Autism spectrum disorder classification using machine learning with factor analysis

Disha Devidas Nayak[1,2], Seema Shedole[1,3], Archana Mathur[4]
[1]Visvesveraya Technological University, Belagavi, India
[2]Department of Artificial Intelligence and Machine Learning, NMAM Institute of Technology, Nitte Deemed to be University, Udupi, India
[3]Department of Computer Science in Ramaiah Institute of Technology, Bangalore, India
[4]Department of Artificial Intelligence and Data Science, Nitte Meenakshi Institute of Technology, Bangalore, India

Article Info

Article history:

Received Nov 21, 2023
Revised Feb 5, 2025
Accepted Mar 15, 2025

Keywords:

Autism spectrum disorder
Correlation analysis
Factor analysis
Machine learning
Pearson correlation

ABSTRACT

Due to the complexity and heterogeneity of autism spectrum disorder (ASD), diagnosis and categorization have attracted a lot of interest. To improve the robustness of ASD classification across the toddler age group, this work proposes an integrated strategy that integrates machine learning approaches with factor analysis and correlation validation. Benchmark dataset representing toddlers used to test this strategy's efficiency. To first find the latent variables behind the ASD features in each dataset, factor analysis is used. We intend to capture the shared variance between variables and lower the dimensionality of the initial feature space by identifying these latent components. The subsequent machine-learning classification models used the retrieved components as input features. To validate the categorization results, correlation analyses were carried out in addition to factor analysis. The associations between the latent components discovered by factor analysis and the diagnostic labels were examined using Pearson correlation, a measure of linear association. The results highlight the method's potential to improve diagnostic precision and shed light on the intricate connections between characteristics and diagnostic labels on the autism spectrum for toddlers.

Corresponding Author:

Disha Devidas Nayak
Department of Artificial Intelligence and Machine Learning, NMAM Institute of Technology
Nitte Deemed to be University
Nitte, India
Email: disha.dn.2@gmail.com

1. INTRODUCTION

Owing to the complex and varied character of the illness, the diagnosis and classification of autism spectrum disorder (ASD) have attracted a great deal of attention. ASD exhibits a wide range of symptoms and changes, making it difficult to classify the disorder accurately and consistently across age groups [1], [2]. ASD poses a formidable challenge in its diagnosis and classification owing to its intricate and heterogeneous nature. The condition encompasses a broad array of symptoms spanning social interaction, communication, behavior, and sensory processing domains, thereby complicating efforts to accurately and consistently categorize it across age cohorts [1]. ASD is typified by deficits in social reciprocity, manifesting as difficulties in discerning social cues, maintaining eye contact, interpreting facial expressions, and cultivating interpersonal relationships. Afflicted individuals often exhibit a propensity towards solitary pursuits, alongside a notable impediment in sharing emotions or interests with others [2]. Communication impairments in ASD range from delayed language acquisition to outright verbal autism. Such behaviors often serve as mechanisms for self-regulation or sensory modulation. These deficits significantly impact academic performance, adaptive functioning, and autonomy in daily activities [3]. While some individuals may exhibit amelioration in symptomatology with targeted interventions, others may endure persistent challenges into adulthood, necessitating ongoing support and accommodations [4]. Classifying ASD using machine learning models presents several challenges due to the inherent complexity and heterogeneity of the condition. Moreover, identifying relevant features or variables for ASD classification is non-trivial. Selecting the most discriminative features while avoiding overfitting or underfitting is a significant challenge, especially given the high dimensionality of many ASD datasets. Data collected for ASD research may vary in quality and reliability, leading to noise and confounding factors in the dataset. Despite the complexities involved, machine learning holds great promise for advancing our understanding of ASD and facilitating early diagnosis and personalized interventions. The research problem we aim to address in the manuscript encompasses the following inquiries:

- Can latent features be extracted from existing ASD datasets?
- Are these extracted features sufficiently reliable to enhance the performance of machine learning algorithms in classification tasks?
- Can empirical evidence from the cross-correlation of features substantiate the research findings?

To enhance the consistency and accuracy of ASD classification, this study proposes an integrated methodology leveraging machine learning techniques alongside factor analysis and feature correlation. The toddler dataset serves as the basis for evaluating the effectiveness of this combined approach, showcasing improved classification performance across diverse age cohorts [5]. Factor analysis, a sophisticated statistical method in data analysis [6], discerns latent variables underlying complex patterns of correlations among observed variables. To validate the correlations identified during factor analysis, Pearson correlation analysis is employed as a validation tool [7]. This ensures the robustness of the correlations formed and reinforces the reliability of the factor analysis results. The integration of machine learning, factor analysis, and correlation validation signifies a comprehensive analysis to enhance diagnostic accuracy. By these techniques, the study provides a framework for ASD classification, promising improved understanding and detection of the disorder.

The immediate need to address the complex issues raised by the diagnosis and categorization of ASD serves as the driving force behind the effort [8]. A special set of challenges. Traditional diagnostic techniques frequently struggle to account for this diversity, potentially resulting in incorrect diagnoses and delays in necessary intervention. It is commonly recognized that machine learning techniques have the power to completely transform medical diagnostics, including the classification of ASD [9]. However, applying machine learning to ASD classification directly without considering the underlying difficulties may lead to less-than-ideal results [10]. Factor analysis provides a method for addressing the shared variance and multidimensionality seen in ASD data. Factor analysis can give the data a more meaningful representation by finding latent components within the features, potentially improving classification precision. Additionally, the reliability and robustness of the classification process are improved with the addition of correlation validation techniques like Pearson correlation.

The goal of this research is to close the knowledge gap between machine learning methods and the complexity of ASD categorization. This research aims to achieve two crucial goals by proposing an integrated strategy that combines machine learning, Factor analysis, and correlation validation: first, to improve the accuracy of ASD classification across different age groups, and second, to shed light on the complex relationships between the diagnostic labels and the various features that define ASD. In the end, the results of this study not only enhanced the field of ASD diagnosis but also led to a better comprehension of the underlying traits of the illness. Our contribution is:

- Analyze the correlation between features of ASD patients.
- Investigate the working of factor analysis for extracting crucial features of the autistic patients.
- Explore the working of machine learning algorithms on the derived features.

The remainder of the manuscript is structured as follows: We begin with a review of relevant literature concerning previous studies on Autistic patients and their associated characteristics. Section 3 outlines the methodology employed to conduct factor analysis on the features of Autistic patients. In section 4, we detail the experimentation conducted on the ASD dataset, incorporating factor analysis and its impact on the classification of ASD patients. Finally, section 5 concludes our manuscript.

2. RELATED WORK

Different machine-learning methods were used in this work by Uddin *et al.* [7] to identify ASD across a range of age groups. The objective of the study was to compare the performance of various classifiers, including random forest (RF) [11], support vector machine (SVM) [12], multinomial naive Bayes (MNB) [13], gaussian naive Bayes (GNB), Bernoulli naive Bayes (BNB), and quadratic discriminant analysis (QDA), to determine the most effective method for ASD identification. Another work by Qureshi *et al.* [14] helps early detection and intervention, successfully reducing the escalation of autism-related difficulties, and lowering the price tag connected to a delayed diagnosis. A solution suggested by [15] of using computer aids for accurate predictions outperforms the existing methods. Its ability to accommodate prediction for many age groups and provide a thorough comparative review of various machine learning algorithms distinguishes it from other approaches and highlights its potential to considerably enhance the field of autism screening and diagnosis [16]. Research by Hyde *et al.* [17] offers a thorough literature review with a specific focus on the use of supervised machine learning algorithms to explore ASD. The results highlight the significant benefits and usefulness of using supervised machine learning in ASD research. The research also draws attention to some drawbacks, including the need for labeled data, how model complexity affects interpretability, and the processing requirements imposed by sophisticated machine learning models.

Vakadkar *et al.* [18] deals with the time-consuming process of evaluating behavioral characteristics of ASD, which is made more challenging by overlapping symptoms and the absence of a quick and reliable diagnostic test. The authors suggest an automated ASD prediction approach that uses basic behavior sets taken from diagnosis datasets to address this issue. Logistic regression (LR) [19] demonstrated the highest level of accuracy in predicting ASD among the five tested models. The model's performance was limited by the insufficient number of samples in the dataset. A critical study by Zamit *et al.* [20] reveals that artificial intelligence (AI) is being used exhaustively for ASD screening due to the significant rise in the frequency of ASD, based on a dataset of 2090 papers from the Scopus database. This study undertakes a bibliometric analysis of AI-powered ASD screening research. The study's time spans from 2010 to 2021. A 23-fold rise in scientific production from 2010 to 2021, with the bulk (62.54%) published between 2019 and 2021, as revealed by the investigation, which also shows a spectacular annual growth rate of 33.05%. The costs associated with ASD are high, hence early identification is necessary to reduce these effects. The research emphasizes the need for quick and accurate ASD screening techniques to help patients, healthcare providers, and individuals make educated choices about the clinical diagnosis as proposed by the authors Thabtah *et al.* [19]. This article presents a novel machine-learning architecture designed specifically for autism screening in adolescents and adults due to the restricted screening-associated datasets, which mostly focus on genetics.

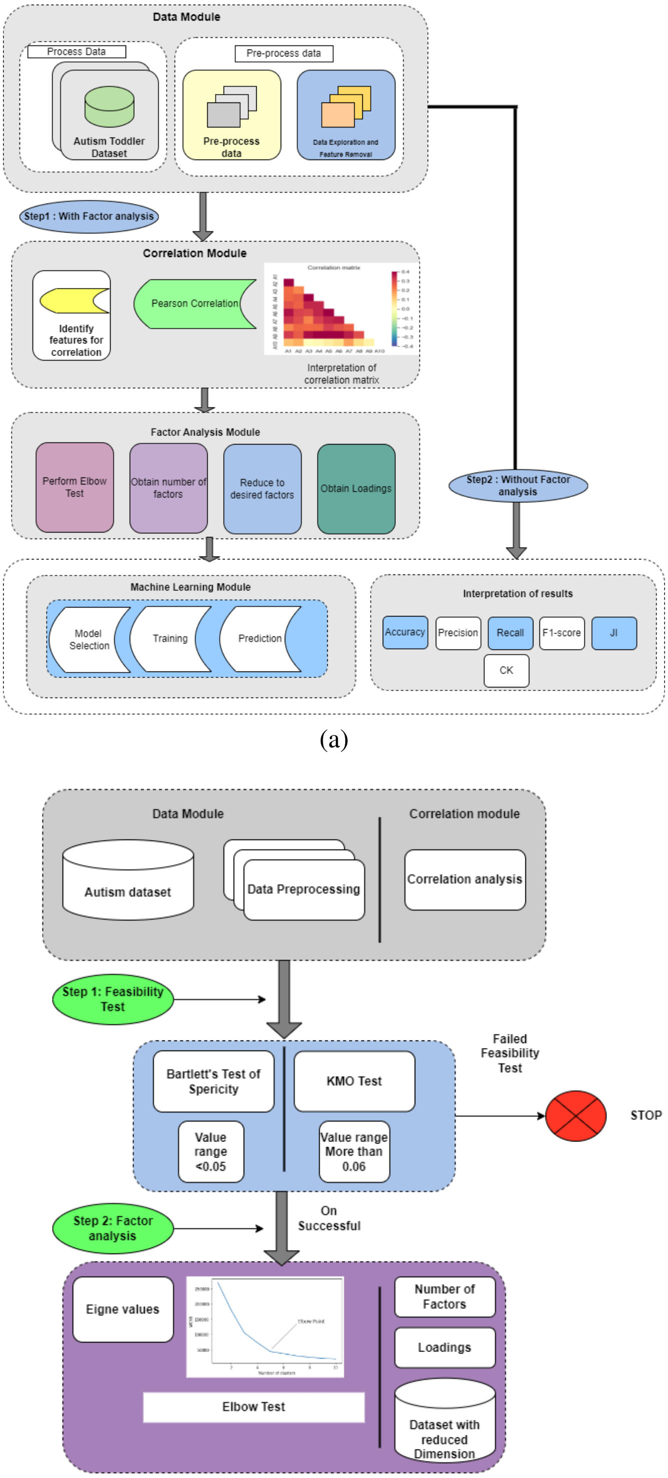
According to Mareeswaran and Selvarajan [21] a deep learning model is used for predicting ASD using sentiment analysis of social media text collected from Twitter. The proposed model is a hybrid bidirectional long short-term memory (BiLSTM) with an attention mechanism, which employs n-gram feature extraction and Adam optimizer to improve the prediction and training speed. Successfully discriminating patterns in textual content that the model associates with common behaviors seen in individuals on the autism spectrum, the this achieves a validation accuracy 98% which is an improvement over other models such as convolutional neural network (CNN) and classic long short-term memory (LSTM). It shows that this method could help make the correct clinical decision sooner by using deep learning and sentiment analysis for early ASD detection.

3. METHOD

The methodology diagram is reflected in Figure 1. As illustrated in Figure 1(a), the proposed architecture leverages correlation analysis and factor analysis. Figure 1(b) shows the steps for the factor analysis alone. Through this process, we extract essential factors that will subsequently inform the application of various machine learning and deep learning algorithms, facilitating comprehensive performance analysis. Steps:

- We obtain our dataset from the UCI repository
- The dataset is analyzed for any missing values and categorical encoding and feature scaling have been done.
- Applied Pearson correlation analysis to observe, if there exists any correlation between data
- Identify the correlation score of every feature with another
- Apply factor analysis which involves the removal of correlated features using the factor analysis approach

– Machine learning algorithms are applied to the new dataset to analyze its performance.



(a)

(b)

Figure 1. The detailed methodologies: (a) the overall architecture of the research; and (b) the factor analysis on the toddler dataset

3.1. Correlation analysis

The correlation analysis [22] is a statistical measure to evaluate the strength and direction of one variable concerning another variable. Pearson correlation analysis is used to identify the relationship among variables. Given two variables X and Y with n data points, the Pearson correlation coefficient (r) is calculated in (1):

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \tag{1}$$

where r is the Pearson Correlation coefficient, x_i= i^{th} feature vector, $\bar{x}$= mean of values in x variable, y_i= y variable sample, and $\bar{y}$=mean of values in y variable

3.2. Factor analysis

Factor analysis as described in Figure 1(b), is a statistical method to model the relationships between observed variables and underlying latent factors. It aims to uncover the latent structure or dimensions that explain the observed correlations or covariances among the variables. Given a matrix X of observed variables with dimensions n (number of observations) × p (number of variables), the goal of factor analysis is to find two matrices, F (factor loadings) and L (unique variances), such that:

$$X = F * L^T + E \tag{2}$$

where, X is the observed data matrix (n × p), F is the factor loading matrix (n × k), k is the number of latent factors, L is the unique variance matrix (p × p), representing the uniqueness of each variable, E is the matrix of error terms (n × p). The primary assumption in factor analysis is that the observed variables are linear combinations of a smaller number of latent factors, and the unique variances and error terms are uncorrelated. The goal of factor analysis is to estimate the factor loading matrix F and the unique variance matrix L that best explains the covariance structure of the observed variables.

4. EXPERIMENTAL RESULTS

In this research endeavor, we have executed the entire workflow within the Jupyter Notebook environment, with the Anaconda Navigator version 2023.01-1. The dataset presented challenges with missing values; it was handled using the imputation method. Addressing categorical variables, we employed specific modules from the sci-kit-learn library for their appropriate encoding. To understand the relationships among features, we utilized Pearson correlation from the sci-kit-learn library, conducting a comprehensive analysis of correlations within the dataset. Additionally, we introduced a dimensionality reduction technique, namely factor analysis, utilizing the scikit-learn factor analyzer module. The performance of different machine learning modules is tested without applying the factor analysis and by applying the factor analysis approach.

4.1. Dataset description

The ASD dataset, in the Kaggle Repository is a benchmark dataset and does not require any consent of patients to use it for the experimental analysis. This dataset is a well-curated collection that offers insightful information about the features and characteristics connected to ASD at various developmental stages. The Toddler ASD dataset is a thorough compendium of characteristics that are relevant to young children in the toddler age range. This dataset makes it easier to look at the patterns and predictors that could reveal toddlers' early signs of ASD. The characteristics included in this dataset cover a variety of socio-communicative behaviors, sensory sensitivity, and motor skills, all of which are essential to comprehending the potential ASD signs at this developmental stage. The total number of instances are 1054, it includes 18 attributes including the class variable. The dataset has 728 positive instances and 326 negative instances. The dataset has no missing values, categorical encoding is required.

4.2. Analysis of correlation using - Pearson correlation analysis

Initially, we conducted Pearson correlation analysis using the Pingouin (open-source) package and elbow test as in the Figure 2. Pearson analysis helped us determine if there were any correlations among the features. The outcomes are detailed in Table 1, supplementary file and illustrated in Figure 2(a).

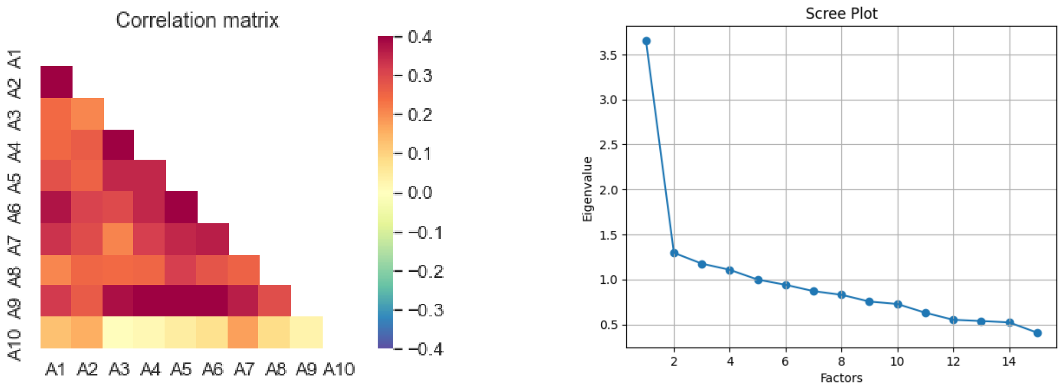
The observations from the experiments conducted using Pearson correlation analysis provide insights into the relationships among the parameters of the autism dataset. According to the Pearson correlation analysis

A1 shows a strong correlation with A2, A3 exhibits a strong correlation with A4, A4 demonstrates a strong correlation with A9, A5 displays a strong correlation with A9. These findings suggest significant associations between specific parameters within the dataset. After identifying correlations between certain factors, our objective is to leverage factor analysis to extract the most significant features from the dataset. This process allows us to uncover the fundamental constructs or components driving the observed relationships, enabling a more concise representation of the data and facilitating further analysis and interpretation.

4.3. Factor analysis test to identify the number of factors

We conducted a factor analysis on the autism toddler dataset using the FactoAnalyzer module from the sklearn library. The Kaiser criterion was applied to assess the viability of forming factors based on eigenvalues derived from the dataset's features. Factors with eigenvalues exceeding 1 were considered for further analysis. To validate our approach, we employed the elbow test, depicted in Figure 2(b), which revealed a significant drop in the curve after two points. This suggests we can utilize a maximum of two factors for our dataset. These loadings are essential for interpreting the relationships identified through factor analysis. Factor loadings close to +1 or -1 indicate strong relationships between observed variables and the derived factors. The factor loadings extracted from our experiment are detailed in Table 2 of the supplementary file. To determine the optimal number of factors, an elbow test (see Figure 2(b)) was performed. analysis of the elbow plot indicates a noticeable decline in values from 3.5 to 1.5, suggesting that a maximum of two factors can be formed (refer to Table 2 in the supplementary file for the loadings).

Utilizing these loadings, we can derive values for each feature and condense the dataset into only two factors, amalgamating information from all other features. It has been observed that Age_Mons, Sex, Ethnicity, Jaundice, Family_mem_with_ASD has not much impact on forming factors. Hence it has been removed from correlation analysis. In summary, the high loading factors are used as Factor 1 $->$ [A3, A4, A5, A9] and Factor 2 is formed using A1, A2.



(a) (b)

Figure 2. Toddler dataset: (a) correlation analysis for toddler dataset; few features are strongly correlated; the correlated features are eliminated via factor analysis; and (b) Elbow test for identifying the number of factors; as evident from the test, there are two prominent factors to be used for this studied

4.4. Analysis of the performance of different models without/with Factor analysis

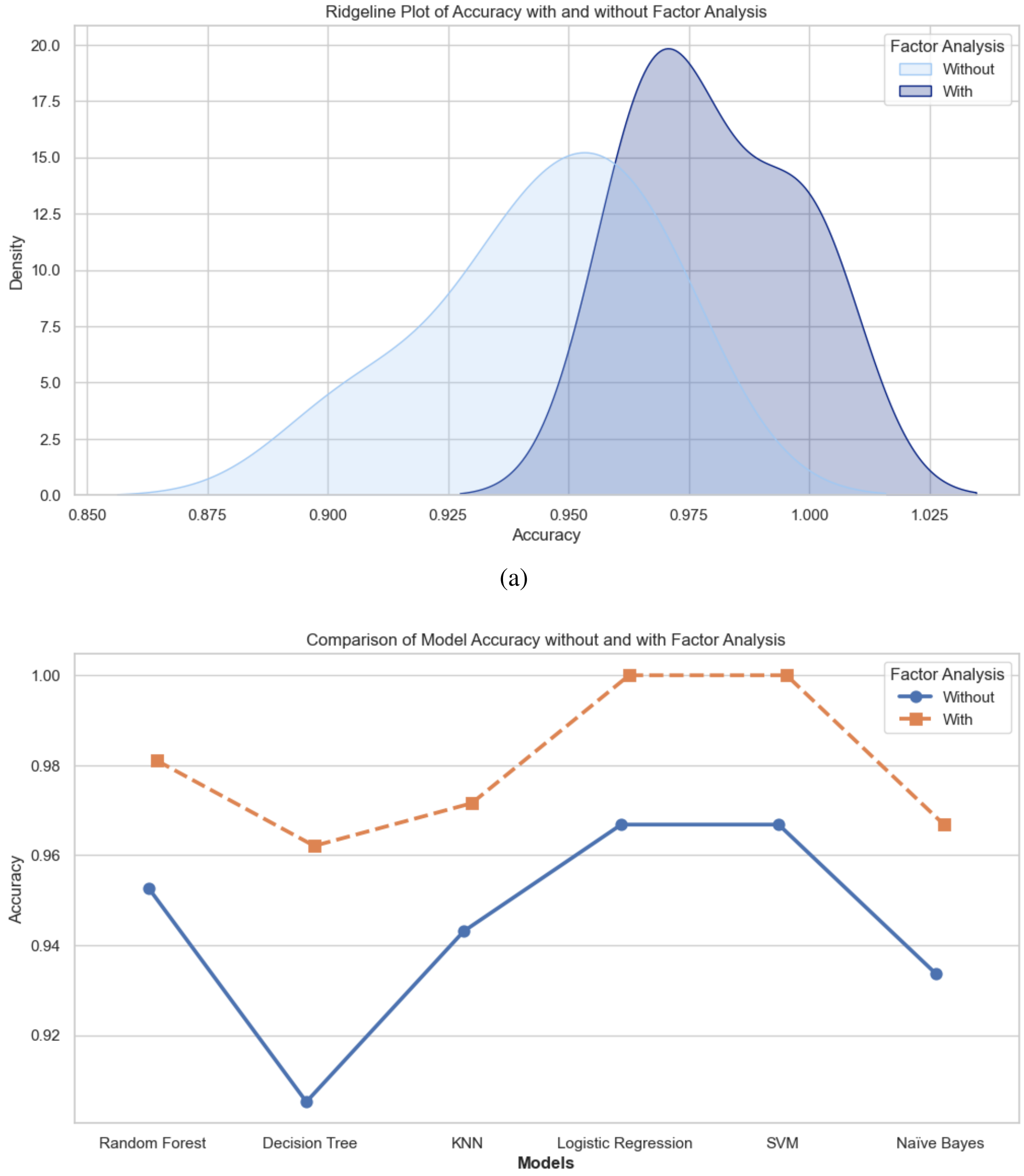
The features received from the factor analysis are fed into k-nearest neighbors (KNN), NB, decision tree (DT), RF, LR, and SVM. The results are presented in Tables 1-3 and along with Figures 3(a) and 3(b).The application of factor analysis enhances the performance of machine learning models for classifying toddlers with autism and this is evident from the tables. LR and SVM outperformed all the remaining classifiers when factor analysis was applied to the dataset. The results are updated in the form of accuracy, precision, recall, Jaccard Index (JI) [23], [24] and Cohen Kappa Coefficient (CK) [25], [26] for each classifier. These findings demonstrate significant improvements in classification accuracy and overall results after the application of factor analysis.

Table 1. Performance evaluation using different machine learning models-without factor analysis

Algorithms		No	Yes	Accuracy	Macro avg	Weighted Avg
RF	Precision	0.912281	0.967532	0.952607	0.939907	0.952607
	Recall	0.912281	0.967532	0.952607	0.939907	0.952607
	F1-Score	0.912281	0.967532	0.952607	0.939907	0.952607
	Jaccard Index	0.910525				
	Cohen’s kappa coefficient:	0.8798131				
DT	Precision	0.813559	0.940789	0.905213	0.877174	0.906419
	Recall	0.842105	0.928571	0.905213	0.885338	0.905213
	F1-Score	0.827586	0.934641	0.905213	0.881113	0.905721
	Jaccard Index	0.830993311				
	Cohen’s kappa coefficient:	0.762253521				
KNN	Precision	0.857143	0.97973	0.943128	0.918436	0.946614
	Recall	0.947368	0.941558	0.943128	0.944463	0.943128
	F1-Score	0.9	0.960265	0.943128	0.930132	0.943985
	Jaccard Index	0.895097929				
	Cohen’s kappa coefficient:	0.860403573				
NB	Precision	0.90566	0.943038	0.933649	0.933649	0.932941
	Recall	0.842105	0.967532	0.933649	0.904819	0.933649
	F1-Score	0.872727	0.955128	0.933649	0.913928	0.932868
	Jaccard Index	0.876312978				
	Cohen’s kappa coefficient:	0.827935694				
LR	Precision	0.941176	0.979021	0.966825	0.960099	0.967004
	Recall	0.955224	0.972222	0.966825	0.963723	0.966825
	F1-Score	0.948148	0.97561	0.966825	0.961879	0.96689
	Jaccard Index	0.936195371				
	Cohen’s kappa coefficient:	0.923759872				
SVM	Precision	0.941176	0.979021	0.966825	0.960099	0.967004
	Recall	0.955224	0.972222	0.966825	0.963723	0.966825
	F1-Score	0.948148	0.97561	0.966825	0.961879	0.96689
	Jaccard Index	0.936195371				
	Cohen’s kappa coefficient:	0.923759872				

Table 2. Performance evaluation using different machine learning models with factor analysis

Algorithms		No	Yes	Accuracy	Macro avg	Weighted Avg
RF	Precision	0.970149	0.986111	0.981043	0.97813	0.981043
	Recall	0.970149	0.986111	0.981043	0.97813	0.981043
	F1-Score	0.970149	0.986111	0.981043	0.97813	0.981043
	Jaccard Index	0.962894486				0.981043
	Cohen’s kappa coefficient:	0.956260365				
DT	Precision	0.915493	0.985714	0.962085	0.950604	0.963417
	Recall	0.970149	0.958333	0.962085	0.964241	0.962085
	F1-Score	0.942029	0.971831	0.962085	0.95693	0.962368
	Jaccard Index	0.927806272				
	Cohen’s kappa coefficient:	0.913895123				
KNN	Precision	0.942029	0.985915	0.971564	0.963972	0.97198
	Recall	0.970149	0.972222	0.971564	0.971186	0.971564
	F1-Score	0.955882	0.979021	0.971564	0.967452	0.971674
	Jaccard Index	0.945119526				
	Cohen’s kappa coefficient:	0.934910026				
NB	Precision	0.928571	0.985816	0.966825	0.957194	0.967639
	Recall	0.970149	0.965278	0.966825	0.967714	0.966825
	F1-Score	0.948905	0.975439	0.966825	0.962172	0.967013
	Jaccard Index	0.963155481				
	Cohen’s kappa coefficient:	0.952450704				
LR	Precision	1	1	1	1	1
	Recall	1	1	1	1	1
	F1-Score	1	1	1	1	1
	Jaccard Index	1				
	Cohen’s kappa coefficient:	1				
SVM	Precision	1	1	1	1	1
	Recall	1	1	1	1	1
	F1-Score	1	1	1	1	1
	Jaccard Index	1				
	Cohen’s kappa coefficient:	1				



(a)

(b)

Figure 3. The model perform better with factor analysis is depicted here: (a) results of different machine learning models with and without factor analysis; and (b) model comparison of with and without factor analysis

Table 3. Performance evaluation using different machine learning models with factor analysis

Results with and without factor analysis		
Model	Accuracy with factor analysis	Accuracy without factor analysis
RF	0.952607	0.981043
DT	0.905213	0.962085
KNN	0.943128	0.971564
NB	0.933649	0.966825
Logistic Regression	0.966825	1
SVM	0.966825	1

5. RESULTS AND DISCUSSION

In this study, we demonstrated significant improvements in ASD classification accuracy while incorporating factor analysis and Pearson correlation validation, compared to previous studies that solely employed machine-learning techniques. Earlier work often struggled with high-dimensional data and overlapping symptoms limiting accuracy with different machine-learning models. In contrast, our integrated approach effectively reduced dimensionality and highlighted crucial latent features, enhancing classification performance. The brief comparison of techniques concerning our proposed model in comparison with previous work is as in the Table 4.

Table 4. Comparison of techniques from previous study to proposed model

SL. NO	Feature / technique	Kosmicki *et al.* [8]	Qureshi *et al.* [14]	Vakadkar *et al.* [18]	Zamit *et al.* [20]	Thabtah *et al.* [19]	Proposed work
1	Dimensionality reduction	Not used	Not used	Not used	Not used	Not used	Factor analysis
2	Correlation analysis	Not used	Not used	Not used	Not used	Not used	Pearson Correlation
3	Age group targeted	Mixed	Mixed	Mixed	Mixed	Adolescents/Adults	Toddlers
4	Hypothesis testing for feature significance	Not used	Not used	Not used	Not used	Not used	Hypothesis 1 and 2 validated with respect to factors

6. CONCLUSION

Our research proposes an integrated method to enhance the accuracy and robustness of ASD categorization across diverse age groups. This approach combines machine learning techniques with Factor analysis and correlation validation. We applied this methodology to a benchmark dataset representing toddlers and compared outcomes with and without factor analysis. Our findings indicate that the classification process was notably improved when factor analysis was utilized, emphasizing the impact of dimension reduction achieved through this method. Additionally, we employed Pearson correlation analysis to test Hypothesis #1, revealing significant relationships among several features. This investigation provided valuable insights into the intricate connections between traits and diagnostic labels within different age groups on the autism spectrum, shedding light on correlations between distinct aspects. Notably, our results demonstrate the importance of eliminating highly correlated features to enhance classification accuracy. Hypothesis #2, which suggested that certain characteristics (such as Age_Mons, Sex, Ethnicity, Jaundice, and Family_mem_with_ASD) were not significant in creating components based on loading values, was successfully refuted. Instead, we identified two crucial factors closely related to the most critical characteristics in the autism dataset. In our study, dimensionality reduction using factor analysis was applied to eliminate the most correlated feature. As we focused solely on the Autism-Toddler dataset future research will explore the performance of our approach on the Autism Adolescents and Adults datasets. Moreover, given the observed correlations among data, we recognize the need to modify the scoring methodology for calculating autism scores. Our forthcoming work will propose a new scoring methodology using the PSO-CES framework, leveraging a particle swarm optimization algorithm with the Constant Elasticity Substitution function.

FUNDING INFORMATION

This research work is not funded by any of the agencies.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Disha Devidas Nayak	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓			
Seema Shedole				✓	✓		✓			✓		✓	✓	
Archana Mathur				✓	✓		✓		✓	✓		✓	✓	

C	: **C**onceptualization	I	: **I**nvestigation	Vi	: **Vi**sualization
M	: **M**ethodology	R	: **R**esources	Su	: **Su**pervision
So	: **So**ftware	D	: **D**ata Curation	P	: **P**roject Administration
Va	: **Va**lidation	O	: Writing - **O**riginal Draft	Fu	: **Fu**nding Acquisition
Fo	: **Fo**rmal analysis	E	: Writing - Review & **E**diting		

CONFLICT OF INTEREST STATEMENT

Authors declare there is no conflict of interest.

DATA AVAILABILITY

The dataset is taken from UCI Repository. It is publicly available.

REFERENCES

[1] M. I. Al-Hiyali, N. Yahya, I. Faye, Z. Khan, and K. Alsaih, "Classification of BOLD FMRI signals using wavelet transform and transfer learning for detection of autism spectrum disorder," in *2020 IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES)*, IEEE, Mar. 2021, pp. 94–98. doi: 10.1109/IECBES48179.2021.9398803.

[2] C.-H. Min and A. H. Tewfik, "Novel pattern detection in children with autism spectrum disorder using iterative subspace identification," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, 2010, pp. 2266–2269. doi: 10.1109/ICASSP.2010.5495885.

[3] C. Lord, M. Elsabbagh, G. Baird, and J. Veenstra-Vanderweele, "Autism spectrum disorder," *The Lancet*, vol. 392, no. 10146, pp. 508–520, Aug. 2018, doi: 10.1016/S0140-6736(18)31129-2.

[4] D. H. Geschwind and P. Levitt, "Autism spectrum disorders: developmental disconnection syndromes," *Current Opinion in Neurobiology*, vol. 17, no. 1, pp. 103–111, Feb. 2007, doi: 10.1016/j.conb.2007.01.009.

[5] D. Granpeesheh, D. R. Dixon, J. Tarbox, A. M. Kaplan, and A. E. Wilke, "The effects of age and treatment intensity on behavioral intervention outcomes for children with autism spectrum disorders," *Research in Autism Spectrum Disorders*, vol. 3, no. 4, pp. 1014–1022, Oct. 2009, doi: 10.1016/j.rasd.2009.06.007.

[6] M.-B. Posserud, A. J. Lundervold, M. C. Steijnen, S. Verhoeven, K. M. Stormark, and C. Gillberg, "Factor analysis of the autism spectrum screening questionnaire," *Autism*, vol. 12, no. 1, pp. 99–112, Jan. 2008, doi: 10.1177/1362361307085268.

[7] M. J. Uddin *et al.*, "An integrated statistical and clinically applicable machine learning framework for the detection of autism spectrum disorder," *Computers*, vol. 12, no. 5, Apr. 2023, doi: 10.3390/computers12050092.

[8] J. A. Kosmicki, 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*, vol. 5, no. 2, pp. e514–e514, Feb. 2015, doi: 10.1038/tp.2015.7.

[9] D. Bone, M. S. Goodwin, M. P. Black, C.-C. Lee, K. Audhkhasi, and S. Narayanan, "Applying machine learning to facilitate autism diagnostics: Pitfalls and promises," *Journal of Autism and Developmental Disorders*, vol. 45, no. 5, pp. 1121–1136, May 2015, doi: 10.1007/s10803-014-2268-6.

[10] A. S. Mohanty, K. C. Patra, and P. Parida, "Toddler ASD classification using machine learning techniques," *International Journal of Online and Biomedical Engineering (iJOE)*, vol. 17, no. 7, pp. 156–171, Jul. 2021, doi: 10.3991/ijoe.v17i07.23497.

[11] J. Ali, R. Khan, N. Ahmad, and I. Maqsood, "Random forests and decision trees," *International Journal of Computer Science Issues*, vol. 9, no. 5, pp. 272–278, 2012.

[12] T. Joachims, "Making large scale SVM learning practical," *Technical reports*, 1998, doi: 10.17877/DE290R-5098.

[13] S. Xu, Y. Li, and Z. Wang, "Bayesian multinomial Naïve Bayes classifier to text classification," in *MUE/FutureTech*, 2017, pp. 347–352. doi: 10.1007/978-981-10-5041-1_57.

[14] M. S. Qureshi, M. B. Qureshi, J. Asghar, F. Alam, and A. Aljarbouh, "Prediction and analysis of autism spectrum disorder using machine learning techniques," *Journal of Healthcare Engineering*, vol. 2023, no. 1, Jan. 2023, doi: 10.1155/2023/4853800.

[15] C. L. Alves *et al.*, "Diagnosis of autism spectrum disorder based on functional brain networks and machine learning," Scientific Reports, vol. 13, no. 1, May 2023, doi: 10.1038/s41598-023-34650-6.

[16] K. S. Omar, P. Mondal, N. S. Khan, M. R. K. Rizvi, and M. N. Islam, "A machine learning approach to predict autism spectrum disorder," in *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, IEEE, Feb. 2019, pp. 1–6. doi: 10.1109/ECACE.2019.8679454.

[17] K. K. Hyde *et al.*, "Applications of supervised machine learning in autism spectrum disorder research: A review," *Review Journal of Autism and Developmental Disorders*, vol. 6, no. 2, pp. 128–146, Jun. 2019, doi: 10.1007/s40489-019-00158-x.

[18] K. Vakadkar, D. Purkayastha, and D. Krishnan, "Detection of autism spectrum disorder in children using machine learning techniques," *SN Computer Science*, vol. 2, no. 5, Sep. 2021, doi: 10.1007/s42979-021-00776-5.

[19] F. Thabtah, N. Abdelhamid, and D. Peebles, "A machine learning autism classification based on logistic regression analysis," *Health Information Science and Systems*, vol. 7, no. 1, Dec. 2019, doi: 10.1007/s13755-019-0073-5.

[20] I. Zamit, I. H. Musa, L. Jiang, W. Yanjie, and J. Tang, "Trends and features of autism spectrum disorder research using artificial intelligence techniques: a bibliometric approach," *Current Psychology*, vol. 42, no. 35, pp. 31317–31332, Dec. 2023, doi: 10.1007/s12144-022-03977-0.

[21] M. A. Mareeswaran and K. Selvarajan, "Predicting autism spectrum disorder through sentiment analysis with attention mechanisms: a deep learning approach," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 37, no. 1, pp. 325-334, Jan. 2025, doi: 10.11591/ijeecs.v37.i1.pp325-334.

[22] J. Cohen, P. Cohen, S. G. West, and L. S. Aiken, *Applied multiple regression/correlation analysis for the behavioral sciences*, 3rd Edition. New York: Routledge, 2013. doi: 10.4324/9780203774441.

[23] Z. Wang, X. Ning, and M. B. Blaschko, "Jaccard metric losses: optimizing the Jaccard index with soft labels," *arXiv-Computer Science*, Feb. 2023.

[24] L. da F. Costa, "Further generalizations of the Jaccard index," *arXiv-Computer Science*, Oct. 2021.

[25] İ. Doğan and N. Doğan, "Evaluation of cohen kappa coefficient and distinguishability for binary data: a simulation study," *Turkiye Klinikleri Journal of Biostatistics*, vol. 14, no. 3, pp. 190–198, 2022, doi: 10.5336/biostatic.2022-89212.

[26] S. M. Vieira, U. Kaymak, and J. M. C. Sousa, "Cohen's kappa coefficient as a performance measure for feature selection," in *International Conference on Fuzzy Systems*, IEEE, Jul. 2010, pp. 1–8. doi: 10.1109/FUZZY.2010.5584447.

BIOGRAPHIES OF AUTHORS

Disha Devidas Nayak currently working as Assistant Professor II at NMAM Institute of Technology, Nitte Deemed to be University. She received her Masters of Technology from M S Ramaiah Institute of Technology and her current research is in machine learning and autism spectrum disorders. Her area of interest lies in foundations of machine learning, deep learning, and prompt engineering techniques. She can be contacted at email: disha.dn.2@gmail.com.

Seema Shedole is Professor in Computer Science and Engineeering Department, M S Ramaiah Institute of Technology, Bangalore, India. She obtained her Ph.D. from Visveswaraya Technological University, Belgaum. Her area of research is in machine learning and bioinformatics. Her research interests include machine learning, data analytics, virtual reality, and augmented reality. She is a member of ACM, and ISTE. She has published over 40 technical papers published in reputed Indian and international conferences and journals. She has many book chapters to her credit. She has reviewed papers of journals and conferences. She was part of the Technical Program Committee and the Advisory Committee of many conferences. She can be contacted at email: seemas@msrit.edu.

Archana Mathur currently working as a Professor at Nitte Meenakshi Institute of Technology. She received her Masters in Engineering from Bangalore University and her Ph.D. is in Machine Learning and Scientometrics. She has worked as research assistant at Indian Statistical Institute, Bangalore. Her area of interest lies in foundations of machine learning, deep learning, and optimization techniques. She can be contacted at email: mathurarchana77@gmail.com.