Detection of Autism Spectrum Disorder in Toddlers using Machine Learning

Dr. Indu John
Department of Computer Science and Engineering
Amal Jyothi College of Engineering
Kanjirapally, Kerala, India
indujohn@amaljyothi.ac.in

Abdul Musawir
Department of Computer Science and Engineering
Amal Jyothi College of Engineering
Kanjirapally, Kerala, India
abdulmusawir2024@cs.ajce.in

Gauri Santhosh
Department of Computer Science and Engineering
Amal Jyothi College of Engineering
Kanjirapally, Kerala, India
gaurisanthosh2024@cs.ajce.in

Glady Prince
Department of Computer Science and Engineering
Amal Jyothi College of Engineering
Kanjirapally, Kerala, India
gladyprince2024@cs.ajce.in

Jesna Susan Reji
Department of Computer Science and Engineering
Amal Jyothi College of Engineering
Kanjirapally, Kerala, India
jesnasusanreji2024@cs.ajce.in

***Abstract*—The aim of this study is to identify toddlers at risk for Autism Spectrum Disorder (ASD) early on by developing a web-based tool that uses the machine learning method logistic regression. Our approach emphasises the vital need of early intervention because it recognises the lifelong impact of ASD on language development, speech, cognitive, and social skills, especially when symptoms appear during the first two years of life. Respondents to nominal questions are asked to provide a score that indicates the probability of Autism Spectrum Disorder. Using toddler datasets, our study demonstrates the efficacy of logistic regression in producing precise predictions with little characteristics. The study contributes to the larger objective of improving the diagnostic process by highlighting the importance of early discovery in reducing the long-term impacts of ASD. Crucially, this method is presented as a quick and affordable substitute for clinical testing, providing an invaluable tool for enhancing diagnostic accuracy in cases with toddler ASD.**

***Keywords*—Autism Spectrum Disorder, Logistic Regression, Machine Learning, Early Detection, Toddler Diagnosis.**

I. INTRODUCTION

When it comes to neurodevelopmental problems, early identification is essential to improving the effectiveness of therapeutic techniques. Since autism spectrum disorder (ASD) has a significant impact on speech, language acquisition, and cognitive abilities, prompt identification requires creative solutions. In order to identify ASD in toddlers, this study uses a machine learning approach that especially makes use of logistic regression. The aim is to optimise the diagnostic procedure by removing superfluous features and determining the minimum number of attributes that make a substantial contribution to the categorization. The model asks nominal questions of carers and uses logistic regression to quantify the answers in order to classify people into ASD or non-ASD groups. This strategy presents a viable path for successful and economical diagnosis while also addressing the challenges brought on by the time-consuming and costly nature of clinical testing. It also supports the urgent need for early action. Motivated by the success of machine learning algorithms in other fields, like educational attention quantification, our work compares with the use of blink patterns to measure attentiveness. Blinks were found to be a useful indicator of attention status, indicating changes in concentration and tiredness, in a study aimed at optimizing pupils' concentration. Similar to this, our approach for detecting ASD makes use of logistic regression to identify trends in carers' answers, which helps to produce a binary classification of ASD or non-ASD. This study

DOI: 10.5281/zenodo.12542285

aims to fill a crucial gap in the field of neurodevelopmental disorders by developing a reliable and useful tool for early ASD detection by reducing the number of characteristics while increasing the predictive accuracy of the model.

II. LITERATURE SURVEY

In [1] after a careful comparison and assessment using eval- uation matrices, the most accurate approach was determined to be logistic regression. Moreover, we are employing the dataset presented in this work to strengthen the validity of our analysis. The AQ-10 questionnaire method for assessing Autism Spectrum Disorder (ASD) traits in children has been adopted [2] with the description of the questions. Addition- ally, the overall research design and methodology have been structured following the principles outlined in the same paper, specifically incorporating feature selection, data preprocessing. The features encompassing the steps for constructing a logistic regression model suitable for clinical applications and effective model evaluation methods, as detailed in [3] can be seamlessly adopted into our research framework. The incorporation of behavioral characteristics, emotional and social attributes, as well as speech and language indicators into our study draws inspiration from the insights provided in [4].

A comprehensive examination of data pre-processing in ma- chine learning, encompassing various challenges encountered across different problem domains is studied [5]. Specifically, it addresses two pivotal aspects within the pre-processing phase:
(i) addressing issues associated with data, and (ii) outlining the optimal steps for conducting data analysis with a best- practice approach. This paper [6] incorporates the principles of machine learning, including algorithms and statistical models, applied in everyday tasks, spanning fields like data mining and predictive analytics. This study [7] enlights with a thor- ough examination of forty-five research that use supervised machine learning to treat Autism Spectrum Disorder (ASD), with a particular emphasis on text analysis and classification techniques. The goal is to provide direction for the creation of statistically, computationally, and clinically sound metho for mining ASD data. The favourable results obtained in [8] indicate the potential usefulness of such a framework for A screening, and the feature exploration and logistic regression for predictive analysis could be a valuable method Li et al. [9], extracted 6 personal characteristics from 851 subjects and performed the implementation of a cross validation strategy for the training and testing of the M models. This was used to classify between patients with a without ASD, respectively. In most cases, it can usually be identifed in its preliminary stages, but the major bottleneck lies in the subjective and tedious nature of existing diagno procedures. As a result, there is a waiting time of at least 13 months, from the initial suspicion to the actual diagnosis. The diagnosis takes many hours, and the continuously growing demand for appointments is much greater than the peak capacity of the country’s pediatric clinics [10].

III. PROPOSED METHODOLOGY

The methodology that has been suggested for the early diagnosis of Autism Spectrum Disorder (ASD) in toddlers is based on a methodical process that uses logistic regression as the main machine learning tool. Carers are asked a series of nominal questions intended to elicit answers that will act as in- put features for the logistic regression model, which starts the diagnostic process. These carefully considered questions aim to gather data on social interactions, communication styles, sensitivity, and other critical behavioural qualities linked to ASD. The replies serve as the foundation for training the logistic regression model, which is characterised by binary outcomes or numerical values. A critical component of the approach is feature selection, which improves the effectiveness and interpretability of the model. The process of removing redundant and less relevant characteristics makes sure that the final set of attributes that the logistic regression model uses is clear and powerful. In order to achieve a balance between prediction accuracy and attribute minimization, which enables a more efficient and clinically meaningful diagnosis, careful feature curation is essential. Using this revised dataset as training, the logistic regression model is able to classify individuals into discrete groups based on whether they are likely to have ASD or not.

Furthermore, the approach offers a prompt and economical replacement, thereby addressing the practical issues with clinical testing. This method, which uses logistic regression as the foundation of machine learning, is in line with the general objective of utilising cutting-edge technology to enhance the diagnosis procedure, especially when it comes to neuro developmental abnormalities in young children. The emphasis on logistic regression highlights the technique’s effective- ness in binary classification tasks as well as its potential to significantly aid in the early detection of ASD, providing opportunities for prompt intervention and support. Setting out on a comprehensive process, the subsequent actions are made to ensure a precise diagnosis of Autism Spectrum Disorder:

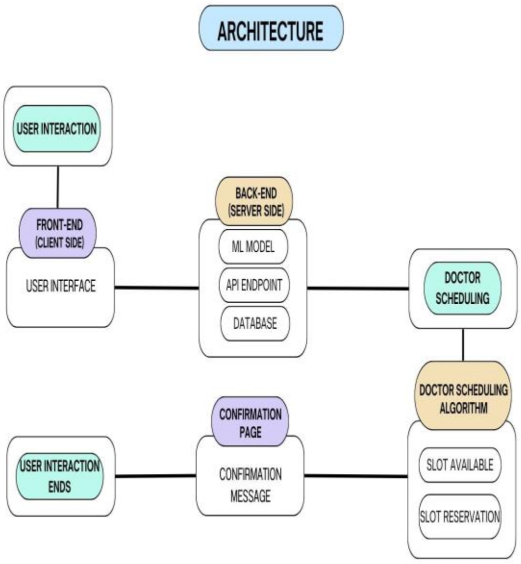

Fig. 1. Architecture Diagram

- **Data collection**: Gather historical and current data to create a comprehensive dataset by obtaining datasets from Dataset: Dataset link here. Accessed 1 Oct 2023.
- **Dataset Description:**
  - The dataset utilized for this study consists of records from 1054 toddlers.
  - Data attributes include responses to the Q-Chat-10 questionnaire, age of toddlers, sex, ethnicity, history of jaundice, family history of ASD, and test comple- tion details.
  - All data attributes were carefully curated to ensure relevance and accuracy in predicting ASD traits.
- **Data Preprocessing:**
  - Analysis of missing values: After thorough inspec- tions, the dataset was found to be free of missing values.
  - Coding categorical variables: To ensure interoper- ability with machine learning algorithms, string and Boolean data types were encoded suitably.
  - Engineering features: The Q-Chat-10 answers were converted to binary values in accordance with the given instructions. Furthermore, toddler age was kept as a numerical element.
  - Data splitting: To preserve class balance, the dataset was split using a stratified method into training and testing sets.
- **Model Selection and Training:**
  - Evaluation of machine learning models: A num- ber of classification techniques, such as Logistic Regression(LR), Naive Byes(NB), Support Vector Machines (SVM), K-Nearest Neighbors(KNN) and Random Forest Classifier (RFC) were evaluated by comparing their accuracies and F1 scores.
  – Training models: The best option turned out to be logistic regression since it was easy to understand, effective, and had good performance indicators.

TABLE I
VARIABLE IN DATASET CORRESPONDING Q-CHAT-10
TODDLER FEATURES

Variable	Corresponding Q-chat-10-Toddler Features
A1	Does your child look at you when you call his/her name?
A2	How easy is it for you to get eye contact with your child?
A3	Does your child point to indicate that s/he wants something? (e.g., a toy that is out of reach)
A4	Does your child point to share interest with you? (e.g., pointing at an interesting sight)
A5	Does your child pretend? (e.g., care for dolls, talk on a toy phone)
A6	Does your child follow where you're looking?
A7	If you or someone else in the family is visibly upset, does your child show signs of wanting to comfort them? (e.g., stroking hair, hugging them)
A8	Has your child spoken their proper first words yet?
A9	Does your child use simple gestures? (e.g., wave goodbye)
A10	Does your child stare at nothing with no apparent purpose?

TABLE II
FEATURES OF Q-CHAT-10-TODDLER SCREENING

Feature	Type	Description
Age	Number	Toddlers (months)
Score by Q-chat-10	Number	1-10 (Less than or equal to 3: no ASD traits; ¿ 3: ASD traits)
Sex	Character	Male or Female
Ethnicity	String	List of common ethnicities in text format
Born with jaundice	Boolean	Whether the case was born with jaundice (Yes or No)
Family member with ASD history	Boolean	Whether any immediate family member has a PDD (Yes or No)
Who is completing the test	String	Parent, self, caregiver, medical staff, clinician, etc.
Why are you taking the screening?	String	User input textbox
Class variable	String	ASD traits or No ASD traits (automatically as- signed by the ASDTests app)

- **Model Integration into Web Application:**
  - A web application was developed with the Flask framework to enable user interaction.
  - Pickle was used to serialise the learned

  logistic regression model, which was then easily included into the Flask application.
  - Front-end development: To create an easy-to-use user interface for effective data entry, HTML, CSS, and JavaScript were used.
  - Form-based questionnaire: Users are asked to provide extra demographic data in Addition to answering ten important questions (A1 through A10) from the Q-Chat-10 questionnaire.
  - Classification output: Based on the responses submit- ted, the machine learning model divides the user into ASD and non-ASD groups at the time of submission.
- **Validation and Performance** Evaluation:
  - Metrics for evaluating the performance of the logistic regression model were calculated, including accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC-ROC).
  - Validation of the model’s predictions using a different test dataset was done to make sure the model was reliable and resilient.
  - Results interpretation: The relevance of early identification and intervention was emphasised by placing the categorization outcomes within the framework of ASD screening.
- **Feedback Mechanism and Future Enhancements:**
  - Integration with Firebase: Firebase is used to handle user comments and queries, increasing user engage- ment and yielding insightful data for further revi- sions.
  - Iterative improvements to the web application and machine learning model will be guided by user feedback and continuing research endeavours. This is an example of continuous improvement.
- **Doctor Appointment Booking System:**
  - Seamless Integration: The website gives caregivers an easy-to-use interface for making doctor appointments straight through the platform, giving them a handy way to take care of their child’s medical requirements.
  - Real-Time Availability: Patients can choose appropriate appointment times that work with their schedule by viewing the real-time availability of medical specialists in childhood development and autism spectrum disorders.
  - Simplified Process: The appointment booking system easily integrates with the scheduling platforms of healthcare providers, guaranteeing effective coordi- nation and communication between medical professionals and caregivers.
  - Improved Accessibility: The website lowers obstacles to receiving specialized treatment by providing online appointment booking services, especially for families who live in rural places or have logistical difficulties.
  - Privacy and Security: Strict protocols are put in place to safeguard user data privacy and guarantee adherence to healthcare laws, upholding confidence in the booking procedure.

A reliable and precise classification system for diagnosing Autism Spectrum Disorder in individuals is ensured by this methodical approach.

IV. RESULTS

The model effectively predicts the likelihood of autism in toddlers, facilitated by a user-friendly website that sim- plifies the process for guardians or caregivers to complete questionnaires and receive predictions. Through feature selec- tion, attributes such as ethnicity were identified and removed from the dataset, resulting in improved accuracy. Further- more, after experimenting with various machine learning al- gorithms(Naive Bayes (NB), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Random Forest Classifier (RFC)), linear regression emerged as the most accurate choice based on metrics such as accuracy, precision, and recall. However, the unavailability of real-time data has impacted the results. Nevertheless, this project serves as a foundational step towards a larger and more scalable initiative, with numerous opportunities for additional features and advancements. Given the rising prevalence of autism among children today, intuitive web applications like ours play a crucial role in facilitating easy prediction. In the future, this project can be expanded to offer additional healthcare resources or necessary interventions for toddlers identified as at-risk for autism.

TABLE III
A COMPARISON OF THE APPLIED ML MODELS

ML Models	Accuracy	F1 Score
LR	97.15%	0.98
NB	94.79%	0.96
SVM	93.84%	0.95
KNN	90.52%	0.93
RFC	81.52%	0.88

V. *DISCUSSION*

A. Interpretation of results

The dataset provided by Dr. Fadi Thabtah, consisting of 1054 instances and 18 attributes, served as the foundation for our project. Upon initial analysis, we observed a noticeable decline in the model's accuracy when all attributes were included. To mitigate this issue and enhance the model's performance, we employed rigorous feature selection techniques. These techniques enabled us to identify and eliminate attributes that contributed minimally to the predictive outcome. By refining the dataset in this manner, we witnessed a remark- able improvement in the model's accuracy, with the accuracy rate soaring from an initial 82% to an impressive 96%. This significant enhancement underscores the critical role of feature selection in optimizing machine learning models for robust and reliable predictions.

B. Comparison with existing literature

After rigorously testing five machine learning algorithms—Logistic Regression (LR), Naive Bayes (NB), Sup- port Vector Machine (SVM), K-Nearest Neighbors (KNN), and Random Forest Classifier (RFC)—using the dataset, our analysis revealed compelling insights. Notably, among these algorithms, linear regression emerged as the frontrunner, demonstrating the highest accuracy rate. This observation underscored the superior performance of linear regression in accurately predicting outcomes based on the dataset's attributes. Consequently, we confidently selected linear regression as the optimal model for our project, leveraging its robust predictive capabilities to drive meaningful insights and outcomes. This strategic decision aligns with our objective of harnessing advanced machine learning techniques to deliver impactful solutions in the domain of autism detection and intervention.

C. Limitations

One notable limitation encountered during the course of this project was the scarcity of real-time data, which posed a significant challenge to the accuracy and effectiveness of our model. Despite our efforts to access relevant data from clinics and institutions specializing in the care of autistic toddlers, we encountered widespread reluctance among stakeholders to share their datasets. This reluctance stemmed from various factors, including privacy concerns, institutional policies, and logistical constraints. As a result, our model's performance was constrained by the absence of real-time data, limiting its ability to accurately predict outcomes in real-world scenarios. Despite this limitation, we remain committed to addressing this challenge and exploring alternative strategies to acquire and integrate real-time data into our model, thereby enhancing its accuracy and relevance in clinical settings. By overcoming this barrier, we aim to further validate and refine our model, ensuring its effectiveness in accurately identifying and sup- porting toddlers with autism spectrum disorder.

D. Future research

a notable limitation encountered during this project was the scarcity of real-time data. Many clinics and institutions involved in the care of autistic toddlers exhibited reluctance in sharing data, which constrained the model's accuracy as it was not tested with real-time data. Future endeavors will focus on gathering additional real-time data and incorporating information about autism and its management strategies. Additionally, enhancements to the user interface are planned to ensure that the website serves as a comprehensive resource for accurate autism detection and caregiving guidance. Through these efforts, we aim to make a meaningful impact in supporting individuals and families affected by autism spectrum disorder.

E. practical application

The increasing prevalence of autism diagnoses among toddlers in recent years highlights the critical need for tools like our website. Lifestyle factors such as excessive screen time, poor dietary habits, parental stress, and limited real-world interaction may contribute to the development of autism in toddlers. Our website, powered by a highly accurate machine learning model, offers a valuable solution for predicting autism in this vulnerable population. Its intuitive and user-friendly interface ensures accessibility for guardians and caregivers, providing them with essential information and resources for effectively managing toddlers with autism spectrum disorder. Through our platform, we aim to offer support and guidance to families navigating the complexities of autism diagnosis and care, ultimately enhancing the well-being and quality of life for affected individuals.

VI. *CONCLUSION*

ASD Predict stands as an innovative tool in the early detection of autism. Harnessing the power of logistic regression through sophisticated machine learning techniques, it has been shown to be accurate in predicting the likelihood of Autism Spectrum Disorder (ASD) in young children This innovation is poised to revolutionize pediatric health care

The project not only holds great promise for improving the lives of children with ASD but also highlights the critical importance of early intervention for developmental problems. By enabling health care professionals and caregivers to identify potential developmental problems early, this tool opens the door to timely intervention, ultimately resulting in results and quality of life occurs for individuals with ASD

Looking ahead, the applicability of ASD Predict expands beyond what was originally available. Its success highlights the transformative impact of machine learning on health care and paves the way for future innovations in the early detection and intervention of a wide range of developmental disorders. As we continue to refine and expand this technology, we are moving closer to a future where every child, regardless of their developmental challenges, has the opportunity to succeed.

ACKNOWLEGMENT

The authors of this paper are, Jesna Susan Reji, Gauri Santhosh, Glady Prince, and Abdul Musawir. We would like to thank our project supervisor Dr. Indu John for her guidance and support throughout the duration of this project.

REFERENCES

[1] Kaushik Vakadkar, Diya Purkayastha, Deepa Krishnan, (2021). " Detec- tion of Autism Spectrum Disorder in Children Using Machine Learning Techniques", 386, 5–7.

[2] Md Delowar Hossain , Muhammad Ashad Kabir , Adnan Anwar and Md Zahidul Islam, "Detecting autism spectrum disorder using machine learning techniques An experimental analysis on toddler, child, adoles- cent and adult datasets", (2021) 9:17, 3-5

[3] Maren E. Shipe, Stephen A. Deppen, Farhood Farjah, Eric L. Grogan, An Overview of Constructing Prediction Models for Clinical Applications Using Logistic Regression, J Thorac Dis 2019;11(Suppl 4):S576-S580.

[4] Brooke Ingersoll, David Z. Hambrick, The relationship between the broader autism phenotype, child severity, and stress and depression in parents of children with autism spectrum disorders Elsevier 1–15 Volume 5, Issue 1, 2019 Pages 340-343

[5] Kiran Maharana , Surajit Mondal, Bhushankumar Nemade: A review: Data pre-processing and data augmentation techniques Global Transi- tions Proceedings 3 (2022) 91–99.

[6] Batta Mahesh, Machine Learning Algorithms - A Review, International Journal of Science and Research (IJSR) (2018) 381-383.

[7] Kayleigh K. Hyde, Marlena N. Novack, Applications of Supervised Machine Learning in Autism Spectrum Disorder Research, 2019.

[8] A machine learning autism classifcation based on logistic regression analysis, Thabtah et al. Health Inf Sci Syst (2019) 7:12

[9] Parikh MN, Li H, He L. Enhancing diagnosis of autism with optimized machine learning models and personal characteristic data. Front Comput Neurosci. 2019. https://doi.org/10.3389/fncom. 2019.00009.

[10] Kosmicki JA, Sochat V, Duda M, Wall DP. Searching for a minimal set of behaviors for autism detection through feature selection-based machine learning. Transl Psychiatry. 2015. https://doi. org/10.1038/tp.2015.7.