5. SUPERVISED LEARNING MODEL TO PREDICT GENE SUSCEPTIBILITY TO ASD

A genetic predisposition or susceptibility to ASD is an increased likelihood of developing it based on the genetic makeup of a person. 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 led to the recognition of hundreds of genes containing thousands of variants. These 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 methodologies for quantitative assessment of ASD risk genes with co-occurring mutations which will provide a clear understanding of their relevance to ASD. This chapter 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.

Methodology

It is not only significant to identify the triggering mutations but also to evaluate the potential risk conferred by each individual genetic variant, given the faster pace of ASD candidate gene discovery. Hence this work focuses on developing machine learning model for classifying genes based on their susceptibility to ASD. The problem is devised as a multi-class classification problem and solved using supervised learning techniques. Gene susceptibility prediction model is built by integrating the cumulative strength of evidence for each ASD associated gene and its variant. The process is divided into three functional parts namely feature extraction and dataset creation, model building and performance evaluation. The architecture of the gene susceptibility prediction model is depicted in Fig.5.1.

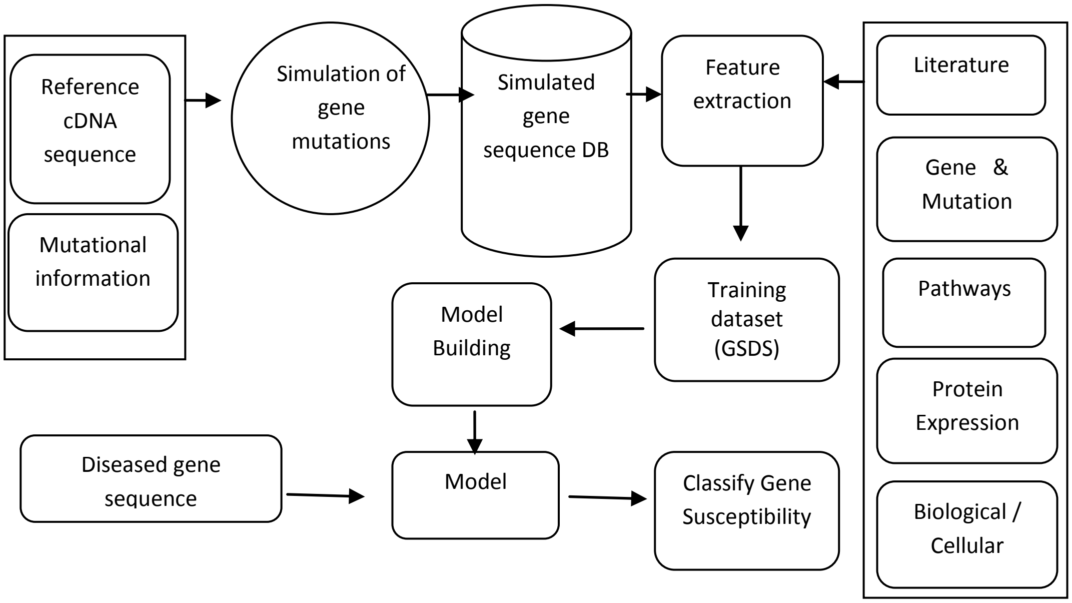


Fig. 5.1 Architecture of ASD Gene Susceptibility Prediction Model

The prioritization of candidate genes is based on the cumulative strength of evidence for each ASD-associated gene and its variant. The reference cDNA sequences and mutational information are retrieved from OMIM and SFARI database respectively. R coding is used to induce genetic mutations and the simulated gene sequences are stored in a corpus by considering ten ASD genes with four types of mutations as described in Chapter 3. Human Gene Module of SFARI Gene is probed to identify a variety of pathways and biological signatures. ASD genes considered for this work are examined in the freely available Molecular Signatures Database, MSigDB [96]. It is a collection of annotated gene sets for use with Gene Set Enrichment Analysis (GSEA) software. The Compute Overlaps tool in the investigate gene sets category uses the hypergeometric distribution to examine the overlap between ASD gene set and other gene sets. Gene symbol is used as gene identifier to import the gene set. Each individual record included in the corpus is manually annotated with 25 standardized descriptors providing a level of gene-mutation detail. 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: The millions of biomedical abstracts provided by PubMed, characterize an enormous amount of knowledge that can be mined. There is ample co-occurrence of gene and ASD terms in scientific texts available in PubMed database. The MeSH terms used to identify the evidences were Autism, Autism Spectrum Disorder, ASD, Intellectual Disability, developmental delay, mental disorders, neurodevelopmental disorders. To assess the strength of evidence linking a gene to ASD, a score is computed by summing the cumulative evidences generated from the biomedical literature.

Intrinsic Gene / Protein properties: Intrinsic gene properties like count of exons, protein length, conserved domains also provide a clue about a possible relevance for hereditary disorders because these properties differ statistically between disease genes and those which are not involved in the disease. During evolution, changes at specific positions of an amino acid sequence in the protein have occurred in a way that preserves the functional properties of that region of 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. Early termination of the protein is very often associated with disease so genes with nonsense mutants are automatically moved up in the list of possible suspects with a score of 3. Missense, frameshift mutations which alter the protein sequence without destroying it, may or may not be disease associated and so get a score of 2. Silent mutations which do not alter the protein sequence receive a score of 1. Variants with a population frequency <1 % are rare and those found in the general population at a frequency of ≥1 % 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 XIII. A gene is assigned a score of 1 for each interaction and 0 otherwise.

Table XIII Protein - Protein Interaction Pathways Associated with Genes and their P-Values

	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	✓	✓	✓			✓

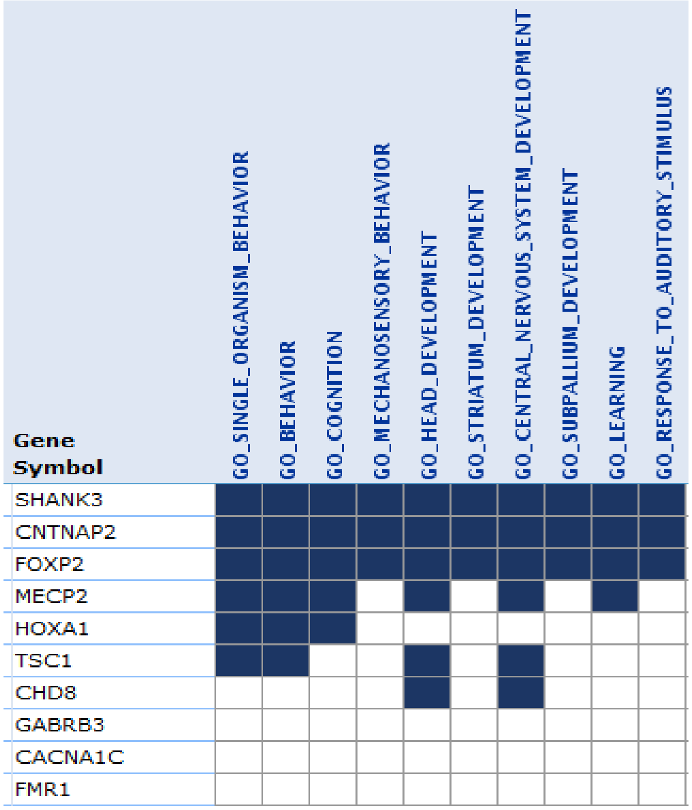
Protein Differential Expression: If genes known to be involved in a disease are significantly expressed in a particular tissue, then the presence or absence of that tissue in candidate genes may be a meaningful criterion for their evaluation. The gene expression in tissues at brain, cortex, cerebellum, nervous system associated with ASD is considered. Each gene receives a score of 1, if the gene is most highly expressed in these tissues 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: The biological function of ASD susceptible genes converge to disrupt neuronal function in brain regions that support language, social cognition and behavioral flexibility resulting in the phenotypes commonly associated with ASD. The genes are investigated to compute overlaps in the biological processes and their p values are analyzed. The top 10 enriched pathways are chosen from the computation of overlaps with KEGG gene sets and are shown in Table XIV. These pathways were sorted by p-values regarding significance of overlaps. The pathways like single organism behavior with p-value 6.68E-11 and behavior with p-value 3.93E-10 stand out with their markedly stronger 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.5.2. 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.

Table XIV 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	The specific behavior of a single organism in response to external or internal stimuli.	6	0.0156	6.68E-11
Behavior	516	The internally coordinated responses of whole living organisms to internal or external stimuli.	6	0.0116	3.93E-10
Cognition	251	It includes the mental activities associated with thinking, learning, and memory.	5	0.0199	1.15E-09
Mechanosensory behavior	12	Behavior that is dependent upon the sensation of a mechanical stimulus.	3	0.25	1.63E-09
Head development	709	The biological process whose specific outcome is the progression of a head from an initial condition to its mature state.	6	0.0085	2.63E-09
Striatum development	16	The development of striatum from initial formation to mature state. The striatum is a region of the forebrain	3	0.1875	4.15E-09

Central nervous system development	872	The process that controls the growth of the central nervous system from its formation to the mature 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 as the result of experience.	4	0.0305	1.31E-08
Response to auditory stimulus	23	The process that results in change of movement, enzyme production, gene expression, etc. as a result of an auditory stimulus.	3	0.1304	1.31E-08



Gene Symbol	GO_SINGLE_ORGANISM_BEHAVIOR	GO_BEHAVIOR	GO_COGNITION	GO_MECHANOSENSORY_BEHAVIOR	GO_HEAD_DEVELOPMENT	GO_STRIATUM_DEVELOPMENT	GO_CENTRAL_NERVOUS_SYSTEM_DEVELOPMENT	GO_SUBPALLIUM_DEVELOPMENT	GO_LEARNING	GO_RESPONSE_TO_AUDITORY_STIMULUS
SHANK3	■	■	■	■	■	■	■	■	■	■
CNTNAP2	■	■	■	■	■	■	■	■	■	■
FOXP2	■	■	■	■	■	■	■	■	■	■
MECP2	■	■	■		■		■		■	
HOXA1	■	■	■							
TSC1	■	■			■		■			
CHD8					■		■			
GABRB3										
CACNA1C										
FMR1										

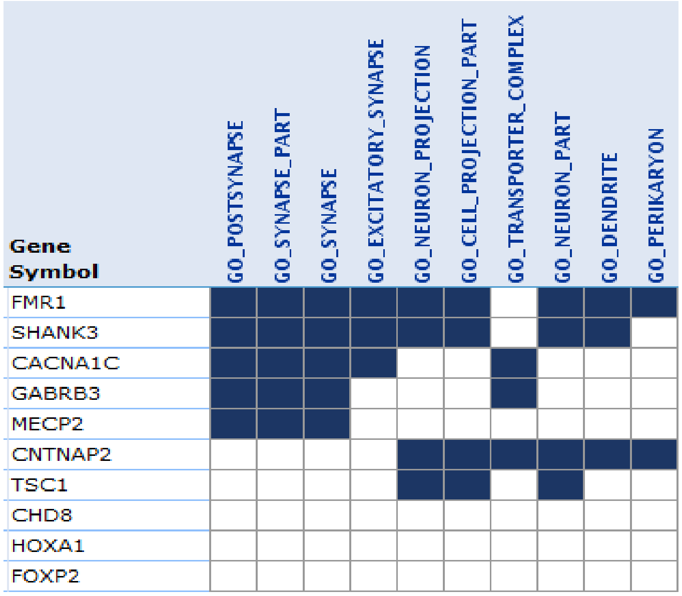
Fig. 5.2 Overlap Matrix of Genes and Biological Processes

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 chosen from the computation of overlaps with KEGG gene sets and are shown in Table XV. 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.5.3. 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.

Table XV 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	Any constituent part of a synapse, the junction between a nerve fiber of one neuron and another neuron or muscle fiber	5	0.0082	9.67E-08
Synapse	754	The junction between a nerve fiber of one neuron and another neuron; the site of interneuronal communication.	5	0.0066	2.76E-07
Excitatory synapse	197	This is the synapse in which an action potential in the presynaptic cell elevates the probability of an action potential happening in the postsynaptic cell.	3	0.0152	9.11E-06

Neuron projection	942	A prolongation or process extending 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	Any constituent part of a neuron, the basic cellular unit of nervous tissue	4	0.0032	1.05E-04
Dendrite	451	A neuron projection that signals from other neurons and conducts a nerve impulse towards the axon or the cell body.	3	0.0067	1.07E-04
Perikaryon	108	The part of the cell soma (cell body) that excludes the nucleus.	2	0.0185	2.43E-04



Gene Symbol	GO_POSTSYNAPSE	GO_SYNAPSE_PART	GO_SYNAPSE	GO_EXCITATORY_SYNAPSE	GO_NEURON_PROJECTION	GO_CELL_PROJECTION_PART	GO_TRANSPORTER_COMPLEX	GO_NEURON_PART	GO_DENDRITE	GO_PERIKARYON
FMR1	■	■	■	■	■	■		■	■	■
SHANK3	■	■	■	■	■	■		■	■	
CACNA1C	■	■	■	■			■			
GABRB3	■	■	■				■			
MECP2	■	■	■							
CNTNAP2					■	■	■	■	■	■
TSC1					■	■		■		
CHD8										
HOXA1										
FOXP2										

Fig. 5.3 Overlap Matrix of Gene and Cellular Components

For each gene sequence in the corpus, a consolidated score is calculated by summing the various features for each individual variant of an ASD-implicated gene leading to a clear understanding of 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.

Features	**Count**	**Features**	**Count**
Publications	1	Biological Processes	1
Exon count	1	Cellular Components	1
Protein altered	1	Substitution scores	5
Mutation length	1	Protein differential expression	1
Amino acid observed	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 for the gene sequence shown in Fig.3.4 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       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 have been built using supervised machine learning algorithms namely Decision trees, SVM and MLP. The models are trained and tested using GSDS dataset. Various metrics such as precision, recall, F-measure, accuracy, specificity and ROC area are used to evaluate the performances of the models.

In the concluding phase, 10 - fold cross-validation technique is applied and the predictive performance of the three models are evaluated using various metrics such as precision, recall, F - measure, accuracy, specificity and ROC area.

Experiment and Results

In this experiment the training dataset GSDS which includes 1000 instances of ten types of ASD genes involving four types of genetic mutations has been used to build the classifiers. Standard supervised classification algorithms namely decision tree induction, multilayer perceptron and SVM are implemented to build the classifiers in Scikit learn. Ten fold cross-validation technique is used to estimate their predictive performance. The results obtained from the classifiers are analysed through precision, recall, F- measure, accuracy, specificity and ROC area which is tabulated in Table XVI.

Table XVI Performance of Gene Susceptibility Classifiers

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

The comparative analysis shows that SVM achieves high accuracy of 82% than decision tree and MLP with 76% and 80% accuracy respectively. SVM 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 SVM with a value of 0.87 whereas it is 0.81 for MLP. The F-measure value for SVM is 0.82 which is higher than MLP with 0.79. The Kappa statistic value of MLP and decision tree are 0.715 and 0.697 respectively which are comparatively less than that of SVM with 0.824. The performance analysis of gene susceptibility prediction model is depicted in Fig.5.4 to Fig.5.9.

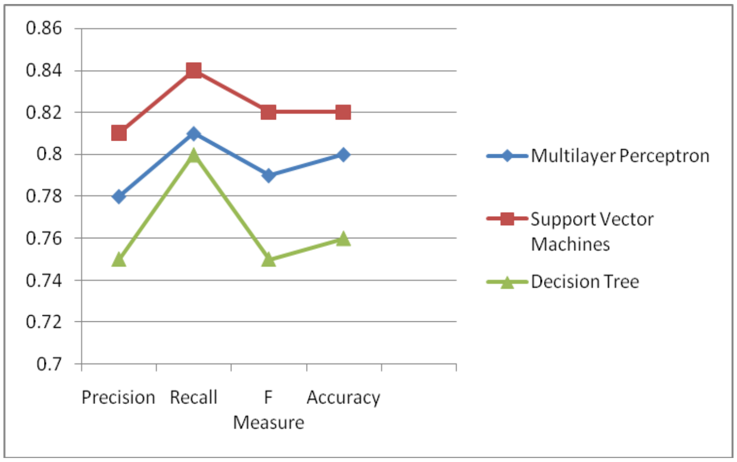


Fig. 5.4 Precision, Recall, F - Measure, Accuracy of Gene Susceptibility Classifiers

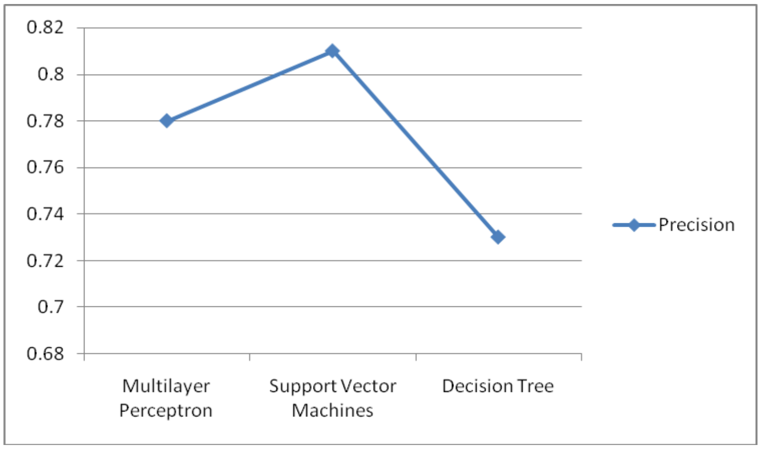


Fig. 5.5 Precision of Gene Susceptibility Classifiers

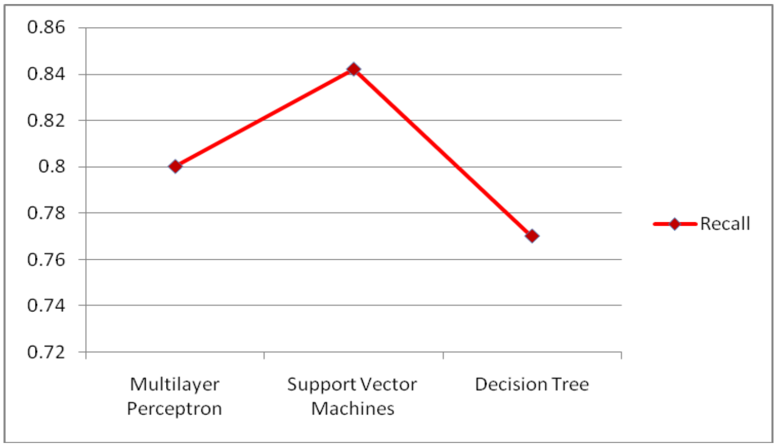


Fig. 5.6 Recall of Gene Susceptibility Classifiers

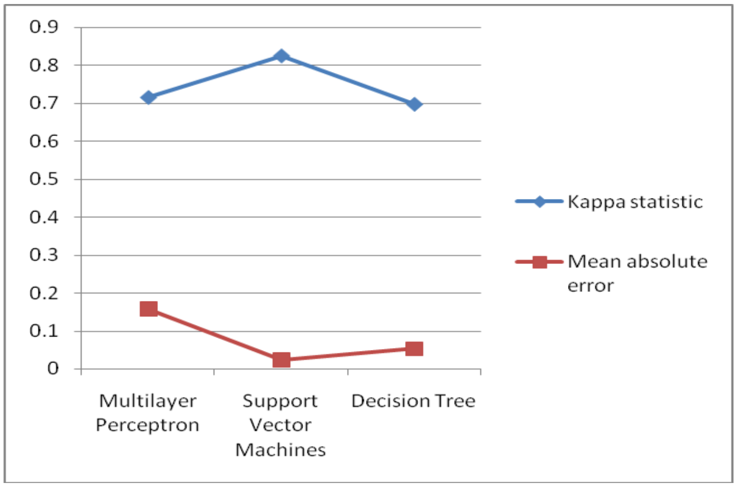


Fig. 5.7 Performance Metrics of Gene Susceptibility Classifiers

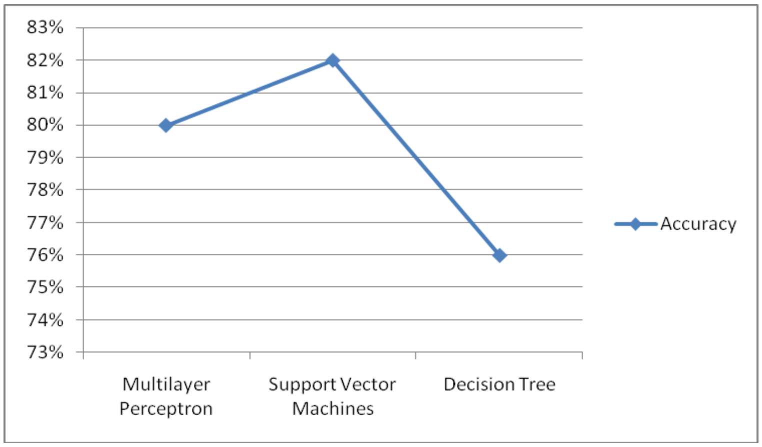


Fig. 5.8 Accuracy of Gene Susceptibility Classifiers

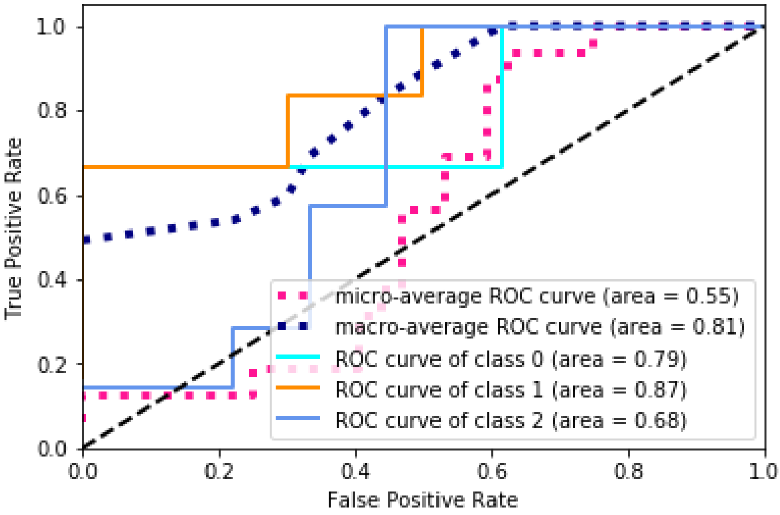


Fig. 5.9 ROC Curve of SVM Classifier

The comparative performance of the classifiers shows that SVM outperforms the other models. The curve of precision and recall is elevated for SVM whereas it is lesser steep for MLP and decision tree. The mean absolute error for SVM is least whereas for MLP it remains high. The count of correctly classified instances for SVM is comparatively high than the other two classifiers. ROC curve depicted in Fig.5.9 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 distinctly higher scores than all other genes whereas GABRB3 and CNTNAP2 are in medium risk category. CACNA1C, HOXA1 and TSC1 fall in the less susceptible class which is depicted in Fig.5.10 and Fig.5.11.

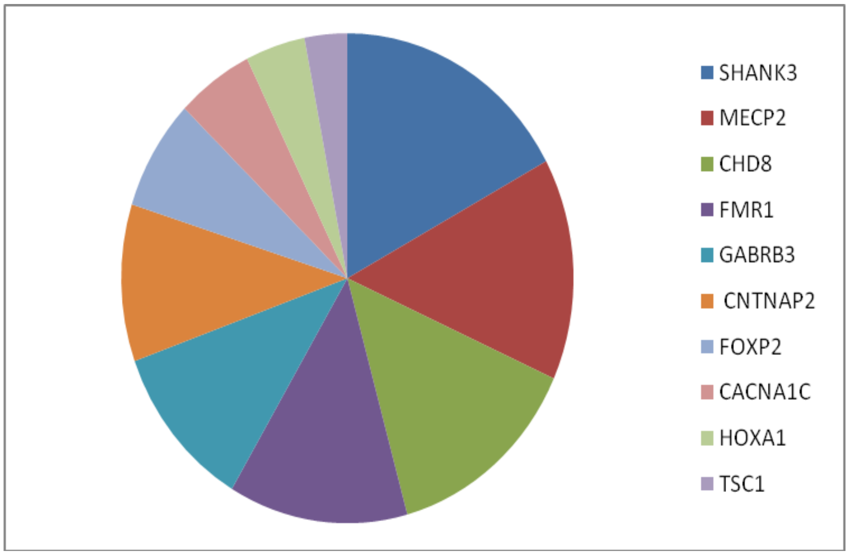


Fig. 5.10 Pie Chart Showing the Gene Susceptibility

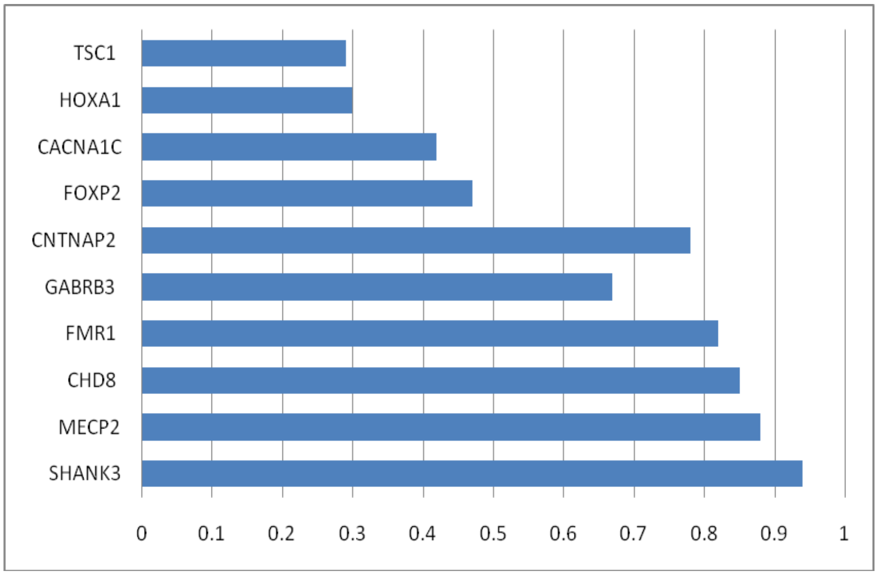


Fig. 5.11 Genes with Susceptibility Score

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 primary regulator of postsynaptic density due to its ability to form complexes with postsynaptic receptors, signal molecules, and cellular skeleton. As synaptic alterations are involved in etiology of the disorder, various mutations in SHANK3 contribute to the derangement in cognitive development and communication. Mutations in MeCP2 gene that encode a protein functioning as a general transcriptional receptor is accountable for Rett Syndrome. MeCP2 which is categorized as highly susceptible to ASD has an important role in regulation of neuronal activity. It has been suggested that MeCP2 mutations may be a risk factor for autism by causing dendritic differentiation in cortex. FMR1 associated with autism secondary to Fragile X syndrome affects synaptic plasticity and growth of synaptic connections between neural cells due to mutations. CHD8 has a major role in autism tendency and is associated with early phase of brain stem development in autism etiology. The scoring process done in this work verifies the above facts and is in correlation with the same.

Findings

The SVM 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.

SUMMARY

This chapter demonstrated the modeling of the gene susceptibility prediction problem. The implementation of supervised machine learning techniques for classifying the genes based on a scoring process was elaborated. The performance analysis of three independent models built with GSDS dataset was also illustrated in this chapter with tables and charts. The interpretations and the findings drawn from the experimental analysis were summarized. The contemporary deep learning approach for ASD gene identification, their susceptibility prediction and mutation recognition will be discussed in the forthcoming chapters.