Prediction Of Gene Susceptibility To Autism Spectrum Disorder Using Deep Architectures

V. Pream Sudha, Dr. M.S.Vijaya

Abstract: A genetic predisposition or susceptibility to Autism Spectrum Disorder (ASD) is an increased likelihood of developing it based on the genetic makeup of a person. The multiple variants found in each gene have their own probability of associated risk and so the major problem lies in the systematic evaluation of their functional significance to ASD. Hence it is essential to develop methods for quantitative evaluation of ASD candidate genes with co-occurring mutations which will provide a clear understanding of their relevance to ASD. This research work deals with the development of a discriminative model for prioritization of candidate genes considering mutations in them and to classify them based on their predisposition to the disorder. The model for gene susceptibility prediction is built by integrating the combined potential of substantiation for each ASD linked gene and the related mutations. In this research work gene susceptibility prediction is modelled as a pattern classification problem and deep learning techniques are employed to build the models. The performance evaluation of these models proves that Long Short Term Memory (LSTM) based gene susceptibility prediction model has shown better performance.

Index Terms : Autism Spectrum Disorder, Gene susceptibility prediction, Long Short Term Memory, Pattern classification

1 INTRODUCTION

A genetic predisposition is caused due to specific genetic variations that are inherited from a parent and contribute to the development of the disorder. The exploration for genetic factors that are fundamental to ASD has enabled the identification of numerous genes with abundant of mutations. The methods for ranking genes depend on functional annotations [1-2], properties of network [3-5] and gene expression data [6 -7]. The databases of functional annotations and pathways that are mostly used are GO and KEGG. Databases of ASD causing genes have been formed [8-9] to summarize the list of candidate genes behind this disorder. In recent times steps have been taken to formulate methodologies for evaluation of ASD causing genes [10-16]. A joint score for each ASD linked gene is computed by merging the indicative scores generated and then used for prediction in this study.

2 METHODOLOGY

In our previous work, decision tree based model was developed to classify the ASD genes [17] and deep learning was employed to predict ASD using codon encoding [18]. It is important to recognize the triggering mutations and also to assess the probable risk of each genetic variant to discover the deleterious genes. Hence this work focuses on developing machine learning model for classifying genes based on their susceptibility to ASD. The problem uses simulated gene sequences to solve the multi-class classification problem using deep learning techniques such as BRNN, LSTM and GRU to build model that predicts gene susceptibility to ASD. The process is divided into four functional parts namely establishment of corpus, extraction of features, building the model and evaluation of model performance. The architecture of the gene susceptibility prediction model is shown in Fig.1. In the first phase reference cDNA sequences and mutational information are retrieved from OMIM and SFARI database respectively. Genetic mutations for ten ASD genes with four types of mutations are induced using R script and the simulated gene sequences are stored in a corpus. SFARI Gene database especially the Human Gene Module is investigated to spot a range of pathways and biological signatures. ASD genes considered for this work are examined in the Molecular Signatures Database (MSigDB). It consists of Gene Set Enrichment Analysis (GSEA) software with a compilation of annotated gene sets. The overlap between ASD gene set and other gene sets is found using compute Overlaps tool that employs the hypergeometric distribution. Twenty five features that provides the information regarding gene-mutation is mapped to every record included in the corpus. These include multiple attributes of gene, mutation, conserved protein domains, inheritance pattern, the type of variant, gene expression profiles and pathway interactions. The various descriptors considered for scoring the gene - mutation instance are described below.

Biomedical Literature

There are huge volumes of biomedical extracts provided by PubMed, providing a massive amount of knowledge that can be mined. In PubMed database, there is plenty of co-occurrence between gene and ASD terms in the scientific literature. The MeSH terms used to identify the evidences were Autism, Autism Spectrum Disorder, ASD, Intellectual Disability, developmental delay, mental disorders and neuro-developmental disorders. A candidate ASD gene is assigned a score which is the sum of the cumulative evidences generated from the biomedical literature.

Intrinsic Gene / Protein properties

Gene characteristics like count of exons, protein length, conserved domains vary between genes causing ASD and those which are not implicated in the disease and hence provide a hint about the likely significance for hereditary

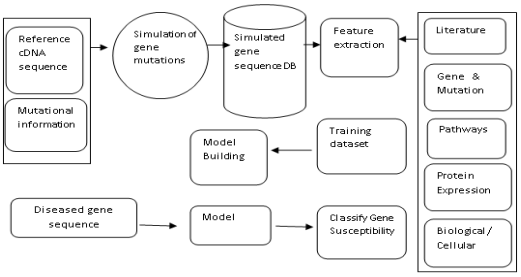

Fig. 1 *Architecture of ASD Gene Susceptibility Prediction Model*

disorders. During evolution, the functional properties of a specific region of the protein is preserved though changes have occurred at specific positions of an amino acid sequence in the protein. If conserved domains are less in a gene it indicates that it is more vulnerable to diseases. The gene SHANK3 has 10 conserved domains.

Mutation Properties

Structural variants which differ in from patient to patient also play a role in determining the vulnerability of the gene to the disease. The features considered here are mutation length, type and inheritance pattern. A score of 3 or 2 or 1 is assigned based on the type of mutation. A score of 3 is given to genes that are affected by nonsense mutations which are high in risk as untimely dismissal of the protein is frequently related with disease. Genes affected by the missense and frameshift mutations are assigned a score of 2 as the disease association is uncertain. The protein sequence is not altered by silent mutations and so the respective genes receive a score of 1. Variants found with a population frequency less than one percent are rare and those found in the population at an occurrence greater than one percent are common. Rare variants get a score of 2 whereas common variants receive 1. The inheritance pattern of mutation whether familial or de novo or unknown is explored and a score of 3 or 2 or 1 is assigned respectively.

Protein - Protein Interaction Pathways

The participation of candidate genes in known disease-associated pathways is investigated. The protein–protein interactions analyses of genes was done to study the enrichment of proteins involved in KEGG pathways like axon guidance, neuronal system, interaction of neurexin and neuroligin at synapses, DNA methylation, chromatin remodeling factors and synaptic transmission as shown in Table 1. A gene is assigned a score of 1 for each interaction and 0 otherwise.

Table 1

Protein - Protein Interaction Pathways Associated with Genes and Their P-Values

Gene	*Axon guidance*	*Neuronal system*	*Interaction of neurexin and neuroligin*	*DNA methylation*	*Chromatin remodeling factors*	*Synaptic transmission*
FMR1		√	√	√	√	
MECP2		√	√	√	√	
TSC1				√		√
CACNA1C	√	√			√	
SHANK3	√	√	√	√		√
CHD8	√	√	√		√	
FOXP2		√				√
CNTNAP2		√	√		√	√
GABRB3		√	√	√		
HOXA1	√	√	√			√

Protein Differential Expression

The incidence or nonexistence of a tissue in candidate genes may be an important factor for their assessment, if genes identified to be implicated in a disease are significantly expressed in that particular tissue. The gene expression in tissues at brain, cortex, cerebellum, nervous system associated with ASD is considered. If a gene is exceedingly expressed in these tissues it receives a score of 1, and the cumulative score of the gene is computed. SHANK3 gene is expressed in all these tissues and hence its score is 4.

Biological Processes

ASD susceptible genes affect the biological function of brain part responsible for social cognition and behavioral flexibility. The genes are investigated to compute overlaps in the biological processes and their p values are analysed. The calculation of overlaps with KEGG gene sets and the top 10 enriched pathways is shown in Table 2. The p-values of these pathways were sorted to obtain significance of overlaps. The pathways like single organism behavior with p-value 6.68E-11 and behavior with p-value 3.93E-10 have high statistical significance. Cognition, learning, response to auditory stimulus, neurological system processes are some of the biological processes with p -values that are considered for this work and the overlap matrix is shown in Fig.2. Each gene which shows an overlap in anyone of these processes receives its associated p-value and the cumulative score is calculated. SHANK3, CNTNAP2 and FOXP2 receive a score of 5.66 E-08 whereas GABRB3, CNTNAP2 and FMR1 receive a score of 0.

Cellular Component

GO Cellular components like Synapse, transporter complex, dendrite, neuron spine, cation channel complex are considered for this work as they show a significant p-value. The top 10 enriched cellular components are shown in Table 3. The cellular components related to neurons are enriched for the genes and the top two components are postsynapse pathway with p-value 8.93E-09 and synapse part with p-value 9.67E-09. Neuron projection, transporter complex, dendrite, perikaryon are the other cellular components that are considered for this work and the overlap matrix is shown in Fig.3. A gene with an overlap with anyone of these processes gets a score of its respective p-value. The score for FMR1 gene is 5.71E-04 whereas for FOXP2 it is 0. For each gene sequence in the corpus and for every single variant of the corresponding gene, a consolidated score is determined by adding the feature values described above. This gives a lucid insight about their relevance to the disorder. This score determines the gene susceptibility to the disorder. The various above attributes and their respective feature count are summarized below.

Publications	1	Biological Processes	1
Exon count	1	Cellular Components	1
Protein altered	1	Substitution scores	5
Mutation length	1	Amino acid observed	1
Protein differential expression	1	Pathways	6
Amino acid			
expected	1	Conserved domains	1
Common / Rare Variant	1	Alteration type	1
Inheritance pattern	1	Score	1

Total number of features extracted for a sequence – 25 A

sample record is presented below.

122	2	1731	1	8	20
1	2	3	-1	-6	-0.4
-3	-4 -2	.2	5.66E-08	2.94E-04	
4	1	1	1	1	0
1	1833.4				

The instances are assigned class label low when the score is less than 0.5 indicating that the gene is less vulnerable to the disorder. Similarly the instances are assigned class label medium when the score lies between 0.5 and 0.8. A gene with a score greater than 0.8 has elevated susceptibility to ASD and the class label high is assigned. Min - max normalization is done to all the feature values. Thus the Gene Susceptibility dataset (GSDS) with 1000 feature vectors of dimension 25 is developed for this work.

In the second phase, three independent gene susceptibility prediction models are built using deep learning techniques such as DNN, BRNN and LSTM. These three models are constructed with an input layer, two hidden layers with eight memory units and an output layer. A DNN receives inputs and generates a depiction of the experimental patterns existing in the data in every layer. The BRNN integrates information from the backward and forward pass. The recurring connections from the previous time steps distinguish the two hidden layers and in the second layer the recurrent connections are flipped, where activation is sent backwards along the sequence. An LSTM network consists of memory cell which is smarter than a classic neuron and consists of input, forget and output gate. Removal of information that are not required by LSTM is done by the forget gate. The inputs which are very long in the past do not have much influence on the network. The output gate modulates whether the output of the linear unit is to be broadcasted to the network. Two such LSTM units are stacked and the output of one unit is presented as input to the other. The fully connected dense output layer has three neurons for predicting gene susceptibility to ASD. The activation function used in the output layer is softmax that permits the network to gain knowledge and output the probable class values.

Table 2

Biological Components Associated with Genes and Their P-Values

Gene Set Name	*No. of Genes in Gene Set (K)*	*Description*	*No. of Genes in Overlap (k)*	*k/K*	*p-value*
Single organism behavior	*384*	*An organism's reaction to exterior or internal stimuli*	*6*	*0.0156*	*6.68E-11*
Behavior	*516*	*The synchronized reply of organisms to internal or external stimuli*	*6*	*0.0116*	*3.93E-10*
Cognition	*255*	*The intellectual activities connected with thinking and learning*	*5*	*0.0199*	*1.15E-09*
Mechanosensory behavior	*12*	*Reaction to a mechanical stimulus.*	*3*	*0.25*	*1.63E-09*
Head development	*709*	*The process involving the development of a head from a preliminary condition to its fully grown state.*	*6*	*0.0085*	*2.63E-09*
Striatum development	*16*	*The development of striatum in the forebrain from initial formation to mature state.*	*3*	*0.1875*	*4.15E-09*
Central nervous system development	*872*	*The process that manages the growth of the central nervous system from its creation to the grown up structure.*	*6*	*0.0069*	*9.03E-09*
Subpallium development	*22*	*This process controls the development of the subpallium from its early state to fully grown structure.*	*3*	*0.1364*	*1.14E-08*
Learning	*131*	*The process whose outcome is an adaptive behavioral change that comes from experience.*	*4*	*0.0305*	*1.31E-08*
Response to auditory stimulus	*23*	*The process that results in change of movement, enzyme production due to auditory stimulus.*	*3*	*0.1304*	*1.31E-08*

TABLE 3

CELLULAR COMPONENTS ASSOCIATED WITH GENES AND THEIR P-VALUES

Gene Set Name	*No.of Genes in Gene Set (K)*	*Description*	*No. of Genes in Overlap (k)*	*k/K*	*p-value*
Postsynapse	*378*	*The part of a synapse that is part of the post-synaptic cell.*	*5*	*0.0132*	*8.93E-09*
Synapse part	*610*	*Part of a synapse, the intersection between the nerve strand of a neuron and another*	*5*	*0.0082*	*9.67E-08*
Synapse	*754*	*The site of contact for interneurons.*	*5*	*0.0066*	*2.76E-07*
Excitatory synapse	*197*	*The synapse that increases the chance of an action potential happening in the postsynaptic cell due to an act in the presynaptic cell*	*3*	*0.0152*	*9.11E-06*
Neuron projection	*942*	*A continuation that broadens from a nerve cell, e.g. an axon or dendrite.*	*4*	*0.0042*	*3.34E-05*
Cell projection part	*946*	*The essential part of a cell projection that extends from a cell e.g. a flagellum*	*4*	*0.0042*	*3.39E-05*
Transporter complex	*321*	*A protein complex facilitating transport of molecules between cells.*	*3*	*0.0093*	*3.91E-05*
Neuron part	*1265*	*Component of a neuron*	*4*	*0.0032*	*1.05E-04*
Dendrite	*451*	*A neuron projection that signals from other neurons and carries*	*3*	*0.0067*	*1.07E-04*

		a nerve impulse to the axon			
Perikaryon	*10 8*	*The part of the cell body that eliminates the nucleus.*	*2*	*0.01 85*	*2.43E-04*

Various hyperparameters such as batch size, epochs, dropout, learning rate and optimizer are considered and are fine tuned so that the accuracy and efficiency of the models are improved. Training and testing of the models are done using GSDS dataset. In the concluding phase, 10 - fold cross-validation technique is applied and the predictive performance of the three models is evaluated using various metrics.

3 EXPERIMENT AND RESULTS

In this experiment the training dataset GSDS which includes 1000 instances consisting of ten ASD gene types that involves 4 kinds of mutations has been used to build the classifiers. Deep learning techniques such as BRNN, LSTM and GRU are implemented to build the classifiers in Scikit learn. The predictive performance of classifiers is found using 10 fold cross-validation. The results are analysed through precision, recall, F- measure, accuracy, log loss and ROC area which is tabulated in Table 4 - Table 6.

Table 4

Epochwise Accuracy of Gene Susceptibility Classifiers

Epochs	*DNN*	*BRNN*	*LSTM*
50	*77.9%*	*73.9%*	*75.7%*
100	*76.5%*	*74.1%*	*78.1%*
150	*78.3%*	*76.4%*	*79.5%*
200	*80.3%*	*78.2%*	*80.1%*
250	*80.4%*	*80.9%*	*82.0%*

From the results in Table 4, it is seen that LSTM offers performance improvement for gene susceptibility classification when compared to the other tasks. The model achieves an accuracy of 82% for this task which is higher than the other two tasks. The experiments are conducted for epochs 50, 100, 150, 200, 250 and the accuracy of all three models is seen to increase steadily as epochs increase. The DNN gene prediction model attains 77.9% at 50 epochs and reaches 80.4% at 250 epochs. Initially at 50 epochs, the BRNN prediction model accomplishes 73.9% and later at 250 epochs arrives at 80.9%. The accuracy of LSTM gene susceptibility prediction model starts with 75.7% in the initial epoch reaches 80.1% in 200 epochs and finally attains 82% at 250 epochs. The epochwise log loss of these three models is shown in Table 5.

Table 5

Log loss of Gene Susceptibility Classifiers

Epochs	*DNN*	*BRNN*	*LSTM*
50	*1.9954*	*1.8950*	*1.7952*
100	*1.4477*	*1.7816*	*1.4614*
150	*0.9810*	*0.9830*	*1.2830*
200	*0.9438*	*0.9488*	*1.0489*
250	*0.8159*	*0.9459*	*0.7811*

From the results it is seen that the log loss associated with LSTM model for classifying the predisposition of genes to the disorder is 0.7811 which is comparatively less when compared to that of DNN and BRNN with 0.8159 and 0.9459 respectively. The DNN gene susceptibility prediction model has an initial log loss of 1.9954 whereas the BRNN and LSTM models have 1.8950 and 1.7952 loss at 50 epochs. As the epochs steadily increase, the log loss reduces and the models show better performance by reducing its misclassifications. A steady decrease in the log loss of gene susceptibility prediction model is noticed from 50 to 250 epochs. The performance comparison of these three models is shown in Table 6.

Table 6

Performance of Gene Susceptibility Classifiers

Classifier	*DNN*	*BRNN*	*LSTM*
Precision	*0.78*	*0.73*	*0.81*
Recall	*0.81*	*0.76*	*0.84*
F Measure	*0.79*	*0.75*	*0.82*
Accuracy	*80%*	*76%*	*82%*
Kappa statistic	*0.715*	*0.697*	*0.824*
Mean absolute error	*0.1576*	*0.0525*	*0.0233*
Correctly classified instances	*402*	*382*	*410*
Specificity	*0.88*	*0.85*	*0.92*
Mathew correlation coefficient	*0.81*	*0.79*	*0.87*
ROC Area	*0.74*	*0.71*	*0.77*

The comparative analysis shows that LSTM achieves high accuracy of 82% than DNN and BRNN with 76% and 80% accuracy respectively. LSTM shows promising results with respect to precision and recall values which are 0.81 and 0.84. The Mathew correlation coefficient value is also comparatively high for LSTM with a value of 0.87 whereas it is 0.81 for DNN. The F-measure value for LSTM is 0.82 which is higher than DNN with 0.79. The Kappa statistic value of DNN and BRNN are 0.715 and 0.697 respectively which are comparatively less than that of LSTM with 0.824. The performance analysis of gene susceptibility prediction model is depicted in Fig.2 to Fig.6. The comparative performance of the classifiers shows that LSTM outperforms the other models. The curve of precision and recall is elevated for LSTM whereas it is lesser steep for DNN and BRNN. The mean absolute error for LSTM is least whereas for DNN it remains high. The count of correctly classified instances for LSTM is comparatively high than the other two classifiers. ROC curve depicted in Fig.8 shows that class 2 has a high area under ROC curve of 0.87 whereas class 0 has it low with 0.79. The macro - average ROC curve area is 0.81 whereas the micro – average area is 0.55. The scoring process done for each gene sequence has categorized genes into three major classes. The genes SHANK3, MECP2, CHD8 and FMR1 had markedly elevated scores whereas GABRB3 and CNTNAP2 are in medium risk category. CACNA1C, HOXA1 and TSC1 are classified as less susceptible.

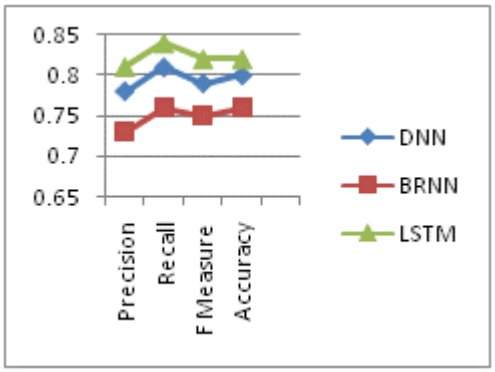

Fig.2 *Precision, Recall, F - Measure, Accuracy of Gene Susceptibility Classifiers*

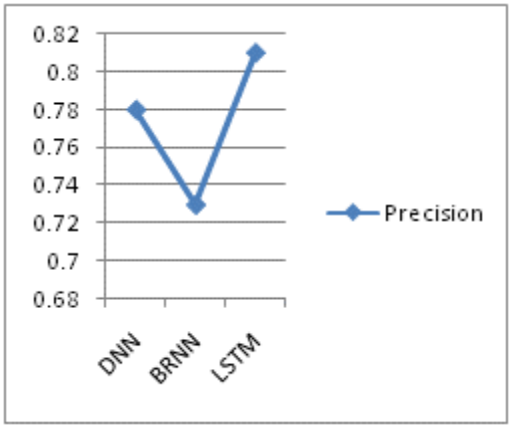

Fig. 3 *Precision of Gene Susceptibility Classifiers*

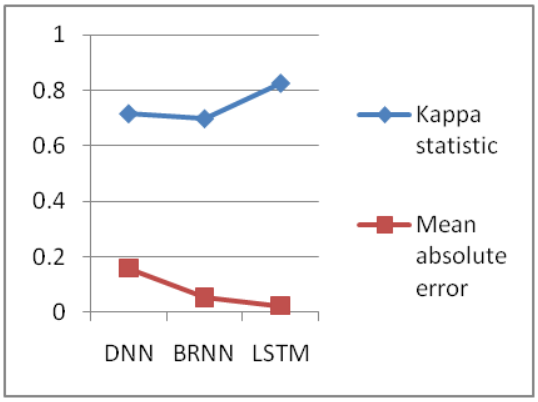

Fig. 4 *Performance Metrics of Gene Susceptibility Classifiers*

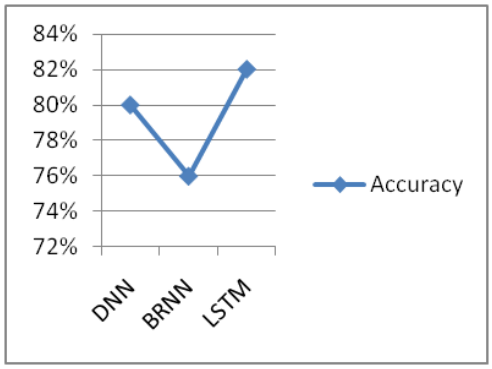

Fig.5 *Accuracy of Gene Susceptibility Classifiers*

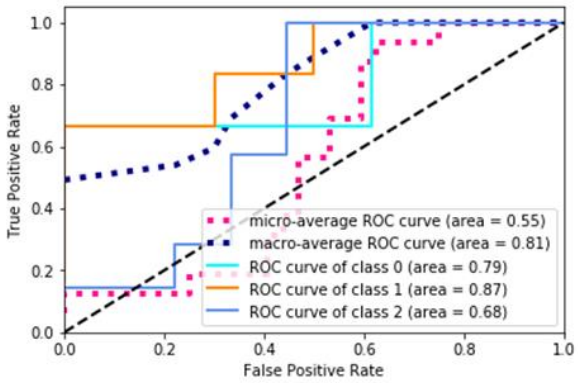

Fig. 6 *ROC Curve Of LSTM Classifier*

It is found that the experimental results correlate with the gene scores of AutDB which is a standard information portal for autism research and there are various factors that validate this fact. SHANK3 which is in the high risk category is the main controller of postsynaptic density as mutations in SHANK3 contribute to the disruption of cognitive development and communication. Mutations in MeCP2 gene that encode a protein acting as a general transcriptional receptor is the reason for Rett Syndrome. MeCP2 has a significant role in directing the neuronal activity and is classified as highly susceptible to ASD. MeCP2 mutations originate dendritic differentiation in cortex that may be a risk factor for autism. FMR1 mutations affect synaptic plasticity and development of synaptic connections between neural cells. CHD8 is a key player in autism and is involved in the beginning stage of brain stem development. The scoring process done in this work verifies the above facts and is in correlation with the same.

4 FINDINGS

The LSTM model developed is promising in discriminating the predisposition of the gene to ASD and is found to outperform the other two methods. The scheme of combining gene-mutation specific features along with evidences from biological processes and pathways for building the classifier is found to be helpful for the learning model to identify the gene susceptibility to ASD. The genes are assigned scores depending on their evidences for their predisposition to ASD. The results of the scoring process showed that four genes namely SHANK3, MECP2, FMR1 and CHD8 are highly susceptible to ASD. The nonsense,

missense and frameshift mutations in the genes which are rare and familial also contribute to the increase in the score. Similarly enrichment of genes in biological pathways, cellular components, pathway interactions is also a major factor for the elevated score. The model can be generalized to predict gene susceptibility to any genetic disorder provided the corresponding pathways and biological processes are considered.

5 CONCLUSION

The modeling of identifying gene susceptibility to ASD as a multi - class classification problem is demonstrated in this work. This work represents a significant step of progress from identifying ASD genes to finding the comparative strength of involvement with ASD for each gene taken for study employing data-driven methods. Deep learning techniques were implemented for classifying the genes based on a scoring process. The performance analysis of three independent models built with GSDS dataset shows that LSTM is a promising frontier for ASD gene susceptibility identification.

6 REFERENCES

[1] Krishnan, A., Zhang, R., Yao, V., Theesfeld, C.L., Wong, A.K., Tadych, A., Volfovsky, N., Packer, A., Lash, A. and Troyanskaya, O.G., 2016. Genome-wide prediction and functional characterization of the genetic basis of autism spectrum disorder. Nature neuroscience, 19(11), p.1454

[2] Ritchie GR, Dunham I, Zeggini E, Flicek P. Functional annotation of noncoding sequence variants. Nat Methods. 2014;11(3):294–6.

[3] Kou, Y., Betancur, C., Xu, H., Buxbaum, J.D. and Ma'Ayan, A., 2012, May. Network-and attribute-based classifiers can prioritize genes and pathways for autism spectrum disorders and intellectual disability. In American Journal of Medical Genetics Part C: Seminars in Medical Genetics (Vol. 160, No. 2, pp. 130-142). Hoboken: Wiley Subscription Services, Inc., A Wiley Company.

[4] Hormozdiari, F., Penn, O., Borenstein, E. and Eichler, E.E., 2015. The discovery of integrated gene networks for autism and related disorders. Genome research, 25(1), pp.142-154.

[5] Liu L, Lei J, Sanders SJ, Willsey AJ, Kou Y, Cicek AE, Klei L, Lu C, He X, Li M, et al. DAWN: a framework to identify autism genes and subnetworks using gene expression and genetics. Mol Autism. 2014;5(1):22.

[6] Latkowski, T. and Osowski, S., 2015. Computerized system for recognition of autism on the basis of gene expression microarray data. Computers in biology and medicine, 56, pp.82-88

[7] Liu, L., Lei, J., Sanders, S.J., Willsey, A.J., Kou, Y., Cicek, A.E., Klei, L., Lu, C., He, X., Li, M. and Muhle, R.A., 2014. DAWN: a framework to identify autism genes and subnetworks using gene expression and genetics. Molecular autism, 5(1), p.22.

[8] Banerjee-Basu, Sharmila, and Alan Packer. SFARI Gene: an evolving database for the autism research community. Disease Models & Mechanisms 3.3-4 (2010): 133-135. Web. https://gene.sfari.org/

[9] Liberzon A, Birger C, Thorvaldsdottir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst. 2015 Dec 23;1(6):417-425.

[10] Adie EA, Adams RR, Evans KL, Porteous DJ, Pickard BS. SUSPECTS: enabling fast and effective prioritization of positional candidates. Bioinformatics. 2006; 22(6):773–4. [PubMed: 16423925]

[11] Aerts S, Lambrechts D, Maity S, Van Loo P, Coessens B, De Smet F, Tranchevent LC, De Moor B, Marynen P, Hassan B. Gene prioritization through genomic data fusion. Nat Biotechnol. 2006; 24(5):537–44. others. [PubMed: 16680138]

[12] Wu C, Zhu J, Zhang X (2012) Integrating gene expression and protein-protein interaction network to prioritize cancer-associated genes. BMC Bioinformatics 13: 182. PMID: 22838965

[13] Hulsegge I, Woelders H, Smits M, Schokker D, Jiang L, Sorensen P (2013) Prioritization of candidate genes for cattle reproductive traits, based on protein-protein interactions, gene expression, and textmining. Physiol Genomics 45: 400–406.

[14] Zhang SW, Shao DD, Zhang SY, Wang YB (2014) Prioritization of candidate disease genes by enlarging the seed set and fusing information of the network topology and gene expression. Mol Biosyst 10:1400–1408. doi: 10.1039/c3mb70588a PMID: 24695957

[15] Jiang, Jing, Wan Li, Binhua Liang, Ruiqiang Xie, Binbin Chen, Hao Huang, Yiran Li, Yuehan He,Junjie Lv, Weiming He, and Lina Chen. A Novel Prioritization Method in Identifying Recurrent Venous Thromboembolism-Related Genes, PLoS ONE, 2016.

[16] Wen, Ya, Mohamad J. Alshikho, and Martha R. Herbert. Pathway Network Analyses for Autism Reveal Multisystem Involvement, Major Overlaps with Other Diseases and Convergence upon MAPK and Calcium Signaling, PLoS ONE, 2016.

[17] Pream Sudha V, Vijaya M.S, Decision Tree Based Model for the Classification of Pathogenic Gene Sequences Causing ASD, Smart Trends in Information Technology and Computer Communications, Springer Nature America, Inc, 2018

[18] Pream Sudha V, Vijaya M.S, Deep Learning Based Prediction of Autism Spectrum Disorder using Codon Encoding of Gene Sequences, International Journal of Engineering and Advanced Technology, 2019