*Predicting autism using most impactful gene expressions in the autism spectrum.*

Syed Shaheryar Qadir   
*Big Data Analytics  
School of Science, Technology & Engineering*  
*St. Thomas University*Miami Gardens, USA  
sqadir@stu.edu

*Abstract*— Data driven approach to identify and solve problems is evident in the medical field like never before, which is why autism intervention researches such as this paves way towards the advancement of new treatments and precision medicine best suited for specific cases. This analysis looks to predict the probability of autism by identifying the most impactful genomic symbols that make up the immune system to uncover trends of growth or deterioration of gene expressions while cross validating across six prediction models.

Keywords— autism, gene expression, spectrum disorder, neuronal induction, prediction, precision medicine, homo sapiens, RNA genes.

# Introduction

**Which genetic markers can help in comprehending not just autism as a whole but also of the developmental stage, particular biology or behavioral needs of each person on the autism spectrum?**

The motivation for this study is to develop a knowledge-guided bioinformatics model which identifies convergent paths in patients with Autism Spectrum Disorder (ASD). This study revolves around human species in which neurons are derived from patients at two different time points to study the progression of genome levels for the subjects under study. This study will help in working towards the advancement of enhancing lives today and expedite a spectrum of solutions for the generation of tomorrow. Perhaps, improve future researchers evaluate new treatments based on the genome progress shown by these patients under study. This study is critically important as we have already entered an era where genetic data and biological markers allow the use of precision medicine best suited for a specific case.

The purpose is to find and promote solutions, across the spectrum and better assist with the diagnosis of idiopathic ASD. It is a widely understood fact that autism in a child is difficult to diagnose at the early ages of development and most children do not show signs of autistic behavior until they are at least two or three years old. [1] By implementing this study the gene expressions can be compared with the ones of a health person and an individual who had ASD. This in turn would help in looking at the neurodevelopmental disorders and understanding at an early stage such as to what can be done to rectify the genetic markers.

The approach is to identify the names of the genomic symbols which make up the immune system so that an exploratory data analysis can be performed on that sub-class of the data. Furthermore, build classifier models and use machine learning techniques such as Support Vector Machine, Artificial Neural Network, Naïve Bayes etc. to predict ASD in an individual with similar levels of gene expression, and finally present the results in a dashboard built using Power Bi along with accuracies of the models built.

# BACKGROUND

As mentioned before that the topic of this study is not something that has not been researched upon previously, the background section of this paper uncovers some interesting approaches carried out by different researchers [2],[3], [4], [5], [6] and [7].

The art of predicating and predicting underlying disease has been a topic of interest for a while now which if identified at the right time can help in identifying risk factors and or biomarkers for neurodegenerative diseases. With the evolution of science and medicine, a way has been paved for the modern science interacting with various disciplines to utilize a structured set of ideas using machine learning algorithms and predictive modeling. It has been found that Autism spectrum disorder (ASD) is a heterogeneous spectrum of disorders with a commonness of one in every 60 to 70 youngsters [1]. The core indications fluctuate in seriousness, behavioral deficits in social correspondence and interaction, dull practices, and confined social interests and as a result the word Autism is coined as a “spectrum disorder” due to the exceptional heterogeneity caused by these core symptoms. [2].

A previous study performed by Nelson [4] collected data from infants who were classified into three groups based on their original family recruitment group. The objective of this study was to determine a positive or negative diagnosis of ASD in infants based on the original family recruitment group. Infants belonging to the low risk controls (LRC) family recruitment group had at least one older sibling and no first-degree relatives with a known developmental disorder, the base of data collection was done from giving a questionnaire to perform a screening. Consequently, infants who were put into the high risk for ASD (HRA) recruitment group had an elder sibling with a strong diagnosis of ASD (not due to a known genetic disorder; e.g. Fragile X syndrome). The elder siblings all had expert clinical community diagnoses, which were confirmed by a member of the study staff using the Social Communication Questionnaire (SCQ). Moreover, the prediction of the binary diagnostic outcome, either ASD or not-ASD, was computed from the nonlinear features mentioned above using a leave-one-out cross validation procedure. This study uses the nonlinear values that were computed from the relative short segments of a resting state EEG signal that contained information to be used as biomarker for the individual profiles which would as a result enable an early prediction of a future outcome of ASD, this can be identified as the goal of this study. It is evident that this study focused on classifying data and using a cross validation method to predict autism which in turn makes room for a more data driven approach using neuronal induction to monitor gene growth which can be summed up as the gist of this research.

A similar study was done by Levin that involved studying 3-month-old infants who belonged to such families which have a high familial risk for autism [5]. This study compared the first 3 years of those infants pertaining to familial risk for autism and deduced the basis for declaring them high risk for autism (HRA) if they had at least an elder sibling with a community diagnosis of ASD. This study used SPSS and MATLAB to study the median difference in group-averaged power spectrum by using a frequency domain-based bootstrapping algorithm. Although this study points to the potential of brain-bases markers for the identification of neurodevelopmental disorders it doesn’t cover the mechanism underlying its findings.

Another study done by Arpit that differs highly from the previous mentioned approaches utilizes big data and machine learning techniques to predict for an underlying neurological disorder [6]. The main technique used is a machine learning algorithm which focuses on predicting autism by doing a classification on the detailed explanation of the symptoms from which they were suffering. The goal of this study was to perform text classification using machine learning, since the nature of autism is primarily diagnosed through documented behavioral evaluations such as restricted and repetitive behaviors. The machine learning algorithm utilized in this study is Weka which is an open source software consisting of a combination of machine learning algorithms used for data mining. Using a logistic regression model labels and featured were used to predict ASD with an accuracy of 89%. Although this approach utilizes machine learning techniques to predict ASD, it doesn’t delve into the gene expressions and RNA-sequencing of the patients.

Lastly the research done by Shankar titled “Data Autism: An Early Detection Framework of Autism in Infants using Data Science” is a perfect example of data analytics and machine learning techniques applied in the health care field [7]. It primarily focuses on autism in infants characterized by their communications skills, agnostic behaviors and diminished social interaction. This study utilizes the use of a support vector machine classification by developing a framework for early detection of autism in infants which delivers an accuracy of 89%. Most of these approaches have been focused towards analyzing the behavioral changes which can be deduced by interacting with the patient but fail to predict the neurological changes that undergo in the brains of the patient such that they can be identified and corrected via in vitro treatment editing.

All in all, the algorithms used in the previous approaches use familial history or deep learning techniques to identify trends in behaviors and restricted interaction etc. Despite these studies having been done with numerous underlying mechanisms leading to ASD, none of them explicitly focuses on the genome sequencing as a whole.

On the flip side, this small preliminary study focuses on tracking the transcriptional changes along the patients having ASD and the ones who don’t, leading to highlighting the changes that goes along the way of the early stages of human development. For this study a relatively clinically homogeneous patient population with proven idiopathic forms of ASD were studied, patients with known generic disease and recognizable syndrome were distinguished from the patients which do not have any signs of ASD. Moreover, all the test subjects involved in this study were associated with the race of non-Hispanic white males.

The patients were examined over a 135-day course of neuronal differentiation via the Induced pluripotent stem cells (iPSCs) derived from the cortical neurons of the test subjects, simply put the patients were treated with neuronal induction to gain insight into pathways disrupted by ASD during cortical development, and to better understand how these disturbances to cellular function progress as neurons mature.

RNA-Seq was performed on ASD and control neurons at two time points after neural induction (35 and 135 days), then the data was analyzed using multiple approaches. The reason these two time points were chosen was to capture differences in the transcriptome at early stages of development (late neuronal progenitor/early neuron stage (Day 35)) and later stage where there are more mature neurons (Day 135) which lead onto revealing important and functionally validated insights into common processes altered in early neuronal development contribute to ASD pathogenesis.

This study also focuses on the genome expression for the immune system, more towards understanding the function and diversity of the immune system using genome expression and prescribing regulations to the immune response using gene engineering approaches. The data mainly revolves around the RNA sequence obtained from the respective patients, which basically serves as a message-carrier to carry instructions from the DNA to control the process of synthesizing proteins in other words it serves the role of coding, decoding and regulating the expression of genes. The reason why this is so important for this analysis is because naturally RNA molecules are edited much less in the brains of autistic people than people with no signs of ASD. This study focuses on analyzing the RNA sequence found from the patients and identifies the RNA molecules that needed more modification since more RNA editing is good, consider it as a way to make up for the lacking of RNA editing that the brain of an average individual having ASD goes through in the early development stage.

# METHODOLGY AND MODEL DESIGN

The datasets involved in this study are obtained from the website of national center for biotechnology information which provide extensive data for a study done on a total of eleven patients from which five of the patients had no signs of autism whereas the other six counterparts experienced autism spectrum disorder.

As discussed in the background that this analysis builds on top of the analysis done by Derek [1] to track the transcriptional changes on patient derived neurons, gene expressions are stored in two datasets with samples derived over a span of 135 days to track the changes. The day 35 represents the early stages of the neuronal induction whereas the day 135 reflects the same genes with altered gene expressions on a mature stage after performing the RNA sequencing.

Six different models are utilized in this study which are aimed towards predicting the probability of a patient having autism or not. The models used are: Rpart, Support Vector Machine, Random Forest, Logistic Regression, Deep Learning and Naïve Bayes.

The original dataset stored on the NCBI website [2] exists in a GNU zip file format as a table separated value (tsv) file. The dataset contains 57,820 features each expressing a different gene expression pattern for each of the eleven patients under study. Initially the data lists the genes in rows whereas the patients were placed as column names. Upon importing the data in R, the order of the data was flipped by doing a transpose to bring all the features (genes) as the column names and vice versa. Furthermore, an additional column was added which identifies a patient having autism with 1 as the value and the patient with no signs of ASD (control neurons) as 0. The reason this column was added so a correlation can be established between all the 57,820 different independent variables against the manually created column which serves as the dependent variable to see which gene contributes towards the weighted gene co-expression analysis (WGCNA). Table 1 displays the top 5 rows of the raw dataset in which “ASD” refers to the patient having autism whereas “WASD” refers to patient without autism.

Table 1: Raw Dataset

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Symbol** | **ASD 1** | **ASD 2** | **ASD 3** | **ASD 4** | **ASD 5** | **ASD 6** | **WASD 1** | **WASD 2** | **WASD 3** | **WASD 4** | **WASD 5** |
| ENSG00000000003.10 | 2023 | 2687 | 6580 | 5558 | 1450 | 3236 | 12473 | 7163 | 3214 | 1980 | 2604 |
| ENSG00000000005.5 | 8 | 12 | 6 | 139 | 100 | 3 | 31 | 12 | 46 | 2 | 10 |
| ENSG00000000419.8 | 816 | 1660 | 2002 | 2189 | 932 | 1073 | 1371 | 928 | 854 | 640 | 1126 |
| ENSG00000000457.9 | 372 | 486 | 1205 | 1015 | 420 | 577 | 1022 | 528 | 704 | 776 | 650 |
| ENSG00000000460.12 | 301 | 487 | 646 | 709 | 273 | 471 | 456 | 298 | 567 | 461 | 316 |
| ENSG00000000938.8 | 2 | 0 | 4 | 14 | 3 | 0 | 0 | 5 | 9 | 2 | 0 |

The dependent variable is identified as “Autism”, whereas the independent variables are gene expressions for “day 35” and “day 135”.

By using a scrapper, the name, description, synonyms and location were linked with the raw dataset such that an identification of genes can be obtained as can be seen from table 2. For the sake of simplicity, the table only displays ASD1 along with the scrapped data and symbol.

Table 2: Scrapped data merged to raw dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Symbol** | **ASD1** | **Name** | **Description** | **Synonyms** | **Location** |
| ENSG00000152583.8 | 26236 | SPARCL1 | SPARC-like 1 (hevin) [Source:HGNC Symbol;Acc: | MAST9, PIG33, SC1 | Chromosome 4: 88,394,487-88,452,213  reverse strand. |
| ENSG00000148053.11 | -257 | NTRK2 | neurotrophic tyrosine kinase, receptor, type 2 [Source:HGNC Symbol;Acc: | GP145-TrkB, TRKB, trk-B | Chromosome 9: 87,283,466-87,638,505 forward strand. |
| ENSG00000120885.15 | 30913 | CLU | clusterin [Source:HGNC Symbol;Acc: | APO-J, APOJ, CLI, CLU1, CLU2, KUB1, NA1/NA2, SGP-2, SGP2, SP-40, TRPM-2, TRPM2 | Chromosome 8: 27,454,434-27,472,548  reverse strand. |
| ENSG00000202198.1 | 673 | RN7SK | RNA, 7SK small nuclear [Source:HGNC Symbol;Acc: | 7SK | Chromosome 6: 52,860,418-52,860,748 forward strand. |
| ENSG00000131095.7 | 30608 | GFAP | glial fibrillary acidic protein [Source:HGNC Symbol;Acc: | FLJ45472 | Chromosome 17: 42,982,376-42,994,305  reverse strand |

Before performing the exploratory data analysis (EDA) a challenge faced was that there was not enough computational power to handle all 57,820 unique gene levels. So, a method of calculating the average gene expression was used to identify the top 20 genes that went through the most change from day 35 to day 135 by subtracting the averages of day 35 from day 135. Figure 1 and 2 display a tree map to display hierarchical data using nested rectangles focusing on the names and the synonyms for the respective genes.

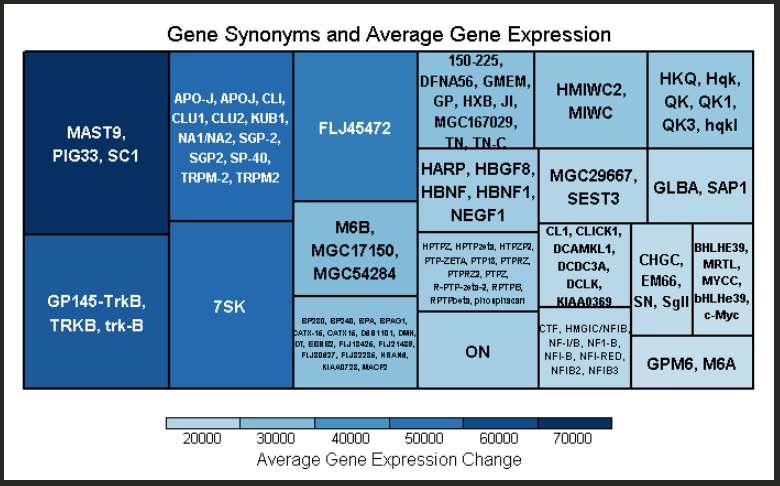


Figure 1: Gene synonyms and average gene expression

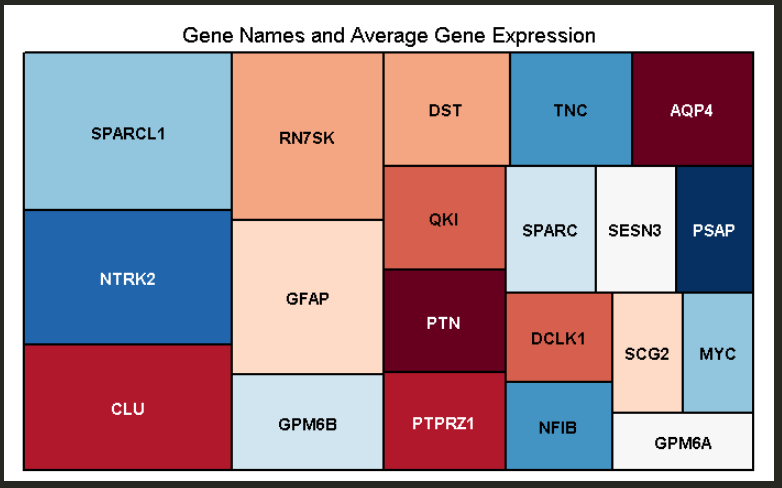


Figure 2: Gene names and average gene expression

To look into the details another approach was utilized which involved creating subsets of the original dataset with one having patients with ASD and patients not having ASD. Figure 3 shows the top 20 genes that went through a change but nonetheless doesn’t distinguish among patients with autism. Therefore, figure 4 and 5 were generated using the subsets created of the raw dataset to identify the most impacted genes after the neuronal induction for patients with and without ASD.

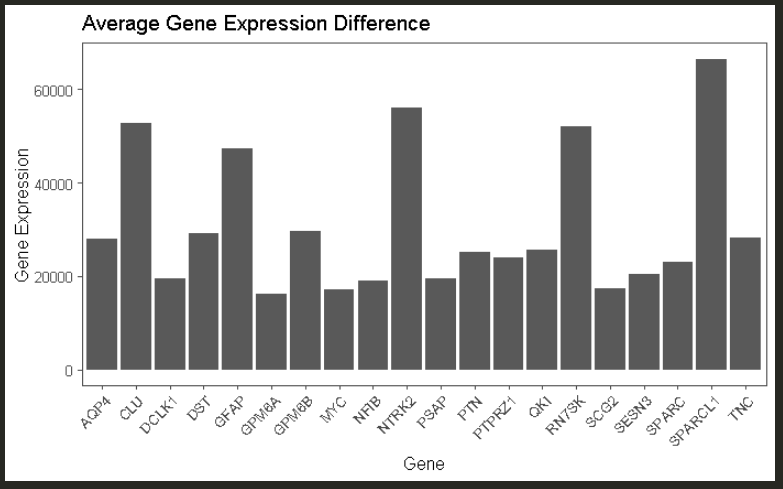


Figure 3: Average gene expression difference

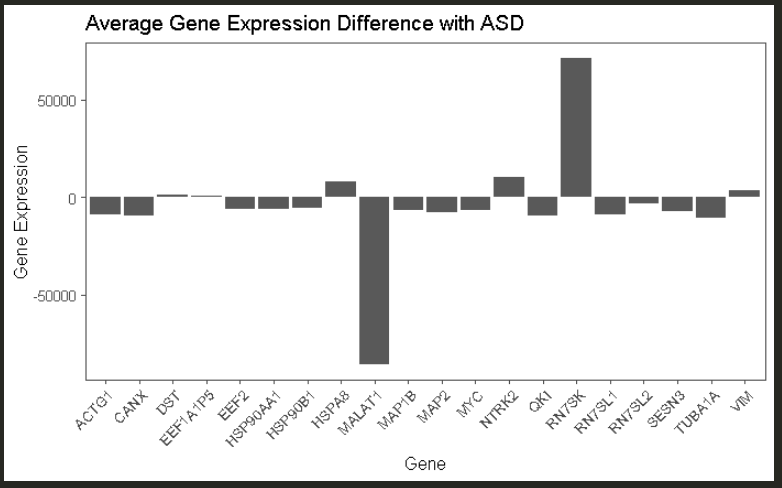


Figure 4: Average gene expression difference with ASD

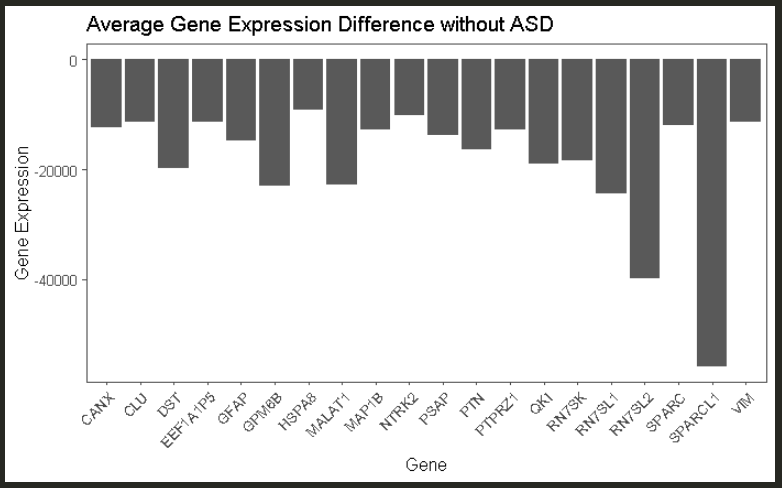


Figure 5: Average gene expression without ASD

Figure 6 looks to visualize the most impacted gene from the tree map, “SPARCL1” which is a protein coding gene. The 0 and 1 refers to patient having no signs of ASD and having ASD respectively.

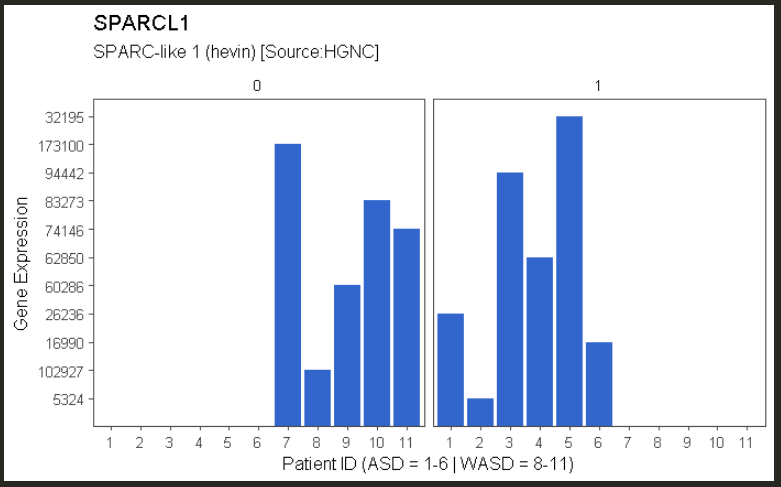


Figure 6: The most impacted gene obtained from treemap

The purpose of figure 7 is to gain an initial overview of the data via SPLOM (scatter plot of matrices) for top 10 genes.

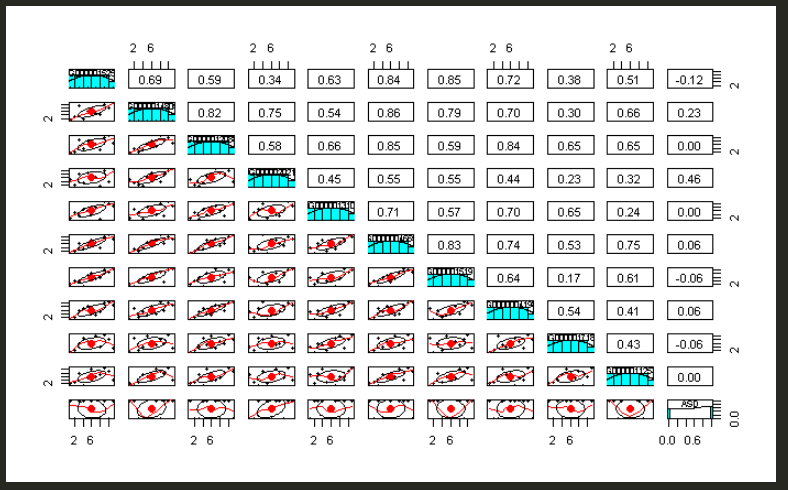


Figure 7: Scatter plot of matrices for top 10 genes

Coming over to the main models of the analysis, the models were built on a data which is split in the ratio of 0.75. The set.seed utilized was 123 which is a way of randomly splitting the data such that the same results can be replicated using the same set.seed but, in an effort to better the results each of the models were ran 25 times and the accuracies which are reported are a mean of the accuracies for each model.

The analysis adds on a tukey and anova test performed on the identified genes which is a multiple comparison procedure to find the interaction among the independent variable is mutually statistically significant. The overlap in the both datasets can be visualized using powerbi to give a better visual representation of the findings. The final dataset will be pushed to mysql hosted on the local machine using rsqls from which it can be fed to powerbi.

Lastly, as mentioned above that the raw data is in tsv format which basically is a form of structured data and covers the 5Vs of Big data. One sample of blood culture was able to uncover 57,820 features of underlying genes which covers the velocity and volume aspect of the 5Vs. The purpose of finding and promoting solutions means of better assisting with the diagnosis of idiopathic ASD answers the value aspect. As far as the veracity is concerned the analysis utilizes the data gathered by DeRosa [1] in a controlled environment by University of Miami under Millar school of medicine institutional review board approval protocol. The IRB protocols guarantee that all experiments were performed in compliance with the guidelines and regulations of the institutional biosafety committee at the University of Miami.

The libraries used from R include: devtools, RSQLS, tidyverse, h2o, kernlab, randomForest, caret, deeplearning, rpart, e1071, caTools, readxl, ggthemes.

# RESULTS

In order to perform the analysis a new dataset was created which takes in the top 20 symbols identified during the EDA and adds columns of corresponding gene expressions along with patient id and the column of autism containing values of 0 and 1. The head of the dataset can be seen in table 3 below.

Table 3: Dataset created for analysis

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Symbol** | **Name** | **X35** | **Patient** | **Autism** | **X135** |
|  |  |  |  |  |  |
| ENSG00000041982.10 | TNC | 249 | ASD1 | 1 | 20157 |
| ENSG00000046653.10 | GPM6B | 11567 | ASD1 | 1 | 16322 |
| ENSG00000105894.7 | PTN | 1662 | ASD1 | 1 | 13346 |
| ENSG00000106278.7 | PTPRZ1 | 21146 | ASD1 | 1 | 14232 |
| ENSG00000112531.12 | QKI | 44680 | ASD1 | 1 | 22681 |

The first model built can be visualized using figure 8 that reflects on the linear regression of day 35 and day 135 genes which can be used to predict autism.

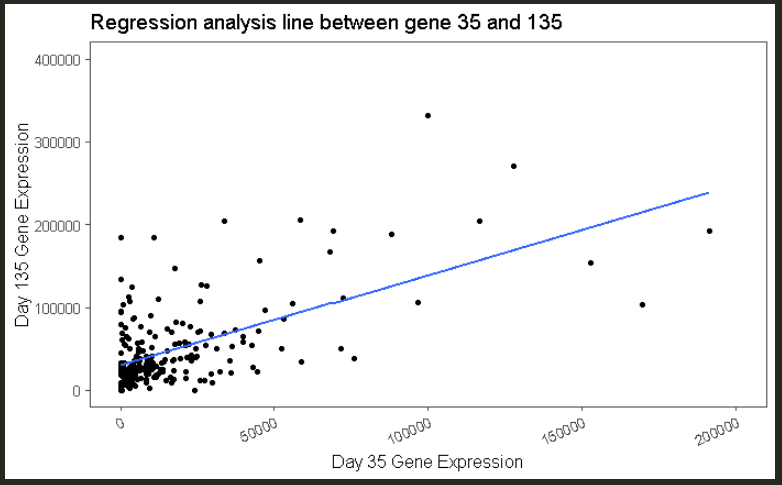


Figure 8: Regression analysis between day 35 and day 135

For the smallest concentration of gene expression, the variables impacting the model the most belong to the data pertaining to day 135 as can be seen from the figure 9. The final split contains 10 leaves which leaves us with the probability of 0 (no ASD) and 1 (ASD).

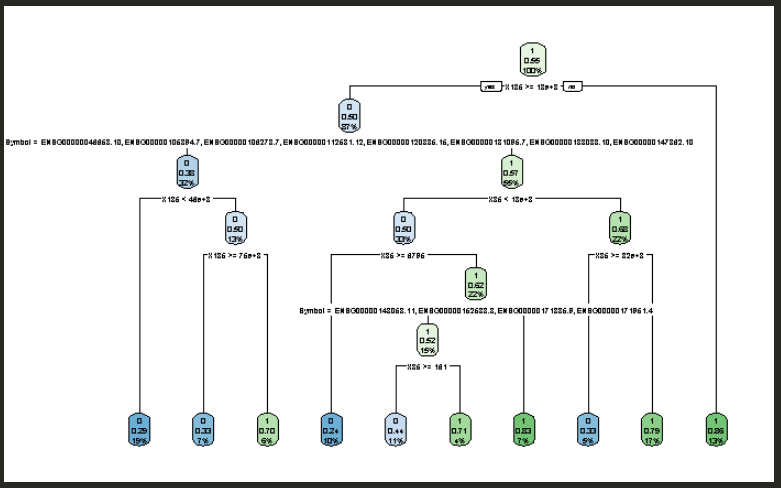


Figure 9: Rpart plot decision tree

The accuracy obtained using the six different models can be seen from figure 10. The highest value of accuracy was obtained using Support Vector Machine, Random Forest and Naïve Bayes respectively.

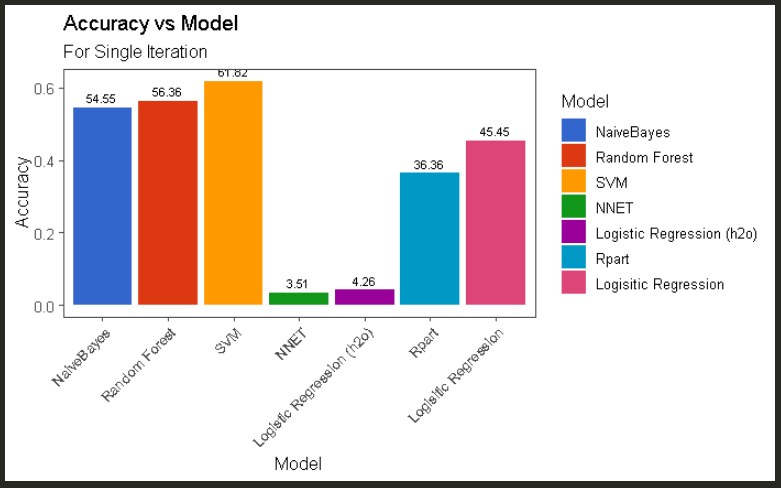


Figure 10: Accuracy of each model with single iteration

The accuracies obtained through each iteration is expressed in figure 11 which clearly shows how different iterations can result in fluctuating accuracies.

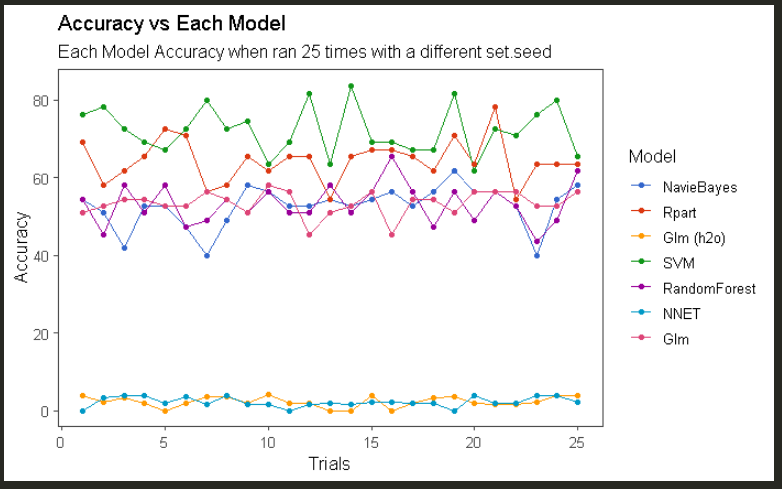


Figure 11: Accuracy of each model with 25 different set.seed

# DISCUSSION

The neural network and the logistic regression utilize h2o which is a scalable machine learning platform it can be used in R interface to crunch large datasets in less time. However, in this study h2o was used for two models neuralnet and glm, and for both models h2o delivered an accuracy of less than 5% which is the only reason an addition glm model was performed with the base R library to look at the difference. This is something that could be studied in detail for future reference as to why this anomaly ridden accuracy was generated using h2o.

So far, all the results that have been presented in this paper are visualization based which aren’t of much value if not backed statistically. So, in order to check for the significant difference between the models, anova was performed on the accuracies of each of the models obtained after changing the seed and getting different samples for training and testing. The p-value came out to be less than alpha which suggests that there is a significant difference in the accuracies of the different models.

Since they are significantly difference, now in order to find out the best model statistically speaking tukey test was performed and the results are displayed in table 4. According to tukey the best model to predict is SVM since it is significantly different from the rest of the model and the p value is less than alpha (0.05).

Table 4: Tukey test for the prediction

|  |  |
| --- | --- |
| **Combinations** | **p adj** |
| Rpart-NavieBayes | 0 |
| Glm (h2o)-NavieBayes | 0 |
| SVM-NavieBayes | 0 |
| RandomForest-NavieBayes | 0.992742 |
| NNET-NavieBayes | 0 |
| Glm-NavieBayes | 0.9956602 |
| Glm (h2o)-Rpart | 0.0000001 |
| SVM-Rpart | 0 |
| RandomForest-Rpart | 0 |
| NNET-Rpart | 0 |
| Glm-Rpart | 0 |
| SVM-Glm (h2o) | 0 |
| RandomForest-Glm (h2o) | 0 |
| NNET-Glm (h2o) | 1 |
| Glm-Glm (h2o) | 0 |
| RandomForest-SVM | 0 |
| NNET-SVM | 0 |
| Glm-SVM | 0 |
| NNET-RandomForest | 0 |
| Glm-RandomForest | 0.9999977 |
| Glm-NNET | 0 |

Visually speaking there is a clear change in figure 10 and figure 12, by running the model for 25 times and averaging the accuracies it was made sure that the data was randomized and that the model is not just giving results on one lucky randomization that gives high results. The final reported accuracy for the SVM increased by 10% after 25 iterations.

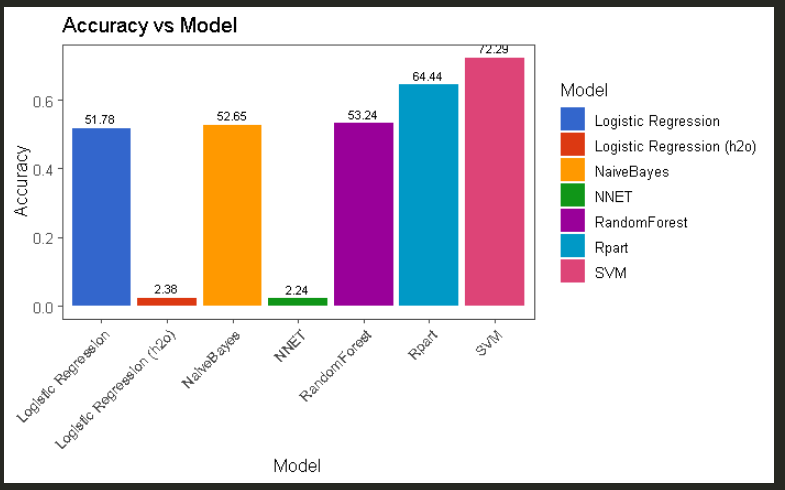


Figure 12: Average accuracies of each model with 25 iterations

# CONCLUSION

In today’s data-driven world, employers, lenders, marketers, educators, and many others are able to obtain a bounty off of information about individuals. These parties may then use data to identify those with future risks and prospects, and make adverse decisions concerning them. The same could be said for the medical field and the enormous potential that underlies in research for predicating diseases. Efforts should be taken by the governing body of medical association to focus on such disorders, work on them for the betterment of the future.

The most accurate model for prediction of ASD was obtained using SVM with an accuracy of 72% and it was backed by anova and tukey test to affirm its dominance.

To validate this preliminary study further, a substantial greater number of patient-specific iPSC-derived neuron must be generated in future investigation of larger cohorts.

##### References

1. Centers for Disease Control and Prevention. Morbidity and Mortality Weekly Report: Surveillance Summaries. Prevalence of Autism Spectrum Disorder Among Children Aged 8 Years - Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States 2010 Retrieved from: https://www.kaptest.com/shsat/what-is-the-shsat
2. Fifth Edition of the Diagnostic and Statistical Manual of Mental Disorders. American Psychiatric Association; 2013.
3. Kim YS, Fombonne E, Koh YJ, Kim SJ, Cheon KA, Leventhal BL. A comparison of DSM-IV pervasive developmental disorder and DSM-5 autism spectrum disorder prevalence in an epidemiologic sample. J Am Acad Child Adolesc Psychiatry. 2014;53(5):500–8.
4. Bosl, W.J., Tager-Flusberg, H. & Nelson, C.A. EEG Analytics for Early Detection of Autism Spectrum Disorder: A data-driven approach. *Sci Rep* **8,** 6828 (2018)
5. Levin, A.R., Varcin, K.J., O’Leary, H.M. *et al.* EEG power at 3 months in infants at high familial risk for autism. *J Neurodevelop Disord* **9,** 34 (2017).
6. Arya, Arpit, "Predicting Autism over Large-Scale Child Dataset" (2015). Master's Projects. 452 (2015).
7. Shankar V.G., Sisodia D.S., Chandrakar P. (2020) DataAutism: An Early Detection Framework of Autism in Infants using Data Science. In: Sharma N., Chakrabarti A., Balas V. (eds) Data Management, Analytics and Innovation. Advances in Intelligent Systems and Computing, vol 1016. Springer, Singapore.

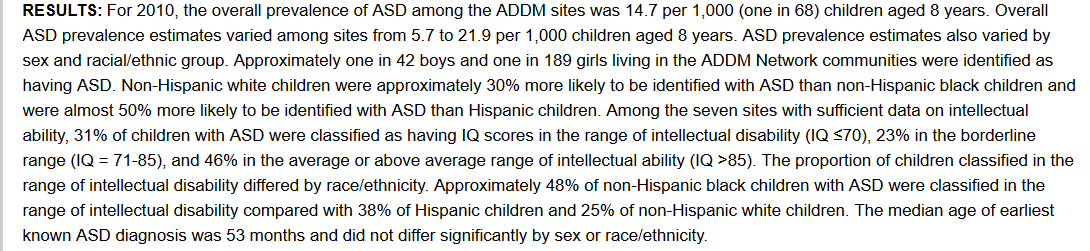


Figure 10: Reference 1

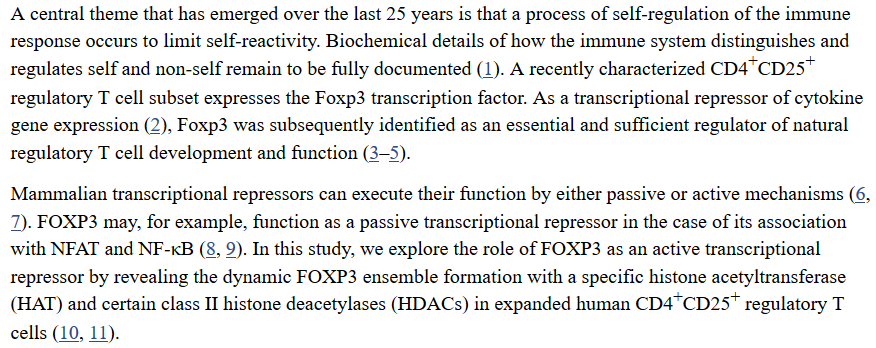


Figure 2 Reference 2

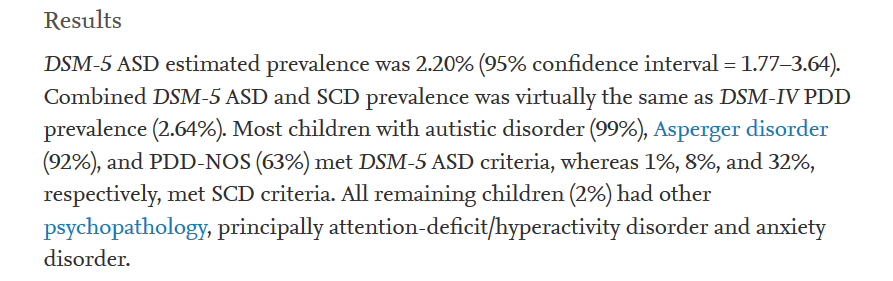


Figure 3 Reference 3

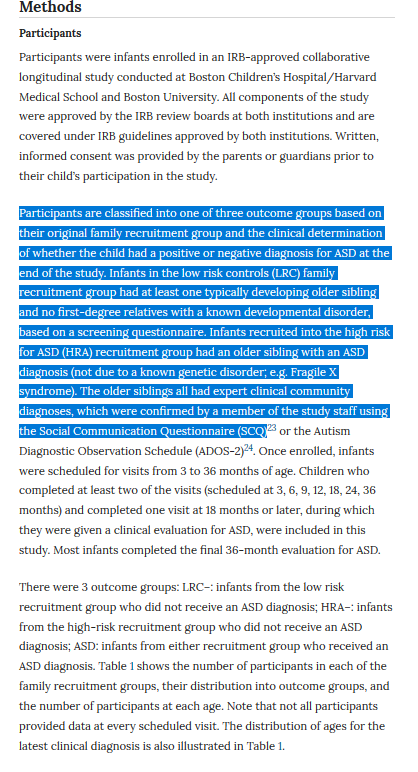


Figure 4 Reference 4

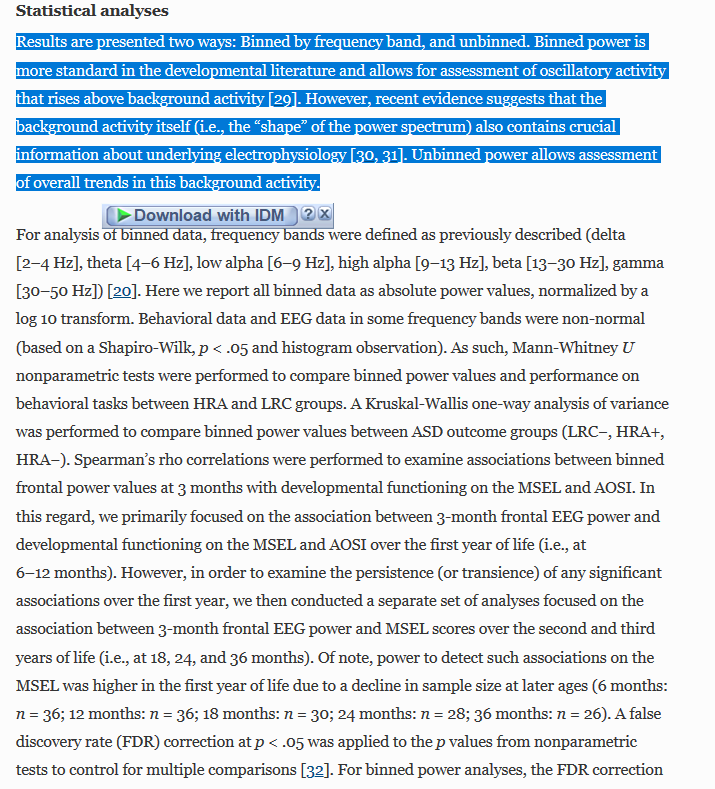


Figure 5 Reference 5

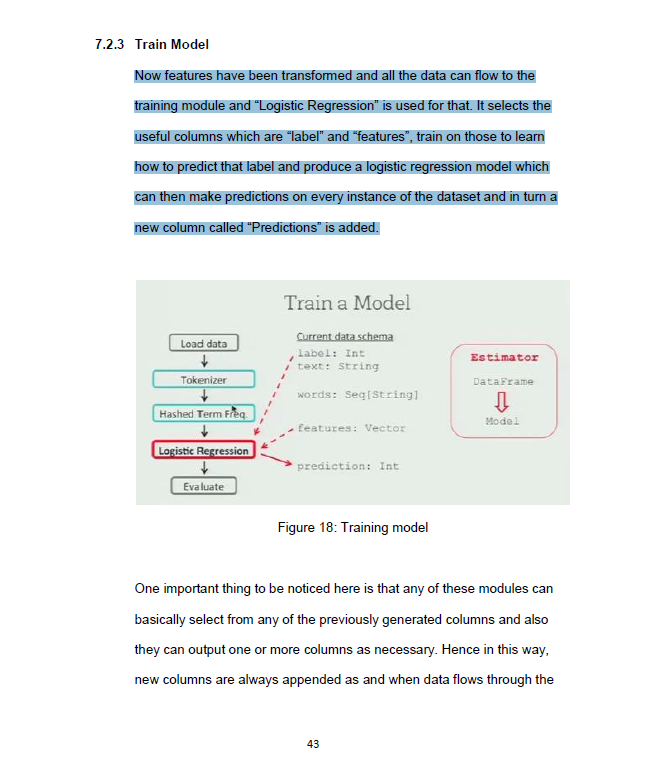


Figure 6 Reference 6

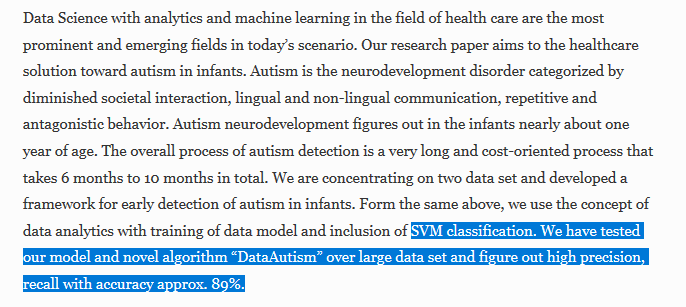


Figure 7 Reference 7