engineering proceedings

Proceeding Paper

Machine Learning-Based Classification of Autism Spectrum Disorder across Age Groups [†]

Resmi Karinattu Reghunathan [1,*], Poornima Nanjagoundan Palayam Venkidusamy [1], Raju Gopalakrishna Kurup [1], Bindu George [2] and Neetha Thomas [3]

[1] Department of Computer Science, CHRIST University, Bengaluru 560029, India; poornima.nv@christuniversity.in (P.N.P.V.); raju.g@christuniversity.in (R.G.K.)
[2] Department of Computer Science, Nirmala College Muvattupuzha, Muvattupuzha 686661, India; tessgeorgesh@gmail.com
[3] Department of Computer Science, Santhigiri College, Vazhithala, Thodupuzha 685583, India; neethathomas@santhigiricollege.com
[*] Correspondence: resmi.kr@christuniversity.in

[†] Presented at the 2nd Computing Congress 2023, Chennai, India, 28–29 December 2023.

Abstract: Autism Spectrum Disorder (ASD) is a complex neurodevelopmental condition that has gained significant attention in recent years due to its increasing prevalence and profound impact on individuals, families, and society as a whole. In this study, we explore the use of different machine learning classifiers for the accurate detection of ASD in children, adolescents, and adults. Furthermore, we conduct feature reduction to identify key features contributing to ASD classification within each age group using Cuckoo Search Algorithm. Logistic Regression has the highest accuracy compared to the other two models.

Keywords: Autism Spectrum Disorder; machine learning; classifiers; cuckoo search

check for updates

Citation: Reghunathan, R.K.; Venkidusamy, P.N.P.; Kurup, R.G.; George, B.; Thomas, N. Machine Learning-Based Classification of Autism Spectrum Disorder across Age Groups. *Eng. Proc.* **2024**, *62*, 12. https://doi.org/10.3390/engproc2024062012

Academic Editors: Geetha Ganesan, Xiaochun Cheng and Valentina Emilia

Published: 15 March 2024

1. Introduction

Autism Spectrum Disorder (ASD) is a neurodevelopmental disorder characterized by a variety of behavioral and developmental abnormalities. A person with ASD will experience lifelong effects on their ability to interact and communicate with others [1]. Since its symptoms frequently appear in the first two years of life, autism is considered as a "behavioral disease" and can be diagnosed at any age. Experts claim that the ASD issue begins in childhood and lasts through adolescence and old age. The disease ASD also has an impact on how the human brain develops. Typically, a person with ASD cannot interact socially or have a discussion with others.

The effects of ASD on a person's life typically last throughout the rest of their lives. It is important to remember that this illness could occur as a result of both hereditary and environmental causes. This condition's symptoms can appear at almost three years of age and may continue for the rest of one's life. Although a patient with this condition cannot be totally cured, the effects can be temporarily reduced if the signs are caught early. Researchers believe that ASD may be linked to human genetics, though they have not definitively identified the precise underlying factors.

The major goal of this research is to improve the diagnosis of autism by developing a machine learning system that makes use of various machine learning algorithms to create an autism predictive model with the highest level of accuracy. The solution is to provide a very accurate predictive model that can predict whether an individual (adolescent, child, or adult) has ASD or not. The goal is to employ a standard approach for diagnosing autism and convert it into a machine learning model that can use medical data to generate predictions and observations, and lead to better solutions for identifying ASD as early as possible in the future.

2. Literature Review

This Section presents some of the studies related to ASD using machine learning. In the study [2], the authors used the Autism Spectrum Questions (AQ) to create models for classifying ASD. They employed Least Absolute Shrinkage and Selection Operator (LASSO) and Chi-square to identify the most relevant features from the AQ dataset. Subsequently, they applied three supervised machine learning algorithms, Logistic Regression (LR), Random Forest, and K-Nearest Neighbors, utilizing K-fold cross-validation for robust evaluation. The results indicated that Logistic Regression achieved the highest accuracy rate, reaching 97.541%. This impressive performance was achieved by selecting 13 essential features-based Chi-square method.

Deshpande et al. [3] used functional MRI (fMRI) to examine how individuals with autism and normally developing controls differ in terms of the causal influence of one brain area on another (effective connectivity) during Theory-of-Mind (ToM) tasks. The participants include 15 high-functioning people with autism and 15 typically developing people who served as controls. The SVM classifier distinguished between people with autism and typically developing controls, with a maximum accuracy of 95.9%.

Duda et al. [4] investigated the potential of machine learning in accurately and swiftly differentiating between Attention Deficit Hyperactivity Disorder (ADHD) and Autism Spectrum Disorder (ASD) using data from the Social Responsiveness Scale. The study uses 65 behavioral features with the maximum accuracy of 96.5%. A feature selection wrapper using swarm intelligence to perform ASD diagnosis on the UCI ML repository is presented in [5]. The study is based on the hypothesis that an ML model can achieve superior classification accuracy with a minimal subset of features. The results support this idea, showing that only 10 of the 21 essential traits in the ASD dataset are necessary to distinguish between patients with ASD and those without it. Surprisingly, using these ideal feature subsets, the technique produces an accuracy range of 92.12% to 97.95% on average.

In [6], the early signs of ASD in children are identified. The experiment was conducted on UCI data of children using different classifiers, and the results showed that Logistic Regression achieved the highest accuracy among the models, offering a promising approach to aid in the early detection of ASD. Convolutional Neural Network (CNN)-based prediction models were applied to UCI data in [7]. After addressing the missing data and applying machine learning models, the results highlight the superiority of Convolutional Neural Network (CNN)-based prediction models, achieving remarkably high accuracy rates of 98.30%, 96.88%, and 99.53% for ASD screening in children, adolescent, and adult populations, respectively. In [8], federal learning is applied to achieve 98% and 81% accuracy for ASD child and adult datasets, respectively. A detailed review about ASD is presented in [9,10].

3. Proposed Methodology

The proposed method is shown in Figure 1 and includes data preprocessing, feature reduction, model evaluation, and ASD prediction.

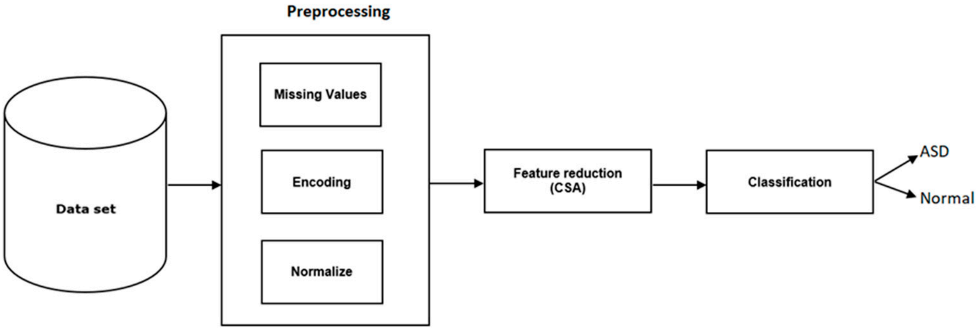

Figure 1. Proposed system for autism prediction.

3.1. Preprocessing

The autism dataset is first preprocessed to remove missing values and encode categorical attributes. The dataset contains some missing values in individual features especially in terms of gender, country, ethnicity, etc., and the different types of attributes. Preprocessing is applied on the dataset for handling missing values and categorical attributes. Binary label encoding is used for four features in the dataset. For example, the attributed gender is either male or female. This is converted to numeric value 0 for a female and 1 for a male. The dataset includes data collected from 89 countries. Each country is represented in alphabetic order from 1 to 89, and the missing country in the dataset is represented as 90. The dataset includes a total of 14 ethnicity and is represented by 14 values used in an alphabetic order, and the missing value is represented as 15. The preprocessing step applied in the dataset is shown in Table 1, and the data before and after preprocessing is shown in Table 2.

Table 1. Rule applied for missing and encode categorical values.

Attribute Gender		**Attribute Born with Jaundice**		**Attribute Family Member with PDD**		**Attribute Usage of Screening App before**		**Attribute Country of Residence**		**Attribute Ethnicity**		**Attribute Who is Completing the Test**	
String	String	String	Value	String	Value	String	Value	String	Value	String	Value	String	Value
M (male)	1	Yes	1	Yes	1	Yes	1	Afghanistan	1	Asian	1	Health care professional	1
F (female)	0	No	0	No	0	No	0	Albania	2	Black	2	Others	2
								?	90	?	15	?	6

? is represented as missing value in the databse.

Table 2. Data before and after preprocessing.

Data before Preprocessing																				
1	1	0	1	0	0	1	1	0	1	35	m	Asian	yes	yes	Albania	yes	6	‘18 and more’	Self	yes
1	1	0	0	1	1	0	1	0	0	40	f	?	no	no	Egypt	no	2	‘18 and more’	?	NO
Data after Preprocessing																				
1	1	0	1	0	0	1	1	0	1	35	1	1	1	1	1	a	6	1	5	1
1	1	0	0	1	1	0	1	0	0	40	0	15	0	0	10	0	2	1	6	0

? is represented as missing value in the databse.

3.2. Cuckoo Search Algorithm (CSA)

CSA algorithm is used for feature reduction. Using the Cuckoo Search Algorithm (CSA) for feature selection in the context of Autism Spectrum Disorder (ASD) research can be a promising approach to improve the accuracy and efficiency of data analysis. The algorithm for cuckoo search is provided below. The parameters used for cuckoo search include population = 20, stopping criterion = 100, probability of abandoning a nest = 0.25, and scale factor for leavy flight = 0.6 (Algorithm 1).

Algorithm 1 Cuckoo Search Algorithm

```
1. Get preprocessed autism dataset as input
2. Initialize population of solutions (nests)
3. Evaluate fitness of each nest
4. Choose a cuckoo randomly from the autism Dataset
5. Generate a new solution (features) by modifying the cuckoo's solution
6. Evaluate the fitness of the new combination of features
7. Implement the CSA to search for optimal feature subsets.
8. while (stopping criterion not met) Repeat the following steps:
    i. Levy Flight Generation: Use Levy flights to generate new solutions.
    ii. Evaluate New Solutions: Assess the fitness of the new solutions.
    iii. Replace Solutions: Replace less fit solutions with better ones.
    iv. Abandon Solutions: Occasionally replace some solutions with new random solutions
        (exploration).
    v. Evaluate fitness of nests
9. return best solution found
```

After feature reduction, the output label (ASD or normal) is predicted using different classification methods. Each classifier's accuracy is evaluated and compared.

4. Experimental Results and Discussion

The performance of the proposed approach was evaluated on ASD datasets from the UCI database and implemented in MATLAB.

4.1. Dataset Description

In this research, three publicly accessible ASD datasets from the UCI database were utilized, which are relevant for the clinical diagnosis of ASD at various ages. Dataset description is shown in Table 3. Children (age between 4 and 11 years), adolescents (age between 12 and 17 years), and adults (age above 18+ years) are the three age groups represented in the datasets. The dataset includes a total of 21 features with 10 behavioral features and 10 individual features. The individual features are related to personal information which includes age, ethnicity, gender, born with jaundice, country, etc., and the behavioral features are related to the screening questions. The data were collected by using a survey in nations across the world through a mobile application called ASD Tests.

Table 3. ASD dataset.

Sl.No	Name of Dataset	No. of Features/Attributes Including Class Label	Missing Values	Number of Instances/Records	Type of Attributes
1	ASD adult dataset [11]	21	Yes	704	Categorical, binary, and continuous
2	ASD child dataset [12]	21	Yes	292	Categorical, binary, and continuous
3	ASD adolescent dataset [13]	21	Yes	104	Categorical, binary, and continuous

4.2. Classification Methods

The classification model classifiers such as Logistic Regression (LR), K-Nearest Neighbors (KNN), and Support Vector Machine (SVM) are used for classification.

4.2.1. Logistic Regression

It is one of the most popular machine learning algorithms used primarily for binary classification tasks. It uses a logistic function to find the optimal curve to fit the data points.

4.2.2. K-Nearest Neighbors (KNN)

This algorithm is a straightforward and intuitive machine learning method employed for both classification and regression purposes. It operates as a non-parametric, instance-based approach that makes predictions on how closely data points in a particular dataset resemble one another. The experiment was conducted on different K values, and the maximum accuracy was obtained with K = 10.

4.2.3. Support Vector Machines (SVM)

The primary utilization of Support Vector Machines (SVM) is for both multiclass and binary classification tasks. Its core objective is to identify an optimal decision boundary that effectively segregates data points into distinct classes, all the while maximizing the separation margin between these classes.

4.3. Result and Discussion

We applied three ML models for evaluation. Accuracy is calculated for all the models using the following equation:

$$\text{Accuracy} = \frac{\text{True Positive}(\text{TP}) + \text{True Negative }(\text{TN})}{\text{True Positive}(\text{TP}) + \text{False Positive}(\text{FP}) + \text{True Negative}(\text{TN}) + \text{False Negative }(\text{FN})} \quad (1)$$

The accuracy of various ML models on the ASD datasets is shown in Table 4 and Figure 2. According to the results, compared to other models in the available dataset, linear regression has the highest accuracy. Table 5 provides a comparative analysis with the prior research concerning ASD.

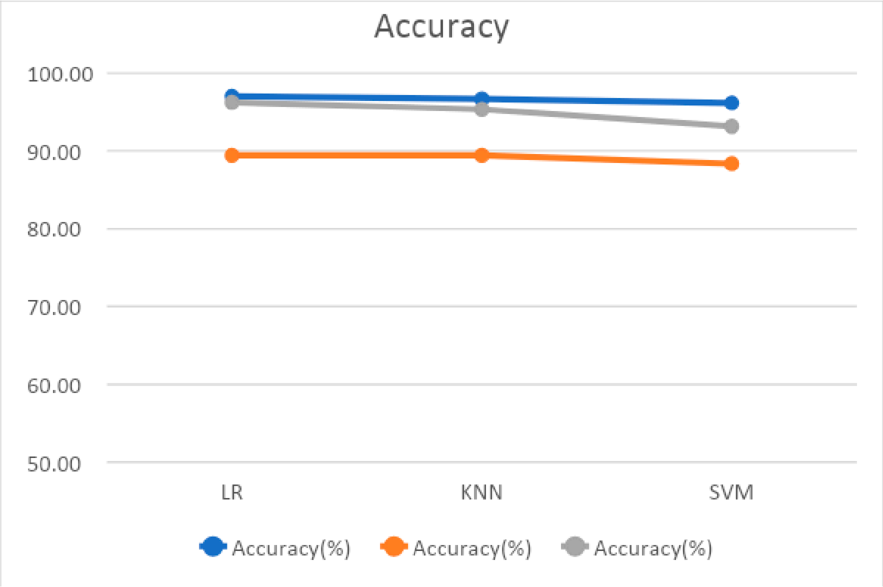

Figure 2. Accuracy using different classifiers.

Table 4. Accuracy of different age groups of ASD dataset.

Datasets		LR	KNN	SVM
Adult	Accuracy	97.01	96.87	96.16
	Precision	0.996	0.970	0.973
		0.908	0.966	0.931
	Recall	0.963	0.988	0.975
		0.989	0.915	0.926

Table 4. *Cont.*

Datasets		LR	KNN	SVM
Adolescent	Accuracy	89.42	89.42	88.36
	Precision	0.857	0.875	0.837
		0.919	0.906	0.909
	Recall	0.878	0.854	0.868
		0.905	0.921	0.904
Child	Accuracy	96.23	95.32	93.15
	Precision	0.979	0.942	0.958
		0.945	0.966	0.906
	Recall	0.947	0.942	0.907
		0.979	0.962	0.957

Table 5. Comparison with the existing methods.

Reference	Method	Dataset	Accuracy
[2]	LASSO and Chi-square	AQ Dataset	97.54%
[3]	Functional MRI	Functional MRI images	95.90%
[4]	Behavioral features	data from the Social Responsiveness Scale	96.50%
[5]	Minimal subset of features	UCI dataset	92.12% To 97.75%
[6]	CNN	ASD child	98.30%
		ASD adolescent	96.88%
		ASD adult	99.53%
[8]	Federal learning	ASD child	98%
		ASD adult	81%
Proposed Method	Feature reduction using cuckoo search	ASD child	96.23%
		ASD adolescent	89.42%
		ASD adult	97.01%

Since the authors used their different methods and datasets, the results in Table 5 are not comparable. This research can be enhanced by the usage of deep learning techniques, more datasets, and more features. Since the authors of [6] used CNN, their achieved accuracy is very high.

5. Conclusions

In this study, three publicly available ASD screening datasets offered by the UCI machine learning repository were used to detect Autism Spectrum Disorder (ASD) using several ML models. This study evaluated different machine learning models for the accurate and robust classification of ASD across various age groups, from early childhood to adulthood. The findings and insights from this research contribute to a deeper understanding of ASD diagnosis, offering potential benefits to clinicians, researchers, and individuals on the autism spectrum. In order to increase the system's robustness and overall performance, future research should concentrate on large datasets, enhancing feature selection methods, and using deep learning strategies that combine CNNs and classification.

Author Contributions: Conceptualization, R.K.R. and B.G.; methodology, B.G. and R.K.R.; software, N.T. and P.N.P.V.; validation, R.K.R., R.G.K. and P.N.P.V.; formal analysis, R.K.R.; investigation, R.K.R.; resources, B.G.; data curation, N.T.; writing—original draft preparation, R.K.R. and N.T.; writing—review and editing, R.G.K. and R.K.R.; visualization, B.G.; supervision, R.G.K.; All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are available at [11–13].

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Kang, J.; Han, X.; Song, J.; Niu, Z.; Li, X. The identification of children with autism spectrum disorder by SVM approach on EEG and eye-tracking data. *Comput. Biol. Med.* **2020**, *120*, 103722. [CrossRef] [PubMed]
2. Abdullah, A.A.; Rijal, S.; Dash, S.R. Evaluation on machine learning algorithms for classification of Autism Spectrum Disorder (ASD). *J. Phys. Conf. Ser.* **2019**, *1372*, 012052. [CrossRef]
3. Deshpande, G.; Libero, L.E.; Sreenivasan, K.R.; Deshpande, H.D.; Kana, R.K. Identifcation of neural connectivity signatures of autism using machine learning. *Front. Hum. Neurosci.* **2013**, *7*, 670. [CrossRef]
4. Duda, M.; Ma, R.; Haber, N.; Wall, D.P. Use of machine learning for behavioral distinction of autism and ADHD. *Transl. Psychiatry* **2016**, *6*, e732. [CrossRef]
5. Vaishali, R.; Sasikala, R. A machine learning based approach to classify Autism with optimum behavior sets. *Int. J. Eng. Technol.* **2018**, *7*, 18.
6. Vakadkar, K.; Purkayastha, D.; Krishnan, D. Detection of Autism Spectrum Disorder in Children Using Machine Learning Techniques. *SN Comput. Sci.* **2021**, *2*, 386. [CrossRef] [PubMed]
7. Raj, S.; Masood, S. Analysis and Detection of Autism Spectrum Disorder Using Machine Learning Techniques. *Procedia Comput. Sci.* **2020**, *167*, 994–1004. [CrossRef]
8. Farooq, M.S.; Tehseen, R.; Sabir, M.; Atal, Z. Detection of autism spectrum disorder (ASD) in children and adults using machine learning. *Sci. Rep.* **2023**, *13*, 9605. [CrossRef] [PubMed]
9. Hirota, T.; King, B.H. Autism Spectrum Disorder: A Review. *JAMA* **2023**, *329*, 157–168. [CrossRef] [PubMed]
10. Salari, N.; Rasoulpoor, S.; Rasoulpoor, S.; Shohaimi, S.; Jafarpour, S.; Abdoli, N.; Khaledi-Paveh, B.; Mohammadi, M. The global prevalence of autism spectrum disorder: A comprehensive systematic review and meta-analysis. *Ital. J. Pediatr.* **2022**, *48*, 112. [CrossRef] [PubMed]
11. Thabtah, F. Autism Screening Adult. *UCI Mach. Learn. Repos.* **2017**. [CrossRef]
12. Thabtah, F. Autistic Spectrum Disorder Screening Data for Children. *UCI Mach. Learn. Repos.* **2017**. [CrossRef]
13. Tabtah, F. Autistic Spectrum Disorder Screening Data for Adolescent. *UCI Mach. Learn. Repos.* **2017**. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.