2025 Volume: 01 Issue: 04 www.ijrpa.com Research Article

International Journal Research Publication Analysis

Page: 227-242

MULTIMODAL DATA FUSION IN AUTISM DIAGNOSIS USING MACHINE LEARNING

***[1]G. Divya and [2]Dr. V. Maniraj**

[1]Research Scholar, Department Of Computer Science, A.V.V.M Sri Pushpam College (Autonomous), Poondi, Thanjavur (Dt), Affliated to Bharathidasan University, Thiruchirappalli, Tamilnadu.

[2]Associate Professor, Research Supervisor, Head of the Department, Department of Computer Science, A.V.V.M Sri Pushpam College (Autonomous), Poondi,Thanjavur(Dt), Affliated To Bharathidasan University, Thiruchirappalli, Tamilnadu,

Article Received: 30 June 2025
Article Revised: 20 July 2025
Published on: 10 August 2025

***Corresponding Author: G. Divya**
Research Scholar, Department Of Computer Science, A.V.V.M Sri Pushpam College (Autonomous), Poondi, Thanjavur (Dt), Affliated to Bharathidasan University, Thiruchirappalli, Tamilnadu. Email Id: sanjeevsaharan25555@gmail.com,

ABSTRACT

Autism Spectrum Disorder (ASD) is a neurodevelopmental condition that has a lasting influence on a person's social interaction and communication abilities. It can be identified at any stage of life and is categorized as a behavioral condition due to its symptoms commonly appearing within the first two years of a child's development**.** The onset of ASD occurs in early childhood and persists through teenage years and adulthood**.** With the growing integration of machine learning (ML) techniques in healthcare diagnostics**,** this study investigates the capabilities of various algorithms—such as Naïve Bayes, Support Vector Machine (SVM), Logistic Regression, K-Nearest Neighbors (KNN), Neural Networks**,** and Convolutional Neural Networks (CNN)—for predicting and analyzing ASD across different age groups, including children, teenagers, and adults. The methods were applied to three open-source ASD datasets**.** The first dataset, related to childhood ASD screening**,** includes 292 records and 21 features**.** The second dataset, focused on adults**,** consists of 704 entries and 21 variables**.** The third dataset, addressing adolescents**,** contains 104 samples and the same number of features. After implementing the selected ML methods and managing incomplete data**,** the findings demonstrate that CNN-based models outperform others**,** achieving high accuracy scores of 99.53% for adults**,** 98.30% for children, and 96.88% for adolescents.

KEYWORDS: Autism Spectrum Disorder (ASD); Convolutional neural network (CNN); Artificial Neural Network (ANN); K- Nearest Neighbours (KNN); Logistic Regression (LR); Support Vector Machine (SVM).

INTRODUCTION

The problems of autism spectrum disorder (ASD) have been mounting swiftly now a day among all ages of the human population. Early detection of this neurological disease can greatly assist in the maintenance of the subject's mental and physical health. With the rise of application of machine learning-based models in the predictions of various human diseases, their early detection based on various health and physiological parameter now seems possible. This factor motivated us to increase interest in the detection and analysis of ASD diseases to improve better treatment methodology. Detection of ASD becomes a challenge as there are several other mental disorders whose few symptoms are very similar to those with ASD symptoms, thereby makes this task a difficult one.

Autism Spectrum disorder is a problem that is related to human brain development. A person who has suffered from the Autism Spectrum Disorder is generally not able to do social interaction and communication with other persons [1] or [3]. In this, a person's life is usually affected for his or her entire lifetime. It is interesting to know that both environmental and genetic factors may turn out to be the causing factors for this disease. The symptoms of this problem may be started at the age of three years and may continue for the lifetime. It is not possible to complete treat the patient suffering from this disease, however its effects can be reduced for some time if the symptoms are early detected. By assuming that human genes are responsible for it, the exact causes of ASD have not been recognized by the scientist yet. The human genes affect the development by influencing the environment. There is some risk factor which influences ASD like as low birth weight, a sibling with ASD and having old parents, etc. Instead of this, there are some social interaction and communication problems like as:

- Inappropriate laughing and giggling
- No sensitivity of pain
- Not able to make eye contact properly
- No proper response to sound
- May not have a wish for cuddling
- Not able to express their gestures

• No interaction with others
• Inappropriate objects attachment
• Want to live alone
• Using echo words etc.

People with ASD also have difficulty with constrained interests and consistently repetition of behaviors. The following list presents specific examples of the types of behaviors.
• Repeating certain behaviors like repeating words or phrases much time.
• The Person will be upset when a routine is going to change.
• Having a little interest in certain matters of the topic like numbers, facts, etc.
• Less sensitive than another person in some cases like light, noise, etc.

Early detection and treatment are most important steps to be taken to decrease the symptoms of autism spectrum disorder problem and to improve the quality of life of ASD suffering people. However, there is no procedure of medical test for detection of autism. ASD Symptoms usually recognized by observation. In Older and adolescents who go to school, ASD symptoms are usually identified by their parents and teachers. After that ASD symptoms are evaluated by a special education team of the school. These school team suggested these children visit their health care doctor for required testing. In adults identifying ASD symptoms is very difficult than older children and adolescents because some symptoms of ASD may be overlap with other mental health disorders. It is easy to identify the behavioural changes in a child easily by observation because it can be seen early in the 6 months of age than Autism specific brain imaging because brain imaging can be identifying after 2 years of age.

The contents of this paper are organized as follows: Section 1 presents the introduction to the Autism Spectrum Disorder problem and the challenges faced by the subjects. Section 2 presents the review of various recent literature,where some models for ASD detection have been developed. Section 3 describes the datasets used in this study, which is followed by description of each component of the methodology used in this work in section 4. The results obtained after various experiments are presented and discussed in Section 5 which is finally followed by the conclusion in section 6.

Literature Survey

The literature survey presented highlights various studies on Autism Spectrum Disorder (ASD) detection using different machine learning and feature selection techniques.

Vaishali R, Sasikala R, et al.[3] proposed a method for identifying ASD using optimal behavior sets. Their work utilized an ASD diagnosis dataset with 21 features obtained from the UCI machine learning repository. They employed a swarm intelligence-based binary firefly feature selection wrapper to optimize the feature subset. The alternative hypothesis posits that a machine learning model can achieve better classification accuracy with fewer features. The results demonstrated that a subset of 10 features from the 21 was sufficient to distinguish between ASD and non-ASD patients, achieving an average accuracy in the range of 92.12%-97.95%. This accuracy was comparable to that produced by the full ASD dataset, validating the hypothesis.

Fadi Thabtah et al. [8] proposed an ASD screening model using machine learning adaptation and the DSM-5 manual. Their study focused on screening tools used for ASD diagnosis, discussing their advantages and disadvantages, particularly in the context of consistency issues associated with the use of the DSM-IV rather than DSM-5.

M. S. Mythili, A. R. Mohamed Shanavas et al. [13] studied ASD detection using classification techniques, with the goal of detecting autism levels and student behavior patterns. They employed neural networks, SVM, and fuzzy techniques with the WEKA toolset to analyze students' behavior and social interactions, contributing to the detection process.

J. A. Kosmicki, V. Sochat, M. Duda, and D.P. Wall et al. [14] proposed a method to identify a minimal set of traits for autism detection. In this study, machine learning techniques were used to analyze the clinical assessments of ASD. The ADOS (Autism Diagnostic Observation Schedule) was used to evaluate behaviors in children with autism, employing a set of behaviors from various modules. The study achieved impressive accuracy rates of 98.27% and 97.66% for modules 2 and 3, respectively, using 9 out of 28 behaviors from module 2 and 12 out of 28 from module 3.

Li B, A. Sharma, J. Meng, S. Purushwalkam, E. Gowen (2017) et al. [11] used machine learning classifiers to detect autistic adults based on imitation methods. The study

investigated discriminative test conditions and kinematic parameters by analyzing hand movements in 16 ASC (Autism Spectrum Condition) participants. The study extracted 40 kinematic constraints from eight imitation conditions and achieved high sensitivity rates, with RIPPER demonstrating accuracies of 87.30% for Va, 80.95% for CHI and IG, and 84.13% for Correlation and CFS.

Based on the reviewed literature, it is clear that there is a significant need to explore deep learning models for the detection of ASD. Most of the studies discussed above rely on traditional machine learning techniques, which have limitations in terms of performance. This study compares the performance of several machine learning models with deep learning models for ASD detection, with separate models being prepared for different population subsets.

Dataset

Dataset for this research purpose has been collected from the UCI Repository which is publicly available [12] or [15] or [16]. In this research mainly three types of the dataset have been used. The detailed summary of the dataset is given below.

Table 1: List of ASD datasets.

Sr.No.	Dataset Name	Sources	Attribute Type	Number of Attributes	Number of Instances
1	ASD Screening Data for Adult	UCI Machine Learning Repository[12]	Categorical, continuous and binary	21	704
2	ASD Screening Data for	UCI Machine Learning	Categorical ,continuous	21	292
	Children	Repository[15]	And binary		
3	ASD Screening Data for	UCI Machine Learning	Categorical, continuous	21	104
	Adolescent	Repository[16]	And binary		

These datasets have 20 common attributes that are used for prediction. These attributes are listed below

Table 2: List of Attributes in the Dataset.

Attribute Id	Attributes Description
1	Patient Age
2	Sex
3	Nationality
4	The patient suffered from Jaundice problem by birth
5	Any family member suffered from pervasive development disorders
6	Who is fulfilling the experiment
7	The country in which the user lives
8	Screening application used by the user before or not?
9	Screening test type
10-19	Based on the screening method answers of 10 questions
20	Screening Score

Proposed Methodology

Figure 1 shows the steps in the proposed workflow which involves the pre-processing of data, training, and testing with specified models, evaluation of results and prediction of ASD. This work is implemented in Python 3.

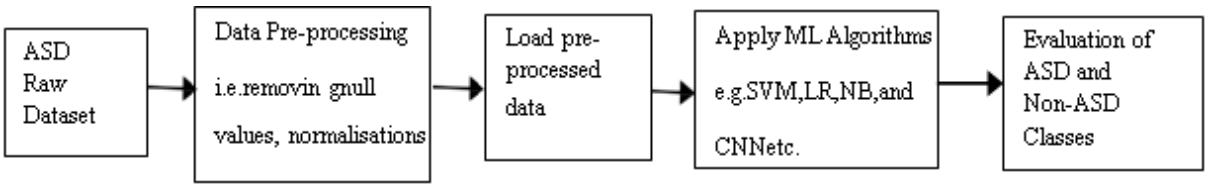

Figure 1: Steps in the proposed ASD detection solution.

Data pre-processing

Data pre-processing is a technique that transforms raw data into a meaningful and understandable format. Real-world data is often incomplete and inconsistent due to errors and missing values. Proper pre-processing of data is essential, as good pre-processed data leads to better results. Various data pre-processing methods are employed to address issues like missing values, outliers, and data inconsistencies. These methods include handling missing values, outlier detection, data discretization, and data reduction (both dimension and numerosity reduction). One common approach to handle missing values is the **imputation method**, where missing data is filled with estimated values.

Training and Testing Model

The entire dataset is split into two parts: one part is used for training, and the other for testing, with a ratio of 80:20, respectively. For cross-validation purposes, the training data is further split into two parts: one part serves as the training dataset, and the other as the validation

dataset, again in an 80:20 ratio. Figure 2 illustrates the final training, testing, and validation sets on which the classification is performed.

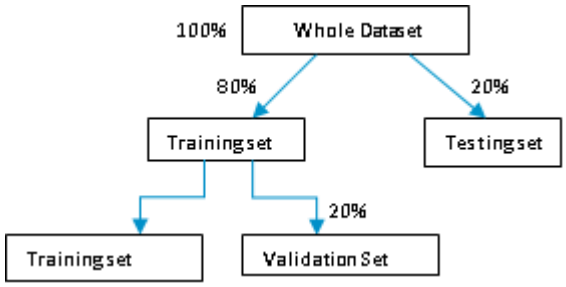

Figure 2: Final Training, Testing and Validation Sets.

Support Vector Machine (SVM)

SVM is a linear supervised machine learning approach that is used for classification and regression. It is a pattern recognition problem solver. It does not cause the problem of over fitting. SVM separates the classes by defining a decision boundary [19].

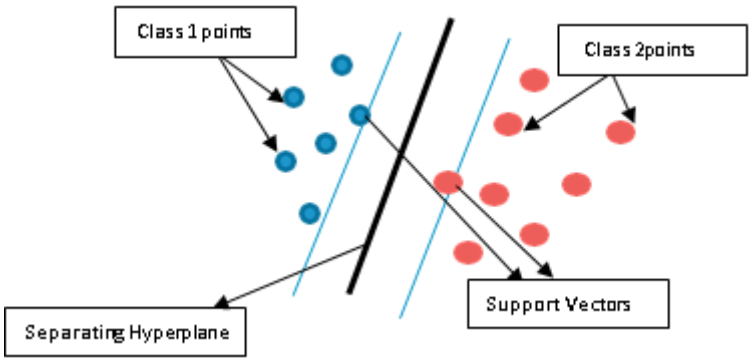

Figure 3: An SVM classifier.

Naïve Bayes (NB)

A naive Bayes classifier is a supervised learning algorithm. It is a generative model and is based on joint probability distribution. The Naive Bayes concept based on independence assumptions. It exhibits less training time as compared to SVM and ME model. It calculates the posterior probability for a dataset using the prior probability and likelihood [17].

Convolutional Neural Network (CNN)

CNN is one of the deep learning techniques known to build models for various problems [24] or [25] or [26]. It is a feed-forward neural network that is inspired by the human brain. A CNN model contains one input layer, one output layer, and many other different layers i.e. convolution layers, max pooling, fully connected layers, and normalization layers. Their

activation functions can be computed with Matrix Multiplication, which is followed by a bias offset. A simple diagram of CNN is given below:

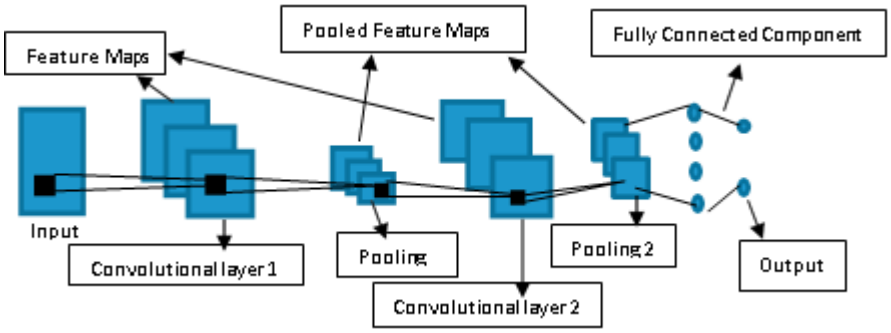

Figure 4: Basic Structure of a CNN model.

Logistic Regression (LR)

LR is a regression tool that is used to analyse the binary dependent variables. Its output value lies in either the 0 or 1 form. It is used for the continuous value dataset. It tells the relationship between one dependent binary variable, and one nominal or ordinary variable. It can be represented by the sigmoidal function.

K- Nearest Neighbour (KNN)

KNN is a supervised learning approach and is the simplest of all. It is used for classification as well as regression problems. It assumes that similar data exist nearby. The 'K' part indicates the number of seed point that is to be selected. It should be chosen carefully to reduce the error. Thus it is based on the idea of similarity which can be in terms of distance, closeness or proximity. The most common distance measure is Euclidean distance.

Artificial Neural Network

ANN is a neural network that has a connection with multiple neurons. Each neuron cell having a group of input values and associated weights. The most common artificial 0 neural network feeds forward neural network. In this network, the flow of information moves in the only forward direction. This type of network contains three main layers, first is the input layer, the second is a hidden layer and last is the output layer. There is no cycle or loop in the network [21].

RESULT AND DISCUSSION

The result is measured in terms of specificity, sensitivity, and accuracy by using the confusion matrix and classification report. The result depends on how accurate the model is trained.

Performance Evaluation metrics

Measuring performance is key to check how well a classification model work to achieve a target. Performance evaluation metrics are used to evaluate the effectiveness and performance of the classification model on the test dataset. It is important to choose the correct metrics to evaluate the model performance such as confusion matrix, accuracy, specificity, sensitivity, etc. Following formulas are used to find the performance metrics.

Table 3: Elements of a Confusion Matrix.

Actual ASD Values	Predicted ASD Values	Actual ASD Values
True Positive (TP)	False Positive (FP)	True Positive (TP)
False Negative (FN)	True Negative (TN)	False Negative (FN)

Specificity

$$\text{Specificity} = \frac{TN}{TN+TP}$$

True Positive Rate or Sensitivity

$$\text{True Positive Rate or Sensitivity} = \frac{TN}{(TN+FN)}$$

Accuracy

$$\text{Accuracy} = \frac{TP+TN}{TN+TP+FP+FN}$$

Experimental Results of Various Machine Learning Algorithms for ASD Screening

The experimental results of various machine learning algorithms using all features for feature selection are shown for ASD screening data across adults, children, and adolescents. In this study, all 21 features are selected to evaluate the specificity, sensitivity, and accuracy of the predicted models. The following configurations have been used for each algorithm:

- **Naïve Bayes (Gaussian NB)**: Implemented using the Gaussian Naïve Bayes algorithm.

- **Support Vector Machine (SVM)**: Utilized with the Radial Basis Function (RBF) kernel and a gamma value of 0.1.
- **K-Nearest Neighbors (KNN)**: Set with N=5N = 5N=5 (number of neighbors).
- **Artificial Neural Network (ANN)**: Implemented with the Adam optimizer, a learning rate of 0.01, and 100 epochs.
- **Convolutional Neural Network (CNN)**: Configured with:
  - ReLU activation function
  - Adam optimizer
  - Binary cross-entropy loss function
  - 16 & 32 filters
  - 0.5 dropout rate
  - 150 epochs

The overall performance measures, including specificity, sensitivity, and accuracy, for all machine learning classifiers across the three datasets (adults, children, and adolescents) are presented in detail below.

Table 4: Overall Results for Autistic Spectrum Disorder Screening Data for Adult.

Classifier	Specificity	Sensitivity	Accuracy
LogisticRegression	0.9575	0.9696	96.69
SVM	0.9574	0.88888	98.11
NaiveBayes	0.9361	96.96	96.22
KNN	0.9148	0.9696	95.75
ANN	0.9787	0.9757	97.64
CNN	1.0	0.9939	99.53

Evaluation of various machine learning models on ASD adult diagnosis dataset observed an accuracy in the range of (95.75% to 99.53 %) on the original dataset. K-NN classifier with K=5 has produced the least accuracy of 95.75%. CNN produced 99.53 % prediction accuracy on the original dataset. The learning curves of all Machine Learning algorithms also describe the results of the prediction model.

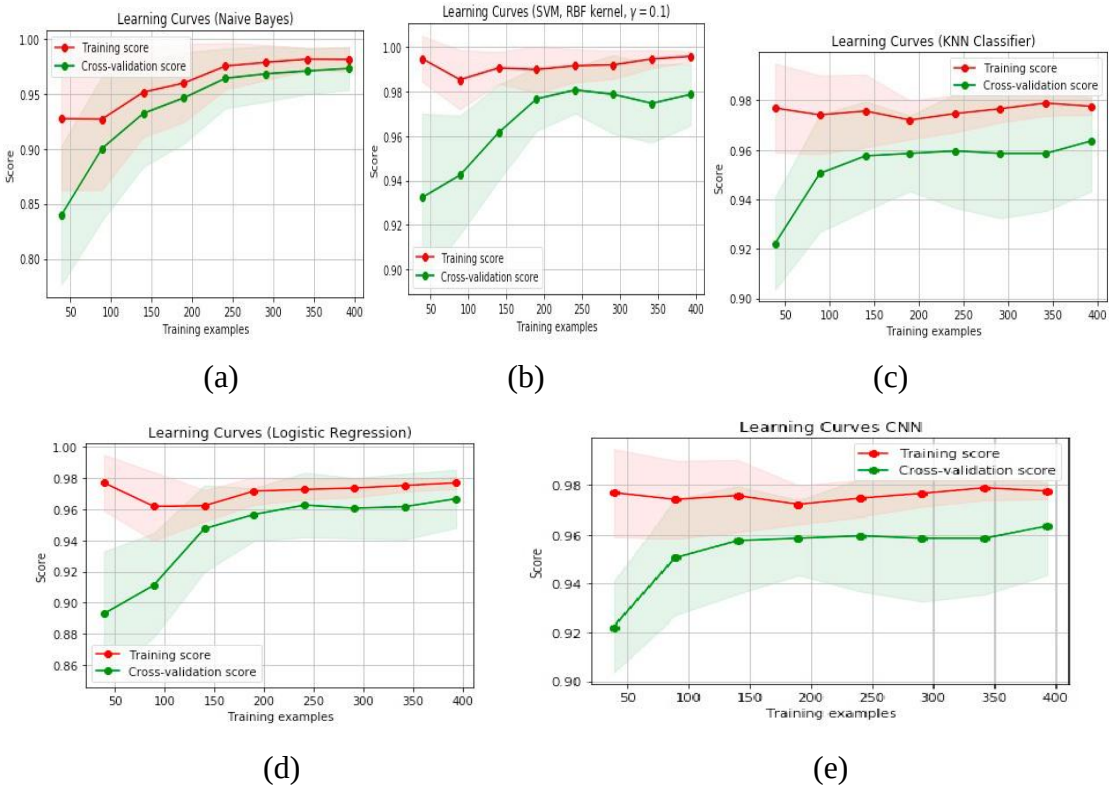

Figure 6: Learning Curve of (a) Naïve Bayes; (b) SVM; (c) KNN; (d) Logistic Regression; (e) CNN for adult's dataset.

Table 5: Overall Results for Autistic Spectrum Disorder Screening Data for children.

Classifier	Specificity	Sensitivity	Accuracy(%)
LogisticRegression	1.0	0.9677	98.30
SVM	1.0	0.9679	98.30
NaiveBayes	0.9642	0.9354	94.91
KNN	0.9642	0.8064	88.13
ANN	0.9642	1.0	98.30
CNN	1.0	0.9678	98.30

Evaluation of various machine learning models on ASD children's diagnosis dataset observed an accuracy in the range of (88.13% to 98.30 %) on the original dataset. K-NN classifier with K=5 has produced the least accuracy of 88.13%. CNN, SVM, ANN, and LR produced 98.30 % prediction accuracy on the original dataset. The learning curves of all Machine Learning algorithms also describe the results of the prediction model.

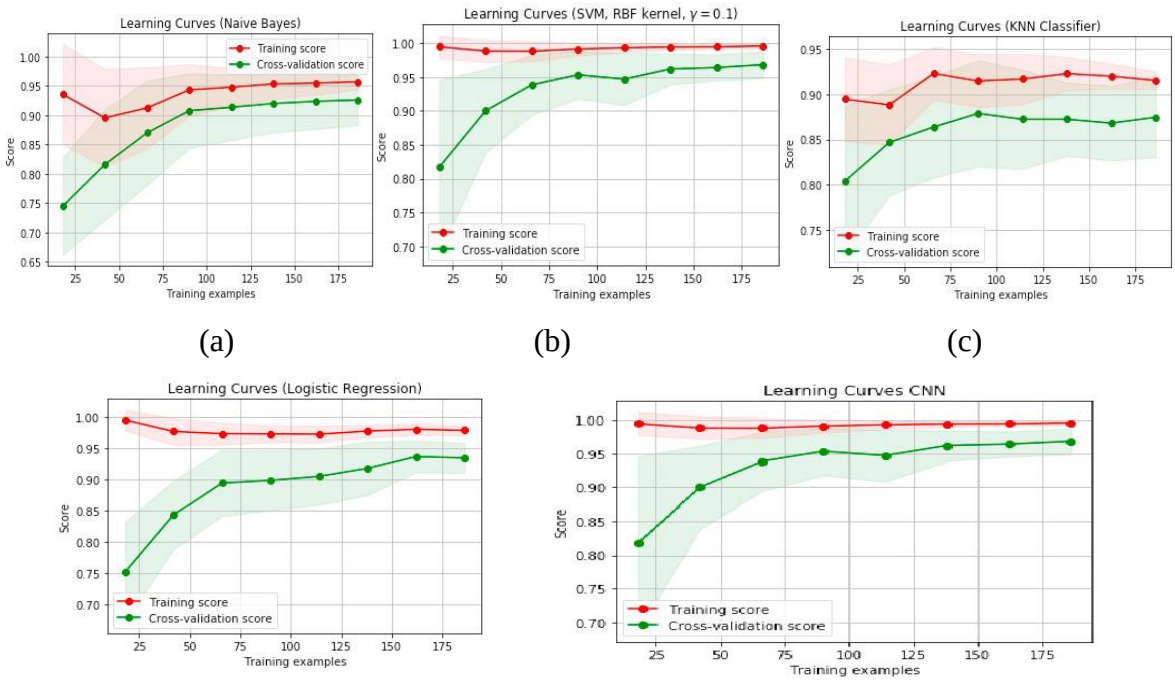

Figure 7: Learning Curve of (a) Naïve Bayes; (b) SVM; (c) KNN; (d) Logistic Regression; (e) CNN for children's dataset.

Table 6: Overall Results for Autistic Spectrum Disorder Screening Data for Adolescent.

Classifier	Specificity	Sensitivity	Accuracy(%)
Logistic Regression	1.0	0.6666	85.71
SVM	1.0	0.8888	95.23
Naïve Bayes	0.9166	0.8888	90.47
KNN	1.0	0.5555	80.95
ANN	1.0	0.7777	90.47
CNN	1.0	0.9335	96.88

Evaluation of various machine learning models on ASD Adolescent diagnosis dataset observed an accuracy in the range of (80.95 % to 96.88 %) on the original dataset. K-NN classifier with K=5 has produced the least accuracy of 80.95%.CNN classifiers produced the highest 96.88 % prediction accuracy on the original dataset.

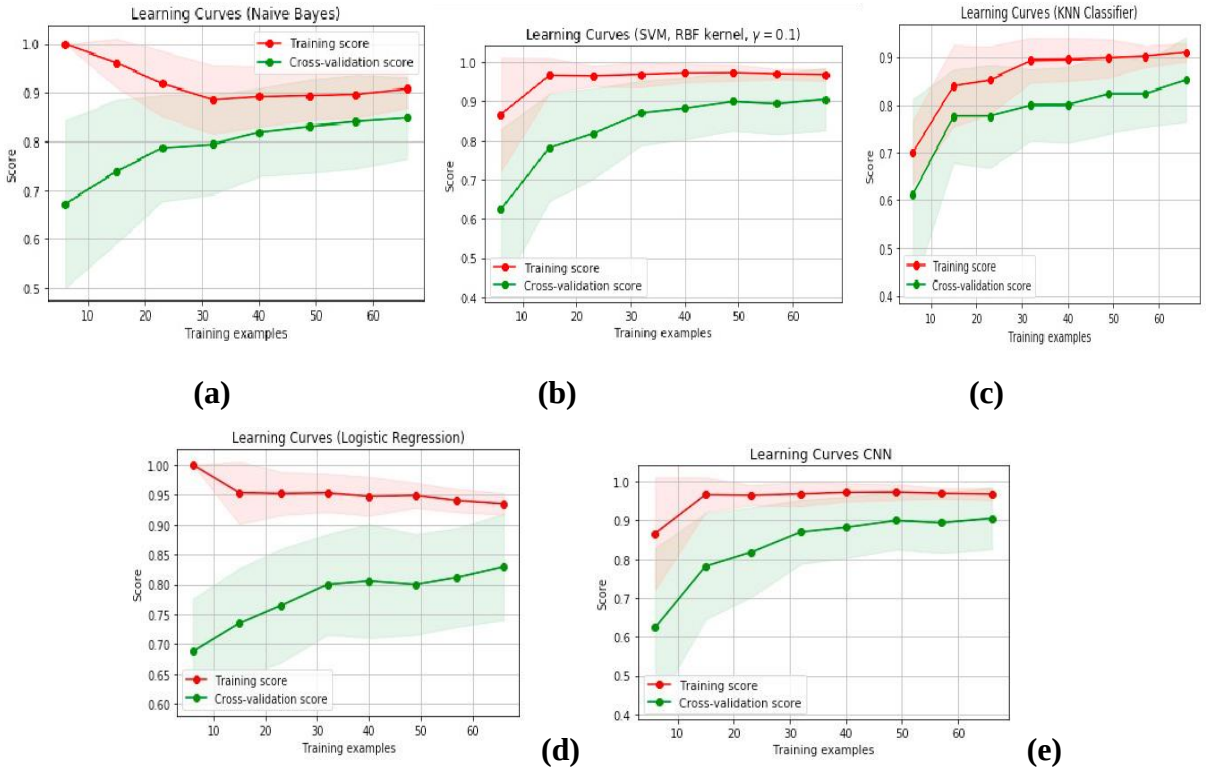

Figure 8: Learning Curve of (a) Naïve Bayes; (b) SVM; (c) KNN; (d) Logistic Regression; (e) CNN for adolescent's dataset.

CONCLUSION

In this work, detection of Autism Spectrum Disorder was attempted using various machine learning and deep learning techniques. Various performance evaluation metrics were used to analyze the performance of the models implemented for ASD detection on non-clinical dataset from three sets of age groups viz. Child, Adolescents and the Adult. When comparing the result with another recent study [3] on this problem got a better result of the CNN classifier instead of SVM with including all its features attributes after handling missing values. In this work after handling missing value, both the SVM and CNN based models show the same accuracy of prediction of about 98.30 % for ASD Child dataset. However for the remaining two other datasets, the CNN based model was able to achieve highest accuracy result than all the other considered model building techniques, These results strongly suggest that a CNN based model can be implemented for detection of Autism Spectrum Disorder instead of the other conventional machine learning classifier suggested in earlier researches.

Table 7: Comparison of results with existing methods [3] on Autistic Spectrum Disorder Screening Data for Children.

Model	Accuracy before handling missing values	Accuracy after handling missing values
Support Vector Machine	97.95	98.30
Artificial Neural Network	97.60	98.30
Convolutional Neural Network	Not implemented	98.30

REFERENCES

1. Thabtah, Fadi. "Machine learning in autistic spectrum disorder behavioral research: A review and ways forward." *Informatics for Health and Social Care*, 2018, 1-20.
2. Thabtah, Fadi, Firuz Kamalov, and Khairan Rajab. "A new computational intelligence approach to detect autistic features for autism screening." *International Journal of Medical Informatics*, 117: 112-124, 2018.
3. Vaishali, R., and R. Sasikala. "A machine learning based approach to classify Autism with optimum behaviour sets." *International Journal of Engineering & Technology*, 7(4): 18, 2018.
4. Constantino, John N., Patricia D. Lavesser, Y. I. Zhang, Anna M. Abbacchi, Teddi Gray, and Richard D. Todd. "Rapid quantitative assessment of autistic social impairment by classroom teachers." *Journal of the American Academy of Child & Adolescent Psychiatry*, 46(12): 1668-1676, 2007.
5. Bone, Daniel, Matthew S. Goodwin, Matthew P. Black, Chi-Chun Lee, Kartik Audhkhasi, and Shrikanth Narayanan. "Applying machine learning to facilitate autism diagnostics: pitfalls and promises." *Journal of Autism and Developmental Disorders*, 45(5): 1121-1136, 2015.
6. Wall, Dennis Paul, J. Kosmicki, T. F. Deluca, E. Harstad, and Vincent Alfred Fusaro. "Use of machine learning to shorten observation-based screening and diagnosis of autism." *Translational Psychiatry*, 2(4): e100, 2012.
7. Wall, Dennis P., Rebecca Dally, Rhiannon Luyster, Jae-Yoon Jung, and Todd F. DeLuca. "Use of artificial intelligence to shorten the behavioral diagnosis of autism." *PLOS ONE*, 7(8): e43855, 2012.
8. Thabtah, Fadi. "Autism spectrum disorder screening: machine learning adaptation and DSM-5 fulfillment." In *Proceedings of the 1st International Conference on Medical and Health Informatics*, pp. 1-6, ACM, 2017.
9. Bone, Daniel, Chi-Chun Lee, Matthew P. Black, Marian E. Williams, Sungbok Lee, Pat Levitt, and Shrikanth Narayanan. "The psychologist as an interlocutor in autism

spectrum disorder assessment: Insights from a study of spontaneous prosody." *Journal of Speech, Language, and Hearing Research*, 57(4): 1162-1177, 2014.

10. Thabtah, Fadi. "ASD Tests. A mobile app for ASD screening." *www.asdtests.com* [accessed December 20th, 2017].
11. Li, Baihua, Arjun Sharma, James Meng, Senthil Purushwalkam, and Emma Gowen. "Applying machine learning to identify autistic adults using imitation: An exploratory study." *PLOS ONE*, 12(8): e0182652, 2017.
12. Thabtah, Fadi Fayez. "Autistic Spectrum Disorder Screening Data for Adults." https://archive.ics.uci.edu/ml/machine-learning-databases/00426/, 2017.
13. Mythili, M. S., and AR Mohamed Shanavas. "A study on Autism spectrum disorders using classification techniques." *International Journal of Soft Computing and Engineering (IJSCE)*, 4: 88-91, 2014.
14. Kosmicki, J. A., V. Sochat, M. Duda, and D. P. Wall. "Searching for a minimal set of behaviors for autism detection through feature selection-based machine learning." *Translational Psychiatry*, 5(2): e514, 2015.
15. Thabtah, Fadi Fayez. "Autistic Spectrum Disorder Screening Data for Children." https://archive.ics.uci.edu/ml/machine-learning-databases/00419/, 2017.
16. Thabtah, Fadi Fayez. "Autistic Spectrum Disorder Screening Data for Adolescents." https://archive.ics.uci.edu/ml/machine-learning-databases/00420/, 2017.
17. John, George H., and Pat Langley. "Estimating continuous distributions in Bayesian classifiers." In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pp. 338-345, Morgan Kaufmann Publishers Inc., 1995.
18. Quinlan, J. R. "Program for machine learning." C4.5, 1993.
19. Keerthi, S. Sathiya, Shirish Krishnaj Shevade, Chiranjib Bhattacharyya, and Karuturi Radha Krishna Murthy. "Improvements to Platt's SMO algorithm for SVM classifier design." *Neural Computation*, 13(3): 637-649, 2001
20. Platt, J. C. "Fast training of Support Vector Machines using sequential minimal optimization." In *Advances in Kernel Methods*, pp. 185-208, 1999.
21. Pal, Sankar K., and Sushmita Mitra. "Multilayer perceptron, fuzzy sets, and classification." *IEEE Transactions on Neural Networks*, 3(5): 683-697, 1992.
22. Aha, David W., Dennis Kibler, and Marc K. Albert. "Instance-based learning algorithms." *Machine Learning*, 6(1): 37-66, 1991.

23. Gotham, Katherine, Susan Risi, Andrew Pickles, and Catherine Lord. "The Autism Diagnostic Observation Schedule: revised algorithms for improved diagnostic validity." *Journal of Autism and Developmental Disorders*, 37(4): 613, 2007.
24. Masood, Sarfaraz, Abhinav Rai, Aakash Aggarwal, Mohammad Najmud Doja, and Musheer Ahmad. "Detecting distraction of drivers using convolutional neural network." *Pattern Recognition Letters*, 2018.
25. Masood, Sarfaraz, Adhyan Srivastava, Harish Chandra Thuwal, and Musheer Ahmad. "Real-time sign language gesture (word) recognition from video sequences using CNN and RNN." In *Intelligent Engineering Informatics*, pp. 623-632, Springer, Singapore, 2018.
26. Masood, Sarfaraz, Harish Chandra Thuwal, and Adhyan Srivastava. "American Sign Language character recognition using convolution neural network." In *Smart Computing and Informatics*, pp. 403-412, Springer, Singapore, 2018.