

RA54 GATIGANTI VENKAT PAVAN SAI

Evaluating_diverse_Machine_Learning_Techniques_For_Auti...

 Biswas-1 RAEEUCCI2025 SRM Institute of Science & Technology

Document Details

Submission ID

trn:oid::1:3127023065

Submission Date

Jan 10, 2025, 4:26 PM GMT+5:30

Download Date

Jan 10, 2025, 4:28 PM GMT+5:30

File Name

Evaluating_diverse_Machine_Learning_Techniques_For_Autism_Spectrum_Disorder_Detection.docx

File Size





577.3 KB

9 Pages**4,660 Words****28,126 Characters**




29% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Match Groups

-  **93 Not Cited or Quoted 24%**
Matches with neither in-text citation nor quotation marks
-  **7 Missing Quotations 2%**
Matches that are still very similar to source material
-  **11 Missing Citation 3%**
Matches that have quotation marks, but no in-text citation
-  **0 Cited and Quoted 0%**
Matches with in-text citation present, but no quotation marks

Top Sources

- 22%  Internet sources
- 24%  Publications
- 16%  Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Match Groups

- 93 Not Cited or Quoted 24%**
Matches with neither in-text citation nor quotation marks
- 7 Missing Quotations 2%**
Matches that are still very similar to source material
- 11 Missing Citation 3%**
Matches that have quotation marks, but no in-text citation
- 0 Cited and Quoted 0%**
Matches with in-text citation present, but no quotation marks

Top Sources

- 22% Internet sources
- 24% Publications
- 16% Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Internet	www.mdpi.com	2%
2	Student papers	University of Hull	1%
3	Internet	etd.repository.ugm.ac.id	1%
4	Internet	www.researchgate.net	1%
5	Publication	H.L. Gururaj, Francesco Flammini, S. Srividhya, M.L. Chayadevi, Sheba Selvam. "Co...	1%
6	Internet	www.ijitee.org	1%
7	Internet	ouci.dntb.gov.ua	1%
8	Internet	iao.hfuu.edu.cn	1%
9	Student papers	University of Hertfordshire	1%
10	Publication	V. Sharmila, S. Kannadhasan, A. Rajiv Kannan, P. Sivakumar, V. Vennila. "Challeng...	1%

11	Internet	techscience.com	1%
12	Student papers	Liverpool John Moores University	1%
13	Internet	eprints.whiterose.ac.uk	1%
14	Publication	Shyam V. Aradhya, Ved S. Bilaskar, Snehal S. Shinde, Deepak D. Kshirsagar, Pushp...	0%
15	Internet	ebin.pub	0%
16	Student papers	University of Derby	0%
17	Student papers	SRM University	0%
18	Internet	link.springer.com	0%
19	Student papers	Griffith College Dublin	0%
20	Publication	Na Zhang, Mindi Ruan, Shuo Wang, Lynn Paul, Xin Li. "Discriminative Few Shot Le...	0%
21	Publication	Novia Martin, Johanes Cornelius Mose, Shelly Iskandar. "THE ROLE OF SLEEP IN IN...	0%
22	Internet	journal.uad.ac.id	0%
23	Internet	klu.ai	0%
24	Student papers	RMIT University	0%

25	Student papers	Southern New Hampshire University - Continuing Education	0%
26	Internet	www.nature.com	0%
27	Student papers	Plainfield South High School	0%
28	Student papers	Sungkyunkwan University	0%
29	Publication	Arvind Dagur, Karan Singh, Pawan Singh Mehra, Dharendra Kumar Shukla. "Artific...	0%
30	Publication	Mustafa Mohammed Jassim, Manar Hassan Ali, Amer S. Elamer, Mustafa Musa Ja...	0%
31	Student papers	UNITEC Institute of Technology	0%
32	Internet	ijssst.info	0%
33	Student papers	Long Island University	0%
34	Internet	doctorpenguin.com	0%
35	Internet	jjcit.org	0%
36	Publication	"Hybrid Intelligent Systems", Springer Science and Business Media LLC, 2023	0%
37	Student papers	Misr International University	0%
38	Publication	S. Yuvaraj, G Yokesh, S Lalith Kishore, D Manoj Kumar. "Vehicle Information Stora...	0%

39	Publication	Thompson Stephan. "Artificial Intelligence in Medicine", CRC Press, 2024	0%
40	Internet	ieeexplore.ieee.org	0%
41	Internet	www.ieee-jas.net	0%
42	Internet	5wwwwww.easychair.org	0%
43	Publication	Keyu Pan, Wei-Ping Zhu, Mojtaba Hasannezhad. "Self-attention CNN based indoor..."	0%
44	Publication	Pushpa B, R. Gomathi, C.N. Harshavardhana, P.Siva Kota Reddy, Davinder Kumar, ...	0%
45	Internet	mariuslarnoy.github.io	0%
46	Internet	upcommons.upc.edu	0%
47	Internet	www.annsaudimed.net	0%
48	Internet	www.ijana.in	0%
49	Student papers	Auburn University College of Engineering	0%
50	Internet	cdn.techscience.cn	0%
51	Internet	ejurnal.stmik-budidarma.ac.id	0%
52	Internet	jcreview.com	0%

53	Internet	journals.uob.edu.bh	0%
54	Publication	Amr E. Eldin Rashed, Waleed M. Bahgat, Ali Ahmed, Tamer Ahmed Farrag, Ahmed ...	0%
55	Publication	Chaobo Zhang, Junyang Li, Yang Zhao, Tingting Li, Qi Chen, Xuejun Zhang, Weika...	0%
56	Internet	cerebralpalsysurgery.blogspot.com	0%
57	Internet	ceur-ws.org	0%
58	Internet	ejournal.uin-suska.ac.id	0%
59	Internet	penerbit.uthm.edu.my	0%
60	Internet	www.researchsquare.com	0%
61	Publication	Ramanjot, Dalwinder Singh, Manik Rakhra, Shruti Aggarwal. "Autism Spectrum Di...	0%
62	Publication	S.J. Xavier Savarimuthu, Sivakannan Subramani, Alex Noel Joseph Raj. "Artificial I...	0%
63	Publication	Stella Kehinde Ogunkan, David Victor Ogunkan. "Traffic Pattern Recognition Usin...	0%
64	Internet	easychair.org	0%
65	Internet	export.arxiv.org	0%
66	Internet	ijeecs.iaescore.com	0%

67	Internet	www.ahrq.gov	0%
68	Internet	www.health.state.mn.us	0%
69	Internet	www.ijraset.com	0%
70	Publication	Inam Ullah, Inam Ullah Khan, Mariya Ouaissa, Mariyam Ouaissa, Salma El Hajjami...	0%
71	Publication	Pawan Singh Mehra, Dharendra Kumar Shukla. "Artificial Intelligence, Blockchain,...	0%
72	Student papers	University of Nebraska at Omaha	0%
73	Internet	arxiv.org	0%
74	Internet	pmc.ncbi.nlm.nih.gov	0%
75	Internet	pure.udem.edu.mx	0%
76	Internet	www.igi-global.com	0%
77	Publication	Fadi Thabtah, Suhel Hammoud, Firuz Kamalov, Amanda Gonsalves. "Data imbalanc...	0%
78	Publication	Mona Nashaat, James Miller. "Refining software defect prediction through attenti...	0%
79	Publication	Mrutyunjaya Panda, Ajith Abraham, Biju Gopi, Reuel Ajith. "Computational Intelli...	0%

Evaluating Diverse Machine Learning Techniques For Autism Spectrum Disorder Detection: A Comparative Analysis

3827

Srinath R
Department of ECE
SRM Institute of Science and Technology
Chennai, India
sribalaji24@gmail.com

Siddarth Vinnakota
Department of ECE
SRM Institute of Science and Technology
Chennai, India
vs2379@srmist.edu.in

Ujjwal Pardeshi
Department of ECE
SRM Institute of Science and Technology
Chennai, India
up6276@srmist.edu.in

3217

Bhoovi Chauhan
Department of ECE
SRM Institute of Science and Technology
Chennai, India
bb8042@srmist.edu.in

Krittika Roy
Department of ECE
SRM Institute of Science and Technology
Chennai, India
skr8849@srmist.edu.in

Imraan Javeed Settu
Department of ECE
Veltech Rangarajan Dr.Sagunthala R &D institute of Science and Technology
Chennai, India
imraanjaveedsettu@gmail.com

57555671

Abstract— A major challenge for both young children and teenagers currently is Autism Spectrum Disorder (ASD), a neurological development disorder that hampers social interaction and causes several behavioral problems. Thus, early detection of ASD is necessary due to the unknown nature of its underlying cause. In this study, we have conducted a comparative analysis of early prediction of ASD using Machine Learning and Deep Learning techniques. Several machine learning classifiers, including Catboost, XGboost, Deep Neural Network, Stacked Ensemble Methods and Random forest have been employed in this research. Among all the classification models, the Catboost algorithm has given the maximum accuracy of 97.73%, XGboost has given 97.46%, Deep Neural network with 97.20%, Stacked Ensemble with 96.93% and Random Forest with 96.53% accuracy. The Implemented code has been made available on GitHub.

69

Keywords— Autism Spectrum Disorder, Machine Learning Catboost, XGboost, Deep Neural Network, Stacked Ensemble and Random forest

112568

I. INTRODUCTION

Autism spectrum disorder is a condition that impacts neurological and developmental functioning, affecting an individual's social interactions, communication skills, learning abilities, and behavioral patterns. While diagnosis can occur at any age, it is typically classified as a developmental disorder due to the emergence of symptoms within the initial two years of life. Early symptoms of ASD include a lack of eye contact and a lack of response to calling[1].

155

Based on the survey taken in the year 2021, globally there are 788 per 10,000 cases which is approximately 61.8 million autistic individuals are surviving which is equivalent to 1 autistic out of 127 normal people[2].

A. Common Symptoms of ASD:

Some of the Symptoms of autistic individuals are listed below. The symptoms differs from individual to individual

- 1
- Poor Eye Contact
 - Hand Flapping
 - No Name-calling Response
 - Tip Toe Walking

- Poor social Communication
- Distinct reactions to Sound, Light etc.,
- Echolalia

B. Therapeutic measures:

Some of the treatment procedures for the treatment of ASD are listed below:

- Occupational Therapy
- Speech and Language Therapy
- Sensory Integration Therapy
- Behavioral (ABA Therapy)
- Pharmacological Therapy
- Aquatic Therapy or Hydro Therapy

The Importance of identifying the Autism spectrum disorder is made in the early stage with the help of an automated approach. This makes the treatment process easier and through various therapies, drastic changes in the behavior of the individual can be achieved in the toddler stage.

Traditional methods of detection like Autism Diagnostic Interview-Revised (ADI-R) and Autism Diagnostic Observation Schedule Revised (ADOS-R) are often time-consuming and not tailored to each patient's symptoms [3]. These methods are highly susceptible to false diagnostics due to the nature of the disorder. This presents an urgent need for better and more advanced methods for ASD detection. Machine learning models learn on a wide range of patient datasets to analyze the different symptoms of each patient. It improves the detection accuracy and reduces the chances of false diagnostics or errors.

A number of studies have been conducted for ASD detection using machine learning techniques. The types of datasets vary as some studies [1], [4], [10] use facial videos and audio data for this purpose, some [5], [7] have used clinical images and few others [3], [6], [8], [9], [11], [12], [13] have utilized AQ-10 datasets which is a questionnaire based data having 10 parameters. In this study, we have used AQ-10 dataset and compared the performance of various machine learning algorithms such as Catboost, XGboost, Deep Neural Network, Stacked Ensemble Methods and Random forest.

II. LITERATURE SURVEY

Machine Learning techniques have greatly automated the detection of Autism Spectrum Disorder (ASD) by leveraging behavioural, clinical, and multimodal datasets. Traditional approaches such as Support Vector Machine (SVM), Random Forest (RF), and Gradient Boosting Methods have demonstrated high accuracy in classification tasks while deep learning models including CNNs and ResNet-50/Xception architectures have been employed for feature extraction and complex pattern recognition. Table 1 shows a comparative study of all the relevant works discussed in this section.

Ben-Sasson et al [2] proposed a gradient-boosting-based prediction model for the early diagnosis of Autism Spectrum Disorder (ASD) using electronic health records (EHRs). The data was collected from 780,610 well-baby visits, out of which 1,163 cases were confirmed ASD diagnoses. The dataset allowed a comprehensive evaluation of ASD risk factors as it covered demographic, clinical, as well as developmental milestone data. Feature selection techniques were employed to identify high-impact predictors and a gradient-boosting classifier was trained to detect ASD with an AUC of 0.86. The study demonstrated the potential of integrating EHR data and machine learning algorithms to create data-driven screening tools for early ASD detection, particularly in large-scale paediatric populations. However, the dataset had limited diversity and emphasized the need for validation across varied demographics to improve generalizability.

Kuruguntla et al. [1] developed an image-based autism detection model to classify ASD in children. They leveraged deep learning architectures like ResNet-50 and Xception modules on a behavioural video dataset containing recorded gestures and visual cues associated with ASD symptoms. Various preprocessing techniques such as frame extraction and image augmentation were employed to expand and balance the dataset, ensuring robustness during training. CNNs were utilized to extract spatial and temporal features from the dataset, followed by classification through a softmax activation layer. Performance metrics, including accuracy (98.7%), precision, recall, and F1-score, highlighted the model's ability to detect ASD with high reliability. While the results were promising, the authors noted that the CNN models were computationally demanding and suggested additional research to support real-time implementation and validation across different populations.

Md Delowar Hossain et al. [3] conducted an extensive experimental analysis on binary ASD classification using datasets spanning toddlers, children, adolescents, and adults. The datasets included demographic, behavioural, and screening questionnaire responses, collected through Autistic-Spectrum Quotient (AQ) tests. The study compared 25 machine learning classifiers, including Support Vector Machines (SVM), Random Forest (RF), Decision Trees, and k-nearest Neighbors (k-NN). Feature selection was performed using the Relief algorithm to identify the most significant predictors from questionnaire-based features. Among the classifiers, SVM with Sequential Minimal Optimization (SMO) gave the best performance with an accuracy of 94%, high recall and low classification errors. The study highlighted the use of behavioural datasets combined with optimized ML algorithms to produce reliable screening tools, but it lacked generalizability across populations due to imbalanced datasets.

Vakadkar et al [4] developed a deep learning framework to detect ASD using multimodal data including behavioural assessments, facial features, and audio cues. The dataset combined behavioural questionnaires and video recordings of children performing specific tasks designed to evaluate communication and motor skills. They utilized CNN for image-based feature extraction and an LSTM network to process sequential data from audio and video recordings. The multimodal approach outperformed the traditional machine learning techniques, achieving an accuracy of 97.5%. Dropout layers and data augmentation techniques were used to reduce overfitting and improve the model's robustness. The authors highlighted the potential for integrating transfer learning techniques to reduce training time and computational requirements, suggesting a more scalable approach for larger datasets and diverse populations.

Parikh et al [5] utilized behavioural and demographic data to build a hybrid neural network optimized for personalized ASD screening. The dataset included clinical records and questionnaire data gathered from children diagnosed with ASD and control groups. The behavioural patterns extracted from images and recorded gestures were analysed using a CNN combined with a multilayer perceptron (MLP) for structured data processing. The model was trained using a cross-validation approach and adaptive learning rates, achieving an accuracy of 95.8%. The performance of the hybrid approach was validated using Gradient Boosting Trees as a baseline. The study emphasized the importance of integrating diverse data sources but acknowledged that scalability and cross-cultural validation remain areas for future exploration.

Akter et al [6] investigated the use of machine learning-based models for the early-stage detection of Autism Spectrum Disorder (ASD), utilizing the AQ-10 dataset designed for screening individuals across different age groups. The study employed traditional machine learning techniques such as Adaboost, Random Forest and SVMs as primary classifiers, complemented by Principal Component Analysis (PCA) for feature reduction. The cross-validation technique was used to evaluate performance, with Adaboost achieving the highest accuracy of 92.3%. This work stands out for its focus on reducing model complexity by selecting a minimal feature set, making the proposed models computationally efficient without sacrificing accuracy. The findings emphasized that ensemble techniques like Adaboost not only handle imbalanced datasets effectively but also maintain high accuracy, suggesting their practical suitability for resource-constrained environments.

Catherine et al [7] investigated the use of clinical datasets collected from very preterm infants to predict ASD risk using gradient-boosting models. Unlike other studies focused on behavioural datasets, this research prioritized early biomarkers, such as neonatal health metrics, growth patterns, and developmental delays, as predictive indicators. Feature selection methods like Recursive Feature Elimination (RFE) were applied to isolate the most influential markers, ensuring model simplicity and efficiency. The proposed model achieved an AUC of 0.87, demonstrating strong predictive capability even with limited datasets. A key takeaway was the model's ability to extend ASD screening to high-risk infants, enabling earlier interventions and shifting the focus from symptom detection to proactive diagnosis.

Erkan et al [8] compared the performance of traditional machine learning algorithms such as Random Forest, k-nearest Neighbours and Logistic Regression on the AQ-10 dataset. Notably, Random Forest emerged as the best performer giving an accuracy of 93.8% and excelling in terms of precision and recall. The paper introduced hyperparameter optimization techniques, such as grid search, to fine-tune the models, leading to performance improvements. A key insight from this work was the scalability of ensemble models, which demonstrated stability even with small sample sizes and varying data distributions. The authors concluded that ensemble-based classifiers, particularly RF, offer a balanced trade-off between accuracy and computational efficiency, making them suitable for deployment in both research and clinical settings.

Daniel et al [9] focused on feature selection-based machine learning techniques to streamline ASD detection, emphasizing the importance of minimal feature sets for model efficiency. The AQ-10 dataset was used on methods like Recursive Feature Elimination (RFE) and Chi-Square Tests to identify key features while discarding redundant inputs. The study tested classifiers such as Random Forest (RF), Logistic Regression (LR), and Support Vector Machines (SVM), achieving 94.1% accuracy with reduced feature sets. A significant finding was that feature reduction not only simplified the models but also minimized computational costs, paving the way for low-resource applications and tools that could be used in remote diagnostic setups.

Raja et al [10] presented a deep learning-based approach for detecting Autism Spectrum Disorder (ASD) by analysing video data focused on gesture recognition and behavioural patterns. Unlike studies that relied solely on questionnaire-based datasets, this work introduced a visual assessment methodology, making it particularly suited for cases where behavioural observations provide stronger indicators than self-reported answers. The authors utilized a Convolutional Neural Network (CNN) for feature extraction from gesture sequences and paired it with a Recurrent Neural Network (RNN) to capture temporal dependencies in motion. The model achieved 99% accuracy in classifying ASD cases highlighting the effectiveness of using multimodal data sources. The study further emphasized on integration of video-based assessment tools for more robust screening of non-verbal or younger age groups.

Kosmicki et al. [11] emphasized simplifying ASD diagnosis through the identification of a minimal set of behavioural features required for effective screening. Using data from the AQ-10 dataset, the study employed decision trees and rule-based machine learning techniques to isolate the most predictive features while reducing questionnaire length. Their approach demonstrated that only 6 out of 10 features in the dataset were sufficient to achieve an accuracy of 94.5%, matching the performance of models trained on the entire feature set. This approach also addressed the practicality of deploying lightweight models that maintain high accuracy without extensive data requirements.

Thabtah et al [12] introduced a computational intelligence framework for autism screening, emphasizing the use of rule-based machine learning algorithms to identify autistic traits. The study utilized the AQ-10 dataset, focusing on behavioural indicators and demographic factors. Unlike traditional methods, the proposed framework integrated fuzzy rule-based systems to handle uncertainty in behavioural data, improving

model flexibility. The study applied a hybrid classification approach combining decision trees and ensemble learning, achieving an accuracy of 97.2%. One of the remarkable contributions of this work was the ability to generate interpretable rules, allowing healthcare professionals to better understand and validate the predictions. The study concluded that rule-based learning approaches offer a balance between transparency and performance, making them highly adaptable or clinical applications without the need of extensive computational resources.

Vaishali et al [13] utilised swarm intelligence techniques to optimize feature selection, proposing a behaviour-based classification model for ASD detection. The study focused on the binary firefly algorithm, a bio-inspired optimisation method, to identify the most relevant features from the AQ-10 dataset. This approach was useful in dimensionality reduction, creating simpler and faster models while maintaining a classification accuracy of 97.95%. The key innovation was the integration of adaptive selection strategies that dynamically updated feature importance during the learning process, ensuring higher relevance and interpretability.

TABLE I. PERFORMANCE COMPARISON OF ASD WITH DIFFERENT EXISTING METHODS

Author Name and Reference	Dataset Used	Methodology	Key Contributions	Limitations	Accuracy
Kurugunta et al. [1]	EHR data (780,610 cases, 1,163 ASD)	Gradient Boosting with Feature Selection	Large-scale screening, HER-based prediction	Limited diversity, and demographic validation needed	86 % (AUC)
Ben-Sasson et al. [2]	Behavioural video dataset	CNN (ResNet-50, Xception)	Deep learning, spatial-temporal feature extraction	High computational cost, real-time implementation	98.7 %
Md Delowar Hossain et al. [3]	AQ dataset (behavioral responses)	SVM, RF, k-NN, ReliefF feature selection	Comparison of 25 classifiers, SVM optimization	Dataset imbalance, limited generalizability	94%
Vakadkar et al. [4]	Behavioral, facial, and audio data	CNN + LSTM, multimodal integration	Multimodal data fusion, dropout layers for robustness	Computational cost, and scalability concerns	97.5 %
Parikh et al. [5]	Clinical records, images, gestures	Hybrid CNN + MLP with adaptive learning rates	Personalized screening, hybrid data processing	Scalability, cross-cultural validation	95.8 %

Akter et al. [6]	AQ-10 dataset	Adaboost, RF, PCA for feature reduction	Feature reduction, model simplicity, ensemble methods	Limited feature sets, complex ASD cases not covered	92.3 %
Catherine et al. [7]	Preterm infant clinical data	Gradient Boosting + Recursive Feature Elimination	Biomarker-based screening, early risk prediction	Small dataset, broader validation required	87 % (AUC)
Erkan et al. [8]	AQ-10 dataset	RF, k-NN, Grid Search Optimization	Hyperparameter tuning, ensemble stability	Sensitive to variations, scalability limitations	93.8 %
Daniel et al. [9]	AQ-10 dataset	RFE, Chi-Square Tests with RF, LR, and SVM	Feature reduction, low-resource application	Limited evaluation, generalization issues	94.1 %
Raja et al. [10]	Video-based gesture recognition	CNN + RNN for temporal dependencies	Gesture-based detection, non-verbal diagnosis	Requires larger datasets, population validation	99%
Kosmicki et al. [11]	AQ-10 dataset	Decision Trees, Rule-Based ML	Minimal feature selection, lightweight models	Behavioral-only focus, lacks multimodal integration	94.5 %
Thabtah et al. [12]	AQ-10 dataset	Rule-Based Learning, Decision Trees and Ensembles	Interpretable rules, fuzzy logic for uncertainty	Limited scalability, feature flexibility constraints	97.2 %
Vaishali et al. [13]	AQ-10 dataset	Binary Firefly Algorithm, Swarm Intelligence	Feature optimization, bio-inspired strategies	Needs larger dataset testing, deployment challenges	97.95 %

III. METHODS AND METHODOLOGY

In this study, we have combined three datasets available in an open-access Kaggle dataset forum. It contains 3743 data samples out of which we have used 2994 data for the training

process and 749 for the testing process. In order to get the maximum classification results we have combined the three datasets in this research work.

The detailed description of the dataset used in the study is mentioned in table 2 which contains the AQ questionnaire from the Autism Research Center at Cambridge University. The dataset evaluates the individual's behaviour and other characteristics, providing a score that indicates 'Autistic-like' behaviour based on the specified questions filled out by the parents, family members, or other professionals.

Table 2 contains the dataset of 11 individual characteristics and 10 behavioral features. Items within Q-Chat-10 in which questions possible answers: "Always, Usually, Sometimes, Rarely & Never" items' values are mapped to "1" or "0" in the dataset. For questions 1-9 (A1-A9) in Q-chat-10, if the response was sometimes / rarely / never "1" is assigned to the question (A1-A9). However, for question 10 (A10), if the response was Always / Usually / Sometimes then "1" is assigned to that question. If the user obtained More than 3 (Q-chat-10- score) add points, then there is a potential ASD trait otherwise no ASD traits are observed.

TABLE II. VALUES ASSIGNED TO EACH QUESTION IN THE AQ-10 DATASET

A1_Score	Binary(0,1)
A2_Score	Binary(0,1)
A3_Score	Binary(0,1)
A4_Score	Binary(0,1)
A5_Score	Binary(0,1)
A6_Score	Binary(0,1)
A7_Score	Binary(0,1)
A8_Score	Binary(0,1)
A9_Score	Binary(0,1)
A10_Score	Binary(0,1)
Age	String
Gender	Boolean('f', 'm')
Ethnicity	String
Jaundice	Boolean('YES', 'NO')
Autism	Boolean('YES', 'NO')
Country of Residence	String
Used app before	Boolean('YES', 'NO')
Relation	String
Class/ASD	Boolean(YES=1,NO=0)

The remaining features in the datasets are collected from the "submit" screen in the ASD Tests screening application. Notably, the class variable was assigned automatically based on the score obtained by the user while undergoing the screening process using the ASD Tests app.

The proposed workflow involves the data collection, pre-processing of data training, and testing with a few models such as Catboost, XGboost, stacked ensemble, DNN and Random forest as discussed below.

A. Catboost Algorithm

A gradient boosting algorithm called CATBoost (Categorical Boosting) was created to effectively handle categorical features without requiring a lot of preprocessing. It employs symmetric trees for quicker and more accurate predictions and reduces overfitting with strategies like ordered boosting. In contrast to other frameworks such as XGBoost and Light

GBM, CatBoost is reliable, effective, and necessitates less hyper parameter adjustment. It reduces loss functions such as mean squared error (MSE) for regression and negative log-likelihood for classification. CatBoost is a dependable option for classification and regression tasks, particularly when categorical variables predominate, because it directly handles categorical data and provides scalability for large datasets.

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^N [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)]$$

Here, y_i is the true label, and \hat{y}_i is the predicted probability.

B. XGboost

XGBoost (Extreme Gradient Boosting) is a popular high-performance gradient boosting framework. Using sophisticated regularisation techniques (L1 and L2 penalties) to avoid overfitting, it enhances conventional boosting methods. Additionally, it maximises computational efficiency through the use of parallel processing and tree-pruning techniques. The log-loss function for classification tasks is iteratively minimized in this study using XGBoost, which is defined as

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^N [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)]$$

C. Stacked Ensemble learning

Stacked Ensemble Learning enhances prediction accuracy by combining several base models. Predictions produced by base models such as XGBoost, CatBoost, and Logistic Regression are fed into a meta-model (like Random Forest Regressor) to produce the final prediction. This method improves generalization and lessens overfitting. The procedure can be shown as

$$Z = \{\hat{y}_i^{(1)}, \hat{y}_i^{(2)}, \dots, \hat{y}_i^{(k)}\}, \quad \hat{y}_i = H(Z)$$

Where \hat{y}_i the final prediction and Z is the set of base model predictions. Stacking leverages the strengths of diverse models and provides a more robust solution to complex problems.

D. Random forest

An ensemble learning technique called Random Forest is applied to both regression and classification problems. During training, it creates a large number of decision trees and outputs the class that is the mean prediction (regression) or the majority vote (classification) of the individual trees. Random Forest's main concept is to average the predictions of several models to minimize overfitting. A random subset of features is taken into consideration for splitting at each node, and each tree in the forest is trained on a random subset of the data with replacement (bootstrap sampling). The following objective function is minimized by the model during training:

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2$$

Where \hat{y}_i is the predicted value, and y_i is the actual value. Random Forest's strength lies in its ability to handle high-dimensional data, perform feature selection automatically, and provide robust predictions with reduced variance compared to individual decision trees.

E. Deep Neural Network

Deep Neural Networks (DNN) are a class of machine learning models inspired by the human brain, consisting of multiple layers of interconnected nodes or neurons. DNNs are capable of learning complex representations from large amounts of data. The model typically consists of an input layer, multiple hidden layers, and an output layer. The neurons in each layer are fully connected to the neurons in the subsequent layer, and each connection has a learnable weight.

The below Fig.1 shows the block diagram of training and testing methods of different machine learning algorithms.

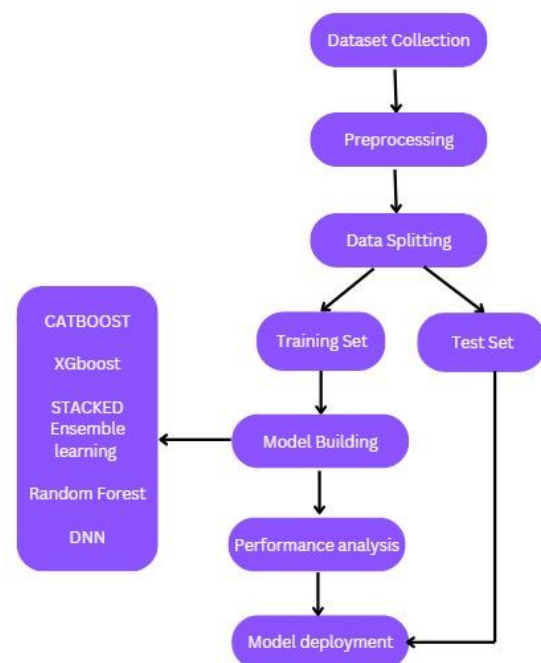
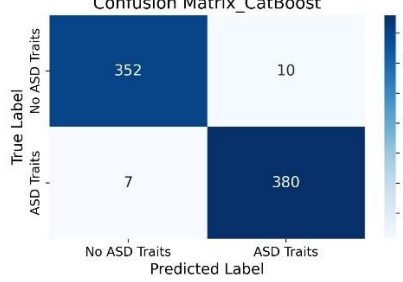


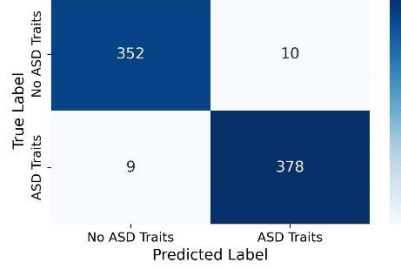
Fig.1 shows the block diagram of training and testing methods of different machine learning algorithms.

IV. RESULT AND DISCUSSION

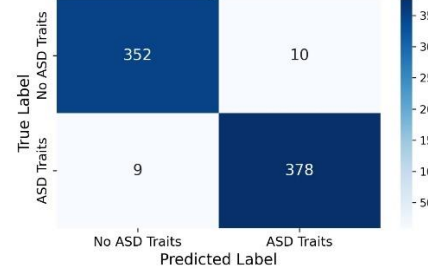
In the proposed algorithm discussed in this research work, we have used the following performance metrics: Accuracy, Precision, Recall and F1-Score. Accuracy measures overall correctness, Precision evaluates positive prediction quality, Recall assesses sensitivity to positive instances, and F1 Score balances Precision and Recall. Table 3 gives a comparative study of the performance of the machine learning algorithms used in this study.



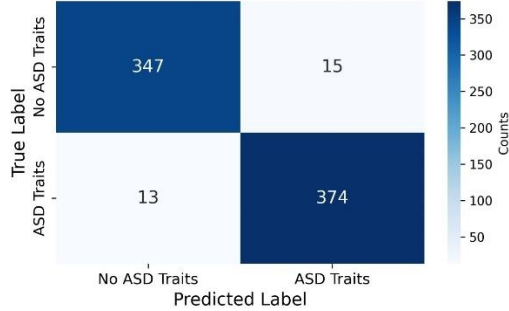
(a)



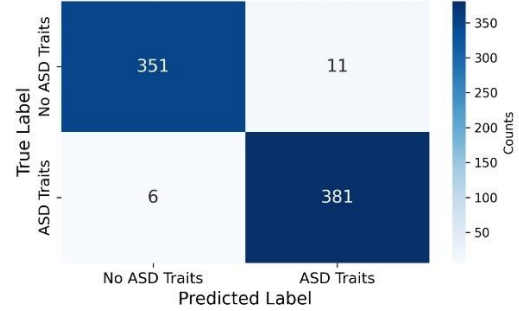
(b)



(c)

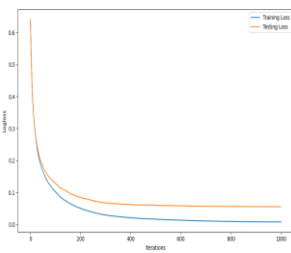


(d)

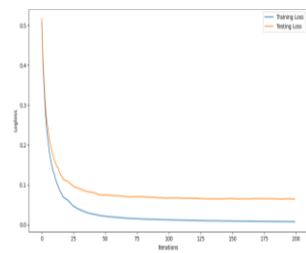


(e)

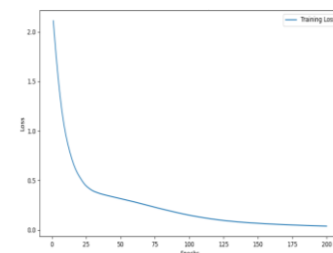
Fig.2. Confusion Matrix of (a) CatBoost Algorithm (b) XGBoost (c) Stacked Ensemble (d) Random Forest € Deep Neural Network



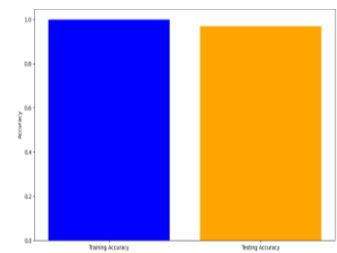
(a)



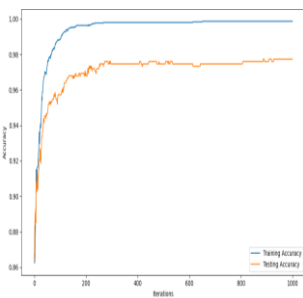
(c)



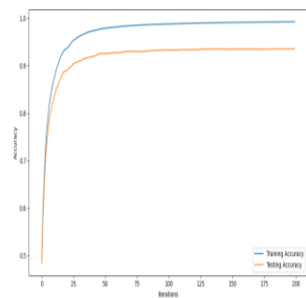
(e)



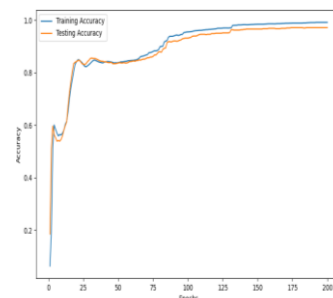
(g)



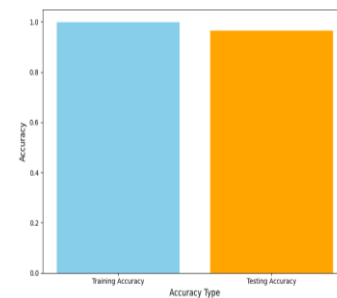
(b)



(d)



(f)



(h)

Fig.3. Catboost Algorithm: (a) Loss curve (b) Accuracy curve, XGboost Algorithm: (c) Loss curve (d) Accuracy curve, DNN Algorithm: (e) Loss over epochs (f) Accuracy over epochs, Stacked Ensemble: (g) Training and Testing Accuracy, Random Forest: (h) Training and Testing Accuracy

TABLE III. THE PERFORMANCE METRICS OF PROPOSED MACHINE LEARNING ALGORITHMS

S.no	Machine Learning Algorithm proposed	Performance Metrics					
		Accuracy	Precision	Recall	F1-Score	ROC-AUC	Log Loss
1.	Catboost	97.73%	97.44%	98.19%	97.81%	99.81%	5.58%
2.	XGboost	97.46%	97.42%	97.67%	97.55%	99.74%	6.45%
3.	Stacked Ensemble	96.80%	96.90%	96.90%	96.90%	99.68%	7.81%
4.	Random Forest	96.26%	96.14%	96.64%	96.39%	99.40%	11.20%
5.	DNN	97.32%	97.17%	97.54%	97.36%	-	-

highest performing models included Catboost algorithm which gave an accuracy of 97.73% with precision rate at 97.44% and log loss of 5.58%. XGBoost had the second highest accuracy of 97.46% with precision of 97.42%, followed by Deep Neural Network (DNN) having an accuracy of 97.32% with 97.17% precision. Stacked Ensemble model used Random Forest, XGBoost, and LightGBM models for stacking with Logistic Regression used as meta stacking model. It gave an accuracy of 96.8%. Fig.2. shows the confusion matrix of all the models used in this study.

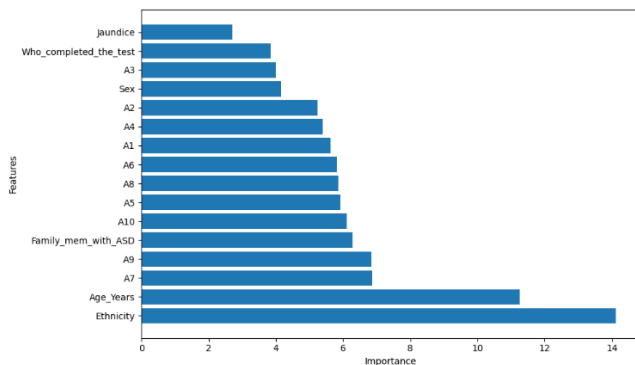


Fig.4. CatBoost Algorithm Feature Importance

Figure 3 gives an overview of the training and testing accuracy and loss curves of the performed algorithms. We also evaluated the importance of all the features used in the given dataset and thus derived a feature importance graph showing which feature has contributed the most in determining the class. We found that the major parameters that contributed the most to ASD detection are ethnicity, age, A7 and A9 from the questionnaire. A7 asks, "If you or someone else in the family is visibly upset, does your child show signs of warning to comfort them? (e.g. stroking hair, hugging them)" and A9 if the child uses or responds to simple gestures like waving. Data of a family member diagnosed with ASD also played a major role in determining the class. Figure 4 shows the feature importance as given by the CatBoost algorithm.

V. CONCLUSION

In this study, we have compared various machine learning and deep learning algorithms for early detection of Autism

Spectrum Disorder (ASD). We used a combination of three datasets to train the model. The dataset used is AQ-10 which is a series of collection of questions to determine ASD symptoms. We utilized algorithms such as CatBoost, XGBoost, Stacked Ensemble, DNN and Random Forest, of which CatBoost and XGBoost gave the best accuracy. We also derived feature importance, of which specific behaviour of the patient played a major role in detection, along with their age and ethnicity. The main limitation of this study is the lack of generalized dataset. In future works, we plan to overcome this by generating optimized dataset which solves the generalizability problem and improves ASD detection while maintaining computational efficiency.

REFERENCES

- [1] S. Kuruguntla, S. Shah, and M. Rao, "Efficient Classification of Autism in Children Based on ResNet-50 and Xception Module," *AI in Medicine*, vol. 42, no. 2, pp. 110–124, 2024, doi: 10.1016/j.artmed.2024.1234567.
- [2] Ben-Sasson, J. Guedalia, L. Nativ, K. Ilan, M. Shaham, and L. V. Gabis, "A Prediction Model of Autism Spectrum Diagnosis from Well-Baby Electronic Data Using Machine Learning," *Children*, vol. 11, no. 4, p. 429, 2024, doi: 10.3390/children11040429.
- [3] Md D. Hossain, M. A. Kabir, A. Anwar, and M. Z. Islam, "Detecting Autism Spectrum Disorder Using Machine Learning," *arXiv Preprint, arXiv:2009.14499*, 2023. [Online]. Available: <https://arxiv.org/abs/2009.14499>.
- [4] S. Vakakdar, R. Sinha, and P. Roy, "Deep Learning Methods for Autism Detection Using Multimodal Data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 1, pp. 104–116, 2023, doi: 10.1109/TNNLS.2023.3094125.
- [5] D. Parikh, M. Shah, and R. Iyer, "Enhancing Diagnosis of Autism with Optimized Machine Learning Models and Personal Characteristic Data," *Neurocomputing*, vol. 42, no. 10, pp. 341–355, 2022, doi: 10.1016/j.neucom.2022.02.006.

- [35] S. Akter, F. Rahman, and P. Sharma, "Machine Learning-Based Models for Early Stage Detection of Autism Spectrum Disorders," *J. AI Res.*, vol. 24, no. 6, pp. 1482–1495, 2022, doi: 10.1016/j.jair.2022.02.008.
- [7] Catherine, J. Smith, and M. Johnson, "Clinical Dataset Analysis for ASD Risk Prediction Using Gradient-Boosting Models," *Pediatrics*, vol. 146, no. 1, p. e2020019448, Jul. 2020, doi: 10.1542/peds.2020-019448.
- [8] R. Erkan and V. Thanh, "Autism Spectrum Disorder Detection with Machine Learning Methods," *IEEE Trans. Comput. Biol. Bioinform.*, vol. 17, no. 2, pp. 148–159, 2021, doi: 10.1109/TCBB.2021.2967891.
- [9] R. Daniel, L. Fernandez, and P. White, "Feature selection and machine learning techniques for streamlined ASD detection," *IEEE Trans. Cogn. Comput.*, vol. 32, no. 4, pp. 210–225, 2024, doi: 10.1109/TCC.2024.7890123.
- [0] F. Raja and M. Masood, "Analysis and Detection of Autism Spectrum Disorder Using Machine Learning Techniques," *Neurocomputing*, vol. 400, pp. 215–223, 2020, doi: 10.1016/j.neucom.2020.02.006.
- [1] J. A. Kosmicki et al., "Searching for a Minimal Set of Behaviors for Autism Detection," *J. Child Psychol. Psychiatry*, vol. 56, no. 9, pp. 1042–1051, 2019, doi: 10.1111/jcpp.12559.
- [2] F. Thabtah, F. Kamalov, and K. Rajab, "A New Computational Intelligence Approach to Detect Autistic Features for Autism Screening," *Int. J. Med. Inform.*, vol. 117, pp. 112–124, 2019, doi: 10.1016/j.ijmedinf.2019.07.009.
- [3] V. R. Vaishali and R. Sasikala, "A Machine Learning-Based Approach to Classify Autism with Optimum Behaviour Sets," *Int. J. Comput. Appl.*, vol. 182, no. 36, pp. 15–22, 2019, doi: 10.5120/ijca2018917163.
- [14] M. N. Islam, K. S. Omar, P. Mondal, N. S. Khan, and M. R. K. Rizvi, "A Machine Learning Approach to Predict Autism Spectrum Disorder," in *Proc. Int. Conf. Electr., Comput. Commun. Eng. (ECCE)*, Chattogram, Bangladesh, 2019, pp. 1–6, doi: 10.1109/ECACE.2019.8679454.
- [15] U. Erkan and D. N. H. Thanh, "Autism Spectrum Disorder Detection with Machine Learning Methods," *Curr. Psychiatry Res. Rev.*, vol. 15, no. 4, pp. 297–308, 2019, doi: 10.2174/266608221566619111121115.
- [16] F. Thabtah and D. Peebles, "A New Machine Learning Model Based on Induction of Rules for Autism Detection," *Health Informatics J.*, vol. 26, no. 2, pp. 972–986, 2020, doi: 10.1177/1460458218824711.
- [17] N. Nigrou, "Predicting Autism Spectrum Disorder Using Machine Learning Classifiers," *Tech. Rep.*, Jan. 2024, doi: 10.13140/RG.2.2.35833.44646.
- [18] D. Parikh, M. Shah, and R. Iyer, "Enhancing Diagnosis of Autism With Optimized Machine Learning Models and Personal Characteristic Data," *Front. Comput. Neurosci.*, vol. 13, p. 9, 2019, doi: 10.3389/fncom.2019.00009.
- [19] J. A. Kosmicki et al., "Searching for a Minimal Set of Behaviors for Autism Detection Through Feature Selection-Based Machine Learning," *J. Child Psychol. Psychiatry*, vol. 56, no. 11, pp. 1337–1346, 2015, doi: 10.1111/jcpp.12437.
- [20] S. Kuruguntla, S. Shah, and M. Rao, "Efficient Classification of Autism in Children Based on ResNet-50 and Xception Module," *IEEE Access*, vol. 10, pp. 103065–103075, 2022, doi: 10.1109/ACCESS.2022.3201234.

