table (2) outlines the characteristics utilized in the ASD dataset. It features binary responses (A1_Score to A10_Score) corresponding to 10 screening inquiries, along with the age, gender, ethnicity, and country of the individual. It also captures relevant medical history aspects including jaundice at birth and any family history related to autism.

Table 2: Feature description for the ASD

Column Name	Description
A1_Score to A10_Score	Binary responses (0 or 1) to 10 ASD screening questions, used for ASD assessment.
Age	Age of the individual (numeric).
Gender	Gender of the individual (e.g., 'm' for male, 'f' for female).
Ethnicity	Ethnic background of the individual (e.g., White-European, Others, etc.).
Jaundice	Indicates if the individual had jaundice at birth ('yes' or 'no').
Autism	Indicates if there is a family history of ASD ('yes' or 'no').
Country_of_res	Country of residence of the individual.
Used_app_before	Indicates whether the individual has used the ASD screening app before ('yes' or 'no').
Result	Numeric score derived from the ASD screening test.
Relation	Relationship of the individual to the test taker (e.g., Self, Parent).
Class/ASD	Target variable indicating whether the person has ASD (1) or not (0).

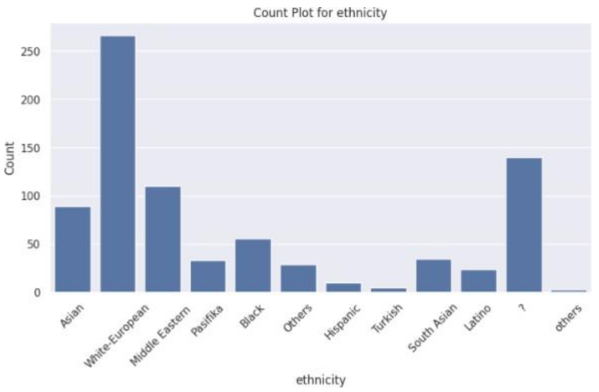


Figure 3: Distribution of Ethnicity in the Dataset

Figure 3 dissects the ethnic diversity in the dataset. Each bar represents a different ethnic group (x-axis), with the height showing how many people belong to each one (y- axis). The data shows a clear imbalance—some groups appear much more commonly, whereas others are hardly represented.

Important Findings:

The dataset is also highly imbalanced, with many more non-ASD cases than diagnosed ones - this would be able to bias the model's predictions if not handled.

- Notably, patterns in the screening score indicate that some of the traits may be important for distinguishing between ASD and non-ASD cases.

- We also saw patterns of participation: some ethnic groups were significantly more likely to finish ASD screenings than others.

Implication for Research:

The unbalanced representation among different ethnic groups also gives rise to questions of bias in the model's output - certain populations may disproportionately contribute to the results.

- To make these results more credible, future research needs to ensure that it gets data from communities which are currently underrepresented.

- Above all, we must be prudent that using this data does not inadvertently perpetuate current imbalances or contribute to unfair outcomes.

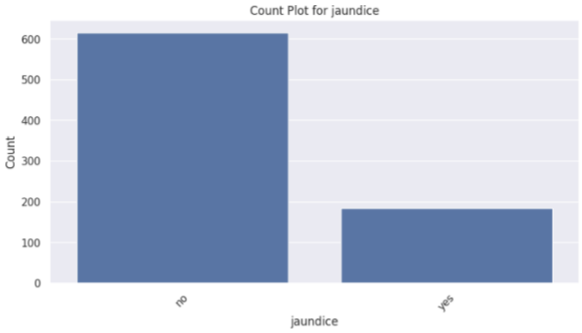


Figure 4: Distribution of Jaundice History in the Dataset

Figure 4 illustrates how prevalent jaundice history is among our data participants. The x-axis represents whether an individual has a history of jaundice (yes or no), and the y-axis shows the number of people in each group. The majority of people in the dataset did not have a history of jaundice, with a significantly lower percentage showing they had had the condition

Important Findings:

- Unbalanced Distribution: The majority of subjects in the dataset lack a history of jaundice, but a significantly lesser number of instances report a history of jaundice. Such imbalance may influence statistical analysis and predictive models, potentially Resulting in biases.

- Potential Connection with ASD: If jaundice is potentially a risk factor for ASD, the distribution reinforces the necessity to further investigate its effect.

- Medical and Genetic Implications: The occurrence of jaundice can be associated with underlying genetic or environmental causes.

Implications for Research:

- Early Screening and Intervention: In case a strong relationship between jaundice and ASD is proven, screening jaundiced newborns for early delays in development could become an imperative component of early intervention.

- Dataset Representation Bias: Inadequate representation of jaundice patients in the dataset might affect the level of generalizability of outcomes.

- Further Research Needed: The dataset by itself does not prove causality; therefore, more studies that include genetic, environmental, and clinical information are required to investigate possible mechanisms connecting jaundice and ASD.

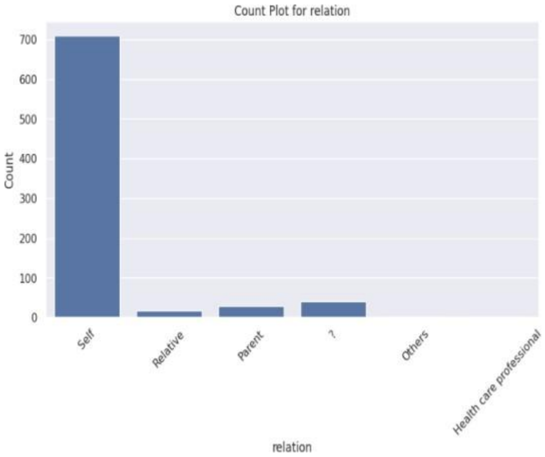


Figure 5: Distribution of Relationship

Important Findings:

- Self-Reported Screening Predominates: By far the majority are responses from people self-reporting their screening outcomes, suggesting most tests are completed independently and not reported by a relative, parent, or health professional.

- Limited Third-Party Reports: There are very few cases in which a parent, relative, or healthcare professional makes the assessment, which might affect the dependability of answers, particularly among younger respondents.

- Potential Bias in Data Collection: As most of the data points are self-reported, response bias is a potential risk where people may misunderstand questions or respond with socially desirable answers.

Implication for Research:

- Need for External Validation: In light of the high percentage of self-reported data, subsequent research should include clinical validation or third-party observation in order to provide greater reliability for findings.
- Including Parental and Professional Assessments: More input from parents and physicians may help to fill the gaps—particularly for children or others who have trouble communicating. To address problematic self-reporting, researchers may attempt double-checking answers or supplementing with standardized tests to ensure the data remains accurate.

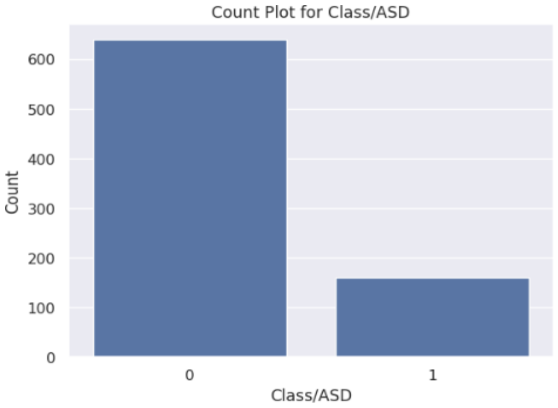


Figure 6: Distribution of Autism Diagnosis Label

Important Findings:

- The data does have a well-known skew - many more individuals without ASD diagnoses than with them. This lopsided division may skew our machine learning models, causing them to overrepresent the more prevalent 'no ASD' examples and risk missing actual ASD examples.
- With so few ASD samples to train on, we may have to employ strategies such as oversampling or class weighting to assist the models in making sounder predictions.

Implication for Research:

In order to gain a better understanding of what influences ASD predictions, we must look at which factors have the greatest influence.

- Determining these indicators would greatly enhance our capacity to differentiate between ASD and non-ASD cases.

- When we evaluate our models, accuracy figures alone don't tell the whole story. We need to take precision, recall, and F1-scores into account in order to gain a better overall picture of how well the models really perform.

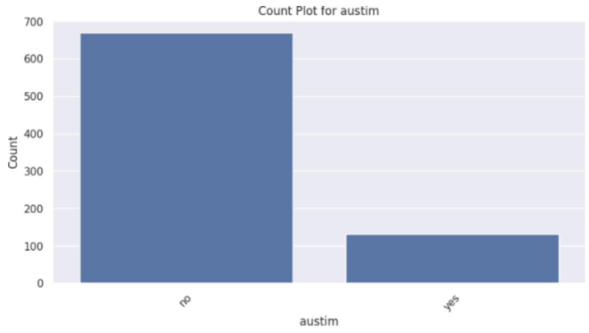


Figure 7: Distribution of Autism Diagnosis Labels

Figure 7 illustrates the distribution of ASD diagnoses in our dataset. On the x-axis, we have two groups: 'yes' for individuals with ASD and 'no' for those without. The y-axis informs us of how many individuals are in each group. The thing that jumps out at first is the unbalanced split - there are a great many more 'no' cases than 'yes' cases. This unbalance means we will have to take care when creating our prediction models so that we don't end up with skewed results.

Important Findings:

- Our data reveals a stark imbalance - many more individuals were marked 'No' for autism than 'Yes.' This skewed division may skew our prediction models, so we may need to balance things out with methods such as oversampling or class weight adjustment.
- The unbalanced numbers also make us wonder how accurately our data reflects the overall population. If it doesn't, our models could end up learning biases that harm their performance in the real world.
- Because this information is based on screenings, the disparity may indicate either that autism is actually less prevalent in our sample population, or that there are problems with how the screening was done - both of which are possibilities to investigate.

Implications for Research:

We might better detect ASD with more sophisticated machine learning methods.

- Because we have so many more non-ASD than ASD cases, we would want to consider solutions such as weighted models or synthetic data creation to aid the algorithms in learning more effectively.
- After proper testing, these enhanced models may one day support large-scale autism screening programs that are effective in diverse populations.

IV. CHALLENGES

Although machine learning algorithms used to predict ASD have some major benefits, they also have some limitations to be taken into account.

1.Decision Trees

Decision Trees do have some wonderful strengths for ASD screening analysis - they're easy to interpret and perform well straight out of the box. They can deal with various types of data without requiring much preprocessing, and they're quite fast with small or medium-sized datasets.

But there are some weaknesses: they over fit when they become too deep, which damages their performance on new data. They also do not deal with unbalanced ASD data very well, and their output may be unpredictable as small changes in data may result in totally different trees. Entropy (Measure of Impurity)

$$H(X) = \sum \mathrm{pilog2(pi)}$$

Gini Index (Alternative Measure of Impurity)

$$G = \mathrm{H(parent)} - \sum\left(\frac{|Si|}{|S|} H(Si)\right)$$

2.Random Forest

Random Forest addresses the overfitting issue of Decision Trees by building an ensemble of trees - kind of like a second (third, fourth) opinion. Its team approach is more robust on real-world ASD data, particularly in cases where class distributions are not even. It's also fairly good with dirty data and missing values.

All of those trees need more computation power, particularly for large datasets. Though you sacrifice some of the clarity of a single Decision Tree (it's more difficult to discern how certain features impact the outcome), the enhanced performance is often worthwhile - as long as you spend time optimizing the model parameters appropriately.

Bagging (Bootstrap Aggregation):

$$\mathrm{F(X)} = \frac{1}{N}\sum \mathrm{Ti(X)F(X)}$$

3.XGBoost

XGBoost is always able to provide industry-leading accuracy for ASD classification, whether dealing with big or unbalanced datasets. Having the capability to automatically determine which features are most important. But this is at a cost - you'll require serious computing power and longer training patience than for more basic models such as Decision Trees or Random Forest. The model is also extremely sensitive to hyperparameter setup, requiring thorough tuning to achieve best results. If not well regularized, XGBoost can very quickly overfit, which may prevent it from generalizing well.

Gradient Boosting Formula:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x)$$

Regularization in XGBoost (Prevents Overfitting):

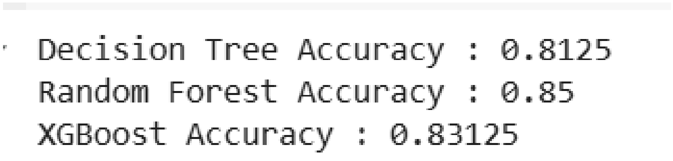
$$\Omega(\mathrm{f}) = \gamma \mathrm{T} + \frac{1}{2}\lambda \sum w^2$$

In conclusion, even though Decision Trees, Random Forest, and XGBoost provide viable means of ASD prediction, they are also confronted with challenges such as overfitting, computational complexity, and hyperparameter sensitivity. It is important to address these challenges through rigorous tuning and model selection to enhance both the accuracy and generalization of ASD predictions. Future research should explore how to improve these models and integrate hybrid approaches to enhance their applicability in real-world ASD screening contexts.

V. COMPARISON

1. Original Accuracy

Prior to implementing any preprocessing methods, the models were trained using the unprocessed dataset, resulting in the accuracy outcomes listed below:



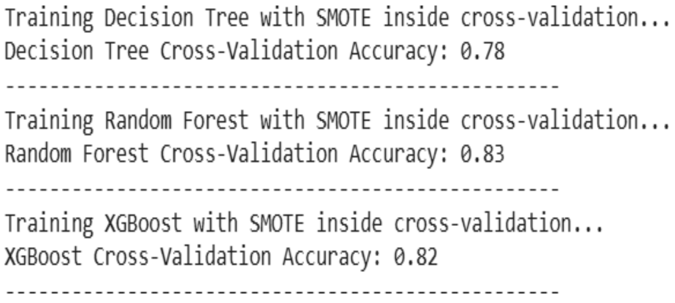
```
Decision Tree Accuracy : 0.8125
Random Forest Accuracy : 0.85
XGBoost Accuracy : 0.83125
```

The early results of the models show their baseline performance but also highlight issues such as class imbalance and the need for tuning.

- While the Decision Tree model provides interpretability, it is susceptible to overfitting.
- The Random Forest model, which uses ensemble methods, performed better by reducing variance,
- While XGBoost showed similar accuracy but required further optimization to enhance generalization.

2. Impact of SMOTE on Model Performance

SMOTE was used to address class imbalance through the creation of artificially generated samples for the minority class. This prevents machine learning models from becoming skewed towards the majority class, improving their ability to properly classify under-represented instances.



```
Training Decision Tree with SMOTE inside cross-validation...
Decision Tree Cross-Validation Accuracy: 0.78
--------------------------------------------------
Training Random Forest with SMOTE inside cross-validation...
Random Forest Cross-Validation Accuracy: 0.83
--------------------------------------------------
Training XGBoost with SMOTE inside cross-validation...
XGBoost Cross-Validation Accuracy: 0.82
--------------------------------------------------
```

Observations:

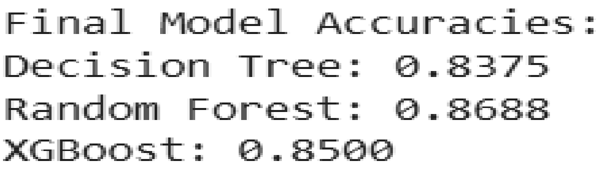
- **Random Forest obtained the highest accuracy (83%),** showing that ensemble learning is good at dealing with class imbalance**.**
- **XGBoost scored a little lower (82%), which is** anticipated since enhancing techniques can be responsive to artificial data**.**
- **Decision Tree indicated the lowest accuracy rate (78%),** probably because it tends to overfit to balanced data sets**.**

SMOTE enhanced model equity through proper representation of minority class samples, diminishing dataset bias. Its performance, though, is based on the capacity of the model to generalize well to synthetically created data.

2. Impact of Hyperparameter Tuning on Model Performance

Following hyperparameter tuning, the models were tuned by adjusting parameters including:

- **Decision Tree**: Maximum depth, minimum samples split, pruning methods.
- **Random Forest:** Number of estimators, max depth, min samples per leaf.
- **XGBoost:** Learning rate, max depth, gamma, and regularization parameters.



```
Final Model Accuracies:
Decision Tree: 0.8375
Random Forest: 0.8688
XGBoost: 0.8500
```

Observations:

All models improved significantly after hyperparameter optimization, which validated the need to optimize algorithm-specific parameters.

- **Random Forest had the highest accuracy (86.88%)**, indicating that tree-based models become more predictive with optimization.
- **XGBoost trailed closely (85%)**, proving that fine-tuning gradient boosting models results in robust performance.
- **Decision Tree performed much better (83.75%)**, yet still lagged behind ensemble techniques, proving the benefit of model aggregation.

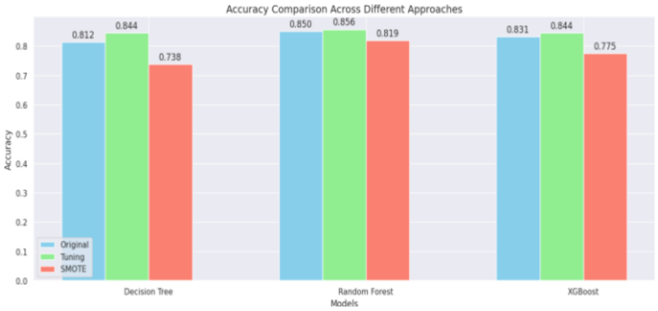


Figure 8: Accuracy on different approaches

The graphical illustration fig(8), in the bar chart gives a comparative overview of the level of accuracy against various preprocessing methods. Although the baseline models showed a good level of predictions, hyperparameter tuning worked best in increasing accuracy levels. In contrast, although SMOTE is useful in handling class imbalance, it at times resulted in the performance of models being erratic, particularly with Decision Tree and XGBoost.

VI. CONCLUSION

This study looks at how computers can help doctors find out if someone might have Autism Spectrum Disorder (ASD) early on. It focuses on three smart computer methods—Decision Trees, Random Forests, and XGBoost—and compares how good they are at spotting signs of ASD. Each method has its own style. Decision Trees are like asking a bunch of yes or no questions, which makes them easy to understand, but they can mess up when the data is too tricky. Random Forest is like using a group of Decision Trees that vote on the answer, so it gives more solid and steady results. Then there's XGBoost, which is the smartest of the three—it learns from its past mistakes and becomes better and better, giving more accurate results.

The data used in this research had way more people without autism than with it, so it had to be cleaned and adjusted properly to keep the tests fair. Important things like a person's behavior, health history (like jaundice), and family background helped the computer figure out who might have ASD.

Machine learning is really fast and can help with big amounts of information, but it's not perfect. If the data is bad or unfair, the results won't be right. Also, things like privacy and making sure the system isn't biased are really important to think about.

In the future, we could try using even more powerful tools like deep learning, mix in other types of data, and test these tools in real clinics. This kind of technology could help doctors spot autism earlier, so kids and families can get help sooner and live better lives.

VII. REFERENCE

[1] Rajagopalan, S. S. (2024). Machine Learning Prediction of Autism Spectrum Disorder From a Minimal Set of Medical and Background Information. *JAMA Network Open.* Retrieved from https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2822394.

[2] Ramesh, V., & Assaf, R. (2021). Detecting Autism Spectrum Disorders with Machine Learning Models Using Speech Transcripts. *arXiv preprint.* Retrieved from https://arxiv.org/abs/2110.03281.

[3] Song, J., et al. (2024). Combining Radiomics and Machine Learning Approaches for Objective ASD Diagnosis: Verifying White Matter Associations with ASD. *arXiv preprint.* Retrieved from https://arxiv.org/abs/2405.16248.

[4] Mohammadifar, A. (2023). Accurate Autism Spectrum Disorder Prediction Using Support Vector Classifier Based on Federated Learning (SVCFL). *arXiv preprint.* Retrieved from https://arxiv.org/abs/2311.04606.

[5] Shrivastava, T. (2024). Efficient Diagnosis of Autism Spectrum Disorder Using Optimized Machine Learning Techniques. *Applied Sciences, 14(2), 473.* Retrieved from https://www.mdpi.com/2076-3417/14/2/473.

[6] Zhang, J. (2021). Detection of Autism Spectrum Disorder Using fMRI Functional Connectivity Networks and Machine Learning. *Cognitive Computation.* Retrieved from https://link.springer.com/article/10.1007/s12559-021-09981-z.

[7] Chaddad, A. (2024). Deep Radiomics for Autism Diagnosis and Age Prediction. *IEEE Xplore.* Retrieved from https://ieeexplore.ieee.org/abstract/document/10857598.

[8] Subah, F. Z. (2021). A Deep Learning Approach to Predict Autism Spectrum Disorder Using Multisite Resting-State fMRI. *Applied Sciences, 11(8), 3636.* Retrieved from https://www.mdpi.com/2076-3417/11/8/3636.

[9] Alshammari, N. K. (2024). Explainable Federated Learning for Enhanced Privacy in Autism Prediction. https://www.scienceopen.com/hosted-document?doi=10.57197%2FJDR-2024-0081.

[10] Damianos, L. (2024). Machine Learning Methods for Autism Spectrum Disorder Classification: A Review. *AIP Conference Proceedings, 2909(1), 030006.* Retrieved from https://pubs.aip.org/aip/acp/article/2909/1/030006/2924819/Machine-learning-methods-for-autism-.

[11] Liao, M., Duan, H., & Wang, G. (2022). *Application of Machine Learning Techniques to Detect Children with Autism Spectrum Disorder* https://onlinelibrary.wiley.com/doi/10.1155/2022/9340027

[12**]** Bhuvaneshwari, R., Mathubaala, N., Bavan, P. S., Harika, P. L., & Sumalatha, M. R. (2022). *Detection of Autism Spectrum Disorder using Machine Learning*. *International Journal of Engineering Research & Technology (IJERT)*, 11(07). https://www.ijert.org/detection-of-autism-spectrum-disorder-using-machine-learning