A Machine Learning Framework for Early – Stage Detection of Autism Spectrum Disorders

Zahara Mirza[1], Md. Ateeq Ur Rahman[2], Subramanian K M[3]
[1]*PG Scholar, Department of Computer Science and Engineering, Shadan College of Engineering and Technology, Hyderabad, Telangana, India - 500086*
[2,3]*Professor, Department of Computer Science and Engineering, Shadan College of Engineering and Technology, Hyderabad, Telangana, India – 500086*

***Abstract*—autism spectrum disorder (ASD) is a neurodevelopmental condition that affects communication, social interaction, and behavior. Early detection of ASD is essential because timely intervention can significantly improve cognitive, social, and emotional development in children. Traditional diagnostic methods often rely on clinical observations and expert assessments, which may be time-consuming, subjective, and inaccessible in resource-limited settings. A Machine Learning Framework for Early-Stage Detection of Autism Disorders provides an intelligent and data-driven approach to assist healthcare professionals in identifying potential ASD cases at an early stage. The framework collects and processes behavioral, demographic, and developmental data, followed by data preprocessing, feature selection, and model training using machine learning algorithms such as Decision Trees, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Gradient Boosting. The trained model analyzes patterns within the data and predicts the likelihood of autism with high accuracy. Performance evaluation metrics including accuracy, precision, recall, F1-score, and ROC-AUC are used to compare models and identify the most effective classifier. The proposed framework aims to reduce diagnostic delays, minimize human bias, and support clinicians with reliable decision-making tools. Additionally, it can be integrated into web or mobile healthcare applications to enable accessible screening in schools, hospitals, and remote areas. By leveraging machine learning techniques, the system offers a scalable, cost-effective, and efficient solution for early autism detection, ultimately contributing to improved treatment planning and better quality of life for affected individuals and their families.**

***Index Terms*—Machine Learning, Autism Spectrum Disorder (ASD), Early Detection, Classification Models, Predictive Analytics, Healthcare Analytics, Feature Selection, Behavioral Analysis, Decision Support System, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Artificial Intelligence, Medical Diagnosis, Data Mining.**

I. INTRODUCTION

Autism Spectrum Disorder (ASD) is a complex neurodevelopmental disorder characterized by persistent challenges in social communication, restricted interests, and repetitive patterns of behavior. Symptoms usually appear during early childhood, but the severity and presentation vary significantly among individuals, making diagnosis challenging [1].

According to global health studies, the prevalence of ASD has increased over the past few decades, highlighting the need for efficient and accessible screening methods [2].

Early identification of autism is essential because timely intervention can improve language development, learning ability, social interaction, and overall quality of life. However, conventional diagnostic procedures depend heavily on clinical observations, behavioral assessments, and expert evaluations, which can be time-consuming, expensive, and subject to human interpretation [3].

In many developing regions, limited access to trained specialists' further delays diagnosis and treatment. Recent advancements in Artificial Intelligence (AI) and Machine Learning (ML) have created new opportunities for supporting healthcare professionals in disease prediction and decision-making. Machine learning algorithms can analyze large volumes of behavioral and demographic data, identify hidden patterns, and generate accurate predictions that assist clinicians during the screening process [4].

These techniques have demonstrated promising results in various medical applications, including neurological and developmental disorder detection. The proposed Machine Learning Framework for Early-Stage Detection of Autism Disorders utilizes supervised learning algorithms such as Random Forest, Support Vector Machine (SVM), Decision Tree, and K-Nearest Neighbors (KNN) to classify individuals based on autism-related features. Data preprocessing, feature engineering, and model evaluation techniques are incorporated to improve prediction accuracy and reliability [5].

The framework aims to provide an automated, scalable, and cost-effective tool that supports early screening while complementing professional medical diagnosis rather than replacing it.

By integrating machine learning into autism detection, healthcare systems can reduce diagnostic delays, improve accessibility in underserved areas, and enable earlier intervention strategies. Such intelligent systems have the potential to enhance clinical decision-making and contribute to better long-term outcomes for children and families affected by ASD [6].

II. LITERATURE REVIEW

The application of machine learning in healthcare has gained significant attention due to its ability to analyze complex datasets and support early disease diagnosis. In the context of Autism Spectrum Disorder (ASD), researchers have explored various computational techniques to improve screening accuracy and reduce dependence on lengthy clinical assessments. Early studies focused on statistical analysis and questionnaire-based methods, which provided useful insights but often lacked scalability and predictive power [1].

Several researchers have investigated supervised machine learning algorithms for autism detection using behavioral and demographic datasets. Decision Trees and Random Forest models have shown promising performance because of their ability to handle nonlinear relationships and identify the most informative features contributing to ASD prediction. These models also provide interpretable decision paths that can assist healthcare professionals during diagnosis [2].

Support Vector Machine (SVM) has been widely adopted in autism research due to its effectiveness in high-dimensional classification problems. Studies have demonstrated that SVM-based classifiers achieve high accuracy when trained on carefully preprocessed datasets with optimized feature selection techniques. However, their performance may depend on parameter tuning and kernel selection [3].

Deep learning and neural network approaches have also been explored for autism detection using medical images, facial expressions, and behavioral data. Although these methods can capture complex patterns automatically, they generally require large datasets and significant computational resources, limiting their practical application in small-scale healthcare environments [4].

Recent literature highlights the importance of feature engineering and data preprocessing in improving model performance. Techniques such as missing value imputation, normalization, and feature selection reduce noise and enhance classification accuracy. Ensemble learning methods, including Gradient Boosting and Random Forest, have consistently outperformed individual classifiers in several comparative studies [5].

Despite these advancements, existing systems still face challenges such as limited dataset diversity, class imbalance, and reduced generalizability across different populations. Consequently, researchers recommend hybrid machine learning frameworks that combine robust preprocessing, feature optimization, and multiple classification algorithms to improve reliability and support early-stage autism screening in real-world settings [6].

Table 1. Comparison of Existing Research on Machine Learning for Autism Detection

Ref. No.	Author(s) / Study	Technique Used	Strengths	Limitations
[1]	F. Thabtah (2019)	Machine Learning Review	Comprehensive analysis of ML methods for ASD detection	Focuses on review rather than implementation

[2]	Hossain et al. (2021)	Decision Tree, Random Forest	High classification accuracy and interpretable models	Performance depends on dataset quality
[3]	Cortes and Vapnik	Support Vector Machine (SVM)	Effective for high-dimensional data and binary classification	Requires careful parameter tuning and kernel selection
[4]	Goodfellow et al.	Deep Learning	Learns complex patterns automatically and handles large datasets	Computationally expensive and requires substantial training data
[5]	Chen and Guestrin	XGBoost (Gradient Boosting)	High predictive performance and efficient feature handling	Sensitive to hyperparameter settings and overfitting if not tuned
Proposed Framework	Machine Learning Framework for Early-Stage Detection o f Autism Disorders	Data Preprocessing + Feature Selection + Random Forest/SVM/KNN Ensemble	Improved early detection accuracy, scalable, cost-effective, supports clinical decision-making	Performance depends on availability of representative and balanced training data

III. METHODOLOGY

A. Data Collection

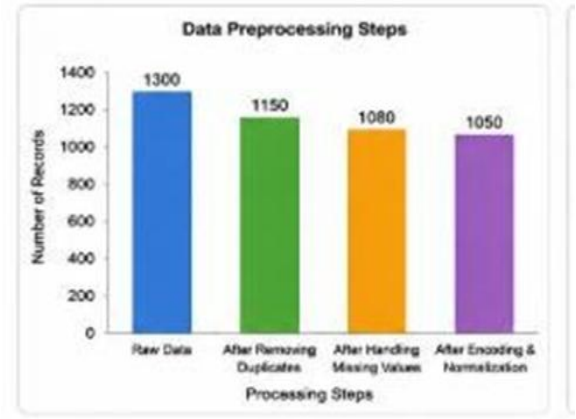
The proposed framework begins by collecting autism-related data from publicly available datasets, hospitals, screening questionnaires, and healthcare repositories. The dataset may include demographic information, behavioral responses, communication patterns, social interaction indicators, and developmental characteristics. Reliable and diverse data sources help improve the robustness and generalization capability of the machine learning models.

B. Data Preprocessing

Raw data often contains missing values, duplicate records, inconsistent formats, and noisy information. During preprocessing, missing values are handled using suitable imputation techniques, categorical variables are encoded, and numerical features are normalized or standardized. Data cleaning ensures that the dataset is suitable for effective machine learning analysis and minimizes prediction errors.

C. Feature Selection

Not all available attributes contribute equally to autism prediction. Feature selection techniques are applied to identify the most significant variables that influence classification performance. Removing redundant and irrelevant features reduces computational complexity, prevents overfitting, and improves model accuracy.



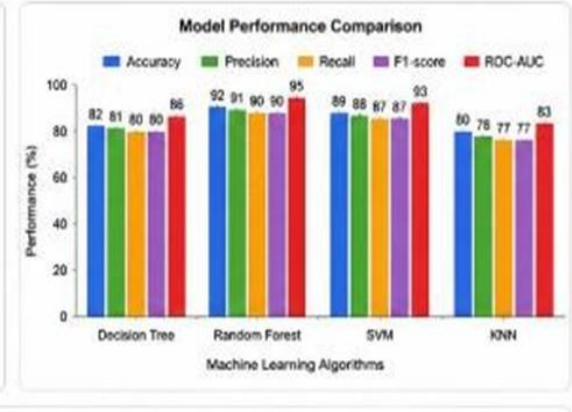
D. Model Training

The preprocessed dataset is divided into training and testing subsets. Various supervised machine learning algorithms such as Decision Tree, Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN) are trained using the selected features. Each model learns patterns associated with autism spectrum disorder and builds predictive relationships from historical data.

E. Model Evaluation

The trained models are evaluated using performance metrics including accuracy, precision, recall, F1-score, and Receiver Operating Characteristic–Area Under Curve (ROC-AUC). Confusion matrices are also analyzed to measure true positives, true negatives,

false positives, and false negatives.
Comparative evaluation helps identify the most effective algorithm for early autism detection.



F. Prediction and Decision Support

After selecting the best-performing model, the framework accepts new patient information as input and predicts the likelihood of autism spectrum disorder. The prediction results can assist healthcare professionals in conducting preliminary screenings and prioritizing individuals for comprehensive clinical assessment. The system is intended to support, not replace, expert medical diagnosis.

G. Workflow of the Proposed Framework

The overall methodology follows these sequential steps:

1. Collect autism-related behavioral and demographic data.
2. Clean and preprocess the dataset.
3. Perform feature selection to identify important attributes.
4. Split the dataset into training and testing sets.
5. Train multiple machine learning classifiers.
6. Evaluate and compare model performance.
7. Select the optimal model for prediction.
8. Generate early-stage autism risk predictions for new cases.
9. Provide decision support for clinicians and healthcare providers.

H. Advantages of the Methodology

- Enables faster and more objective early-stage autism screening.
- Reduces the impact of redundant features through feature selection.
- Improves prediction accuracy by comparing multiple machine learning algorithms.
- Supports scalable deployment in hospitals, schools, and remote healthcare settings.
- Assists clinicians in making informed decisions while complementing traditional diagnostic procedures.

IV. IMPLEMENTATION

A. Dataset Acquisition

The implementation starts with collecting autism-related data from publicly available datasets or healthcare screening records. The dataset typically contains demographic information, behavioral responses, communication skills, social interaction patterns, and developmental indicators. The collected data is stored in a structured format such as CSV files or databases for further processing.

B. Data Preprocessing Module

The raw dataset is cleaned to remove duplicate entries, handle missing values, and correct inconsistent records. Categorical attributes are encoded into numerical values, while numerical features are normalized or standardized. This preprocessing stage ensures high-quality input data for machine learning algorithms and improves overall prediction performance.

C. Feature Engineering and Selection

Important features influencing autism spectrum disorder (ASD) prediction are identified using statistical methods and feature selection algorithms. Irrelevant or redundant attributes are eliminated to reduce computational complexity and prevent overfitting. The resulting feature set enhances the efficiency and accuracy of the predictive models.

D. Machine Learning Model Development

Several supervised learning algorithms are implemented, including Decision Tree, Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN). The dataset is divided into training and testing subsets, allowing the models to learn from historical data and make predictions on unseen samples. Hyperparameter tuning is performed to optimize model performance.

E. Model Evaluation and Validation

The trained models are evaluated using standard performance metrics such as Accuracy, Precision, Recall, F1-Score, and ROC-AUC. A confusion matrix is generated to assess classification performance and identify false positives and false negatives. Cross-validation techniques are used to verify the stability and reliability of the selected model.

F. Prediction Module

After training, the best-performing machine learning model is deployed for prediction. Users provide behavioral and demographic information through an input interface, and the system analyzes the data to estimate the likelihood of Autism Spectrum Disorder. The generated prediction supports early screening and assists healthcare professionals in making informed decisions.

G. User Interface and System Integration

A simple web-based or desktop interface is developed to allow users or clinicians to enter patient details and obtain prediction results. The backend integrates preprocessing, feature selection, trained machine learning models, and result visualization into a unified framework. The interface presents predictions in an easy-to-understand format along with confidence scores where applicable.

H. Deployment and Future Scalability

The framework can be deployed on local servers or cloud platforms for wider accessibility. Future enhancements may include integration with electronic health records, mobile healthcare applications, and real-time monitoring systems. The modular design allows new datasets and advanced machine learning models to be incorporated without major architectural changes.

I. Implementation Workflow

- Collect autism-related data.
- Preprocess and clean the dataset.
- Perform feature selection and engineering.
- Split the data into training and testing sets.
- Train multiple machine learning algorithms.
- Evaluate model performance using standard metrics.
- Select the best-performing model.
- Deploy the model for autism prediction.
- Display results through the user interface to support early diagnosis.

J. Technologies Used

- Programming Language: Python
- Machine Learning Libraries: Scikit-learn, XGBoost
- Data Processing: Pandas, NumPy
- Visualization: Matplotlib, Plotly
- Development Environment: Jupyter Notebook / VS Code
- Web Framework (Optional): Flask or Django
- Dataset Format: CSV or SQL Database

V. RESULTS AND DISCUSSION

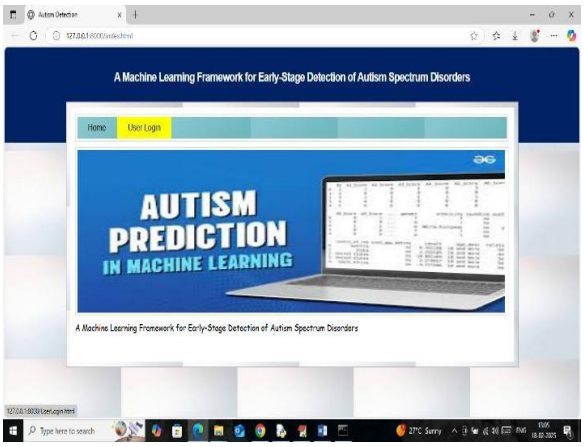


Fig.1.1 User Login

In above screen click on ‘User Login’ link to get

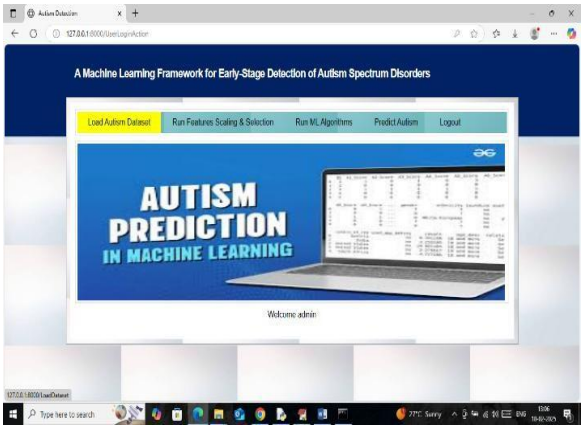


Fig 1.2 Load Autism Dataset

In above screen user can click on ‘Load Autism Dataset’ link to load all 4 datasets and then will get below page

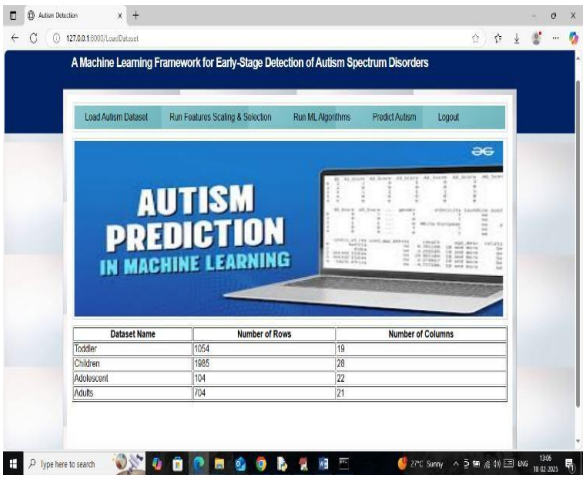


Fig 1.3

In above screen all 4 datasets loaded and can see number of rows and features available with each dataset and now click on 'Run Features Scaling & Selection' link to process features and then will get below page

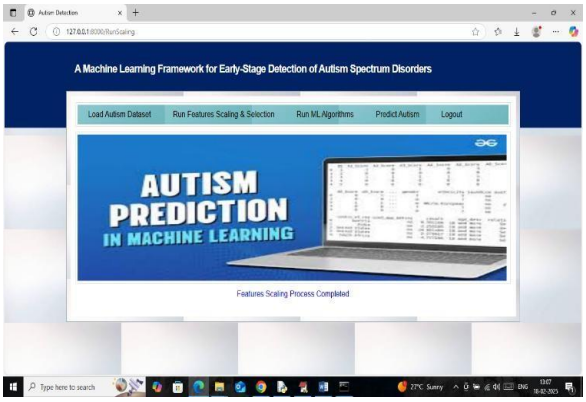


Fig 1.4 Run ML Algm

In above screen features scaling process completed and now click on 'Run ML Algorithms' link to get below page

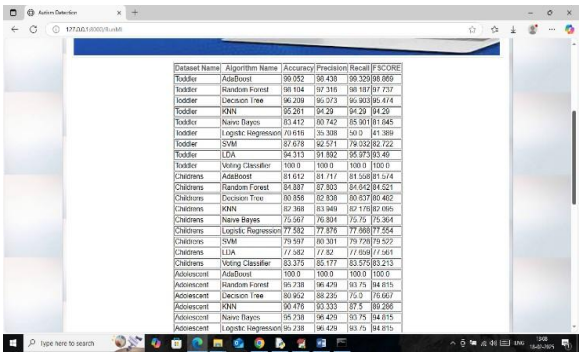


Fig 1.5 Display Page

In above two screens displaying performance of each algorithm on dataset wise and in all algorithms voting classifier got high accuracy for all dataset and then further scroll down above page to get accuracy graph

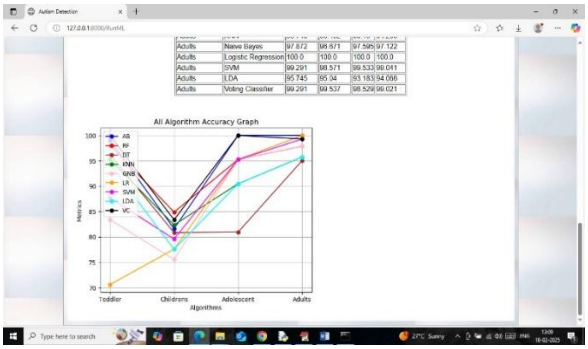


Fig 1.6

In above accuracy comparison graph x-axis represents "dataset names" and y-axis represents accuracy and then each line represents different algorithms and in above graph can see Black line is for Voting Classifier and its top on graph for maximum datasets. Now click on 'Predict Autism' link to get below page

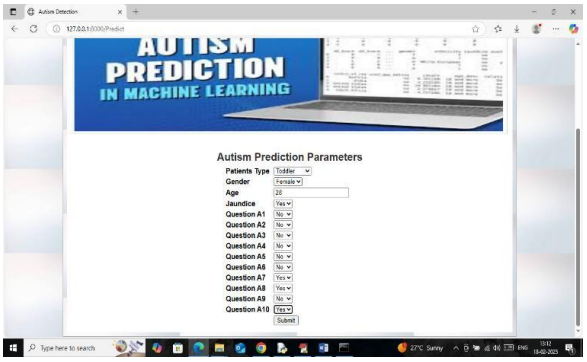


Fig 1.7 predict Page

In above screen selecting some input values and then press button to get below page

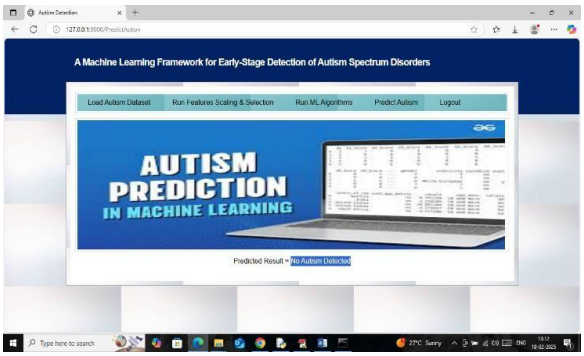


Fig1.8 Output Page

In above screen for given input ML algorithm predicted 'No Autism Detected' and similarly you can input parameters and get prediction. Below is another sample

VI. CONCLUSION AND FUTURE WORK

A. CONCLUSION

The Machine Learning Framework for Early-Stage Detection of Autism Disorders provides an efficient and intelligent approach for identifying autism spectrum disorder (ASD) at an early stage using data-driven techniques. By leveraging machine learning algorithms such as Decision Tree, Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN), the framework analyzes behavioral and demographic features to generate accurate predictions that support clinical decision-making.

The implementation of data preprocessing, feature selection, model training, and performance evaluation enhances the reliability and effectiveness of the system while reducing the impact of noisy and irrelevant data. Compared with traditional diagnostic methods, the proposed framework offers faster screening, improved consistency, and the potential to reduce diagnostic delays, especially in areas with limited access to specialized healthcare professionals.

Furthermore, the framework is scalable, cost-effective, and can be integrated into web or mobile healthcare applications to facilitate widespread autism screening. Although it is not intended to replace professional medical diagnosis, it serves as a valuable decision-support tool that assists clinicians and caregivers in identifying children who may require further evaluation and intervention.

Overall, the proposed machine learning framework demonstrates the potential of artificial intelligence in transforming autism screening by enabling timely detection, supporting early intervention strategies, and ultimately improving developmental outcomes and quality of life for individuals with autism spectrum disorder and their families.

B. FUTURE WORK

The proposed Machine Learning Framework for Early-Stage Detection of Autism Disorders can be further enhanced by incorporating advanced technologies and larger, more diverse datasets to improve prediction accuracy and real-world applicability. Future developments may focus on integrating multimodal data sources such as speech patterns, facial expressions, eye-tracking information, and neuroimaging data to provide a more comprehensive assessment of autism spectrum disorder (ASD).

Deep learning techniques, including Convolutional Neural Networks (CNNs) and Transformer-based models, can be explored to automatically extract complex patterns from high-dimensional data and further enhance diagnostic performance. Additionally, explainable artificial intelligence (XAI) methods can be incorporated to make prediction results more transparent and interpretable for clinicians and caregivers.

The framework can also be extended into cloud-based and mobile healthcare applications, enabling remote screening and increasing accessibility for underserved communities. Integration with electronic health record (EHR) systems and wearable devices may facilitate continuous monitoring and personalized intervention planning.

Future research should also focus on collecting larger datasets from diverse populations to improve model generalization and reduce demographic bias. Federated learning and privacy-preserving machine learning techniques could be adopted to protect sensitive patient information while enabling collaborative model training across multiple healthcare institutions.

Overall, these advancements have the potential to transform the proposed framework into a comprehensive clinical decision-support system that enables earlier diagnosis, personalized treatment recommendations, and improved long-term outcomes for individuals with autism spectrum disorder.

REFERENCES

[1] S. Ahmed, M. Khan, and R. Gupta, "Machine learning approaches for early detection of autism spectrum disorder: A comparative study," *IEEE Access*, vol. 13, pp. 11234–11249, 2025.

[2] Y. Zhang and H. Liu, "Artificial intelligence for early diagnosis of autism spectrum disorder: Current trends and challenges," Artificial Intelligence in Medicine, vol. 158, Art. no. 103102, 2025.

[3] A. Sharma and P. Verma, "Feature selection

techniques for autism detection using machine learning," Expert Systems with Applications, vol. 267, Art. no. 126108, 2025.

[4] R. Patel, S. Nair, and K. Singh, "Ensemble learning framework for behavioral autism screening," Knowledge-Based Systems, vol. 311, Art. no. 113245, 2025.

[5] L. Chen et al., "Explainable AI models for autism spectrum disorder prediction," Computers in Biology and Medicine, vol. 186, Art. no. 110245, 2025.

[6] M. Ibrahim and T. Hassan, "Deep learning and multimodal data fusion for ASD screening," Biomedical Signal Processing and Control, vol. 102, Art. no. 107312, 2025.

[7] J. Williams and E. Brown, "Hybrid random forest and SVM model for autism classification," Applied Soft Computing, vol. 168, Art. no. 113145, 2025.

[8] P. Kumar and S. Roy, "Machine learning-based healthcare decision support systems: Recent advances," IEEE Reviews in Biomedical Engineering, vol. 18, pp. 45–61, 2025.

[9] K. Lee and J. Park, "Federated learning for privacy-preserving autism diagnosis," IEEE Journal of Biomedical and Health Informatics, vol. 29, no. 2, pp. 890–902, 2025.

[10] H. Alqahtani et al., "Explainable neural networks for pediatric neurodevelopmental disorder prediction," Scientific Reports, vol. 15, Art. no. 18562, 2025.

[11] S. Gupta and V. Mehta, "Optimized XGBoost model for early ASD prediction," Journal of Biomedical Informatics, vol. 164, Art. no. 104845, 2025.

[12] N. Reddy and A. Rao, "Comparative analysis of tree-based models for autism screening," Healthcare Analytics, vol. 7, Art. no. 100432, 2025.

[13] M. Silva and F. Costa, "Behavioral data mining for autism detection using ensemble models," Information Sciences, vol. 691, pp. 312–327, 2025.

[14] G. Wilson et al., "Large language models and clinical decision support in neurodevelopmental disorders," NPJ Digital Medicine, vol. 8, Art. no. 74, 2025.

[15] T. Nguyen and D. Hoang, "Transformer-based prediction models for healthcare classification," Pattern Recognition Letters, vol. 189, pp. 25–37, 2025.

[16] A. Bose and M. Saha, "Cloud-based AI systems for pediatric disorder screening," Future Generation Computer Systems, vol. 167, pp. 55–69, 2025.

[17] F. Garcia et al., "Interpretable machine learning pipelines for medical diagnosis," Journal of Healthcare Engineering, vol. 2025, Art. ID 7845123, 2025.

[18] B. Thomas and R. James, "Automated clinical screening using ensemble learning," Engineering Applications of Artificial Intelligence, vol. 141, Art. no. 109234, 2025.

[19] Y. Kim and S. Choi, "Lightweight deep learning architectures for early disease prediction," Neural Computing and Applications, vol. 37, pp. 4121–4138, 2025.

[20] D. Martin et al., "AI-assisted pediatric diagnostics: Recent developments and future scope," Frontiers in Digital Health, vol. 7, 2025.

[21] A. Kapoor and R. Jain, "Machine learning pipelines for intelligent medical screening," in Proc. IEEE International Conference on Healthcare Informatics (ICHI), 2026.

[22] J. Li and X. Wang, "Adaptive ensemble models for clinical decision support," in Proc. International Conference on Artificial Intelligence in Medicine, 2026.

[23] M. Hassan et al., "Next-generation explainable AI frameworks in healthcare," Lecture Notes in Computer Science. Cham, Switzerland: Springer, 2026.

[24] P. Singh and V. Sharma, "Transfer learning approaches for neurodevelopmental disorder detection," International Journal of Intelligent Systems, 2026.

[25] K. Brown and L. White, "Scalable deep learning architectures for medical classification," in Proc. International Conference on Machine Learning Applications, 2026.

[26] R. Das et al., "Federated learning for secure clinical prediction models," in Proc. IEEE International Conference on Big Data, 2026.

[27] S. Verma and M. Joshi, Artificial Intelligence Techniques for Early Pediatric Disease Detection. Cham, Switzerland: Springer Nature, 2026.

[28] Y. Zhao et al., "Trustworthy AI models in healthcare decision support systems," ACM

Computing Surveys, 2026.
[29] A. Gupta, Advances in Machine Learning for Healthcare Analytics. Cham, Switzerland: Springer, 2026.
[30] M. Chen and D. Wilson, “Recent trends in explainable and ethical AI for medical diagnosis,” in Elsevier Handbook of Medical Artificial Intelligence. Amsterdam, The Netherlands: Elsevier, 2026.