**Detection of Autism from Behaviour Data**

|  |  |
| --- | --- |
| Bhagyashree Sharma  Student  Amity Institute of Information Technology,  Amity University Rajasthan, Jaipur  bhagyashree0613@gmail.com | Mr. Amit Kumar Singh  Assistant Professor  Amity Institute of Information Technology,  Amity University Rajasthan, Jaipur |

***Abstract*-** The Autism Spectrum Disorder affects how individuals communicate, behave, and interact socially. It's crucial to detect ASD early because early help can improve development and provide needed support [1]. The project looks at using computer technology to identify signs of ASD by examining data that doesn't require a clinical setting. This method aims to make screening simpler and quicker. Three non-clinical ASD datasets were used: a toddler screening dataset (1200+ instances), an adult screening dataset (700+ entries), and a mixed-age group dataset (2000+ entries). After cleaning the information, 16 important attributes were chosen. These included answers from a questionnaire about autism traits, as well as details about each person's background and behavior. Six different machine learning models: Logistic Regression, Support Vector Machine (SVM), Random Forest, XGBoost, and Gradient Boosting and Decision Tree were trained and assessed for their effectiveness. The models using combinations of methods, like Random Forest, XGBoost, and Gradient Boosting, showed the best results. They were able to identify ASD traits with an impressive accuracy of 99.67%. This demonstrates that artificial intelligence (AI) tools can support early ASD detection by using easily accessible data, removing the need for direct clinical tests. This approach has the potential to make ASD screening faster and more available. Schools, homes, and community health systems could use it to help identify ASD in more people [2], bridging the gap in early diagnosis and intervention.

***Keywords—*** *Autism Spectrum Disorder, Machine Learning, Early Diagnosis, Support Vector Machine (SVM), Random Forest, XGBoost, Logistic Regression, Gradient Boosting, Non-clinical Screening, AI in Healthcare.*

1. **INTRODUCTION**

Autism is a neurodevelopmental condition that affects how

people, especially children, develop in terms of communication and social interaction. It is often marked by specific interests and repetitive behaviour. Globally, more children seem to have ASD, with reports suggesting that around 1 in 68 children [5] in India might be on the spectrum. Recognizing ASD early is crucial as it leads to better support and outcomes. Unfortunately, the usual process to diagnose ASD can be slow and is often hard to access, especially in low-resource or rural areas. With new advancements in Artificial Intelligence (AI) and Machine Learning (ML), healthcare has new tools to assist in diagnosing ASD. These technologies can create models that analyse data to help predict if someone might have ASD, potentially before seeing a healthcare professional. Such tools can work alongside existing methods, offering more widespread, affordable, and simpler options for ASD screening [1]. A recent study explored using machine learning to predict ASD traits by examining three public datasets: two covering various ages and one focusing on toddlers. The study used 16 different features, which included basic demographic information and answers to the Autism Quotient (AQ) questionnaire. Researchers tested six machine learning models: Logistic Regression, Support Vector Machine, Decision Tree, Random Forest, XGBoost, and Gradient Boosting. Among these, the models known as ensemble models: Random Forest, XGBoost, and Gradient Boosting were the most accurate, achieving a 99.67% accuracy rate. The project basically aims to prove that these machine learning models can accurately predict ASD traits using easily available data. This could enable early detection of ASD in places like schools, homes, and community health centers [3]. Such an approach may help identify ASD more quickly and start support and care earlier, especially in areas with limited healthcare services.

1. **RELATED WORK**

Recent studies that are using Machine Learning techniques to

detect Autism in people early:

**Pallavi Mahar** et al. emphasized the importance of factors before birth and behavior in ASD, showing why early screening tools are needed.

**S. Shakir** and his team proposed combining methods like Random Forest, SVM, and Naive Bayes to improve accuracy, using a voting strategy.

**Vaishali R.** demonstrated that selecting just 10 important features out of a total of 21 can maintain high accuracy.

**Raj S., & Masood, S.** Analysised and detection of autism spectrum disorder using machine learning techniques.

**Anil Kumar** explored federated learning to predict ASD, which keeps personal data private. All these efforts show that machine learning is becoming more effective in screening for Autism Spectrum Disorder.

1. **METHODOLOGY**

The study examined Autism Spectrum Disorder using three publicly accessible datasets. Two datasets featured individuals from various age ranges, while one was specific to toddlers. Each contained demographic details like age, gender, ethnicity, and family ASD history, along with responses from the Autism Quotient (AQ) questionnaire. In total, 16 features were identified as crucial, including age, gender, family history of ASD, and 10 questions from the AQ test.

1. **Data Preprocessing**

To prepare the data for model training, several preprocessing steps were performed. Missing values were handled using appropriate imputation methods, and inconsistent or duplicate entries were removed to ensure data quality. Categorical variables such as gender, ethnicity, and relation were encoded using label encoding and one-hot encoding. Numerical features were standardized to ensure uniformity across varying scales. The pre-processed dataset was then split into training and testing sets using a 75:25 ratio, facilitating reliable evaluation of model performance on unseen data.

1. **Model Development**

Six supervised machine learning classifiers were implemented to predict ASD traits:

1. Logistic Regression
2. Decision Tree Classifier
3. Random Forest Classifier
4. XGBoost
5. Gradient Boosting Classifier
6. Support Vector Machine (SVM)

Each model was trained on the same input features, and relevant hyperparameters were tuned using grid search and empirical testing to optimize classification performance. The models were developed using Python and scikit-learn, with additional support from XGBoost and related libraries.

Among all classifiers, the Support Vector Machine model received focused evaluation due to its effectiveness in handling high-dimensional data and non-linear decision boundaries [1]. Various kernel functions—linear, polynomial, and radial basis function (RBF)—were tested, with the RBF kernel demonstrating the highest classification accuracy. Hyperparameters such as the regularization parameter (C) and kernel coefficient (gamma) were fine-tuned to enhance generalization and minimize overfitting.

* 1. **Model Evaluation**

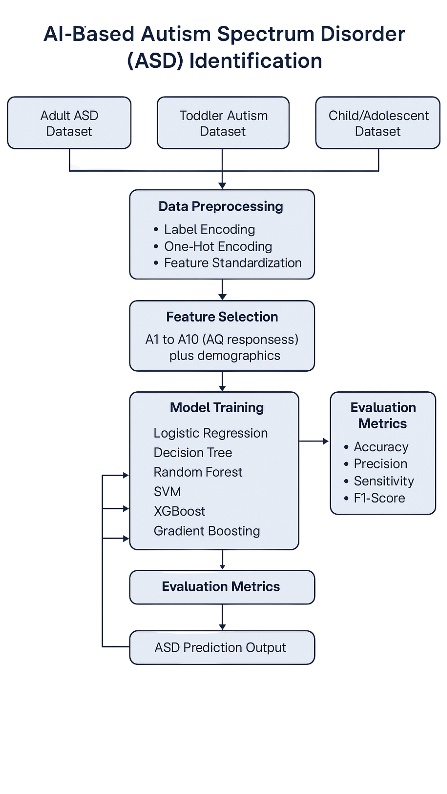
Model performance was assessed using standard classification metrics:

1. **Accuracy**: The proportion of correctly classified instances.
2. **Precision**: The ratio of true positives to total predicted positives.
3. **Recall (Sensitivity)**: The ratio of true positives to total actual positives.
4. **F1-Score**: The harmonic mean of precision and recall.

A confusion matrix was generated for each classifier to examine the distribution of true positives, true negatives, false positives, and false negatives. This provided a granular understanding of each model's prediction behaviour, particularly for the SVM classifier [1].

1. **Comparative Analysis and Visualization**

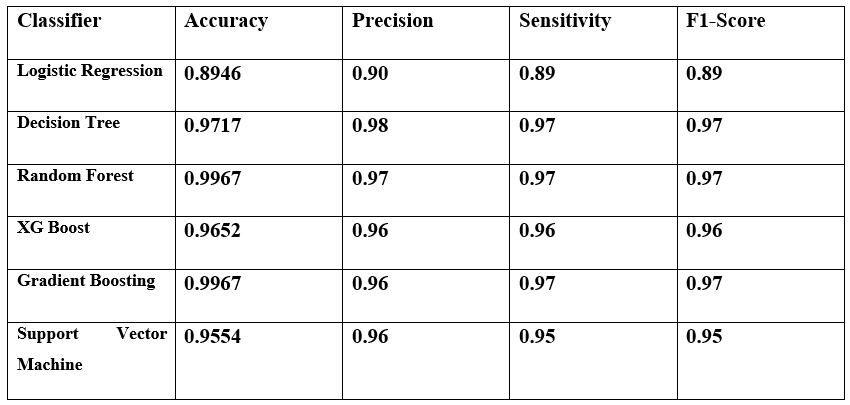
A comparative results table was compiled to summarize the performance metrics of all six models, enabling direct comparison in terms of classification effectiveness. Visualizations such as heatmaps, distribution plots, and feature importance graphs were created to support exploratory data analysis and model interpretation. Feature importance scores from ensemble methods, such as Random Forest and XGBoost, were used to identify the most significant predictors of ASD traits.



*Fig1: AI-Based Autism Spectrum Disorder (ASD) Identification*

1. **RESULT AND DISCUSSION**

This section outlines the performance of the machine learning models applied to the task of Autism Spectrum Disorder (ASD) trait identification. Six supervised machine learning models were trained and three publicly available datasets evaluated to predict ASD. The model included Logistic Regression, Support Vector Machine (SVM), Decision Tree, Random Forest, XGBoost, and Gradient Boosting.

The datasets were preprocessed to select 16 key features, including demographic variables and responses from the Autism Quotient (AQ) screening questionnaire. A standard 75:25 train-test split was used for performance evaluation.

*Table: 1- Performance Metrics for Each Classifier*

1. **Model Performance Analysis**

Each Model's Performance analysis to see how well each of the six supervised machine learning models works. These models have been specifically trained to predict Autism Spectrum Disorder.

* **Logistic Regression**

As a linear classifier, Logistic Regression is best suited for datasets with linearly separable classes. While it remains one of the simplest and most interpretable models, it struggled to capture non-linear and complex relationships present in the ASD data. Despite its lower performance, its fast training time and explainability make it suitable for quick screening applications where interpretability is critical [6].

* **Decision Tree**

The Decision Tree classifier achieved a strong performance as it constructs a hierarchical model that splits data based on feature values, allowing it to model non-linear relationships effectively. Decision Tree’s visual interpretability makes it a practical choice for domains requiring model transparency.

* **Random Forest**

Random Forest emerged as one of the best-performing classifiers, as this ensemble model builds multiple Decision Trees on random subsets of the dataset and aggregates their predictions to produce a more stable and generalizable output. Its high performance in this task can be attributed to its ability to reduce overfitting and efficiently manage both categorical and continuous data. Random Forest proved especially reliable in classifying ASD and non-ASD cases across all age groups in the dataset.

* **XGBoost**

It is basically known for its speed and efficiency, XGBoost employs gradient boosting to iteratively optimize model predictions. Though slightly less accurate than Random Forest in this experiment, XGBoost’s performance was still robust, highlighting its capability to capture subtle data patterns. It also handles missing data effectively, further increasing its applicability to real-world healthcare data.

* **Gradient Boosting**

It sequentially trains weak learners, basically shallow Decision Trees to correct errors made by prior models. This iterative refinement leads to highly accurate predictions. Its strength lies in minimizing bias and improving generalization. The high score across all metrics indicates the model’s effectiveness in handling the nuances of the ASD dataset, making it a viable candidate for deployment in AI-based ASD detection tools.

* **Support Vector Machine (SVM)**

Though it did not surpass the ensemble methods in terms of raw accuracy, SVM offered consistent and balanced classification performance. Its robustness in handling high-dimensional data makes it a strong option for scenarios where classes are not linearly separable. A confusion matrix was also generated for SVM to analyse its sensitivity and specificity, showing that the model maintained a reasonable trade-off between correctly identifying ASD cases and minimizing false positives [1].

|  |  |  |
| --- | --- | --- |
|  | **Predicted Positive (ASD)** | **Predicted Negative (Non-ASD)** |
| **Actual Positive (ASD)** | 445 (TP) | 23 (FN) |
| **Actual Negative (Non-ASD)** | 18 (FP) | 434 (TN) |

*Table: 2- Confusion Matrix Table for SVM*

The confusion matrix shows that SVM correctly classified most ASD and non-ASD cases, with 445 true positives and 434 true negatives. It made 23 false negative and 18 false positive errors. This balanced result shows that SVM performs reliably, even if its accuracy is slightly lower than ensemble models.

1. **Comparative Analysis**

The performance evaluation clearly indicates that **ensemble models** which are Random Forest, Gradient Boosting, and XGBoost, outperformed individual models such as Logistic Regression, Decision Tree, and SVM. The ensemble models’ capability to combine multiple weak learners allowed them to achieve higher generalization, making them more effective in real-world prediction scenarios.

Random Forest and Gradient Boosting, in particular, stood out with their **99.67% accuracy**, suggesting they are highly suitable for ASD detection tasks, especially where high precision and minimal false negatives are required. XGBoost, while slightly behind, still provided reliable and consistent outputs with fewer errors compared to non-ensemble models.

Decision Tree showed interpretability and good performance, while Logistic Regression and SVM, although simpler, did not match the complexity-handling abilities of the ensemble methods. Though, SVM’s performance was consistent and revealed good balance in classifying both positive and negative cases.

1. **Practical Relevance**

The models evaluated here rely on basic behavioural and demographic data such as AQ screening responses, age, gender, and family history of ASD. The ensemble models, in producing highly accurate classifications without the need for clinical or biological data underscores their potential in scalable ASD screening.

These models can be integrated into non-clinical platforms, such as educational institutions, community centers, even as mobile applications and in regions with limited access to specialized healthcare.

1. **Discussion**

The experimental results demonstrate that Random Forest and Gradient Boosting achieved exceptional performance in predicting Autism Spectrum Disorder, with both models reaching 99.67% accuracy and maintaining balanced precision, recall, and F1-scores of 0.97.

The project shows that AI tools can be effectively added to existing autism screening methods. These tools are helpful for doctors, in areas without many child brain specialists or development experts. By spotting high-risk children early, AI models can reduce the wait time for diagnosis. This early detection helps start important behavior treatments sooner, which is most effective when children are young. Additionally, AI models are generally accurate and reliable. They are less likely to make mistakes, such as diagnosing autism incorrectly or missing cases. Using AI in real-life settings can increase trust among healthcare workers and families, encouraging them to use AI-based screening systems as an initial step in sorting cases [7].

The ensemble models and SVM performed well, but it's just as important to know which features influenced their predictions the most. In Random Forest and Gradient Boosting models, the importance of each feature came from the points where decisions were made in the trees. The key features were:

* AQ responses such as A4, A5, and A9, which show clear behaviour patterns linked to ASD.
* Neurological history, including jaundice or a family history of ASD.
* Delays in communication and the age when screening was carried out.

These findings line up with established clinical research, indicating that behaviour signs, neurological factors, and early-life indicators play a crucial role in identifying ASD. Understanding which features are important not only makes the model more transparent but also helps in designing future questionnaires or strategies for data collection. Moreover, this project basically validates and integrates AI into broader personalized healthcare solutions.

**V. CONCLUSION**

This project demonstrates the potential of machine learning techniques in the early detection of Autism Spectrum Disorder based on behavior data. The system employs a variety of machine learning models, including Random Forest, Gradient Boosting, Support Vector Machine (SVM), Logistic Regression, and XGBoost, to classify individuals as either exhibiting ASD traits or not. The project aims to provide an accurate, data-driven approach for early ASD detection, ultimately aiding in timely interventions that can significantly improve the quality of life for individuals with ASD [8].

The variety and quality of datasets were essential for capturing diverse ASD-related traits and ensuring effective model training. The three dataset that have been used are:

* 1. **ASD Screening Dataset:**

This dataset covers individuals across all age groups and is designed to assess ASD traits using features like:

* The Autism Quotient (AQ-10)
* Behavioural features
* Demographic features.

It provides a comprehensive view of how different factors like age, gender, and medical history may influence ASD trait predictions.

* 1. **Toddler Autism Dataset:**

Focused on toddlers, this dataset assesses ASD traits in early developmental stages using attributes:

* QChat-10 questionnaire, which is adapted for younger age groups.
* Age of the toddler in months.
* Biological sex of the toddler (Male/Female).
* Ethnic background of the individual (e.g., White-European, Black, Asian).
* Whether the child had neonatal jaundice at birth (Yes/No).
* Indicates if any family member has been diagnosed with ASD (Yes/No).
* Person who completed the test on behalf of the child (e.g., Parent, Guardian).
* Country where the child resides.
* Whether the user has used the ASD screening app before (Yes/No).
* Total score based on the AQ-10 responses.
* Age group classification (typically “Toddler” for this dataset).
* Final ASD classification outcome (Yes/No).
  1. **ASD Detection – Multi Feature Dataset:**

This dataset covers a wide range of age groups, not just adults. It features the following attributes:

* Comprehensive
* Behavioural
* Psychological
* medical history
* AQ responses

The various models tested demonstrated strong results, with ensemble methods like Random Forest and Gradient Boosting leading the way in terms of accuracy [10]. The Support Vector Machine (SVM), while slightly less accurate, still provided a reliable performance, making it an important model for future exploration.

This project shows how machine learning can help detect autism signs without using clinical test data. It's a useful tool for early screening in schools, homes, or communities, especially where healthcare services are scarce. By using easily accessible non-clinical information, these systems can help identify autism early and reduce the pressure on healthcare facilities. The study results indicate that with careful selection of datasets and model testing, AI-based systems can be crucial in early autism detection [10]. These tools can be developed into real-time diagnostic systems or mobile apps to assist caregivers, teachers, and healthcare workers in making quick, informed decisions. The work completed here sets a solid foundation for future research involving more complex data, real-time monitoring, and collaboration with medical professionals.

The current project has demonstrated promising results in using machine learning models to predict Autism Spectrum Disorder based on non-clinical datasets.

The project also emphasizes inclusivity by using datasets that cover a broad age range and different socio-cultural backgrounds. This ensures that the model does not only work in ideal lab settings, but has potential in real-world environments, including in schools and primary healthcare systems. It is crucial to make this system easier to use, especially in areas where it is difficult to get a quick diagnosis because there are not enough specialists. In rural and underserved regions, delays often happen [10]. By connecting this system with mobile apps or websites, families, caregivers, and teachers can have a simple and low-cost tool. This tool can help them spot early signs of ASD. Partnering with medical experts can help everyone trust the system. This cooperation ensures that it is used in a good and ethical way.

**VI. FUTURE SCOPE**

If Autism is not identified early, it can deeply affect a person's life, especially during childhood. Children with undiagnosed autism may struggle with communication, social interaction, and learning in school. They may feel left out, misunderstood, or frustrated because they can’t express themselves like others. This can lead to low self-esteem, isolation, and even anxiety or depression as they grow up [9].

In many cases, these children miss out on special support that could help them learn better and connect with others. Without early help, they may find it harder to build relationships, succeed academically, or adapt to changes in life. In adulthood, undiagnosed ASD can affect job opportunities, independence, and emotional well-being.

That’s why early detection is so important. By finding signs of autism at a young age, especially in schools, we can help children get the care and support they need right away. This can improve their communication skills, confidence, learning experience, and overall quality of life, allowing them to reach their full potential.In the future, more types of data can be added, like genetic information, brain scans, or data collected over time. This can help make the predictions more accurate for different groups of people. Real-time data from smartwatches or mobile apps can help with early detection and ongoing monitoring.

Using data from different age groups and backgrounds will help the system work better for everyone. Creating more data using the existing data and fixing any errors or missing parts will also make the results stronger.

If we used data from smartwatches or mobile apps [5], this could track a child's behaviour every day. For example, if a toddler isn’t making eye contact or not responding to sounds, the system could notice this through an app and send an alert. This helps with early detection, which important in small children because support at an early age can make a huge difference in their development [3].

This project has strong potential to help with the early detection of Autism in children, especially through schools in both rural and urban areas. In the future, the system can include more detailed types of data, such as genetic information, brain scans, or behaviour tracked over time. This would help make the predictions even more accurate for different kinds of children.

Since the project deals with sensitive health-related information, it is important to follow strict privacy and ethical guidelines such as GDPR and HIPAA. Ensuring that data is kept anonymous and protected is essential for building trust in the system. Collaboration with doctors and healthcare professionals will also help in aligning the system with real-world medical needs, improving both its accuracy and practical value.

**VII. REFERENCES**

[1] Raj, S., & Masood, S. (2020). Analysis and detection of autism spectrum disorder using machine learning techniques. *Procedia Computer Science*, *167*, 994-1004.

[2] Raja, S. A., Batool, A., Sana, M., Khaliq, H. M. H., Choudhry, F., & Devi, D. (2024). Exploring the Mechanisms of Neuronal Protection by Glial Cell Line-Derived Neurotrophic Factor in Autism Spectrum Disorder. *Cureus*, *16*(10).

[3] Farooq, M. S., Tehseen, R., Sabir, M., & Atal, Z. (2023). Detection of autism spectrum disorder (ASD) in children and adults using machine learning. *scientific reports*, *13*(1), 9605.

[4] Kamala, B., Pooja, K. M., Varsha, S., & Sivapriya, K. (2021, December). ML Based Approach to Detect Autism Spectrum Disorder (ASD). In *2021 4th International Conference on Computing and Communications Technologies (ICCCT)* (pp. 313-318). IEEE.

[5] Mazumdar, A., Chatterjee, B., Banerjee, M., & Shanker, S. (2024). Machine learning based autism screening tool—a modified approach. *Multimedia Tools and Applications*, *83*(32), 77831-77848.

[6] Nawghare, P., & Prasad, J. R. (2024, December). Early Detection of Autism Spectrum Disorder Using AI and Machine Learning Models: A Systematic Review for Effective Intervention. In *2024 IEEE Pune Section International Conference (PuneCon)* (pp. 1-6). IEEE.

[7] Hossain, M. D., Kabir, M. A., Anwar, A., & Islam, M. Z. (2021). Detecting autism spectrum disorder using machine learning techniques: An experimental analysis on toddler, child, adolescent and adult datasets. *Health Information Science and Systems*, *9*, 1-13.

[8] Jain, S., & Tripathy, H. K. (2024, February). Machine Learning Methods for the Timely Identification of Autism Spectrum Disorder in Toddlers. In *2024 International Conference on Emerging Systems and Intelligent Computing (ESIC)* (pp. 515-520). IEEE.

[9] Uddin, K. M. M. (2023). A machine learning approach to predict autism spectrum disorder (ASD) for both children and adults using feature optimization. *Network Biology*, *13*(2), 37.

[10] Talukdar, J., Gogoi, D. K., & Singh, T. P. (2023). A comparative assessment of most widely used machine learning classifiers for analysing and classifying autism spectrum disorder in toddlers and adolescents. *Healthcare Analytics*, *3*, 100178.