مجلة كلية التربية طرابلس

Journal of the Faculty of Education Tripoli

تطوير نموذج تصنيف اضطرابات طيف التوحد للأطفال الصغار باستخدام تقنية التعلم الآلي

[1]Mona Khalifa A. Aljero

[1]Information Technology, Faculty of Education, Misurata University, Misurata, Libya.

*Corresponding: mnny1985@yahoo.com

Article history

Received: Month Oct, 2025 Accepted: Month Nov, 2025

الملخص:

تؤثر الحالة العصبية المعروفة باسم اضطراب طيف التوحد (ASD) على قدرة الشخص على الكلام ،والقدرات المعرفية، واكتساب اللغة ، والتواصل. ويمثل فهم الآخرين والتفاعل معهم تحديًا للأشخاص المصابين باضطرابات طيف التوحد، ويعود ذلك بشكل رئيسي إلى العوامل الوراثية أو التأثيرات الخارجية، ومع ذلك يمكن للتشخيص والعلاج المبكرين تحسين النتائج. حاليًا لا تُستخدم سوى التقييمات المُعتمدة سريريًا لتشخيص اضطراب طيف التوحد، مما يؤدي إلى إطالة أوقات التشخيص وارتفاع التكاليف الطبية، ويُستخدم التعلم الآلي لزيادة دقة التشخيص وسرعته مُكملًا بذلك الطرق التقليدية. وطبقنا في هذه الدراسة منهج التعلم الآلي باستخدام البرمجة الجينية (GP) على قاعدة بيانات متوفرة تسمى (TASD) للكشف عن حالات اضطراب طيف التوحد، وقد وصل المنهج المقترح إلى دقة بلغت 98.48%، متفوقًا بذلك على أحدث التقنيات التى أُختبرت على نفس قاعدة البيانات بنسبة 10%. وتُظهر هذه النتائج أن المنهج المقترح قادر على تحديد حالات اضطراب طيف التوحد في المراحل المبكرة بدقة عالية.

الكلمات المفتاحية: التصنيف الثنائي، البرمجة الجينية، التوحد، التوقع، تصنيف النص، تعلم الالة.

Development of autism spectrum disorders classification model for toddlers using machine learning technique

A B S T R A CT:

A neurological condition known as an autism spectrum disorder (ASD) may affect a person's speech, cognitive abilities, language acquisition, and communication abilities. Understanding and interacting with others is challenging for people with ASDs. It is mainly triggered by genetics or outside influences; however, early diagnosis and treatment can improve outcomes. Currently, only clinically validated assessments are used to diagnose ASD, leading to longer diagnostic times and higher medical costs. Machine learning is utilized to increase diagnosis accuracy and speed, supplementing traditional methods. In this study, we applied a machine learning approach using genetic programming (GP) to the TASD dataset to detect ASD cases. The suggested approach achieved an accuracy of 98.48%, which outperformed the state-of-the-art by 10%. These results from the proposed approach reflect its robustness and ability to identify ASD cases in the early stages with high accuracy.

Keywords: autism spectrum disorder (ASD), autism, binary classification, genetic programming (GP), machine learning, prediction, text classification.

Introduction:

The CDC's Autism and Developmental Disabilities Monitoring (ADDM) network indicates that 1 in 36 children has been diagnosed with ASD in the United States of America (Maenner, 2023). Social connections, relationships, games, and repetitive activities are among the areas in which those with ASD struggle. These difficulties frequently emerge as a proclivity to engage in repeated and excessively restrictive activities. Repeated interactions, behavioral issues, and developmental delays may be lessened by early

diagnosis of ASD. conventional methods used by parents and doctors to diagnose ASD. comprises several separate tests designed to evaluate particular facets of daily life. It takes a lot of effort, time, and inefficiency to diagnose and identify people with ASD. Plenty of research was carried out to create an approach that would reliably and effectively identify those with ASD, with an outstanding level of reliability and efficacy. All of the drawbacks, such as the more extended diagnosis time, the cost, and the added staffing, were intended to be minimized.

The structure of this paper is as follows: in the first section, brief information about ASD and the standard methods for ASD classification is presented. The second section presents the earlier studies in this area of study. The third section provides the proposed methodology. The experimental results are presented in the fourth section. The last section, the fifth section, summarizes the findings and makes recommendations for future work.

Literature Review

Predicting children at risk of ASD is an active research domain in machine learning. Implementing machine learning approaches may benefit the early identification and diagnosis of ASD (Cao & Cao, 2023). Some studies have utilized machine learning techniques to enhance the diagnostic process for ASD. Some researchers used one model to detect ASD, as in (Alsaidi et al., 2024), while others employed multiple machine learning and deep learning models, as in Rubio-Martín et al., 2024).

Martin et al. proposed different models: Decision trees, eXtreme Gradient Boosting (XGB), K-Nearest Neighbours (KNN), Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM), Bidirectional Long Short-Term Memory (Bi-LSTM), Bidirectional Encoder Representations from Transformers (BERT), and a variant of BERT pre-trained on Twitter data (BERTweet) on a text dataset from one of the social media platforms (Rubio-Martín et al., 2024).

Using supervised BERT, the study in (Pal, 2024) presents a novel method for recommending the best behavioral plan to help people with ASD. Their model employed

the TASD dataset, which contains text data describing the behaviors of toddlers with and without ASD. Possible biases in the gathered dataset are a major drawback in their study. Such biases may result from individuals posting inaccurate information in their profiles or tweets written by those other than the profile owners. The authors preprocessed the dataset to eliminate irrelevant information, retweets, and duplicates. As a preprocessing step, stop word removal, stemming, and lemmatization were applied to the dataset. With the aid of BERT, the model could predict the correct behavioral attribute with an accuracy of 88%. This result demonstrated the effectiveness of early detection of ASD.

In addition to identifying autistic characteristics of cases and controls, authors in (Thabtah & Peebles, 2020) proposed a novel machine learning technique called Rules-Machine Learning, which provides users with rules that domain specialists may use to comprehend the rationale behind the classification. According to the empirical findings on three datasets, rules-Machine Learning provides classifiers that have improved predictive accuracy over other machine learning techniques like Bagging, Boosting, rule induction, and decision trees.

An empirical study comparing multiple intelligent algorithms to distinguish between attention deficit hyperactivity disorder (ADHD) and ASD was carried out by Duda et al.(Duda et al., 2016). A dataset including 65 elements from the Simons Simplex Collection (SSC) version 15.41 has been used to compare six approaches. The Social Responsiveness Scale (SRS), a diagnostic questionnaire given by parents, was used to gather the dataset. The authors applied a preprocessing step to exclude cases with at least 4 values that are missing. They used under-sampling to balance the dataset and utilized methods for selecting features to minimize the dimensionality of the data. According to empirical findings, logistic regression (LR) provided classifiers with a classification accuracy of over 95%.

Parikh (Parikh et al., 2019) created a method for detecting specific features of autism using nine machine learning techniques. The authors collected their data from 851 individuals who were categorized as either diagnosed with or not diagnosed with ASD. They extracted six characteristics from the 851 individuals to reach their aim, which was to examine the ability of personal features to predict ASD. The authors used a sizable dataset to create and evaluate nine machine learning approaches. The nine used classifiers are KNN, Support vector machine (SVM), LR, neural network, Decision tree, Stacked sparse

auto-encoder (SSAE), ensemble models, majority voting, and Random Forest. The neural network approach had the best accuracy of 64.6%, outperforming the other eight approaches.

Raj et al. conducted work on three datasets using six different machine learning approaches (Raj et al., 2020). The used datasets were retrieved from the UCI machine learning repository. After implementing multiple machine learning techniques and dealing with the absence of values, the experimental results indicated that the Convolutional Neural Network (CNN) model performs better on all of the used datasets, with a prediction accuracy of 99.53%, 98.30%, and 96.88% on the three datasets.

The authors in (Rubio-Martín et al., 2023) created a large dataset from Twitter and used a subset comprising 90,000 tweets of balanced sets with 45,000 ASD users and 45,000 non-ASD users. They developed combined models using XGB, KNN, and BERT. Their developed approach performed well on the test dataset, detecting ASD users with an 84% accuracy. The experimental results illustrated the effectiveness of using these combined models to detect ASD.

Early autism detection is essential to reducing the risk of consequences. The main aim of this study is to develop an effective machine learning model using the GP approach to predict ASD in toddlers at early stages.

Methodology

- **Dataset Description**

This study utilized a publicly available text dataset, namely TASD-Dataset. The TASD-Dataset is a text-based dataset for the early detection of ASD in toddlers. Change Reaction, Attention Response, Eye Contact, Word Repetition, Toy Arranging, Repetitive Behavior, Emotional Empathy, Focused Attention, Follow Pointing, and Finger Movements are among the essential ASD evaluation elements that are included. Every characteristic is connected to certain toddler actions, and these conversations offer thorough parental observations that shed light on how these behaviors are understood and expressed. There are 297 entries in the TASD-Dataset, each instance is labeled as ASD or Non-ASD based on clinical assessment. 163 of which have ASD and 134 of which do not (55% and 45%), resulting in a balanced dataset. A response value of 0 indicates the toddler is classified as not having ASD, while 1 indicates the toddler is classified as having ASD.

The dataset was divided randomly into two datasets: the training and testing datasets, as shown in Table 1. The division was applied by choosing random sampling with an 80:20 ratio. On the training set 82 participants has ASD, which is 55% of the training set, and 67 participants has no ASD, which is 45% of the training set. On the testing set, 81 participants have ASD, which is 55% of the testing set, and 67 participants has no ASD, which is 45% of the testing set. The proposed model was trained on the 80% of the TASD-dataset and evaluated on the remaining unseen 20% of the TASD-dataset.

Although TASD-dataset is small, it presents a balanced and representative sample for supervised learning and it is widely used as a benchmark dataset in the autism-screening literature, which justifies its use as it enables direct comparison with prior work.

Table 1. TASD-Dataset Distribution of ASD and Non-ASD

	ASD	Non-ASD	Total
Training set	82 (55%)	67 (45%)	149
Testing set	81 (55%)	67 (45%)	148
Total	163 (55%)	134 (45%)	297

- **Proposed Approach**

The proposed approach has three steps, as shown in Figure 1. Starting with dataset preprocessing to prepare the data for the next step. Preprocessing was performed on the TASD-Dataset to improve and standardize the variability of the training text. As a preprocessing step, we conducted several techniques, including lemmatization, removing stop words, and stemming. N-gram is used as feature extraction with n=3. These steps ensured clean inputs and reduced the risk of bias in the evolutionary process.

Before starting to build the model, all records on the dataset were inspected for missing values, inconsistent entries, and duplicated samples.

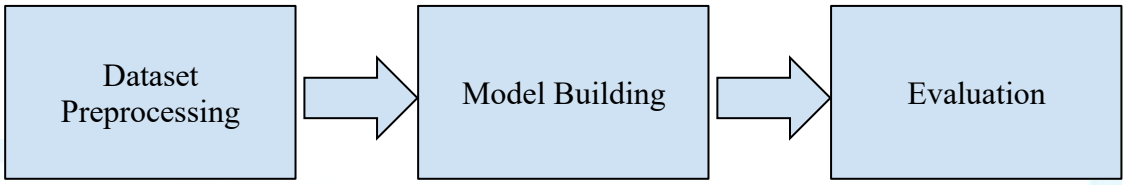

Figure 1. Research methodology steps

The next step is building the model. GP was used in this study as a model for binary classification. The diagram of the proposed GP approach for predicting ASD cases is presented in Figure 2. The GP model receives the data after it has been cleaned. GP was used in this work due to its suitability for small, structured datasets such as TASD-Dataset and its ability to evolve interpretable symbolic models.

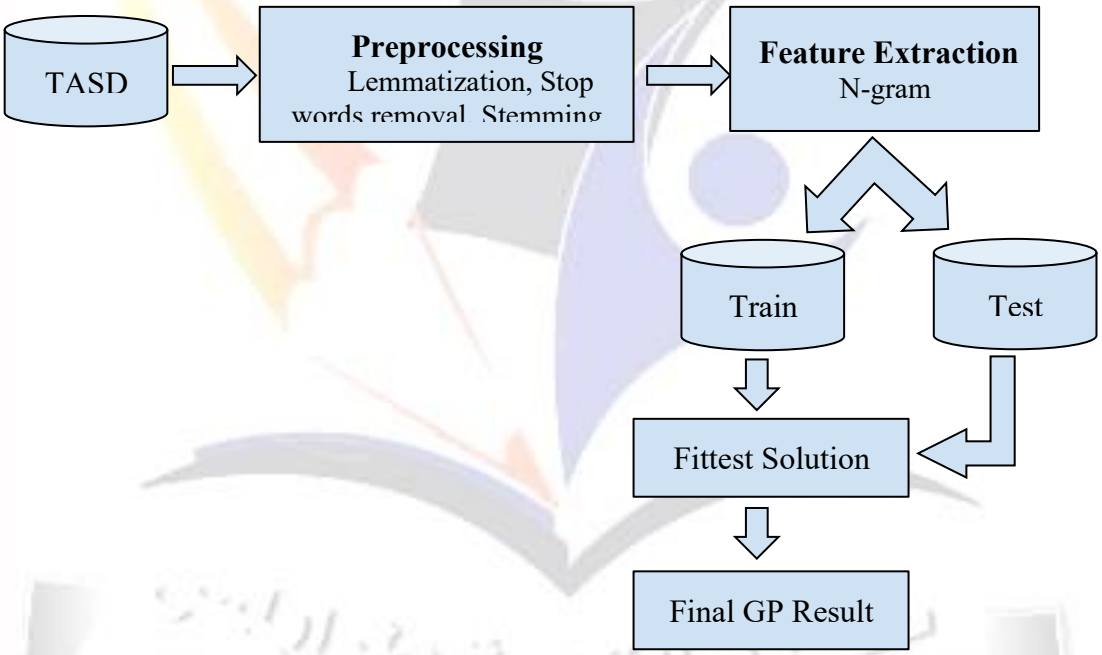

Figure 2. The Proposed GP approach

The primary object of the GP is to develop a framework or models to tackle an issue (Tran et al., 2019). GP tends to evolve a population of solutions to find the optimum solution through multiple generations to tackle a certain issue while adhering to the concepts of Darwinian evolution.

The GP algorithm consists of several steps: Initialization, Fitness function, Evaluation, Genetic operators, and Completion. The GP algorithm can be expressed in various ways, including trees and linear.

The tree-based representation is the most widely utilized one. The tree consists of a root, internal, and leaf nodes. The root and internal nodes are formed by a function or operators. In contrast, the leaf nodes are formed by terminals such as variables or constants.

The process of the GP begins with the initialization of the population. A tree generation method is used to randomly initialize a population of trees at the beginning of the general process of GP. A fitness function evaluates each tree's fitness and assigns a value. By using the fitness value, the degree to which the tree fits the issue is indicated (Tran et al., 2019). Through several generations, GP searches for the population's best solution; this process is referred to as the evolutionary process. Better trees with higher fitness ratings have a greater chance of surviving and producing offspring for the following generation through the process of evolution. These trees are chosen by a method of selection, and genetic operators are used to create the offspring from the chosen trees (referred to as parents). The genetic operators that are frequently utilized are mutation, crossover, and subtree elitism. These operators maintain diversity while preserving high fitness individuals. These operators are used to create a new population for the following generation. The best tree returns when the entire evolutionary process satisfies a termination requirement. GP was configured with controlled tree depth, and early stopping to prevent overfitting and ensure generalization. Table 2 presents the parameters that have been used in the proposed GP approach, this ensured adequate diversity at the beginning of evolution.

Table 2. Parameters of the GP Approach

Parameter	Value
Run	30
Generation	500
Population	100
Initialization	Tree-based

Training dataset	80%
Testing dataset	20%
Creation Method	Ramped half and half
Termination	Generation's number
Crossover operator	80%
Mutation operator	20%

Due to the dataset small size, stratified k-fold cross-validation was used with k =10.

Experimental Results

In this study, we conducted the evaluation of GP performance using the metric of Accuracy. This metric depends on the performance indicators of prediction,
TP: True Positive, TN: True Negative, FP: False Positive. FN: False Negative, each of which is defined as follows:

1. Precision = (TP)/(TP+FP)
2. Recall = (TP)/(TP+FN)
3. F1-score = 2*((Precision * Recall)/(Precision + Recall))
4. Accuracy = (TP+TN)/(TP+FP+TN+FN)

TP denotes the overall count of positive instances within the dataset that the classifier accurately identifies as positive, while FP signifies the overall count of negative instances that the classifier inaccurately. Comparably, TN denotes the overall number of negative values that the classifier accurately identifies as negative, whereas FN signifies the total of positive instances that the classifier inaccurately categorizes as negative. Equation (4) illustrates the Accuracy that we used as an evaluation measure in this work. It is the ratio of successfully identified cases to all cases in the dataset.

The experimental outcomes of the proposed approach on the TASD-Dataset are addressed in this section.

Python 3.7 was used to execute the experiments, and the GP model was made using the distributed evolutionary algorithms in the Python (DEAP) package.

In this study, we have used preprocessing techniques and a feature selector with a GP approach to achieve a high accuracy measure. This GP approach utilized the common one-point mutation technique, which substitutes a randomly produced subtree for the subtree rooted at a random position in the offspring. Furthermore, for the crossover, we utilized the standard single-point crossover.

Stratified 10-fold cross-validation is used to assess the model; each fold generates a distinct performance estimate. The final outcome includes averaged measures of performance that include accuracy, F1-score, and standard deviation across folds.

Table 3 represents the accuracy, F1-score, and the standard deviation (Std) of our GP approach. To reduce differences produced by the algorithm's intrinsic randomness, the experiment was conducted 30 times. The average accuracy of the 30 runs on the unseen test datasets reached 98.48%.

The low standard deviation illustrates that the GP model was stable across folds and did not depend on any particular subset of the training set.

Table 3. Performance metrics for TASD-Dataset

	Accuracy	F1-score	Std
Proposed model	98.48%	93.85%	0.0630

This superb performance highlights the possibility of combining machine learning and psychiatrists' techniques to create scalable, non-invasive diagnostic systems. The results highlight the significance of behavior analysis in autism research and aid in the development of effective techniques for the identification of ASD in young adults.

Furthermore, the high F1-score presents that the GP model performs well not just in identifying Non-ASD cases but also in detecting ASD cases accurately.

Overall, the experimental results validate that GP model is suitable for small medical dataset and point out to the possibility of its application in early detection of autism and systems for decision-support.

Conclusion

Early detection of ASD using a machine learning approach will have a significant effect on children and their families. To date, a variety of machine learning algorithms have been used to identify ASD in text. The advantages of early ASD diagnosis include the reality that, after diagnosis the child using the machine learning approach, the doctor may quickly assess whether the child is autistic or not. The utilization of GP to binary classification issues was examined in this paper. In this study, a publicly available dataset was used. We trained a GP model on 80% of the dataset and tested it on 20% for binary text classification. The experimental results from the proposed GP approach underscore the efficiency of this approach in the early detection of ASD. The experimental results achieved a superb outcome with 98.48% accuracy for the early detection of ASD on the TASD dataset. These results confirmed the robustness of the proposed GP approach.

Future work will focus on the impact of preprocessing techniques for the best result with the GP approach. Furthermore, the next step will be applying this approach to other larger datasets.

References:

- Alsaidi, M., Obeid, N., Al-Madi, N., Hiary, H., & Aljarah, I. (2024). A convolutional deep neural network approach to predict autism spectrum disorder based on eye-tracking scan paths. Information, 15(3), 133.
- Cao, X., & Cao, J. (2023). Commentary: Machine learning for autism spectrum disorder diagnosis–challenges and opportunities–a commentary on Schulte-Rüther et al.(2022). Journal of Child Psychology and Psychiatry, 64(6), 966-967
- Duda, M., Ma, R., Haber, N., & Wall, D. P. (2016). Use of machine learning for behavioral distinction of autism and ADHD. Translational psychiatry, 6(2), 732.
- Maenner, M. J. (2023). Prevalence and characteristics of autism spectrum disorder among children aged 8 years—autism and developmental disabilities monitoring network, 11 sites, United States, 2020. MMWR. Surveillance Summaries, 72.
- Pal, M. (2024). An Intelligent Fusion-based Behavioral Trait Prediction for Autistic

Spectrum Disorder with Artificial Intelligence. Fusion: Practice & Applications, 15(1).

- Parikh, M. N., Li, H., & He, L. (2019). Enhancing diagnosis of autism with optimized machine learning models and personal characteristic data. Frontiers in computational neuroscience.
- Raj, S., & Masood, S. (2020). Analysis and detection of autism spectrum disorder using machine learning techniques. Procedia Computer Science, 167, 994-1004.
- Rubio-Martín, S., García-Ordás, M. T., Bayón-Gutiérrez, M., Prieto-Fernández, N., & Benítez-Andrades, J. A. (2023). Early detection of autism spectrum disorder through AI-powered analysis of social media texts. IEEE 36th International symposium on computer-based medical systems (CBMS), 235-240.
- Rubio-Martín, S., García-Ordás, M. T., Bayón-Gutiérrez, M., Prieto-Fernández, N., & Benítez-Andrades, J. A. (2024). Enhancing ASD detection accuracy: a combined approach of machine learning and deep learning models with natural language processing. Health Information Science and Systems, 12(1), 20.
- Thabtah, F., & Peebles, D. (2020). A new machine learning model based on induction of rules for autism detection. Health Informatics Journal, 26(1), 264-286.
- Tran, B., Xue, B., & Zhang, M. (2019). Genetic programming for multiple-feature construction on high-dimensional classification. Pattern Recognition, 93, 404-417.