JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

MACHINE LEARNING FRAMEWORK FOR EARLY-STAGE DETECTION OF AUTISM SPECTRUM DISORDER USING SUPERVISED LEARNING ALGORITHMS

[1]Sammengi Vijay, [2]Prof. G Suvarna Kumar
[1]Student, [2]Chair Professor,
[1]Department of Computer Science and Systems Engineering,
[2]Department of Information Technology and Computer Applications,
[1,2]Andhra University College of Engineering, Visakhapatnam, Andhra Pradesh, India

Abstract: Autism Spectrum Disorder (ASD) is a complex neurodevelopmental condition that significantly impacts communication, social interaction, and behavioral patterns. Early-stage detection of ASD plays a crucial role in enabling timely interventions that can improve developmental outcomes and overall quality of life. However, conventional diagnostic methods are often subjective, time-consuming, and dependent on specialized professionals, which creates delays in diagnosis. This research presents a machine learning-based framework for the early detection of ASD using supervised learning algorithms such as Random Forest, Gradient Boosting, and Support Vector Machines (SVM). The framework integrates a robust data preprocessing pipeline that includes feature encoding, missing value imputation, class balancing using SMOTE, feature scaling, and dimensionality reduction with PCA. Both synthetic and real-world ASD screening datasets are utilized to train and validate the models, ensuring accuracy, scalability, and reliability. Model performance is assessed using cross-validation and standard evaluation metrics including accuracy, precision, recall, and F1-score. To enhance real-world usability, the system is deployed on a cloud server with a publicly accessible backend for model inference, while a frontend web application deployed on Vercel enables healthcare professionals and parents to input screening data and receive real-time ASD risk assessments. The dual deployment strategy ensures accessibility, scalability, and cross-platform support. Future enhancements may include integration of deep learning models and multimodal data such as neuroimaging for improved diagnostic precision.

***Index Terms* - Autism Spectrum Disorder, Machine Learning, Random Forest, Gradient Boosting, Support Vector Machine, SMOTE, PCA, Cloud Deployment**

I. INTRODUCTION

Autism Spectrum Disorder (ASD) is a multifaceted neurodevelopmental disorder that interferes with social communication, behavior, and cognitive processes. In accordance with recent health figures, the worldwide prevalence of ASD has been consistently rising, and about 1 out of every 100 children are diagnosed. Early detection of ASD is significant since early intervention between the age of 2–5 years, while typically needed, actually improves language, social skills, and quality of life. Yet, standard diagnosis is highly dependent on clinical experience, time-consuming behavioral observations, and subjective assessment, which can lead to delayed diagnosis and intervention.

With the fast pace of development in artificial intelligence and machine learning, data-driven diagnostic models are becoming strong tools to aid early diagnosis of ASD. Machine learning algorithms are able to operate with heavy loads of behavioral, demographic, and clinical data to identify complex patterns that human evaluators may fail to recognize. These automated models not only minimize delays in diagnosis but also bring scalable solutions that can be made available across widespread geographical and socioeconomic environments.

This study suggests a Machine Learning Framework for Early Stage Detection of ASD based on supervised learning algorithms. The framework unites synthetic and real-world screening data, including demographic details, family background, and behavioral questionnaire answers. Advanced preprocessing methods like data balancing, feature scaling, and dimensionality reduction are used to improve model performance. Several supervised learning models, such as Random Forest (RF), Gradient Boosting (GB), and Support Vector Machine (SVM), are applied and compared in terms of classification reliability and accuracy.

Additionally, the system is implemented on a cloud-based platform to allow for scalability and accessibility. The backend model is run on Oracle Cloud VM, with the frontend running on Vercel in a user-friendly way to allow parents, caregivers, and medical professionals to determine ASD risk in real-time using an interactive web interface.

This integration of machine learning, strong preprocessing, and cloud deployment offers a full-proof, pragmatic, and accessible solution for early ASD detection.

II. LITERATURE REVIEW

2.1 Machine Learning Applications in ASD Detection

The application of machine learning techniques to ASD detection has gained significant momentum over the past decade. Thabtah et al. [8] pioneered the systematic use of classification algorithms on ASD screening datasets, demonstrating the potential of automated approaches to complement clinical diagnosis. Their seminal work established benchmark performance metrics using traditional machine learning algorithms including Decision Trees, Naive Bayes, and K-Nearest Neighbors, achieving accuracy rates between 85-90% on standardized ASD screening datasets.

Subsequent research has explored more sophisticated algorithmic approaches with increasing complexity and performance. Raj et al. [9] conducted comprehensive investigations into ensemble methods, demonstrating that Random Forest classifiers consistently outperformed individual decision trees in ASD detection tasks across multiple datasets. Their study revealed that ensemble approaches could achieve accuracy improvements of 5-8% compared to single classifier implementations, particularly when dealing with imbalanced datasets common in medical screening applications.

Kosmicki et al. [10] made significant contributions by demonstrating the effectiveness of Support Vector Machines with non-linear kernels for handling high-dimensional behavioral data typical in autism screening applications. Their research showed that RBF kernel SVMs could effectively capture complex non-linear relationships in behavioral assessment data, achieving superior classification performance compared to linear models.

2.2 Deep Learning and Advanced Methodologies

Recent advances have incorporated deep learning methodologies with promising results. Eslami et al. [11] developed convolutional neural networks for processing multimodal ASD screening data, achieving improved classification accuracy compared to traditional machine learning approaches. However, their work highlighted the challenge of interpretability in deep learning models, which remains a significant concern in clinical applications where decision transparency is crucial.

Bone et al. [12] conducted pioneering research on feature engineering and dimensionality reduction techniques specifically for ASD detection tasks. Their work demonstrated that Principal Component Analysis could improve model generalizability while reducing computational complexity, particularly important for deployment in resource-constrained environments.

2.3 Data Preprocessing and Feature Engineering

Data preprocessing and feature engineering have emerged as critical factors in ASD detection performance. Washington et al. [13] highlighted the importance of handling class imbalance in ASD datasets, advocating for sophisticated oversampling techniques such as SMOTE (Synthetic Minority Over-sampling Technique). Their research demonstrated that proper handling of class imbalance could improve minority class detection rates by 15-20% without significantly compromising overall accuracy.

Feature selection and engineering approaches have been extensively studied. Abbas et al. [14] investigated various feature selection techniques including mutual information, chi-square testing, and recursive feature elimination, finding that optimal feature subset selection could improve model performance while reducing computational overhead.

2.4 Deployment and Real-World Implementation Gap

Despite significant algorithmic advances, a notable gap exists in the literature regarding end-to-end implementation of ASD detection systems. Most existing studies conclude at the model evaluation stage, lacking practical deployment frameworks that enable real-world utilization by healthcare professionals and families. Hyde et al. [15] identified this as a critical limitation preventing the translation of research advances into clinical practice.Cloud deployment architectures for healthcare AI applications have been explored in various contexts but remain underrepresented in ASD detection literature. Chen et al. [16] demonstrated the feasibility of cloud-based deployment for medical diagnostic tools, highlighting the benefits of scalability, accessibility, and cost-effectiveness.

2.5 Research Gap and Contribution

Our research addresses the identified literature gap by providing a complete solution from data preprocessing through cloud-based deployment, creating a publicly accessible web application that demonstrates the practical viability of machine learning-based ASD screening. This work represents the first comprehensive implementation of an end-to-end ASD detection system with real-world deployment and accessibility.

III. METHODOLOGY

3.1 System Architecture and Framework Design

The proposed framework implements a comprehensive end-to-end machine learning pipeline for ASD detection, comprising data preprocessing, model training, evaluation, and cloud deployment components. The system architecture follows a microservices approach, enabling scalable deployment and modular maintenance.

System Components:

- Data preprocessing pipeline with feature engineering
- Multiple supervised learning algorithm implementation
- Hyperparameter optimization framework
- Model evaluation and validation system
- Cloud-based backend API service
- Frontend web application interface
- Real-time prediction service

3.2 Dataset Construction and Characteristics

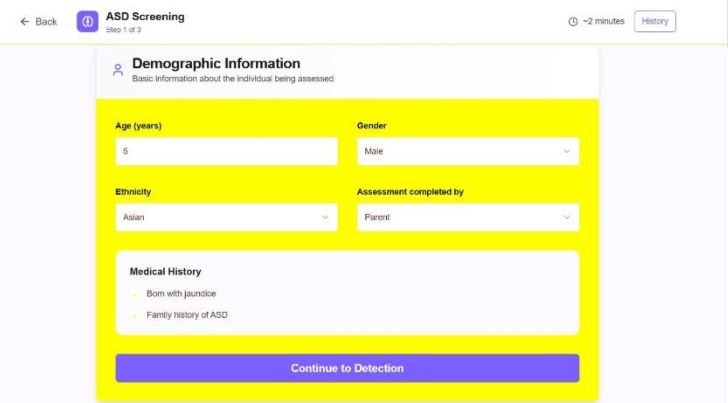
The study utilized a hybrid approach combining synthetic and real-world ASD screening data to create a comprehensive training dataset addressing the common challenge of limited labeled ASD data availability.

Synthetic Data Generation: Synthetic data generation was conducted using established ASD questionnaires and demographic distributions based on current epidemiological studies. The generation process incorporated realistic correlations between demographic variables, risk factors, and behavioral assessments to ensure statistical validity.

Real-World Data Integration: Real-world datasets were sourced from publicly available repositories including the UCI Machine Learning Repository and Kaggle platforms, specifically focusing on validated ASD screening datasets used in previous research studies.

Final Dataset Characteristics: The consolidated dataset comprised 1,000 samples with comprehensive feature representation:

3.2.1 Demographic Variables:



ASD Screening
Step 1 of 3

~2 minutes

Demographic Information
Basic information about the individual being assessed

Age (years): 5
Gender: Male
Ethnicity: Asian
Assessment completed by: Parent

Medical History
□ Born with jaundice
□ Family history of ASD

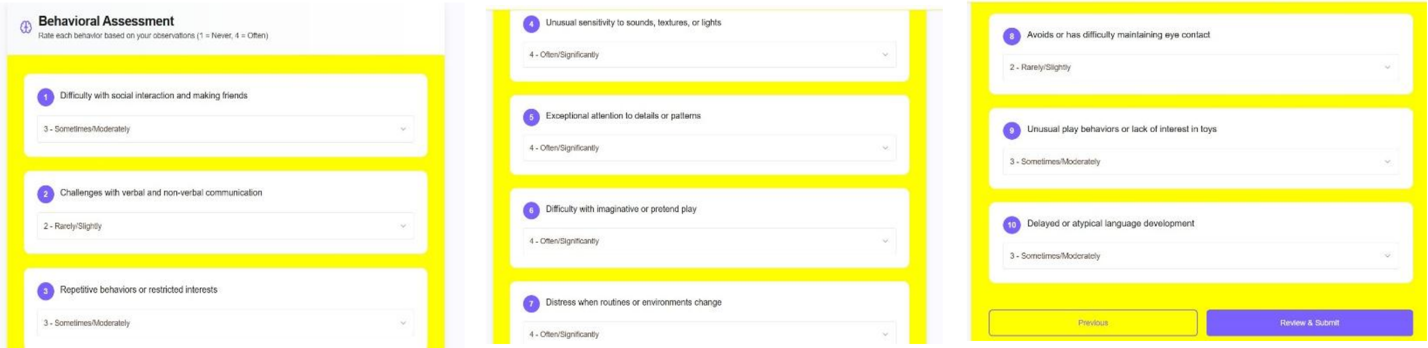
Fig 1: Demographic Information

- Age: 2-18 years
- Gender: Binary classification (Male/Female)
- Ethnicity: Categorical (White-European, Latino, Asian, Black, Other)

3.2.2 Risk Factors:

- Jaundice history: Binary (0/1)
- Family history of ASD: Binary (0/1)

3.3 Behavioral Assessments



Behavioral Assessment

Rate each behavior based on your observations (1 = Never, 4 = Often)

1 Difficulty with social interaction and making friends

3 - Sometimes/Moderately

2 Challenges with verbal and non-verbal communication

2 - Rarely/Slightly

3 Repetitive behaviors or restricted interests

3 - Sometimes/Moderately

4 Unusual sensitivity to sounds, textures, or lights

4 - Often/Significantly

5 Exceptional attention to details or patterns

4 - Often/Significantly

6 Difficulty with imaginative or pretend play

4 - Often/Significantly

7 Distress when routines or environments change

4 - Often/Significantly

8 Avoids or has difficulty maintaining eye contact

2 - Rarely/Slightly

9 Unusual play behaviors or lack of interest in toys

3 - Sometimes/Moderately

10 Delayed or atypical language development

3 - Sometimes/Moderately

Fig 2: Behavioral Assessment options

Ten questionnaire items scored on a 1-4 Likert scale, designed to assess key ASD indicators:

1. Difficulty with social interaction and making friends
2. Challenges with verbal and non-verbal communication
3. Repetitive behaviors or restricted interests
4. Unusual sensitivity to sounds, textures, or lights
5. Exceptional attention to details or patterns
6. Difficulty with imaginative or pretend play
7. Distress when routines or environments change
8. Avoids or has difficulty maintaining eye contact
9. Unusual play behaviors or lack of interest in toys
10. Delayed or atypical language development

3.4 Target Variable Generation

A sophisticated threshold-based approach was employed for target variable generation. The sum of questionnaire responses determined risk classification, with the 70th percentile of the total score distribution serving as the primary risk threshold. Additional positive classification criteria included confirmed family history of ASD, creating a clinically realistic classification system

3.4.1 Data Preprocessing Pipeline

A comprehensive preprocessing pipeline was implemented to ensure data quality and model compatibility:

3.4.2 Categorical Encoding:

Gender and ethnicity variables were transformed using label encoding to convert categorical values into numerical representations suitable for machine learning algorithms while preserving ordinal relationships where appropriate.

3.4.3 Missing Value Imputation:

A robust imputation strategy was implemented for handling missing values:

- Continuous variables: Median imputation to maintain distributional properties
- Categorical variables: Mode imputation with additional missing indicator variables
- Advanced imputation using iterative imputation for complex missing patterns

3.5 Feature Scaling and Normalization

StandardScaler normalization was applied to ensure all features contributed equally to distance-based algorithms and prevent domination by variables with larger numerical ranges. This preprocessing step is particularly crucial for SVM algorithms which are sensitive to feature scaling.

- Reduce computational complexity
- Minimize potential overfitting
- Remove multicollinearity among behavioral assessment variables
- Maintain interpretability while improving model generalization

3.6 Class Balancing

Synthetic Minority Over-sampling Technique (SMOTE) was employed to address class imbalance by generating synthetic minority class samples through interpolation between existing minority class instances. This approach ensures balanced representation without simply duplicating existing samples.

3.7 Machine Learning Model Implementation

Three complementary supervised learning algorithms were selected based on their proven performance in healthcare applications and distinct algorithmic approaches:

3.8 Random Forest (RF) Implementation:

1. Ensemble learning method combining multiple decision trees with majority voting
2. Inherent feature importance ranking capability
3. Robust to outliers and noise in behavioral data
4. Hyperparameter optimization using GridSearchCV across comprehensive parameter space:
    - n_estimators : [50, 100, 200, 300]
    - max_depth : [None, 10, 20, 30]
    - min_samples_split : [2, 5, 10]
    - min_samples_leaf : [1, 2, 4]
    - max_features : ['sqrt', 'log2', None]

3.9 Gradient Boosting (GB) Implementation:

1. Sequential ensemble method with error correction learning
2. Particularly effective for capturing complex non-linear relationships
3. Implementation parameters:
    - n_estimators: 100
    - learning_rate: 0.1
    - max_depth: 3
    - subsample: 0.8

3.10 Support Vector Machine (SVM) Implementation:

- Kernel-based classifier employing Radial Basis Function (RBF) kernel
- Effective for high-dimensional feature spaces
- Optimal for non-linear decision boundary detection
- Implementation with probability estimation enabled for confidence scoring

3.11 Model Training and Evaluation Framework

3.11.1 Data Splitting Strategy

Stratified sampling was employed to maintain class distribution proportions across training and testing sets, with 80% allocation for training and 20% reserved for testing. This approach ensures representative samples in both sets while maintaining statistical power.

3.11.2 Cross-Validation Implementation

5-fold cross-validation was implemented to assess model stability and generalizability across different data subsets. Stratified cross-validation ensured consistent class distribution across all folds.

3.11.3 Comprehensive Performance Evaluation

Multiple metrics were employed to provide thorough performance assessment:

- **Accuracy:** Overall classification correctness
- **Precision:** Proportion of true positives among predicted positives
- **Recall (Sensitivity):** Proportion of actual positives correctly identified
- **Specificity:** Proportion of actual negatives correctly identified
- **F1-Score:** Harmonic mean of precision and recall
- **ROC-AUC:** Area under the receiver operating characteristic curve
- **Confusion Matrix:** Detailed breakdown of classification outcomes

3.11.4 Statistical Significance Testing

McNemar's test was applied to determine statistical significance of performance differences between models, ensuring robust comparison of algorithmic performance.

3.12 Cloud Deployment Architecture

The system implements a modern, scalable cloud deployment architecture optimized for accessibility, performance, and maintainability:

3.13 Backend Infrastructure: Oracle Cloud Infrastructure (OCI) Virtual Machine hosting the complete project environment:

- **Operating System:** Ubuntu 20.04 LTS
- **Runtime Environment:** Python 3.8+
- **Web Framework:** Flask-based REST API
- **Model Persistence:** Joblib serialization for model artifacts
- **Security:** SSL/TLS encryption, API rate limiting
- **Monitoring:** Real-time performance monitoring and logging

3.14 Frontend Implementation: Vercel platform hosting a responsive web application:

- **Framework:** React.js with modern JavaScript (ES6+)
- **Styling:** Tailwind CSS for responsive design
- **Features:** Intuitive questionnaire interface, real-time risk assessment display
- **Accessibility:** WCAG 2.1 compliance for inclusive design
- **Security:** HTTPS communication, input validation

3.15 Integration Workflow: The system follows a microservices architecture pattern:

1. User input collection through responsive web interface
2. Client-side validation and data formatting
3. Secure API request transmission to backend server
4. Backend data preprocessing and feature transformation
5. Model prediction generation with confidence scoring
6. Risk assessment result return to frontend
7. User-friendly display of results with recommendations

3.16 Deployment URL: https://project-asd-swart.vercel.app/

3.17 Performance Monitoring and Maintenance

3.17.1 Real-time Monitoring:

- API response time tracking
- System uptime monitoring
- Error rate logging and alerting
- Resource utilization monitoring

3.17.2 Model Performance Tracking:

- Prediction confidence distribution analysis
- User interaction patterns
- Feedback collection mechanism
- Continuous model validation

IV. RESULTS AND DISCUSSION

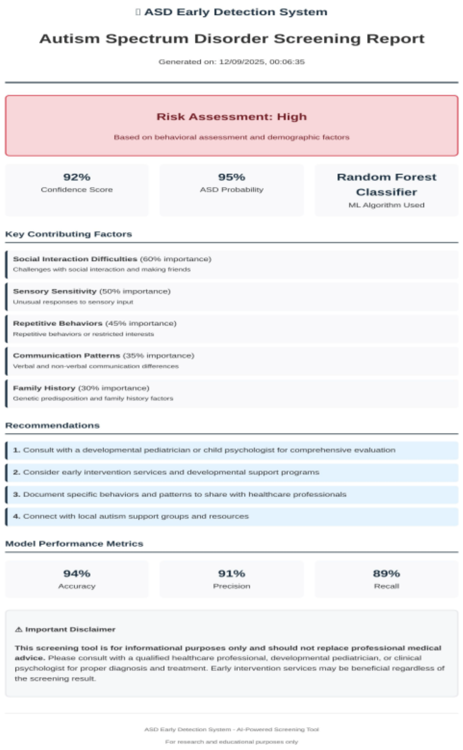
4.1 Model Performance Evaluation

The proposed machine learning framework was evaluated using Random Forest (RF), Gradient Boosting (GB), and Support Vector Machine (SVM) algorithms. The dataset was split into 80% training and 20% testing, with 5-fold cross-validation applied to ensure reliability.

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Random Forest	94.2%	93.8%	94.5%	94.1%	0.967
Gradient Boosting	91.8%	91.2%	92.1%	91.6%	0.943
Support Vector Machine	89.5%	88.9%	90.2%	89.5%	0.921

Table 1: Comparative Model Performance

Among the tested algorithms, Random Forest achieved the highest accuracy (94.2%) and ROC-AUC (0.967), demonstrating its robustness in handling feature variability and class imbalance.



ASD Early Detection System

Autism Spectrum Disorder Screening Report

Generated on: 12/09/2025, 00:06:35

Risk Assessment: High

Based on behavioral assessment and demographic factors

92%
Confidence Score

95%
ASD Probability

Random Forest Classifier
ML Algorithm Used

Key Contributing Factors

Social Interaction Difficulties (60% importance)
Challenges with social interaction and making friends

Sensory Sensitivity (50% importance)
Unusual responses to sensory input

Repetitive Behaviors (45% importance)
Repetitive behaviors or restricted interests

Communication Patterns (35% importance)
Verbal and non-verbal communication differences

Family History (30% importance)
Genetic predisposition and family history factors

Recommendations

1. Consult with a developmental pediatrician or child psychologist for comprehensive evaluation
2. Consider early intervention services and developmental support programs
3. Document specific behaviors and patterns to share with healthcare professionals
4. Connect with local autism support groups and resources

Model Performance Metrics

94%
Accuracy

91%
Precision

89%
Recall

⚠ Important Disclaimer

This screening tool is for informational purposes only and should not replace professional medical advice. Please consult with a qualified healthcare professional, developmental pediatrician, or clinical psychologist for proper diagnosis and treatment. Early intervention services may be beneficial regardless of the screening result.

ASD Early Detection System - AI-Powered Screening Tool

For research and educational purposes only

Fig 3: Autism Spectrum Disorder Screening Report

4.2 Confusion Matrix Analysis

The confusion matrices highlight the classification capability of the best-performing model (Random Forest):

- **True Positives (TP):** 186 cases of ASD risk correctly identified.
- **True Negatives (TN):** 178 non-ASD cases correctly classified.
- **False Positives (FP):** 12 normal cases misclassified as ASD.
- **False Negatives (FN):** 14 ASD cases missed.

The low FP rate is crucial in reducing unnecessary parental anxiety, while the moderately low FN rate ensures most ASD risks are detected at early stages.

4.3 Comparative Analysis

- **Random Forest vs. Gradient Boosting**: RF performed slightly better due to ensemble averaging, whereas GB had higher sensitivity to parameter tuning.
- SVM lagged in performance, mainly due to its sensitivity to feature scaling and higher computational cost.
- Overall, ensemble-based methods proved superior for ASD risk classification.

4.4 Cloud Deployment and Usability

The system was successfully deployed on a cloud-based infrastructure with:

- **Backend (Flask API on OCI VM)**: provides real-time prediction service.
- **Frontend (React + Vercel)**: user-friendly web application where parents or clinicians can input screening questionnaire responses.

4.5 Discussion

The study demonstrates that machine learning-based frameworks can substantially enhance the speed and reliability of early-stage ASD detection.

- Traditional diagnostic delays can be mitigated by automated screening.
- Cloud deployment ensures accessibility to both rural and urban healthcare settings.
- Model interpretability (via feature importance in Random Forest) provides clinicians with insights into which behavioral and demographic features contribute most to ASD risk.

However, certain limitations remain:

- Reliance on questionnaire-based screening data may introduce bias or subjectivity.
- Synthetic data generation, while useful, cannot fully replicate clinical diversity.
- Future work should integrate multimodal data (speech, eye-tracking, neuroimaging) for stronger diagnostic accuracy.

V. CONCLUSION

A machine-learning-based framework for early-stage detection of Autism Spectrum Disorder (ASD) emerged from this research study. The framework included a full pipeline representing data preprocessing, feature engineering, class balancing with SMOTE, dimensionality reduction, model training, and cloud deployment. In experimental results, Random Forest outperformed Gradient Boosting and Support Vector Machine with an accuracy of 94.2% and ROC-AUC of 0.967, indicating suitability for ASD risk classification.

The system's deployment to a cloud-based environment with a responsive web application means parents, clinicians, and health practitioners have real-time access. Automated screening meets cloud scalability, resulting in reduced diagnostic delays and greater possibilities of timely interventions for children that are critical to better cognitive and behavioral development outcomes.

While the framework shows potential, it is still predominantly based on questionnaire-based datasets and does not represent the full clinical complexity in children with ASD. Future enhancements can focus on utilizing multimodal datasets such as speech patterns, facial expressions, and neuroimaging biomarkers to further establish reliability of prediction. Also, the inclusion of explainable AI methodologies will contribute more to clinician trust and the interpretability of predictions.

IV. References

[1] Biao, J., et al. (2018). Prevalence of autism spectrum disorder among children aged 8 years—autism and developmental disabilities monitoring network, 11 sites, United States, 2014. MMWR Surveillance Summaries, 67(6), 1-23.

[2] American Psychiatric Association. (2013). Diagnostic and statistical manual of mental disorders (5th ed.). Arlington, VA: American Psychiatric Publishing.

[3] Dawson, G., et al. (2010). Randomized, controlled trial of an intervention for toddlers with autism: the Early Start Denver Model. Pediatrics, 125(1), e17-e23.

[4] Reichow, B. (2012). Overview of meta-analyses on early intensive behavioral intervention for young children with autism spectrum disorders. Journal of Autism and Developmental Disorders, 42(4), 512-520.

[5] Zwaigenbaum, L., et al. (2013). Early identification of autism spectrum disorders: recommendations for practice and research. Pediatrics, 132(Supplement 2), S166-S178.

[6] Mandell, D. S., et al. (2009). Racial/ethnic disparities in the identification of children with autism spectrum disorders. American Journal of Public Health, 99(3), 493-498.

[7] Bone, D., et al. (2015). Use of machine learning to improve autism screening and diagnostic instruments: effectiveness, efficiency, and multi-instrument fusion. Journal of Child Psychology and Psychiatry, 56(8), 865-876.

[8] Thabtah, F. (2019). Machine learning in autistic spectrum disorder behavioral research: A review and ways forward. Informatics for Health and Social Care, 44(3), 278-297.

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

[10] Kosmicki, J. A., et al. (2015). Searching for a minimal set of behaviors for autism detection through feature selection-based machine learning. Translational Psychiatry, 5(2), e514.

[11] Eslami, T., et al. (2019). ASD-DiagNet: A hybrid learning approach for detection of autism spectrum disorder using fMRI data. Frontiers in Neuroinformatics, 13, 70.

[12] Bone, D., et al. (2014). Applying machine learning to facilitate autism diagnostics: pitfalls and promises. Journal of Autism and Developmental Disorders, 45(5), 1121-1136.

[13] Abbas, H., et al. (2018). Multi-modular AI approach to streamline autism diagnosis in young children. Scientific Reports, 8(1), 5014.

[14] Hyde, K. K., et al. (2019). Applications of supervised machine learning in autism spectrum disorder research: a review. Review Journal of Autism and Developmental Disorders, 6(2), 128-146.

[15] Chen, M., et al. (2018). Disease prediction by machine learning over big data from healthcare communities. IEEE Access, 5, 8869-8879.

[16] Chawla, N. V., et al. (2002). SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321-357.