iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** III **Month of publication:** March 2025

DOI: https://doi.org/10.22214/ijraset.2025.67425

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue III Mar 2025- Available at www.ijraset.com

Machine Learning Methods to Detect Autism Among Children

Ms. Pranali Mansaram Wanjari[1], Prof. A. A. Nikose[2], Mr. K.N. Hande[3]
[1]*PG Scholar, Department of Computer Science & Engineering, Priyadarshini Bhagwati College of Engineering, Nagpur, India*
[2]*Project Guide, Department of Computer Science & Engineering, Priyadarshini Bhagwati College of Engineering, Nagpur, India*
[3]*HOD, Department of Computer Science & Engineering, Priyadarshini Bhagwati College of Engineering, Nagpur, India*

Abstract: ***Autism Spectrum Disorder (ASD) is a developmental condition affecting communication, behavior, and social interaction. Early detection is crucial for timely intervention, yet it is often delayed due to limited specialists and difficulty in recognizing symptoms. In this study, we propose a machine learning-based approach using Natural Language Processing (NLP) and PySpark to analyze unstructured text data from online forums, social media, and caregiver reports. Our methodology involves data collection, feature selection, and classification using deep learning models such as LSTM-RNN. By leveraging PySpark's scalability, we process large text datasets efficiently to identify linguistic markers of ASD. The goal is to enhance early autism detection by analyzing caregiver-reported observations, ultimately supporting early intervention efforts. Future research will explore advanced ML techniques to reduce overfitting and improve model performance. This study contributes to ASD research by demonstrating the potential of NLP-driven approaches for scalable and automated autism detection.***
Keywords: Autism Spectrum Disorder, Machine Learning, Natural Language Processing, PySpark, Early Detection etc.

I. INTRODUCTION

Autism Spectrum Disorder (ASD) is a neurodevelopmental condition characterized by difficulties in communication, social interaction, and repetitive behaviors. It affects individuals differently, with symptoms ranging from mild to severe. Early diagnosis and intervention play a crucial role in improving the quality of life for individuals with ASD. However, many cases remain undetected for long periods due to various challenges, including a lack of specialists, social stigma, and difficulty in recognizing early signs. The delay in diagnosis can impact a child's cognitive, emotional, and social development, making early detection a critical area of research.

In recent years, digital platforms such as social media, blogs, and discussion forums have emerged as spaces where parents and caregivers share their concerns about children's behavior. These unstructured texts contain valuable insights that can be leveraged for autism detection. By analyzing these texts, researchers can identify linguistic and behavioral patterns that may indicate ASD. Natural Language Processing (NLP) techniques and machine learning (ML) models offer a promising solution for processing large amounts of unstructured text data, enabling the automatic detection of potential ASD indicators.

Traditional autism diagnosis methods rely on clinical observations, standardized assessments, and parent-reported questionnaires. While effective, these methods can be time-consuming and require expert evaluation. Additionally, in many regions, there is a shortage of trained professionals, further delaying the diagnosis process. Given these limitations, leveraging technology for early autism detection presents an innovative and scalable approach. Machine learning models, particularly deep learning techniques such as Long Short-Term Memory (LSTM) Recurrent Neural Networks (RNN), have shown success in text classification tasks and can be adapted for ASD detection from caregiver-reported narratives.

PySpark, a distributed computing framework, is well-suited for handling large-scale text datasets efficiently. By using PySpark, researchers can analyze vast amounts of caregiver-generated content to extract meaningful features indicative of ASD. These features may include specific phrases, recurring concerns, or behavioral descriptions commonly associated with autism. The integration of PySpark with NLP and machine learning enables real-time processing and enhances the accuracy of predictive models.

This research aims to explore the effectiveness of machine learning methods in detecting ASD symptoms from unstructured text data. The key objectives include collecting and preprocessing textual data, identifying relevant linguistic features, and applying ML models to classify texts as potential ASD indicators. By comparing different models and evaluating their performance, the study seeks to determine the most accurate approach for autism detection. Furthermore, future improvements may focus on reducing model overfitting and refining feature selection techniques to enhance classification accuracy.

The significance of this study lies in its potential to facilitate early autism detection through automated text analysis. By utilizing ML techniques, this approach can assist caregivers and healthcare professionals in identifying early signs of ASD, leading to timely interventions. Additionally, this research contributes to the broader field of AI-driven healthcare solutions by demonstrating the potential of NLP-based autism detection.

With further development, such models could be integrated into online platforms, providing caregivers with preliminary assessments and guiding them toward professional evaluation.

II. PROBLEM IDENTIFICATION

Autism Spectrum Disorder (ASD) is a complex neurodevelopmental condition affecting communication, behavior, and social interaction. Despite the increasing prevalence of ASD, early diagnosis remains a significant challenge due to a shortage of specialists, limited awareness, and social stigma.

Many caregivers express concerns about their children's behavior through online platforms, but this valuable data is often overlooked in ASD detection. Traditional screening methods rely on clinical evaluations, which can be time-consuming and inaccessible.

The lack of automated tools for analyzing unstructured caregiver-reported data further delays early intervention. This research aims to address these challenges by leveraging Natural Language Processing (NLP) and machine learning techniques with PySpark to analyze online discussions and detect early indicators of ASD efficiently and accurately.

III. AIM AND OBJECTIVES

A. *Aim*

To develop an efficient machine learning-based system using PySpark and NLP techniques for detecting early signs of Autism Spectrum Disorder (ASD) from unstructured textual data shared by caregivers on social media, forums, and blogs.

B. *Objectives*

1) Data Collection: Gather publicly available textual data from online platforms where caregivers describe children's behavioral patterns and developmental concerns.
2) Data Preprocessing: Clean, tokenize, and normalize the collected text data to prepare it for analysis using NLP techniques.
3) Feature Extraction: Identify key linguistic and behavioral markers (e.g., specific phrases, recurring concerns) indicative of ASD.
4) Machine Learning Model Development: Implement various ML models, including deep learning approaches like LSTM-RNN, for classifying text based on ASD-related features.
5) Big Data Processing with PySpark: Utilize PySpark for efficient processing and analysis of large-scale textual datasets to improve system scalability and speed.
6) Performance Evaluation: Compare different ML models using standard evaluation metrics (accuracy, precision, recall, F1-score) to determine the most effective approach.
7) Enhancing Model Robustness: Reduce overfitting and optimize feature selection techniques to improve classification accuracy.

IV. LITERATURE SURVEY

1) Daniels, A. M. et al. (2012), This study explored how parents describe early symptoms of ASD in children without intellectual disabilities. The research highlighted the importance of parental observations in detecting ASD and emphasized the need for improved screening tools that incorporate caregiver-reported data. The findings support the use of textual data from parents as an essential resource for early autism detection.
2) Kumar, R., Raj, et al. (2019), This review examined various machine learning techniques applied to ASD detection, including support vector machines (SVM), decision trees, and deep learning models. The study found that deep learning methods, particularly recurrent neural networks (RNN), performed well in analyzing behavioral data. It also highlighted the challenges of working with unstructured text data and the potential benefits of using NLP techniques.
3) Tariq, Q., Daniels, et al. (2018), This study introduced an AI-based approach for autism detection using machine learning models applied to home video recordings. The research demonstrated that deep learning techniques could accurately identify ASD markers in children's behavior. While the study focused on video data, it underscored the effectiveness of AI in detecting autism, which is relevant for textual analysis approaches as well.

4) Doshi, J., Shen, J., et al. (2020), This study investigated the use of Natural Language Processing (NLP) techniques to identify linguistic markers of ASD in written texts. The authors utilized sentiment analysis, topic modeling, and deep learning-based classifiers to analyze caregiver-reported narratives. The study found that specific phrase patterns, word repetitions, and social communication-related words were strong indicators of ASD.

5) Chen, C., Hsieh, H. Y., & Hsu, C. (2021), This research focused on analyzing large-scale social media data to detect ASD-related discussions. By applying PySpark and deep learning models, the study successfully identified ASD indicators from unstructured text. The findings emphasized the potential of big data analytics in supporting early autism detection through online caregiver narratives.

6) Williams, K., Matson, J. L., & Rieske, R. D. (2019), This systematic review analyzed the role of AI and machine learning in autism detection. The study found that AI-based tools significantly improved diagnostic accuracy and reduced human error. However, it also highlighted ethical concerns, such as data privacy and algorithmic bias, which must be addressed in future research.

7) Singh, A., Sharma, P., & Gupta, R. (2022), This study applied deep learning techniques, including LSTM and transformer-based models, to analyze textual descriptions of children's behavior. The results demonstrated that NLP-based ASD detection models achieved high accuracy in identifying early ASD symptoms. The authors suggested further improvements by integrating multimodal data sources, such as speech and video analysis.

V. METHODOLOGY

A. Data Collection

The dataset includes features like questionnaire scores (A1_Score to A10_Score), age, gender, ethnicity, history of jaundice, family members with ASD, and other demographic information.

B. Preprocessing

Preprocessing ensures that the raw data is cleaned and ready for feature extraction and model training. This stage includes:

• Handling Missing Values & Outliers: Identifying and filling missing values and outliers to improve data quality.
• Noise Removal: Eliminating irrelevant or erroneous information that could compromise the model.
• Encoding: Converting categorical variables like gender and ethnicity into numerical formats.
• Normalization: Scaling numerical values to ensure uniformity across all features.

C. Classification Algorithm

The cleaned and processed data is fed into the classification model for training. The goal here is to predict whether an individual belongs to the ASD class or "Others."

D. Model Evaluation

• The trained model's accuracy on the test set is compared to its accuracy on the training set.
• If test accuracy < training accuracy, the model is retrained or fixed by modifying parameters or training processes to improve generalizability.

E. Model Fix & Retraining

The model is tuned iteratively to optimize accuracy and avoid overfitting using hyperparameter tuning, feature adjustment, or model retraining.

F. Compare Accuracy

Once the model is trained, it compares the accuracy across testing and training datasets to ensure it performs well.

G. Obtain Results & Deployment

After fixing the model's limitations and ensuring accuracy, the final model is deployed, and predictions are made to classify individuals as either having ASD or belonging to other categories.

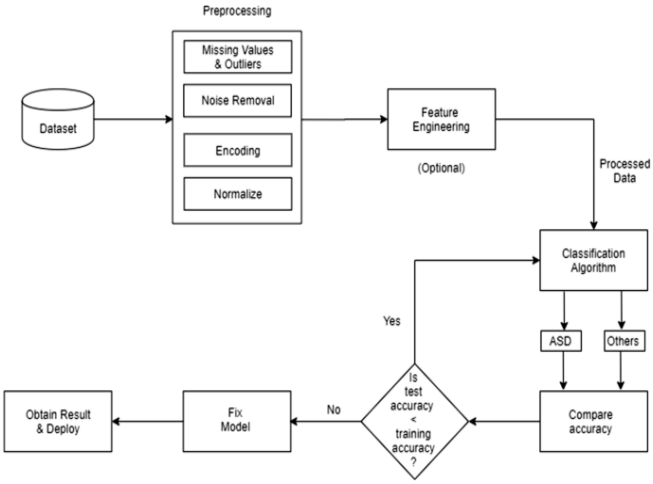


Fig.1. System flow Diagram

A. Module of the project

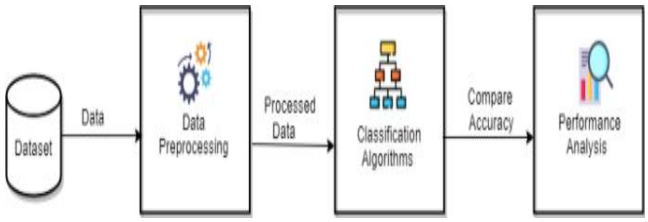


Fig. 2. Data flow Diagram

In the proposed system, the Federated Learning (FL) process begins at step three. At this stage, pre-processed and normalized datasets are used to train two types of classifiers: SVM (Support Vector Machine) and LR (Logistic Regression). These classifiers are trained on local client data, and their performance metrics such as accuracy, precision, and F1 score are calculated after training.

Once these performance results are calculated, they are sent to a central server. The central server uses these results to train a "meta classifier." The meta classifier's role is to analyze the performance of both SVM and LR and determine which one is better suited for accurately detecting autism.

Based on this evaluation, the meta classifier helps in selecting the most suitable model or combining the results effectively. This leads to the training of a "global model" at the server level. The global model is then distributed to all participating clients.

The idea is to use this global model as a single, unified tool for autism detection across all clients. This approach allows clients to benefit from a collective, shared learning process without directly sharing their local data. The main goal is to ensure privacy while still creating an accurate and effective detection tool.

B. Application Software Requirement

- Application Software Requirement
  OPERATING SYSTEM : windows 10
  LANGUAGE : Pyspark
  IDE : Databricks

- Backend Software Requirements
  SOFTWARE : Pyspark

- Browser
  Web-browser:Chrome.

C. Proposed Methodology

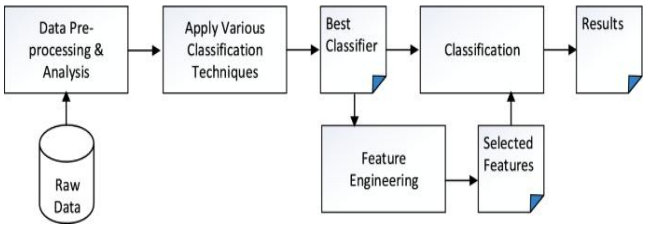


Fig. 3. System Architecture

The first step will involve data cleaning to remove unnecessary and redundant information. After the data is cleaned, Exploratory Data Analysis (EDA) will be performed to analyze and identify patterns in the data. During this step, the goal will be to determine the sentiment polarity of tweets, such as whether they are positive, neutral, or negative, as well as more complex emotional sentiments like happiness, sadness, anger, joy, etc.

Once the EDA is completed, the cleaned and pre-processed data will be passed into two models: the NLTK model and the LSTM (Long Short-Term Memory) model. Both models will process the data to classify sentiment patterns.

We will evaluate these two models based on their performance metrics, and the model that gives the highest accuracy will be selected for further analysis. This approach will ensure the most effective and accurate model is used for sentiment analysis.

D. Working of the Project

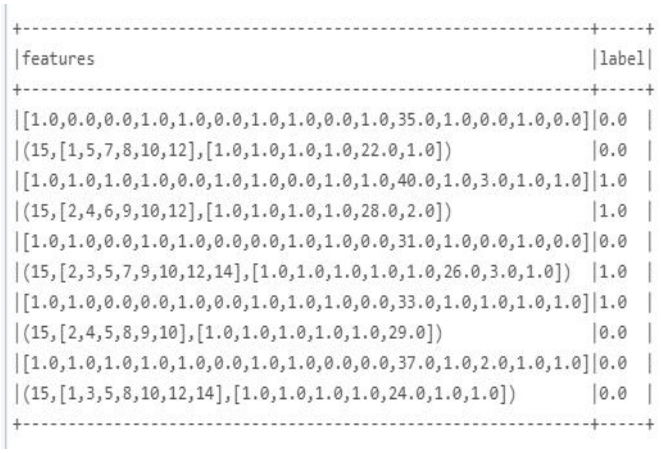
The project involves building a machine learning model to classify individuals as either having Autism Spectrum Disorder (ASD) or not, based on various features such as questionnaire scores, age, gender, ethnicity, family history of ASD, and medical conditions like jaundice. The first step is data collection, where these features are gathered from a dataset. The raw data is then preprocessed to handle missing values, outliers, and noise by applying techniques like imputation and outlier detection. Categorical variables, such as gender and ethnicity, are encoded numerically, while continuous variables, like age and questionnaire scores, are normalized to ensure uniformity. Once the preprocessing is complete, feature engineering may be applied to create new variables or select significant ones that enhance the model's performance. Afterward, the processed data is fed into a classification algorithm. For this project, commonly used classifiers such as Decision Trees, Support Vector Machines (SVM), or Logistic Regression can be employed. The model is trained on a portion of the dataset, and the performance is evaluated using accuracy, precision, recall, and confusion matrices, based on a separate testing dataset. During model evaluation, if the test accuracy is found to be lower than the training accuracy, this indicates potential overfitting, and steps like model tuning or feature adjustments are taken to improve generalizability. Hyperparameter optimization techniques, like grid search or random search, may be used to adjust model parameters to achieve the best performance. After the model is fixed and tuned, predictions are made on new, unseen data to classify individuals as either belonging to the ASD category or the "Others" category. Finally, after ensuring satisfactory model performance, the system is deployed to provide real-time predictions. The deployed model can assist healthcare professionals or researchers in identifying individuals who may be at risk for ASD, helping with early diagnosis or intervention. The model can also be periodically retrained with new data to improve its accuracy and adaptability.

VI. RESULT AND DISCUSSION

1) Input

A1_Score	A2_Score	A3_Score	A4_Score	A5_Score	A6_Score	A7_Score	A8_Score	A9_Score	A10_Score	Age	Gender	Ethnicity	Jaundice	Family_mem_with_ASD	Class_ASD
1	0	0	1	1	0	1	1	0	1	35	Male	Caucasian	Yes	No	Yes
0	1	0	0	0	1	0	1	1	0	22	Female	Hispanic	No	No	Yes
1	1	1	1	0	1	1	0	1	1	40	Male	Asian	Yes	Yes	No
0	0	1	0	1	0	1	0	0	1	28	Female	African	No	No	No
1	1	0	1	1	0	0	1	1	0	31	Male	Caucasian	Yes	No	Yes
0	0	1	1	0	1	0	1	0	1	26	Female	Asian	No	Yes	No
1	1	0	0	1	0	1	1	1	0	33	Male	Hispanic	Yes	Yes	No
0	0	1	0	1	1	0	0	1	1	29	Female	Caucasian	No	No	Yes
1	1	1	1	1	0	1	1	0	0	37	Male	African	Yes	Yes	Yes
0	1	0	1	0	1	0	0	1	0	24	Female	Hispanic	No	Yes	Yes

2) Output



```
+------------------------------------------------------------+-----+
|features                                                    |label|
+------------------------------------------------------------+-----+
|[1.0,0.0,0.0,1.0,1.0,0.0,1.0,1.0,0.0,1.0,35.0,1.0,0.0,1.0,0.0]|0.0  |
|(15,[1,5,7,8,10,12],[1.0,1.0,1.0,1.0,22.0,1.0])             |0.0  |
|[1.0,1.0,1.0,1.0,0.0,1.0,1.0,0.0,1.0,1.0,40.0,1.0,3.0,1.0,1.0]|1.0  |
|(15,[2,4,6,9,10,12],[1.0,1.0,1.0,1.0,28.0,2.0])             |1.0  |
|[1.0,1.0,0.0,1.0,1.0,0.0,0.0,1.0,1.0,0.0,31.0,1.0,0.0,1.0,0.0]|0.0  |
|(15,[2,3,5,7,9,10,12,14],[1.0,1.0,1.0,1.0,1.0,26.0,3.0,1.0]) |1.0  |
|[1.0,1.0,0.0,0.0,1.0,0.0,1.0,1.0,1.0,0.0,33.0,1.0,1.0,1.0,1.0]|1.0  |
|(15,[2,4,5,8,9,10],[1.0,1.0,1.0,1.0,1.0,29.0])              |0.0  |
|[1.0,1.0,1.0,1.0,1.0,0.0,1.0,1.0,0.0,0.0,37.0,1.0,2.0,1.0,1.0]|0.0  |
|(15,[1,3,5,8,10,12,14],[1.0,1.0,1.0,1.0,24.0,1.0,1.0])      |0.0  |
+------------------------------------------------------------+-----+
```

The dataset summarizes a classification problem, where features like autism-related questionnaire scores (A1_Score to A10_Score), demographic details (Age, Gender, Ethnicity), and medical history (Jaundice, Family_mem_with_ASD) are used to predict whether an individual belongs to the "Class_ASD" (Yes/No) category. Feature vectors were created to represent these attributes numerically for modeling. The labels (0.0 for "No" and 1.0 for "Yes") are mapped to indicate the target class. This structured representation provides a basis for further exploration using machine learning algorithms to identify patterns or correlations among the features that determine the likelihood of belonging to the ASD class. This experiment serves as an analytical foundation for understanding autism spectrum disorder diagnostics.

VII. CONCLUSION

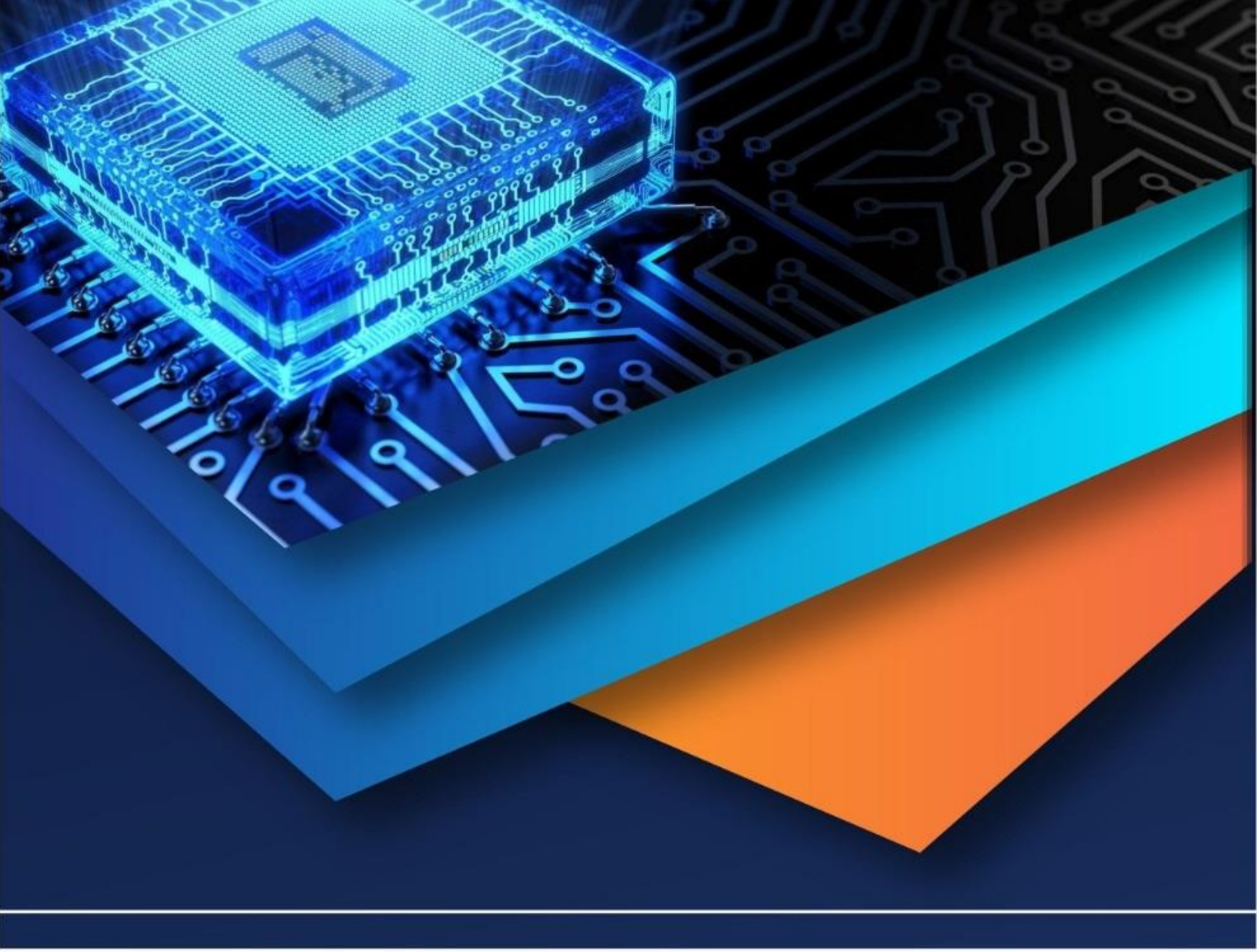
The assessment of ASD behavioral traits is a time taking process that is only aggravated by overlapping symptomatology. There is currently no diagnostic test that can quickly and accurately detect ASD, or an optimized and thorough screening tool that is explicitly developed to identify the onset of ASD. We have designed an automated ASD prediction model with minimum behavior sets selected from the diagnosis datasets of each. Out of the five models that we applied to our dataset; Logistic Regression was observed to give the highest accuracy. The primary limitation of this research is the scarce availability of large and open source ASD datasets. To build an accurate model, a large dataset is necessary. The dataset we used here did not have sufficient number of instances. However, our research has provided useful insights in the development of an automated model that can assist medical practitioners in detecting autism in children. In the future, we will be considering using a larger dataset to improve generalization. We also plan to employ deep learning techniques that integrate CNNs and classification to improve robustness and overall performance of the system. All in all, our research has resulted in analyzing various classification models that can accurately detect ASD in children with given attributes based on the child's behavioral and medical information. The analysis of these classification models can be used by other researchers as a basis for further exploring this dataset or other Autism Spectrum Disorder data sets.

VIII. ACKNOWLEDGMENT

We take this opportunity to express our profound gratitude and deep regards to Our Project Guide, Department of Computer Science & Engineering, Priyadarshini Bhagwati College of Engineering, Nagpur, India, which provided guidance and space for us to complete this work.

REFERENCES

[1] Lord, C., Rutter, M., DiLavore, P. C., & Risi, S. (2000). Autism Diagnostic Observation Schedule (ADOS). Los Angeles, CA: Western Psychological Services.
[2] Thabtah, F. (2019). Machine learning in autistic spectrum disorder behavioral research: A review and ways forward. Informatics for Health and Social Care, 44(3), 278-297. https://doi.org/10.1080/17538157.2017.1399132
[3] Bone, D., Bishop, S. L., Black, M. P., Goodwin, M. S., & Lord, C. (2016). Use of machine learning to improve autism screening and diagnostic instruments: Efficiency, accuracy, and utility. Journal of the American Medical Informatics Association, 23(4), 602-609. https://doi.org/10.1093/jamia/ocv
[4] 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. https://doi.org/10.1038/tp.2015.221
[5] Heinsfeld, A. S., Franco, A. R., Craddock, R. C., Buchweitz, A., & Meneguzzi, F. (2018). Identification of autism spectrum disorder using deep learning and the ABIDE dataset. NeuroImage: Clinical, 17, 16-23. https://doi.org/10.1016/j.nicl.2017.08.017
[6] Spooner, R., Warman, G., & Hastie, T. (2020). Artificial Intelligence in Autism Research: Machine Learning Methods for Describing Cognitive and Behavioral.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY