Grenze International Journal of Engineering and Technology, June Issue

Prediction of Autism Spectrum Disorder using Gene Expression

Ajitha S[1], Nisha Sangavi V[2] and Lavanya M N[3]
[1-3]M S Ramaiah Institute of Technology /MCA, Bangalore, India
Email: ajitha@msrit.edu, {nisha.wrk4, lavanyamn216}@gmail.com

***Abstract*—Autism Spectrum Disorder(ASD) is defined by deficits in social communication and repetitive sensory-motor behaviors that are heavily influenced by both genetic factors and environmental triggers. Genetic factors play a vital role in ASD, although environmental conditions also contribute to the disorder. Rare copy-number variations in DNA are a significant risk factor for ASD. The current diagnosis process for ASD involves clinically based standardized tests, resulting in a lengthy and expensive procedure. To improve diagnosis accuracy and speed, machine learning approaches are being implemented in conjunction with traditional methods. We have implemented different machine learning algorithms, such as Support Vector Machine, Linear Regression Analysis, Naive Bayes, and Random Forest, to classify medical data, determine the association between variables, make real-time predictions and signify attribute relationships in computational biology and genetics research.**

***Keywords*— Autism Spectrum Disorder, Gene expression data, genetic factors, computational biology, Machine Learning, Classification algorithm.**

I. INTRODUCTION

ASD is a complex neurodevelopmental disorder that affects social interaction, communication, and behavior. People diagnosed with ASD have a range of symptoms, which is why it is referred to as a "spectrum" disorder. Some of the classic symptoms include problems in communicating, particularly in social situations, obsessive-compulsive hobbies, and repeated mannerisms. Early and accurate diagnosis of ASD is critical for successful intervention and improving the lives of children with ASD. However, diagnosing ASD can be difficult due to the wide range of symptoms and varying degrees of severity.

ASD is a complex neurodevelopmental complaint that affects interaction with society and causes communication problems and behavior problems that differ from other normal people. Some individuals may have mild symptoms and relatively high functioning, while others may have more significant challenges that require extensive support [9]. Predicting ASD is a challenging task that requires the use of advanced machine learning algorithms and deep learning techniques to accurately identify subtle differences in behavior and social interaction between individuals. Machine learning techniques appear to be a viable alternative for streamlining the entire diagnosis process and assisting families in receiving vital therapies sooner. In recent years, researchers have created a number of machine learning algorithms to predict ASD. These techniques include collecting characteristics from data and then training a machine learning model to categorize the data into one of the categories, such as autistic or non-autistic. The performance of these models has been encouraging, with excellent accuracy.

Grenze ID: 01.GIJET.9.2.429_1

This paper proposes a model to aid in the prediction of ASD in an individual so that a diagnosis can be made and further treatments can be pursued. The dataset used is the Prediction of Autistic Spectrum Disorder is from the SFARI GENE dataset. The datasets provide information on the many aspects that influence the disorder's prediction. For each dataset, machine learning methods such as decision tree, random forest, logistic regression, and support vector classifier are used to discover the best model. Confusion metrics are used to analyze and compare each model from every angle in order to determine which machine learning algorithm is best for the dataset under consideration. The goal is to increase the accuracy and speed of diagnosis, ultimately leading to earlier intervention and better outcomes for individuals with ASD.

II. Literature Review

About 10-20% of instances of autism spectrum disorder (ASD) are likely due to defined mutations, genetic disorders, and de novo copy number variation, with none of these known causes accounting for more than 2% of cases. No evidence supports a single cause of ASD by specific chemicals or disorders. Molecular connections across ASD-related diseases can uncover dysregulated signaling pathways [1]. AutDB is an integrated, disease-driven database model that gathers and annotates data on both small and monogenic candidates for ASD risk. It involves searching, annotating, storing and visualization to identify 26 potential genes. Since 2003, there has been a constant stream of single gene mutations in ASD, and the ratio of reported single-gene alterations to candidates discovered through genetic association studies is changing. It is discovered numerous changes in the CNS's development, including reduced programmed cell death and/or increased cell proliferation, altered cell migration with disturbed cortical and subcortical cytoarchitectonics, abnormal cell differentiation with smaller neuronal size, and altered synaptogenesis [2].

Research findings from genetics and molecular studies indicate that synaptic function impairment is a significant factor in ASD. Genes that regulate synaptic functions have been observed to be altered or mutated in individuals with ASD. This can lead to altered sensory processing, cognitive deficits, hyperactivity, and seizures by disrupting the balance between excitatory and inhibitory neurotransmission [3]. ASD exhibits significant genetic heterogeneity, including both allelic and site heterogeneity. The new results suggesting up to 234 loci contributing to ASD risk39 are likely even an underestimation, according to the exome sequencing research [4]. Linkage analyses have revealed that En2 has emerged as a strong candidate for association with the autism phenotype, a specific genetic mutation in NLGN4 has been identified as being responsible for rare cases of mental retardation and/or pervasive developmental disabilities, and a linkage region on chromosome 17q has been confirmed in independent samples. These findings present opportunities for substantial advancement in the field of autism diagnosis [5]. Synaptic function impairment and disrupted neurodevelopment processes [6]. The discovery of a higher proportion of concordance (sharing the diagnosis) among MZ pairs suggests that genes play a significant role in the etiology of an illness. [8]. In [9-10] authors have addressed ASD using gene expression, and have discussed the importance to detect ASD in nurturing children, CDC released a report citing a prevalence of ASD of 16.8 per 1000 in 2014, a significant increase from 14.7 per 1000 in 2010. a well-trained classification model (based on transfer learning) to detect autism from an image of a child. These gene expression data combined with various deep learning technique showcases heterogeneity in ASD which is lengthy and time consuming process.

From the given information we can infer that ASD exhibits significant genetic heterogeneity, with numerous genes and mutations implicated in its development. Hence we have implemented machine learning techniques to build a prediction model for accurate and efficient ASD Diagnosis.

III. METHODOLOGY

Genomic research has confirmed that autism has a strong hereditary component, with a considerable proportion of its risk associated with rare, high-impact genetic variation. Such variation can include chromosome abnormalities, copy-number variations, and single-nucleotide variations. These studies have also highlighted a remarkable degree of genetic diversity, indicating that both sporadic mutations and rare inherited genetic variations can contribute to autism risk. Moreover, these variations are likely to be dispersed across several genes. The figure [1] depicts the system architecture of our model.

The data was collected from SFARI dataset which contains 1091 values and several attributes including gene-symbol, gene-name, ensemble-id, chromosome, genetic-category, gene-score, syndromic, eagle, and number-of-

reports. Gene-name refers to the name of the gene associated with autism spectrum disorder (ASD), while gene-symbol is a unique symbol assigned to each gene. the dataset's gene-score is divided into 4 different categories

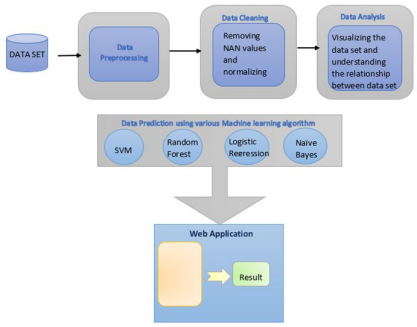

Figure 1: System Architecture

a. S (Syndromic): Mutations associated with a higher risk of ASD and additional characteristics beyond what's required for an ASD diagnosis.
b. 1 (High Confidence): The most rigorous threshold of genome-wide significance. Requires at least 3 de novo mutations and a false discovery rate threshold of <0.1.
c. 2 (Strong Candidate): Genes with 2 reported de novo likely gene-disrupting mutations.
d. 3 (Suggestive Evidence): Genes with a single reported de novo likely gene-disrupting mutation or rare inherited mutations, which have not undergone rigorous statistical comparison with controls.

1. The Genetic – category refers to the classification of a gene into one of the 4 genetic categories.
2. Rare Single Gene Mutation
3. Rare Single Gene Mutation, Functional
4. Rare Single Gene Mutation, Genetic Association
5. Rare Single Gene Mutation, syndromic.

- The chromosome section of the dataset provides information about the specific chromosomal location where a gene can be found.
- Ensembl ID (Ensembl Gene ID): A system used to assign unique identifier codes to genes in different species, consisting of five parts [e.g. ENSG00000183044] where it is represented as ENSG (species) (object type) (identifier) (version).
- Chromosome Number: A field that identifies the chromosome number responsible for the change in the gene, as specified in our dataset.
- Gene Score: A numerical value assigned to a gene that reflects its relative importance or significance in a certain context, such as gene expression analysis. The exact calculation of gene scores can vary depending on the specific analysis and the research question. Gene scores can provide a simplified way to rank genes and prioritize further study or analysis. In our case, the gene score is calculated using de novo assembly.
- De novo assembly: A method for constructing genomes from a large number of (short or long) DNA fragments, with no prior knowledge of the correct sequence or order of these fragments.
- Number of Reports: A field that stores how many similar conditions have been found so far.
- Eagle Score: A specific metric used to measure the similarity of gene expression profiles across different samples. The eagle score is calculated by comparing the expression of a given gene in one sample to the expression of the same gene in another sample and can be used to identify genes that have similar patterns of expression across different conditions or time points.

B. Data Cleaning

The Ensembl Gene ID is a composite identifier that includes both a string prefix and numerical values. However, to facilitate machine learning analyses, the string prefix (ENSG-) has been removed, and only the numerical values are retained in the dataset. Pre-processing steps were conducted to ensure the quality of the dataset is precise, which included replacing all missing values with zeros (NaN to 0), and normalization of the data to ensure a consistent scale across all features. During the data cleaning process, redundant information such as gene symbol and gene-name were eliminated, and only the Ensembl ID was retained as the unique identifier for

genes in the dataset. The Eagle values were kept persistent. The dependent variable in this dataset is the attribute called "Syndromic," which is binary in nature containing values of 0 or 1. A value of 0 indicates a negative association with ASD (autism spectrum disorder), while a value of 1 indicates a positive association with ASD. The figure [2] shows the distribution of the dependent variable in our dataset.

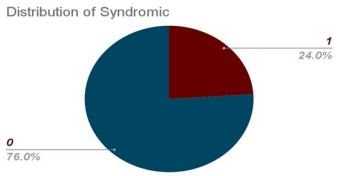

Figure 2: Frequency of syndromic values

Supervised learning is a type of machine learning that can be applied to both classification and regression problems. One popular approach in supervised learning is ensemble learning, where multiple classifiers are combined or grouped together to solve complex problems and improve model performance. In this study, we focus on implementing the Random Forest algorithm, which is an ensemble learning approach that aims to achieve higher performance metrics compared to other algorithms.

Algorithm

Algorithm Random Forest Classifier()

{

Input: Data set with size S which consists on [mxn] values Output: Predicting output “Test set result”

{

1. Data Pre-processing of Dataset D

//Import the Libraries //Import the Dataset

D = Dataset

//Extracting the independent and Dependent variable

X -> Independent variables

Y -> Dependent variable

//Splitting data into training and testing data in 8:2 ratios such as X_train, x_test, y_train, y_test

//future Scaling

Fitting the Random Forest algorithm to the training set

//with entropy as criterion and randomness = 10

classifier=RandomForestClassifier(n_estimator=10, criterion=”entropy”)

//importing RandomForest function to classifier variable. classifier.fit(x_train, y_train)

3.Predicting the Test set result

y_pred -> classifier.predict(x_test) //Output

4. Creating Confusion Matrix for Accuracy

cm -> confusion_matrix (y_test, y_pred) }

C. Data Analysis and Prediction

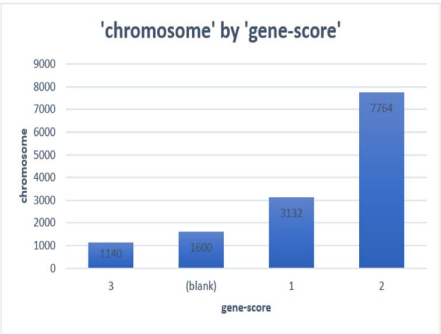

Figure 3:: Gene score on chromosomes

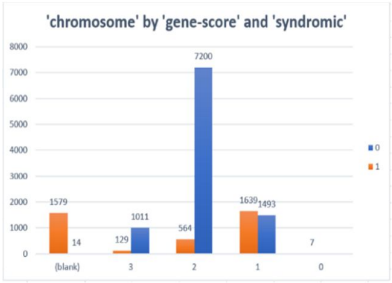

Figure 4:chromosome by gene-score and syndromic

From figure [3] we can understand that chromosome number two has the highest gene score with a value of "2" most of the time, followed by chromosome one with a gene score of "1". This suggests that most of the

chromosomes are either Strong Candidates or High Confidence. The blank values indicate the number of times the gene score is syndromic "s", indicating an uncertain relationship with autism spectrum disorder. In figure [4], the y-axis represents the number of genes, while the x-axis represents the different gene scores. The values 0 and 1 on the x-axis indicate the absence or presence of ASD, respectively. The blank section on the x-axis corresponds to syndromic cases. From the graph, we observe that genes with a gene score of 1 (High Confidence) have a higher tendency to be associated with positive ASD (value 1 on the y-axis) as compared to gene scores 2 and 3. On the other hand, genes with gene score 2 are mostly associated with negative ASD (value 0 on the y-axis). Furthermore, the blank syndromic section shows that there is an equal distribution of positive and negative ASD cases.

After pre-processing the data using "sklearn" preprocessing techniques and replacing all NaN values with "0", the dataset is split into two sets: "test data" and "training data" in a 2:8 ratios. The training data is used to evaluate various machine learning algorithms to determine the best algorithm for predicting the syndrome. The correlation relationship between each attribute is shown in the figure [6], with lighter colors indicating higher correlations between attributes. This figure indicates that the dataset is not overfitted.

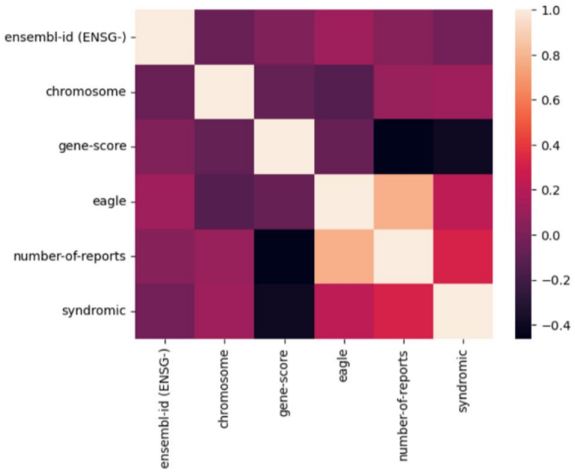

Figure 5: Correlation of Attributes.

IV. Performance Metrics

- *True Positive (TP)*: The number of values predicted is true and is true.
- *False Negative (FN)*: The number of values predicted is false, but is true.
- *True Negative (TN)*: The number of values predicted is false and is false.
- *False Positive (FP)*: The number of values predicted is true, but is false.

Accuracy: It is the ratio of the number of correct predictions to the total number of predictions.

- $Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions}$

- *Sensitivity / Recall:* Sensitivity is the fraction of individuals who will be correctly predicted as positive class. High sensitivity means a greater number of individuals have been correctly predicted to have ASD.

- $Sensitivity = \frac{TP}{TP+FN}$

- *Specificity:* Specificity is the fraction of individuals who will be correctly predicted as negative class. High specificity means a greater number of individuals have been correctly predicted not to have ASD.

- $Specificity = \frac{TN}{TN+FP}$

V. Results

Four distinct machine learning algorithms, including Support Vector Machine, Random Forest, Logistic Regression, and Naive Bayes, have been utilized to analyze the preprocessed dataset. Among all the algorithms,

Random Forest exhibited the highest accuracy, sensitivity, and specificity based on various performance metrics. Therefore, Random Forest has been chosen for further prediction of the syndrome

TABLE I: PERFORMANCE MATRIX OF DIFFERENT ALGORITHMS

Algorithm	Accuracy	Sensitivity	Specificity
SVM	0.7743	0.9728	0.1435
Random Forest	0.831	0.8524	0.7223
Logistic Regression	0.7442	0.7806	0.4347
Naive Bayes	0.7625	0.7968	0.5185

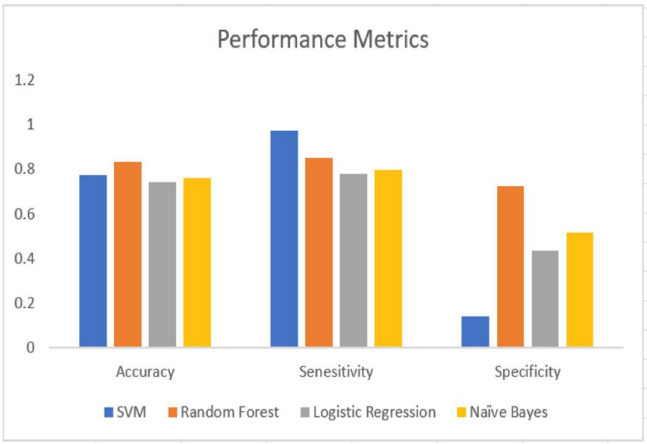

Figure 6: Performance Matrix of different algorithms

We implemented the Django framework to create a local website. The inputs were taken such as patient details like ensemble-id, chromosome number, Gene score and eagle value if not 0 and number of reports which further predict whether the given data is ASD positive or ASD negative.

Figure 7: Home Page

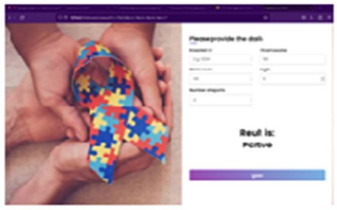

Figure 8:Prediction page for negative value

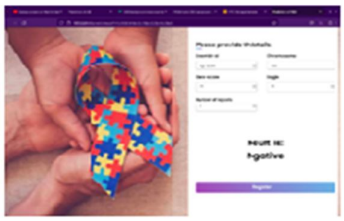

Figure 9:Prediction page for positive values

VI. CONCLUSION

The neurological disease known as autism spectrum disorder (ASD) is characterized by issues with speech, social interaction, and repetitive behaviors. In this work, we used the SFARI gene set dataset, which included 1094 samples with autistic and non-autistic results, from the official website. To analyze our dataset, we used a variety of machine learning techniques, such as Support Vector Machine, Random Forest, Logistic Regression, and Naive Bayes Algorithm. The confusion matrix, a tool for assessing the precision of classification and regression algorithms, was used to gauge the effectiveness of these methods. The findings revealed that the

algorithm's respective levels of accuracy were 77.43%, 83.11%, 74.43%, and 76.26%. These results led us to the conclusion that Random Forest offered the dataset's most dependable accuracy. Therefore, we proceeded further with Random Forest to analyze the dataset and produce the efficient result of autistic positive or negative. Furthermore, the system will save user input as values in a CSV file and count the occurrences of the same ensemble ID while prompting users to enter their locality. The findings can be used on the website for further analysis.

REFERENCES

[1] Abrahams, B., Geschwind, genetics: on the threshold of a Rev Genet 9, https://doi.org/10.1038/nrg2346 https://doi.org/10.1038/nrg2346

[2] Saumyendra N. Basu, Ravi Kollu, Sharmila Banerjee-Basu, AutDB: a gene reference resource for autism research, Nucleic Acids Research, Volume 37, Issue suppl_1, 1 January 2009, Pages D832–D836, https://doi.org/10.1093/nar/gkn835

[3] Scott M. Myers, Thomas D. Challman, Raphael Bernier, Thomas Bourgeron, Wendy K. Chung, John N. Constantino, Evan E. Eichler, Sebastien Jacquemont, David T. Miller, Kevin J. Mitchell, Huda Y. Zoghbi, Christa Lese Martin, David H. Ledbetter, Insufficient Evidence for "Autism-Specific" Genes, The American Journal of Human Genetics, Volume 106, Issue 5,2020 Pages 587-595, ISSN 0002-9297, https://doi.org/10.1016/j.ajhg.2020.04.004.

[4] Chaste P, Leboyer M. Autism risk factors: genes, environment, and gene-environment interactions. Dialogues Clin Neurosci. 2012 Sep;14(3):281-92. doi: 10.31887/DCNS.2012.14.3/pchaste. PMID: 23226953; PMCID: PMC3513682.

[5] Abha SR Gupta,1,2 Matthew W State Correspondence Matthew W. State Child Study Center 230 South Frontage Road, Yale University CT 06520 New Haven, E-mail: matthew.state@yale.edu

[6] Brimacombe et al., 2007 M. Brimacombe, X. Ming, M. Lamendola Prenatal and birth complications in autism Matern. Child Health J., 11 (2007), pp. 73-79 https://link.springer.com/article/10.1007/s10995-006-0142-7

[7] Zoghbi HY, Bear MF. Synaptic dysfunction in neurodevelopmental disorders associated with autism and intellectual disabilities. Cold Spring Harb Perspect Biol. 2012; 4: a009886.

[8] Lepeta K, Lourenco MV, Schweitzer BC, Martino Adami PV, Banerjee P, Catuara-Solarz S, et al. Synaptopathies: synaptic dysfunction in neurological disorders—A review from students to students. J Neurochem. 2016; 138:785–805.

[9] Arun P, Chavan BS. Survey of autism spectrum disorder in Chandigarh, India. Indian J Med Res. 2021 Mar;154(3):476-482. doi: 10.4103/ijmr.IJMR_930_19. PMID: 35345073; PMCID: PMC9131804.

[10] Rastegari M, Salehi N, Zare-Mirakabad F. Biomarker prediction in autism spectrum disorder using a network-based approach. BMC Med Genomics. 2023 Jan 23;16(1):12. doi: 10.1186/s12920-023-01439-5. PMID: 36691005; PMCID: PMC9869547.

[11] Bao, B., Zahiri, J., Gazestani, V.H. et al. A predictive ensemble classifier for the gene expression diagnosis of ASD at ages 1 to 4 years. Mol Psychiatry 28, 822–833 (2023). https://doi.org/10.1038/s41380-022-01826-x.

[12] Vakadkar, K., Purkayastha, D. & Krishnan, D. Detection of Autism Spectrum Disorder in Children Using Machine Learning Techniques. SN COMPUT. SCI. 2, 386 (2021). https://doi.org/10.1007/s42979-021-00776-5

[13] Siddiqui S, Gunaseelan L, Shaikh R, Khan A, Mankad D, Hamid MA. Food for Thought: Machine Learning in Autism Spectrum Disorder Screening of Infants. Cureus. 2021 Oct 12;13(10): e18721. doi: 10.7759/cureus.18721. PMID: 34790476; PMCID: PMC8584605.

[14] Rosenbaum, B. Yamrom, Y.H. Lee, G. Narzisi, A. Leotta, et al. De novo gene disruptions in children on the autistic spectrum Neuron, 74 (2012), pp. 285-299 https://www.sciencedirect.com/science/article/pii/S0896627312003406

[15] Paula, C.S., Ribeiro, S.H., Fombonne, E. et al. Brief Report: Prevalence of Pervasive Developmental Disorder in Brazil: A Pilot Study. J Autism Dev Disord 41, 1738–1742 (2011). https://doi.org/10.1007/s10803-011-1200-6