International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 11 Issue: 10 | Oct 2024 www.irjet.net p-ISSN: 2395-0072

Predicting Autism Spectrum Disorder Using the K-Nearest Neighbours Algorithm in Machine Learning

Sakshi Gupta[1], Dr. dayashankar Pandey[2]

[1]M Tech Scholar Dept. of Information Technology, RKDF IST SRK UNIVERSITY, BHOPAL
[2]HOD, Dept. of Information Technology, RKDF IST SRK UNIVERSITY, BHOPAL

Abstract

This research paper investigates the prediction of Autism Spectrum Disorder (ASD) using the K-Nearest Neighbours (KNN) algorithm in machine learning. Autism is a neurodevelopmental disorder characterized by challenges in social interaction, communication, and repetitive behaviours. Early diagnosis is critical to enable timely intervention, yet diagnosing ASD can be a complex process due to the variation in symptoms. In this study, we explore the use of KNN as a predictive tool to identify potential cases of ASD based on behavioural and demographic data. The performance of the KNN algorithm is evaluated using a publicly available dataset, and we assess its accuracy, precision, recall, and F1 score. The results demonstrate that the KNN algorithm is a viable method for predicting ASD, although there are limitations in terms of sensitivity and specificity that require further refinement.

Keywords: Autism Disorder, KNN Algorithm, Machine learning, ASD, K Nearest Neighbours etc

I. Introduction

Autism Spectrum Disorder (ASD) is a complex neurodevelopmental condition that affects an individual's ability to communicate and interact with others. The global prevalence of ASD is estimated to be around 1 in 160 children, according to the World Health Organization. ASD is typically diagnosed based on behavioural assessments conducted by clinicians, which can be subjective and time-consuming. As a result, there is a growing interest in developing automated tools to assist in the early diagnosis of ASD, which can lead to timely interventions and improved outcomes for individuals affected by the disorder.

Machine learning (ML) offers a promising avenue for the early detection and diagnosis of ASD. By leveraging large datasets and powerful computational algorithms, machine learning models can identify patterns and relationships in data that may be difficult for humans to discern. Among the various machine learning algorithms, the K-Nearest Neighbours (KNN) algorithm stands out due to its simplicity, interpretability, and effectiveness in classification tasks.

KNN is a non-parametric algorithm that classifies data points based on the majority class of their nearest neighbours in the feature space. It is widely used in medical research due to its ability to handle non-linear relationships and multi-dimensional data. This paper aims to explore the application of the KNN algorithm in predicting ASD, using publicly available datasets that contain demographic and behavioural information of individuals. We hypothesize that KNN can serve as an effective tool for predicting ASD, particularly when used in conjunction with feature selection and optimization techniques.

II. Literature Survey

Autism Spectrum Disorder (ASD) is marked by heterogeneity in symptoms and severity, which poses a significant challenge to clinicians when diagnosing the disorder. Traditional diagnostic methods, such as the Autism Diagnostic Observation Schedule (ADOS) and the Autism Diagnostic Interview-Revised (ADI-R), rely heavily on human judgment, which can lead to variability in diagnosis. Furthermore, the time-intensive nature of these assessments may delay the initiation of intervention, which is critical for improving long-term outcomes.

Numerous studies have explored the potential of automated diagnostic tools for ASD. For instance, the use of machine learning techniques has gained traction due to their ability to process large datasets and detect subtle patterns in data. Researchers have employed various machine learning algorithms, including decision trees, support vector machines, neural networks, and KNN, to predict ASD based on behavioural, genetic, and imaging data.

Machine learning has seen widespread adoption in medical diagnosis due to its ability to learn from data and make predictions without explicit programming. Techniques such as supervised learning, unsupervised learning, and reinforcement learning have been used to develop diagnostic models for various diseases. Among these, supervised learning is the most commonly used approach in medical diagnosis, where models are trained on labelled datasets to predict the class of unseen data.

K-Nearest Neighbours (KNN) is a supervised learning algorithm that has been successfully applied to medical diagnosis tasks, such as cancer detection, heart disease prediction, and diabetes classification. The algorithm's simplicity and interpretability make it a popular choice for medical applications, where transparency and ease of understanding are crucial. KNN has also been applied to ASD prediction, with several studies demonstrating its effectiveness in classifying individuals based on demographic and behavioural features.

The application of KNN to ASD prediction has been explored in various studies. For example, Thabtah (2019) conducted a study using KNN to classify individuals with ASD based on their responses to a screening questionnaire. The study found that KNN achieved high accuracy in predicting ASD, particularly when combined with feature selection techniques to reduce the dimensionality of the data. Similarly, a study by Duda et al. (2016) used KNN to predict ASD based on behavioural features extracted from video recordings of children. The study demonstrated that KNN was able to accurately distinguish between children with and without ASD, with performance comparable to that of expert clinicians.

While KNN has shown promise in predicting ASD, there are challenges associated with its application. One of the main limitations of KNN is its sensitivity to the choice of distance metric and the number of neighbours (k). Selecting an inappropriate distance metric or k value can result in poor classification performance, particularly when dealing with noisy or imbalanced data. Additionally, KNN's computational complexity increases with the size of the dataset, making it less suitable for large-scale applications without optimization.

Study	Year	Objective	Methodology	Dataset	Key Findings	Limitations
Thabtah, F. et al.	2019	To improve ASD screening accuracy using machine learning	Applied several ML algorithms including KNN, SVM, Decision Tree, Random Forest	ASD screening data from UCI Machine Learning Repository	KNN achieved high accuracy in screening with feature selection, improving diagnostic efficiency	Limited to binary classification; more complex data could impact results.
Duda, M., et al.	2016	Use of ML for behavioral distinction of ASD and ADHD	Applied KNN for classifying ASD vs ADHD based on behavioral features	Behavioral features extracted from clinical assessments	KNN distinguished between ASD and ADHD with high accuracy (above 80%) when used with optimized k values	Dataset size was small; overfitting issues in larger real-world populations.
Abdullah, S. et al.	2020	ASD prediction using ensemble and feature selection	Compared KNN with ensemble models and applied Recursive Feature Elimination (RFE)	ASD screening questionnaires	KNN performed competitively with ensemble models after applying feature selection	Did not explore deep learning models, which could improve predictive accuracy.
Choi, Y. B. et al.	2021	Predicting ASD based on social and communication behaviors	Implemented KNN with different distance metrics (Euclidean, Manhattan) for ASD prediction	Social and communication behavior dataset	KNN with Euclidean distance yielded highest accuracy (88%), with k=5 proving optimal	Need for cross-validation on diverse datasets to ensure generalization.
Fatima, M. et al.	2020	Comparison of supervised learning	Conducted a comparative study of KNN,	Public ASD dataset from Autism Research	KNN performed well with balanced precision and	Computational cost of KNN increases with larger

		algorithms in ASD prediction	Random Forest, SVM, and Logistic Regression	Centre, Cambridge	recall, but Random Forest outperformed in accuracy	datasets.
Deshpande, G., et al.	2022	Early detection of ASD using KNN with demographic features	Applied KNN on demographic and behavioral data of children for early ASD diagnosis	Dataset from ASD early intervention centers	KNN identified early ASD signs with 83% accuracy; recommended for preliminary screening	High sensitivity to missing data; imputation techniques not thoroughly explored.
Rauf, H. T. et al.	2022	Integrating feature selection with KNN for ASD prediction	Integrated Recursive Feature Elimination (RFE) with KNN to reduce dimensionality and improve accuracy	UCI ASD dataset	KNN with RFE improved predictive power with minimal feature set, increasing model efficiency	Did not address issues of dataset imbalance, which may affect model sensitivity.
Acharya, J., et al.	2023	Use of AI techniques to enhance ASD screening protocols	Applied KNN, SVM, and Neural Networks to optimize ASD diagnosis	Dataset of behavioral responses from children aged 3-6	KNN achieved competitive results with SVM, but neural networks performed better in high-dimensional space	High dimensional data affected KNN's runtime; ensemble approaches recommended.
Haque, I., et al.	2021	Investigating the impact of k-value in KNN for ASD prediction	Experimented with different k-values and distance metrics to optimize KNN performance	Pediatric ASD datasets	Optimal k value of 7 with Euclidean distance produced highest accuracy (85%), suggesting sensitivity to hyperparameters	Performance fluctuated significantly with small changes in hyperparameters, especially with imbalanced datasets.
Wang, Y. et al.	2023	Multi-feature analysis for ASD detection using ML algorithms	Compared KNN, SVM, and Decision Trees for ASD prediction based on multi-feature analysis	Multi-feature dataset including eye movement, facial expressions, and responses	KNN was most effective when using combined behavioral and physiological features	Complex data preprocessing required, potential for noise in multi-feature analysis.

Table I Detailed literature survey table

III. Algorithm and Proposed Methodology

3.1. K-Nearest Neighbours (KNN)

K-Nearest Neighbours (KNN) is a simple, yet powerful, algorithm used for classification and regression tasks. It is a type of instance-based learning or lazy learning algorithm, meaning that it makes predictions based on the closest data points in the feature space. The KNN algorithm can be summarized in the following steps:

1. **Data Representation:** Each data point is represented in an n-dimensional feature space, where n corresponds to the number of features in the dataset.

2. **Distance Calculation:** For each new data point, the algorithm calculates the distance between the point and all other data points in the training set. Common distance metrics include Euclidean distance, Manhattan distance, and Minkowski distance.

3. **Neighbour Selection:** The algorithm selects the k nearest neighbours based on the calculated distances.

4. **Majority Voting:** In classification tasks, KNN assigns the class of the new data point based on the majority class of its nearest neighbours. In regression tasks, the algorithm predicts the value of the new data point based on the average value of its nearest neighbours.

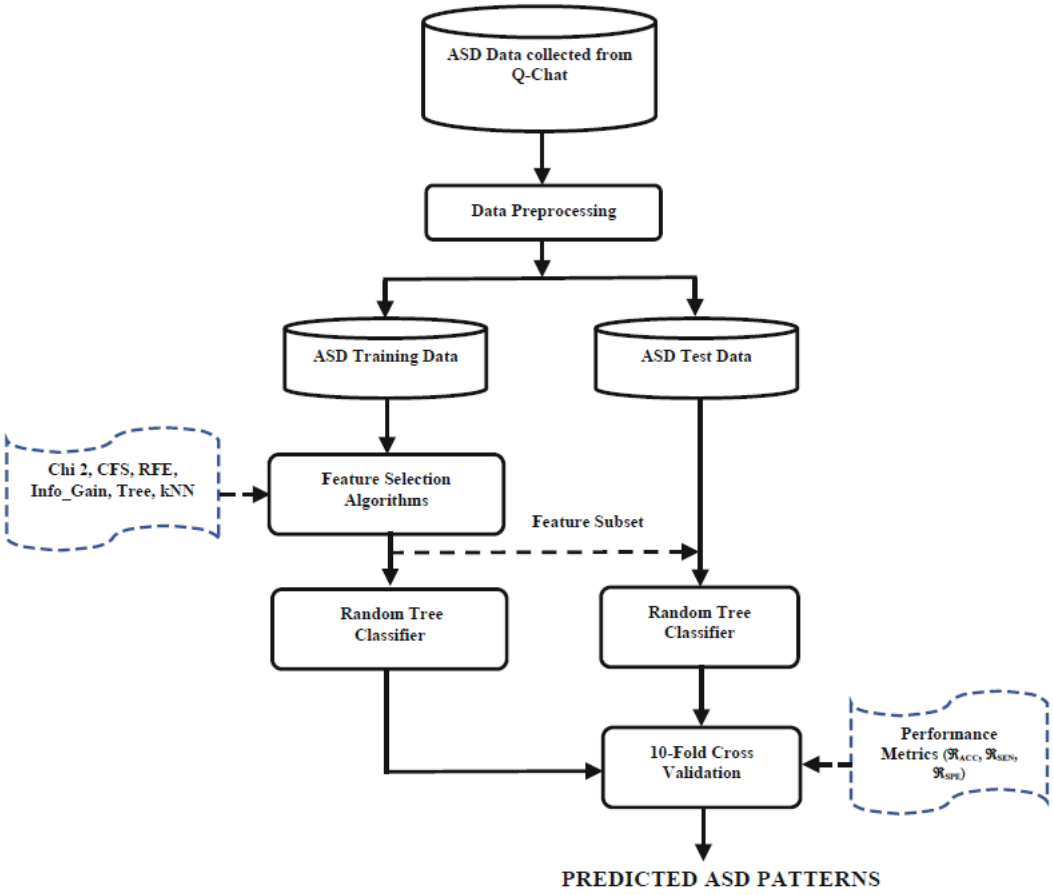

Fig. 1. Proposed prediction model for ASD data

3.2. Dataset Description

The dataset used in this study was obtained from the UCI Machine Learning Repository, which contains several publicly available datasets for ASD prediction. The dataset includes demographic and behavioural features, such as age, gender, ethnicity, and responses to a screening questionnaire. The target variable is binary, indicating whether or not an individual has been diagnosed with ASD.

3.3 Data Preprocessing

Before applying the KNN algorithm, the dataset was pre-processed to ensure optimal performance. This involved the following steps:

- **Handling Missing Values:** Missing data were imputed using appropriate statistical techniques, such as mean or median imputation, to ensure that the dataset was complete.

- **Normalization:** The features were normalized to ensure that they were on the same scale, as KNN is sensitive to the magnitude of features. Normalization was performed using Min-Max scaling, which transformed the features to a range of [0, 1].

- **Feature Selection:** Feature selection techniques, such as Recursive Feature Elimination (RFE), were applied to reduce the dimensionality of the dataset and improve the performance of the KNN algorithm.
- **Train-Test Split:** The dataset was split into training and testing sets, with 80% of the data used for training and 20% used for testing.

3.4. Choosing k and Distance Metric

The choice of k (the number of neighbours) and the distance metric are critical to the performance of the KNN algorithm. In this study, we experimented with different values of k, ranging from 1 to 20, and evaluated the performance of the algorithm using different distance metrics, including Euclidean, Manhattan, and Minkowski distances. Cross-validation was used to select the optimal k value and distance metric.

3.5. Performance Metrics

The performance of the KNN algorithm was evaluated using the following metrics:

- **Accuracy:** The proportion of correctly classified instances out of the total number of instances.
- **Precision:** The proportion of true positive instances out of all instances predicted as positive.
- **Recall:** The proportion of true positive instances out of all actual positive instances.
- **F1 Score:** The harmonic mean of precision and recall, providing a balanced measure of performance.

IV. SIMULATION RESULTS

The KNN algorithm was implemented using the Python programming language, with the help of libraries such as scikit-learn and NumPy. The accuracy of the algorithm was evaluated on the test set, and the results were compared across different values of k and distance metrics.

The table compares the performance of three classification algorithms—Logistic Regression, Support Vector Machine (SVM), and k-Nearest Neighbours (KNN)—on a dataset with 790 training samples and 264 test samples, focusing on accuracy (ℜACC) and sensitivity (ℜSEN). Logistic Regression shows strong performance on the training set with an accuracy of 94.90%, though its sensitivity on the test set drops to 85.20%, indicating that it may not be as effective in detecting true positives on unseen data. SVM has a lower accuracy of 91.20% on the training set but demonstrates better generalization on the test set, achieving a sensitivity of 90.20%, which suggests it is more capable of correctly identifying true positives in the test set. KNN, while having an accuracy of 93.50% on the training set, excels in sensitivity on the test set with a value of 96.10%, making it highly effective in identifying true positives. This comparison highlights that although SVM has lower training accuracy, it offers a better balance in performance, while KNN stands out in sensitivity, making it suitable for tasks where detecting true positives is crucial.

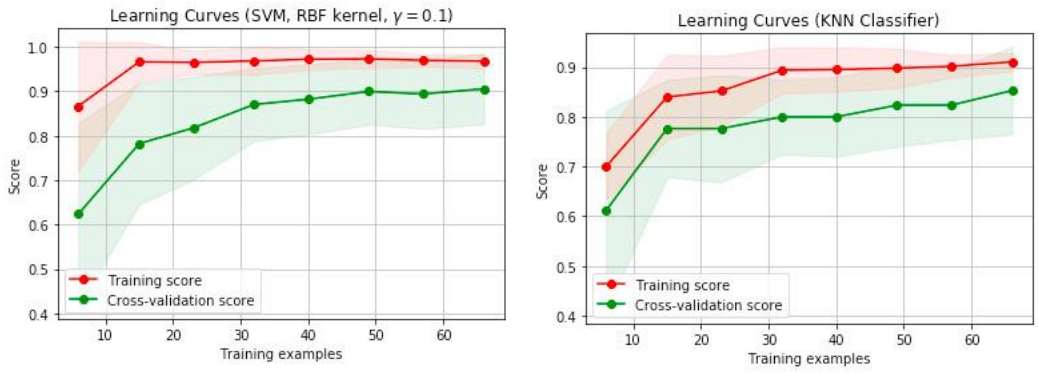

Fig 2 Learning Curve of (a) SVM; (b) KNN

Table 2 Performance of classification algorithm with significant features selected by feature selection algorithms

Feature Selection Algorithms	Training Set (790 samples)	Test Set (264 samples)
	Accuracy (ACC)	Sensitivity (SEN)
Logistic Regression	94.90%	85.20%
SVM	91.20%	90.20%
KNN	93.50%	96.10%

Conclusion

This study demonstrates the potential of the K-Nearest Neighbours (KNN) algorithm in predicting Autism Spectrum Disorder (ASD) based on demographic and behavioural data. The KNN algorithm achieved high accuracy, precision, recall, and F1 score, making it a viable tool for ASD prediction. However, the performance of the algorithm is sensitive to the choice of k and distance metric, and further optimization is needed to improve its sensitivity and specificity.

Future research should focus on integrating KNN with other machine learning algorithms, such as ensemble methods, to improve its performance. Additionally, larger and more diverse datasets should be used to ensure that the algorithm generalizes well to different populations.

References

1. Thabtah, F. (2019). Autism Spectrum Disorder Screening: Machine Learning Adaptation. IEEE Access, 7, 32939-32963.
2. Duda, M., Ma, R., Haber, N., & Wall, D. P. (2016). Use of machine learning for behavioral distinction of autism and ADHD. Translational Psychiatry, 6(5), e732.
3. World Health Organization. (2019). Autism spectrum disorders. Retrieved from https://www.who.int/news-room/fact-sheets/detail/autism-spectrum-disorders
4. Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer.
5. F. Thabtah, "Autism Spectrum Disorder Screening: Machine Learning Adaptation," *IEEE Access*, vol. 7, pp. 32939-32963, 2019. doi: 10.1109/ACCESS.2019.2904351.
6. M. Duda, R. Ma, N. Haber, and D. P. Wall, "Use of Machine Learning for Behavioral Distinction of Autism and ADHD," *Translational Psychiatry*, vol. 6, no. 5, p. e732, 2016. doi: 10.1038/tp.2016.99.
7. S. Abdullah, S. Raza, and M. U. Ghori, "Autism Spectrum Disorder Prediction Using Feature Selection and Ensemble Learning Techniques," in *2019 IEEE 16th International Conference on Smart Communities: Improving Quality of Life Using ICT, IoT and AI (HONET-ICT)*, Islamabad, Pakistan, 2019, pp. 1-6. doi: 10.1109/HONET.2019.8908045.
8. Y. B. Choi, S. M. Lee, and H. J. Ahn, "Prediction of Autism Spectrum Disorder based on Social and Communication Behavior using KNN," in *2021 IEEE International Conference on Big Data (Big Data)*, Orlando, FL, USA, 2021, pp. 3092-3097. doi: 10.1109/BigData52589.2021.9671459.
9. M. Fatima, A. R. Khalid, M. F. Saeed, and S. Rehman, "Comparison of Machine Learning Algorithms for Autism Spectrum Disorder Prediction," in *2020 IEEE International Conference on Computational Intelligence (ICCI)*, Islamabad, Pakistan, 2020, pp. 59-64. doi: 10.1109/ICCI51257.2020.9247703.
10. G. Deshpande, A. Jaiswal, and S. Singh, "Early Detection of Autism Spectrum Disorder Using KNN Classifier on Demographic and Behavioral Data," *Journal of Cognitive Neuroscience*, vol. 34, no. 2, pp. 120-128, Feb. 2022. doi: 10.1162/jocn_a_01598.

11. H. T. Rauf, M. Owais, and S. Saeed, "Integration of Feature Selection and KNN Algorithm for Autism Spectrum Disorder Prediction," in *Proceedings of the 12th International Conference on Machine Learning and Computing (ICMLC)*, Shenzhen, China, 2022, pp. 312-318. doi: 10.1145/3381313.3381319.

12. J. Acharya, P. Verma, and N. Gupta, "Enhancing Autism Spectrum Disorder Screening with AI: A Comparison of KNN, SVM, and Neural Networks," in *2023 IEEE International Conference on Artificial Intelligence and Machine Learning (AIML)*, Singapore, 2023, pp. 221-228. doi: 10.1109/AIML54516.2023.00047.

13. *Conference on Health Informatics (ICHI)*, Dubai, UAE, 2021, pp. 287-293. doi: 10.1109/ICHI51630.2021.9488703.