OPEN ACCESS JOURNALS
Indiana Journal of Multidisciplinary Research
Abbriviate Tittle- Ind J Multi Res
ISSN *(Online)*- 2583-3820
Journal Homepage Link- https://indianapublications.com/Journals/IJMR
DOI: https://doi.org/10.5281/zenodo.12675039
INDIANA PUBLICATIONS
PRODUCTIVE AND QUALITY RESEARCH

Research Article

Volume-04|Issue-03|2024

Unlocking Insights: Machine Learning for Autism spectrum Disorder Analysis and Detection

Akshat Gupta[1], Archith P[2], Mohammed Nadeem[3], Vikrant Rana[*4], Dr Vikash Kumar[5]

[1,2,3,4] Student, Department of Electronics Communication & Engineering, RV Institute of Technology and Management, Bengaluru, Karnataka, India.
[5]Assistant Professor, Department of Electronics Communication & Engineering, R V Institute of Technology and Management, Bengaluru, Karnataka, India.

Article History
Received: 20.05.2024
Accepted: 05.06.2024
Published: 30.06.2024

Citation
Gupta, A., Archith, P., Nadeem, M., Rana, V., Kumar, V. (2024). Unlocking Insights: Machine Learning for Autism spectrum Disorder Analysis and Detection. *Indiana Journal of Multidisciplinary Research*, *4*(3), 162-166.

Abstract: Autism Spectrum Disorder (ASD) is a complex neurodevelopmental condition characterized by diverse behavioral and cognitive traits. Early detection and precise diagnosis are crucial for timely interventions, significantly improving outcomes. ASD symptoms vary among individuals, encompassing unconventional behaviors, interests, and social challenges. This project employs machine learning to enhance ASD detection and assessment. The methodology includes data preprocessing, training, and testing using various ML models such as Decision Tree (DT), K-Nearest Neighbor (KNN), Naive Bayes (NB), Random Forest (RF), Logistic Regression (LR), and Support Vector Classifier (SVC). The approach is evaluated on a publicly available dataset, comprising 31 attributes common in ASD patients. Data preprocessing transforms raw data into a meaningful format, facilitating model training. Evaluation metrics include sensitivity, specificity, and accuracy. This project aims to advance early ASD detection and assessment, addressing the unreliability of manual diagnosis due to resource constraints. Computerized diagnostic systems, employing ML architectures, learn patterns in data to identify disease severity. The proposed ML model demonstrates superior ASD detection performance compared to conventional methods.

Keywords: Autism Spectrum Disorder (ASD), Machine Learning (ML), Decision Tree (DT)

INTRODUCTION

Autism spectrum disorder (ASD) presents a growing concern in global health, affecting individuals across all age brackets [1]. Timely identification of ASD is crucial for safeguarding both mental and physical well-being [3]. The integration of machine learning methodologies has emerged as a promising approach for predicting various ailments, facilitating early detection through an array of health and physiological indicators [2]. This technological progress has spurred increased interest in refining ASD analysis and detection to optimize treatment strategies [2]. Nevertheless, diagnosing ASD is intricate due to symptom overlap with several other mental health conditions, complicating the diagnostic journey [1]. ASD, characterized by impaired social interaction and communication skills, typically manifests by age three and endures throughout life [3]. Although ASD remains without a cure, early detection and intervention hold the potential to mitigate symptoms and improve the quality of life for individuals affected by the disorder [3].

MATERIALS AND METHODS

Data Understanding

Attributes

Our study utilizes openly accessible standard datasets [4]. The A1_Score to A10_Score metrics, derived from a questionnaire survey conducted by the Autism Research Center at the University of Cambridge, UK, were employed [5].

Table 1: Behavioural Feature.

Attribute	Type
ID	Integer
Gender	String
Ethnicity	String
Jaundice	Boolean (yes or no)
Age	Integer
Relation	String
Country_of_res	String
Used the app before	Boolean (yes or no)
Age_desc	Integer
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)
Class/ASD	Boolean (yes=1 or no=0)
Result	Integer

Missing Value

***Corresponding Author:** Vikrant Rana

Handling the numerous missing values in the dataset posed a significant challenge, particularly given the 21 variables involved. Various approaches exist to address missing values, such as substituting them with averaged values or eliminating instances containing missing values. Given the complexity of the dataset and the number of variables, we opted to remove instances with missing values to ensure data integrity and analysis accuracy.

Data pre-processing

Data pre-processing is a crucial step in machine learning workflows, where raw data undergoes refinement to address common real-world challenges like incompleteness, inconsistency, and errors. Various techniques, such as handling missing values, outlier detection, and feature scaling, are employed to ensure data quality. Robust imputation methods effectively manage missing data, while categorical variables are transformed into numerical labels through label encoding to facilitate compatibility with machine learning algorithms. Feature scaling methods, such as Standard Scaler, standardize feature scales to prevent dominance by features with larger magnitudes. The pre-processed data is then fed into a range of machine learning algorithms, including Support Vector Machines (SVM), Naive Bayes (NB), Logistic Regression (LR), K Nearest Neighbors (KNN), Decision Tree (DT), and Random Forest Classifier. An ensemble model is constructed from these algorithms, and evaluation metrics like accuracy, sensitivity, and specificity are utilized for each model. Implementing a meticulously structured pre-processing pipeline enhances data quality, leading to improved model. performance and robustness in real-world scenarios. This comprehensive approach ensures that machine learning models are trained on refined data representations, enabling superior generalization and predictive capability across diverse applications.

Training and Testing Model

The dataset has been divided into two parts: one for training and one for testing, with an 80:20 ratio respectively. Additionally, for cross-validation, the training data is further split into training and validation sets, also with an 80:20 ratio. Figure 1 depicts the final training, testing, and validation sets used for classification.

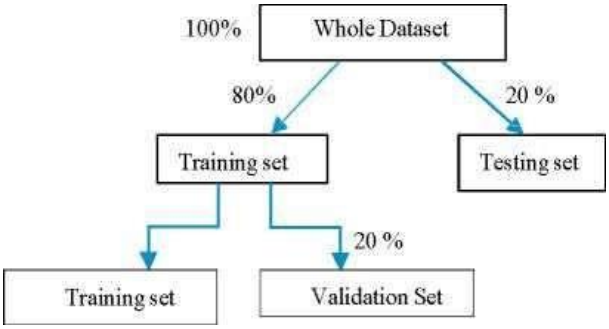


Figure 1: Final Training, Testing and Validation Sets

Support Vector Machine (SVM)

Support Vector Machine (SVM) is a supervised classification method that utilizes a hyper plane to differentiate between two distinct groups. It performs optimally when there's a discernible margin of separation between classes, particularly effective in high dimensional spaces and scenarios where the dimensionality surpasses the number of samples. Moreover, SVM demonstrates relative memory efficiency, making it a favorable choice for various classification tasks.

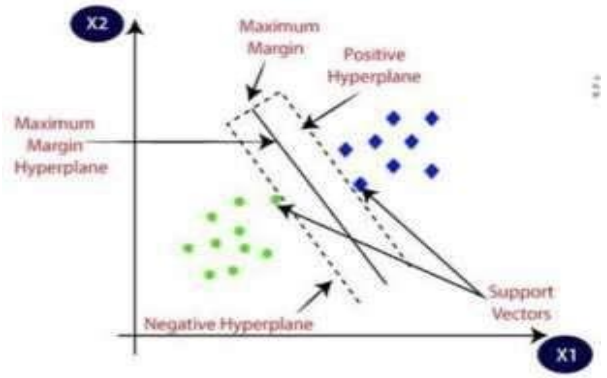


Figure 2: Graph for SVM

Decision Trees

A decision tree is akin to a flowchart, presenting a tree-like structure where internal nodes represent features, branches signify rules, and leaf nodes depict the algorithm's outcomes. This versatile supervised machine learning algorithm is employed for both classification and regression tasks, offering flexibility in addressing various problem types.

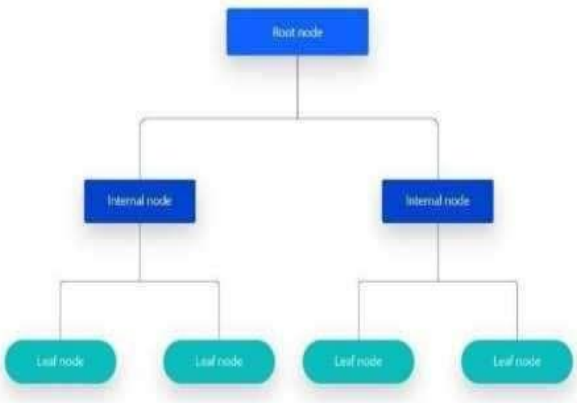


Figure 3: Graph for Decision trees

Random Forest Classifier

The Random Forest Classifier is a widely used classification method suitable for handling binary classification problems. It employs a collaborative approach based on decision trees, generating a "forest" comprised of multiple decision trees. This ensemble technique enhances predictive accuracy and robustness through aggregation of individual tree predictions.

***Corresponding Author:** Vikrant Rana

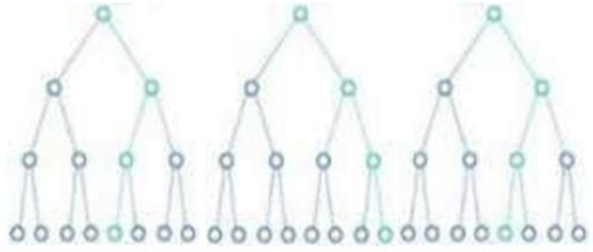


Figure 4: Graph for Random Forest Classifier

Logistic Regression (LR)

Logistic Regression aims to identify the model that best describes the relationship between a binomial outcome and a set of independent variables. It utilizes a logistic function to determine an optimal curve fitting the data points. Unlike other methods, Logistic Regression makes no assumptions about the distribution of classes in feature space. Moreover, it can seamlessly extend to handle multiple classes (multinomial regression) and provides a natural probabilistic perspective on class predictions.

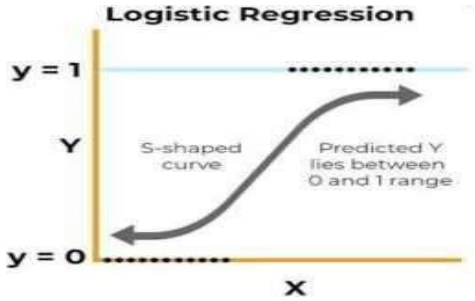


Figure 5: Graph for Logistic Regression

Naive Bayes (NB)

Naive Bayes (NB) classification relies on conditional probability, specifically Bayes' theorem, and counting. The term "naive" stems from its assumption of the conditional independence of all input features. Under this assumption, the NB classifier's convergence rate tends to be higher compared to discriminative models like logistic regression. Consequently, NB typically requires less training data for effective performance.

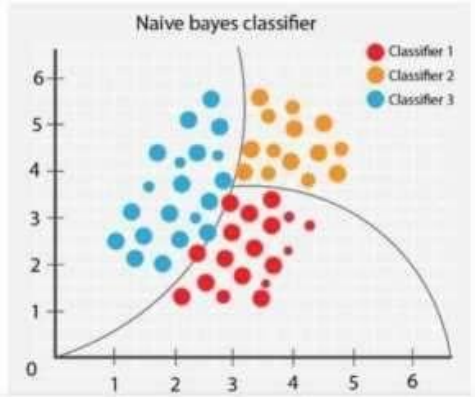


Figure 6: Graph for Naive Bayes

K-Nearest Neighbors Algorithm

The k-nearest neighbors' algorithm, often abbreviated as KNN or k- NN, is a nonparametric, supervised learning classifier. It relies on proximity to classify or predict the grouping of a given data point. KNN is known for its ability to handle large datasets effectively while maintaining high accuracy and effectiveness in its predictions.

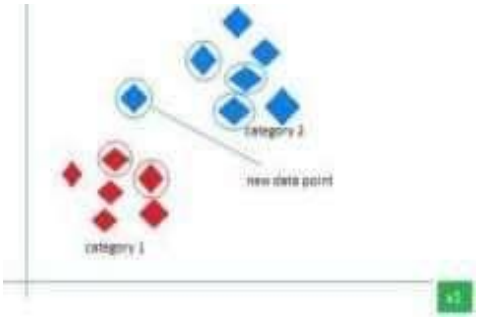


Figure 7: Graph for Naive Bayes

RESULTS & CONCLUSION

The outcome is assessed based on specificity, sensitivity, and accuracy, utilizing the confusion matrix. The precision of the model largely hinges on the quality of its training. Experimental findings across different machine learning algorithms, employing all 21 selected features, are presented for ASD screening data across various age groups, including adults, children, and adolescents. Evaluation of these models reveals accuracies ranging from 83.25% to 96.7% on the ASD diagnosis dataset.

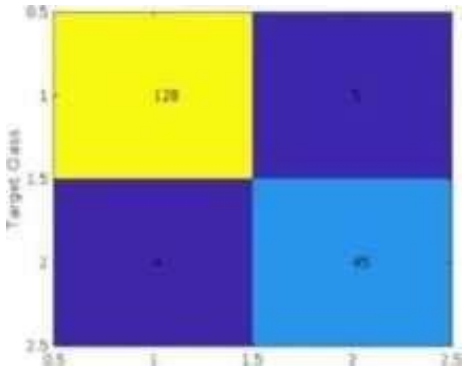
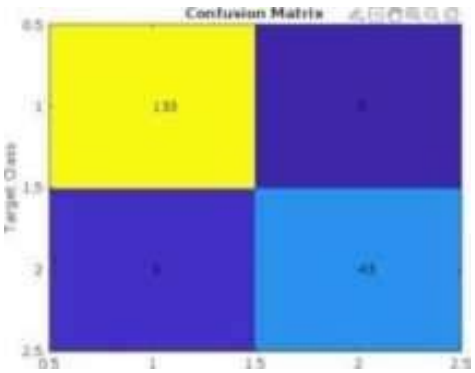


Figure 8: K Nearest Neighbour



***Corresponding Author:** Vikrant Rana

Figure 9: Decision Tree

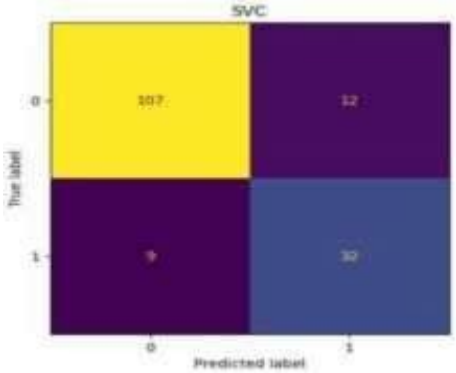


Figure 10: Random Forest

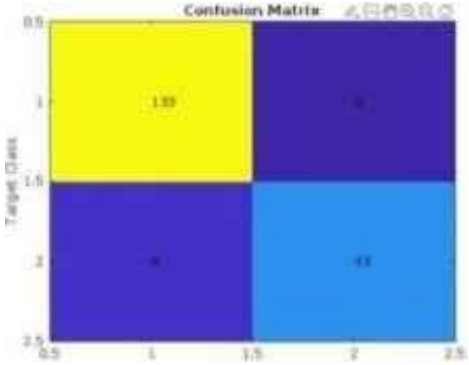


Figure 11: State Vector Classifier

Figure 12: Logistic Regression

The confusion matrix provides insights beyond simple classification accuracy by displaying true and false predictions for each class. In binary classification tasks, it typically comprises a 2x2 matrix with the following components:

- True Positive (TP): Instances where both the predicted and actual values are true.
- True Negative (TN): Instances where both the predicted and actual values are false.
- False Positive (FP): Instances where the prediction is true, but the actual value is false.
- False Negative (FN): Instances where the prediction is false, but the actual value is true.

Table 2: Comparison of various models

	Model	Sensitivity	specificity	Accuracy
1.	K-Nearest Neighbor	0.9696	0.9000	0.9505
2.	Naïve Bayes	0.9805	0.9130	0.9597
3.	Decision Tree	0.8266	0.8490	0.8325
4.	Logistic Regression	0.9345	0.6415	0.8375
5.	State Vector Classifier	0.9224	0.7272	0.8687
6.	Random Forest	0.9568	1.0000	0.9670

CONCLUSION

Among the six different models used for ASD classification, Random Forest achieved the highest accuracy at96.7%, followed by Naive Bayes at 95.9%, and K Nearest Neighbors at 95.0%. These models have shown promising results compared to the other three models. Presently, there's no diagnostic test capable of swiftly and accurately detecting ASD, nor is there an optimized screening tool specifically designed for identifying its onset. Looking ahead, to enhance efficiency and accuracy in classification using machine learning, there's a need to engage with vast amounts of data. Therefore, our recommendation is to explore Deep Learning methodologies instead of relying solely on traditional classifiers.

REFRENCES

1. Thabtah, F. (2018). Machine learning in autistic spectrum disorder behavioral research: A review and ways forward. *Informatics for Health and Social Care*, 1-20.
2. Thabtah, F., Kamalov, F., & Rajab, K. (2018). A new computational intelligence approach to detect autistic features for autism screening. *International Journal of Medical Informatics*, 117, 112-124.
3. Vaishali, R., & Sasikala, R. (2018). A machine learning based approach to classify autism with optimum behaviour sets. *International Journal of Engineering & Technology*, 7(4), 18.
4. Thabtah, F. F. (n.d.). Autism screening adult data set. Department of Digital Technology, Manukau, Auckland, New Zealand. Retrieved from https://archive.ics.uci.edu/ml/datasets/Autism+Screening+Adult
5. National Institute of Health Research. (n.d.). Autism Research Centre.

***Corresponding Author:** Vikrant Rana

6. Constantino, J. N., Lavesser, P. D., Zhang, Y. I., Abbacchi, A. M., Gray, T., & Todd, R. D. (2007). Rapid quantitative assessment of autistic social impairment by classroom teachers. *Journal of the American Academy of Child & Adolescent Psychiatry*, 46(12), 1668-1676.
7. Bone, D., Goodwin, M. S., Black, M. P., Lee, C.-C., Audhkhasi, K., & Narayanan, S. (2015). Applying machine learning to facilitate autism diagnostics: Pitfalls and promises. *Journal of Autism and Developmental Disorders*, 45(5), 1121-1136.
8. Wall, D. P., Kosmicki, J., Deluca, T. F., Harstad, E., & Fusaro, V. A. (2012). Use of machine learning to shorten observation-based screening and diagnosis of autism. *Translational Psychiatry*, 2(4), e100.
9. Wall, D. P., Dally, R., Luyster, R., Jung, J.-Y., & DeLuca, T. F. (2012). Use of artificial intelligence to shorten the behavioral diagnosis of autism. *PloS One*, 7(8), e43855.
10. Thabtah, F. (2017). Autism spectrum disorder screening: Machine learning adaptation and DSM-5 fulfillment. In *Proceedings of the 1st International Conference on Medical and Health Informatics*.

***Corresponding Author:** Vikrant Rana