3. PROBLEM MODELLING

The primary focus of this research is to propose an efficient solution for identifying the ASD causing genes, their susceptibility and mutations. The research problem of identifying ASD causing genes, their susceptibility and mutations is formulated as a multi- class classification problem and suitable solution is derived using traditional machine learning and contemporary deep learning approaches. This chapter describes the strategies applied for problem modeling in order to meet the objectives. The overall framework of the modeling process is depicted in Fig.3.1.

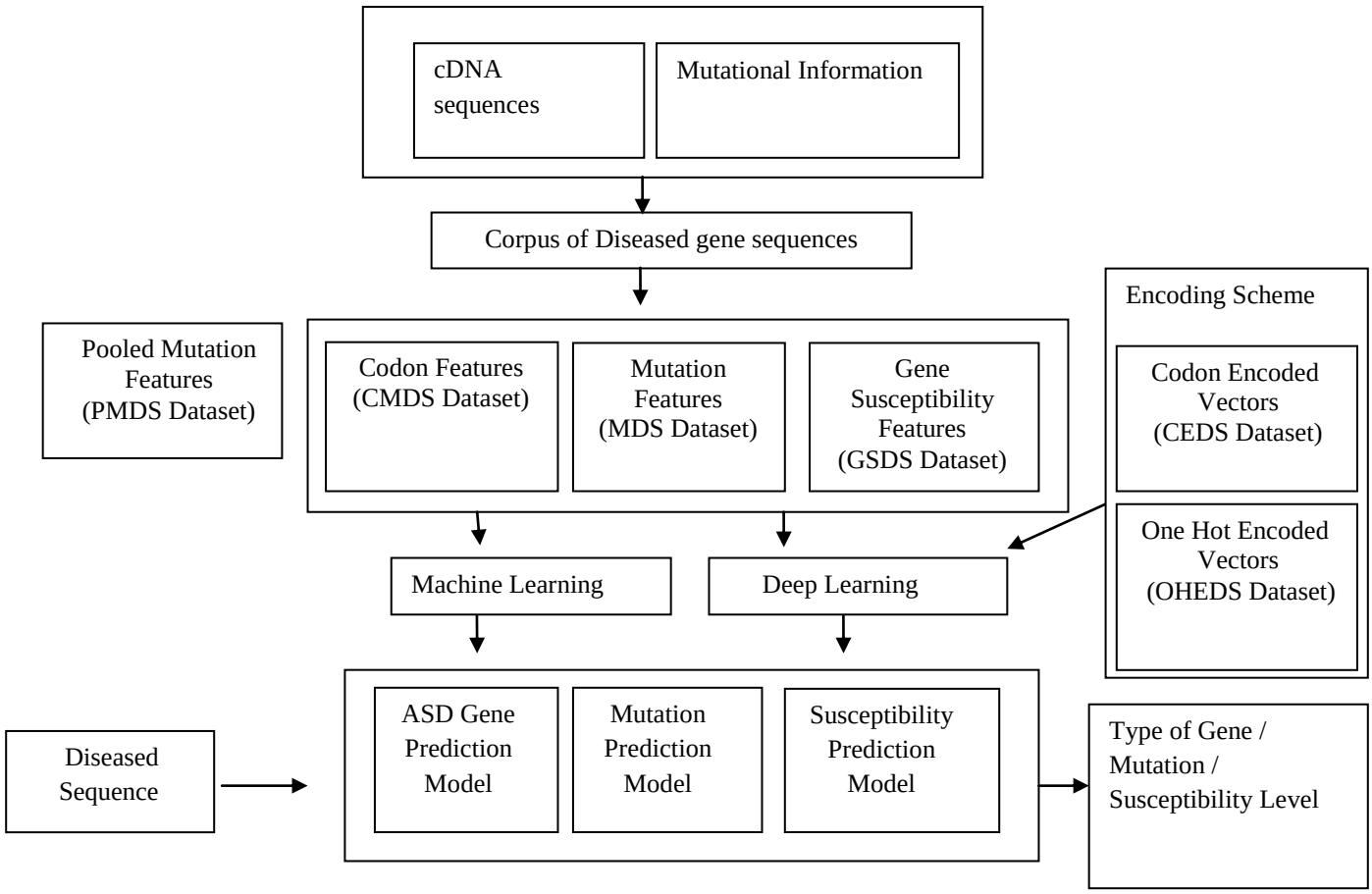

Fig. 3.1 Architecture of the Proposed System

As the mutated gene sequences are not readily available, CDNA sequences of the ASD genes responsible for syndromic and asyndromic ASD are first collected from HGMD database [79]. The mutational information about these genes is collected from SFARI gene database [80]. Diseased gene sequences are then simulated by inducing mutations with the help of this

information and the corpus is built with 1000 mutated gene sequences accounting for ten types of ASD genes and four types of mutations.

The research work is carried out in three stages using conventional machine learning and the contemporary deep learning methods for building the predictive models. In the first stage, the traditional learning approach is employed for building the prediction models and the key idea here is to identify and extract distinctive features from synthetic gene sequences. Various descriptors like codon measures, mutation features and gene susceptibility features are defined and captured from the mutated gene sequences. Four independent datasets with these features are prepared and used in the development of traditional machine learning models to predict the ASD causing genes, their susceptibility and the underlying mutations.

In the second stage, the deep learning approach is employed for building the prediction models and the key idea here is to explore the self- learning capability of deep models and to avoid feature engineering. The same datasets are used for enabling representation learning of the user defined features by different deep architectures and deep models are built to predict the ASD causing genes, their susceptibility and the underlying mutations.

In the last stage, two types of encoding schemes are proposed for deep learning and the key idea here is to exploit the self-learning power of deep learning models by utilizing the gene sequences as raw input data and thereby avoid the time consuming task of feature engineering. Two datasets based on two mapping schemes are developed and utilized for the development of deep learning models to predict the ASD causing genes.

3.1 CORPUS DEVELOPMENT

Data collection refers the systematic process of accumulating and evaluating information on variables of interest to enable answering of research questions, testing hypotheses and evaluating results. The primary focus of data collection is to capture quality evidence that permits analysis to formulate reliable answers to the problems. Accurate prediction of gene sequences is a complicated task as the pattern of the gene sequence varies for every ASD affected individual and the unavailability of diseased gene sequences also poses a challenge. Initially the genes associated with syndromic and asyndromic ASD as discussed in Section 1.3 are examined. The syndromes with ASD prevalence of about 0.5% - 2% given in Table III are

taken for study whereas other syndromes in which ASD prevalence is very rare are not included. The syndromes like Fragile X syndrome, Rett syndrome, Timothy syndrome, Tuberous sclerosis complex, Phelan - McDermid syndrome and the respective 5 causative genes FMR1, MECP2, CACNA1C, TSC1, SHANK3 are taken for the research. The asyndromic ASD symptoms like repetitive behavior, speech and language abnormalities, developmental disabilities, cognitive and behavioral impairments were discussed in Section 1.3.3. The 5 genes responsible for these symptoms CHD8, CNTNAP2, FOXP2, GABRB3, HOXA1 are considered. Table V lists the ten genes that are key players for syndromic and asyndromic ASD considered for the study with their associated behavior.

Table V ASD Causative Genes Taken for Study and their Associated Behavior

Type of ASD	Genes taken for study	Associated Behavior
Syndromic ASD	SHANK3	Impulsivity, social anxiety, biting, obsessive chewing
	TSC1	Behavioral deficits associated with ASD, hyperactivity, epilepsy
	MECP2	Gaze avoidance, limited facial expression, atypical socialization
	FMR1	Social withdrawal behaviors, anxiety, learning disability
	CACNA1C	Restricted and Repititive behaviors, Communication problems
Asyndromic ASD	CHD8	Anxiety, Repetitive behavior
	FOXP2	Speech and language abnormalities
	GABRB3	Unexplained Epilepsy and Intellectual or Developmental Disabilities
	HOXA1	Cognitive and behavioral impairments
	CNTNAP2	Language impairments, Aggression

The reference genes are identified from OMIM database and its corresponding reference gene sequences are downloaded from National Center for Biotechnology Information (NCBI) [81]. Online Mendelian Inheritance in Man (OMIM) is a freely available, complete, trustworthy

compendium of human genes and genetic phenotypes [82]. The full-text, referenced overviews in OMIM contain information on all known mendelian disorders and over 15,000 genes. OMIM highlights the relationship between phenotype and genotype. Each OMIM entry has a summary of a genetically determined phenotype or gene and has numerous links to other genetic databases like DNA and protein sequence, PubMed references and mutation databases. Information in OMIM can be retrieved by queries on OMIM number, disorder, gene name and gene symbol. It is updated on a daily basis and the entries in it contain abundant links to other genetics resources.

The diseased gene sequences will enable geneticists to exactly identify the reason behind the disorder. But the availability of this diseased gene sequences is a challenge and hence it is essential to generate synthetic gene sequences. Mutations are the key players that change the pattern of a gene sequence and so information about mutations are required for this process. The gene mutational information is collected from the Human Gene Mutational Database (HGMD). The Human Gene Mutation Database (HGMD) comprises a broad collection of published germline mutations in nuclear genes that lie behind human inherited disease. The database contains an excess of 203,000 different gene lesions identified in over 8000 genes manually curated from over 2600 journals. There are more than 17000 new mutation entries collected per annum and HGMD is the standard genotype - phenotype repository of heritable mutations. It is widely used by researchers, clinicians, diagnostic laboratories and genetic counsellors, and is an indispensable tool for the annotation of next-generation sequencing data. The public version of HGMD (http://www.hgmd.org) is freely available to registered users from academic institutions and non-profit organizations.

In this work four kinds of mutations such as missense, nonsense, frameshift and silent mutations are taken into account for building the corpus. The corresponding mutation records are identified and captured from the HGMD database by specifying the required information. The sample mutation information retrieved for SHANK3 gene from HGMD is shown in Fig.3.2.

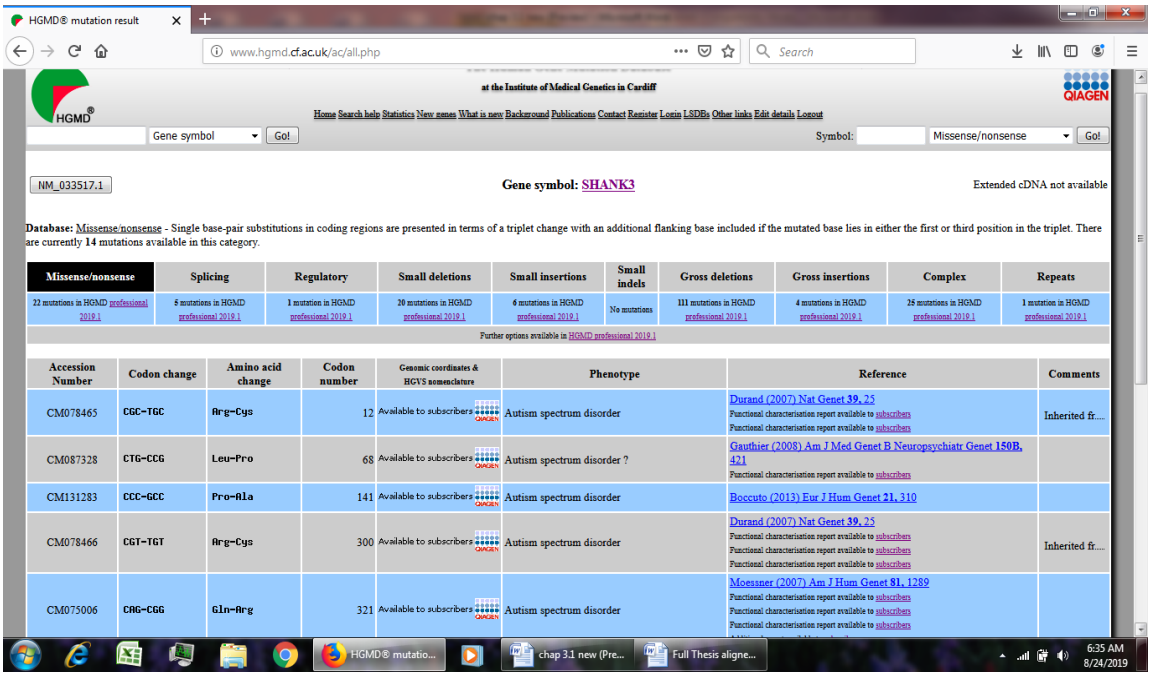

Fig. 3.2 Mutation information retrieved for SHANK3 gene from HGMD

SFARI gene is a growing database for the autism research community that is focused on genes associated in autism susceptibility. SFARI gene is a reliable, inclusive, and dynamic database that identifies the risk genes from the published literature. It enables autism researchers to know about the gene functions in humans and experimental organisms, with links to the primary literature and secondary database. Mutational information are collected from SFARI and HGMD databases.

The raw cDNA sequence is obtained from HGMD and the reference sequences are downloaded from NCBI. Positional cloning is done using R script wherein the nucleotide base is altered based on the mutational information. R coding is executed to identify the nucleotide position and to replace with a different nucleotide specified in the mutational information. The nucleotide base variation in the gene sequences is done based on the mutational information involving Nonsense, Missense, Frameshift and Silent Mutations obtained from SFARI and HGMD database. The positional change of the nucleotide is done in cDNA sequence against the reference gene sequence.

For example, consider the missense mutational information for SHANK3 gene such as CGC > TGC in codon 12. This indicates that the nucleotide C in codon 12 is altered as nucleotide T and hence alters the protein from Arginine to Cysteine.

For the cDNA sequence of the SHANK3 given below

ATGGACGGCCCCGGGGCCAGCGCCGTGGTCGTGCGCGT

the mutation induced in codon 12, replacing the nucleotide C with T, the result of mutated sequence is

ATGGACGGCCCCGGGGCCAGCGCCGTGGTCGTGTGCGT

R script is written to identify the required position to be altered and to induce the mutation in the gene sequence. For SHANK3 gene sequence depicted in Fig.3.3 the mutated gene sequence generated by replacing the nucleotide C with T is illustrated in Fig.3.4.

>ATGGACGGCCCCGGGGCCAGCGCCGTGGTCGTGCGCGTCGGCATCCCGGACCTGCAGCAGAC
GAAGTGCCTGCGCCTGGACCCGGCCGCGCCCGTGTGGGCCGCCAAGCAGCGCGTGCTCTGCGC
CCTCAACCACAGCCTCCAGGACGCGCTCAACTATGGGCTTTTCCAGCCGCCCTCCCGGGGCCGCG
CCGGCAAGTTCCTGGATGAGGAGCGGCTCCTGCAGGAGTACCCGCCCAACCTGGACACGCCCCT
GCCCTACCTGGAGTTTCGATACAAGCGGCGAGTTTATGCCCAGAACCTCATCGATGATAAGCAG
TTTGCAAAGCTTCACACAAAGGCGAACCTGAAGAAGTTCATGGACTACGTCCAGCTGCATAGCA
CGGACAAGGTGGCACGCCTGTTGGACAAGGGGCTGGACCCCAACTTCCATGACCCTGACTCAGG
AGAGTGCCCCCTGAGCCTCGCAGCCCAGCTGGACAACGCCACGGACCTGCTAAAGGTGCTGAA
GAATGGTGGTGCCCACCTGGACTTCCGCACTCGCGATGGGCTCACTGCCGTGCACTGTGCCACA
CGCCAGCGGAATGCGGCAGCACTGACGACCCTGCTGGACCTGGGGGCTTCACCTGACTACAAG
GACAGCCGCGGCTTGACACCCCTCTACCACAGCGCCCTGGGGGGTGGGGATGCCCTCTGCTGTG
AGCTGCTTCTCCACGACCACGCTCAGCTGGGGATCACCGACGAGAATGGCTGGCAGGAGATCCA
CCAGGCCTGCCGCTTTGGGCACGTGCAGCATCTGGAGCACCTGCTGTTCTATGGGGCAGACATG

Fig. 3.3 cDNA sequence of SHANK3 Gene

Codon Change : CGC –TGC
Amino Acid Change : Arg –Cys
Codon number: 12
>ATGGACGGCCCCGGGGCCAGCGCCGTGGTCGTGTGCGTCGGCATCCCGGACCTGCAGCAGACGA
AGTGCCTGCGCCTGGACCCGGCCGCGCCCGTGTGGGCCGCCAAGCAGCGCGTGCTCTGCGCCCTCA
ACCACAGCCTCCAGGACGCGCTCAACTATGGGCTTTTCCAGCCGCCCTCCCGGGGCCGCGCCGGCAA
GTTCCTGGATGAGGAGCGGCTCCTGCAGGAGTACCCGCCCAACCTGGACACGCCCCTGCCCTACCTG
GAGTTTCGATACAAGCGGCGAGTTTATGCCCAGAACCTCATCGATGATAAGCAGTTTGCAAAGCTTC
ACACAAAGGCGAACCTGAAGAAGTTCATGGACTACGTCCAGCTGCATAGCACGGACAAGGTGGCAC
GCCTGTTGGACAAGGGGCTGGACCCCAACTTCCATGACCCTGACTCAGGAGAGTGCCCCCTGAGCC
TCGCAGCCCAGCTGGACAACGCCACGGACCTGCTAAAGGTGCTGAAGAATGGTGGTGCCCACCTGG
ACTTCCGCACTCGCGATGGGCTCACTGCCGTGCACTGTGCCACACGCCAGCGGAATGCGGCAGCAC
TGACGACCCTGCTGGACCTGGGGGCTTCACCTGACTACAAGGACAGCCGCGGCTTGACACCCCTCTA
CCACAGCGCCCTGGGGGGTGGGGATGCCCTCTGCTGTGAGCTGCTTCTCCACGACCACGCTCAGCT
GGGGATCACCGACGAGAATGGCTGGCAGGAGATCCACCAGGCCTGCCGCTTTGGGCACGTGCAGC
ATCTGGAGCACCTGCTGTTCTATGGGGCAGACATGGGGGCCCAGAACGCCTCGGGGAACACAGCCC

Fig. 3.4 Mutated SHANK3 Gene

In each category of ASD genes taken for study 100 synthetic mutated gene sequences are generated and a corpus comprising of 1000 gene sequences for all 10 genes is developed. These synthetic gene sequences are stored as fasta files.

3.2 DESIGN OF FEATURES AND DATASETS

Data preparation is a crucial step as it transforms the initial raw data into a final dataset that is vital for the development of reliable models that have high accuracy and efficiency. Feature extraction is one of the important steps in data analysis, mainly influencing the accomplishment of any machine learning task. Feature engineering plays a vital role in determining the accuracy of the model in traditional machine learning. The key idea of feature engineering in this work is to extract distinguishing features from synthetic gene sequences for building the prediction models.

To facilitate traditional machine learning and to provide appropriate solution for the objectives under consideration, four different datasets have been developed.

Codon Measures Dataset

Identification of harmful genes causing ASD is a complex research problem as numerous genes underlie this disorder. Hence it is essential to develop machine learning models to recognize the diseased genes associated with ASD using gene characteristics. The coding measures are dissimilar in different gene families and hence this trait is a well-chosen descriptor for identifying different gene types. Hence a dataset that includes attributes that describe a gene on different aspects was developed. The study investigated a total of 43 attributes in both intrinsic and extrinsic categories which are the contributing features for representing the mutated gene sequences. The features taken into consideration for gene identification are nucleotide composition, GC content, Rho values of biwords, Z-scores of biwords, Alignment score, Number of exons, Number of donor sites, Number of acceptor sites, CpG percent, ratio of CpG percent / expected.

Intrinsic content sensors use the measures like GC content, frequency of k-mers, base occurrence periodicities based on the content of the sequences to identify a region as protein coding or not. Nucelotide composition and codon composition varies with regard to protein and non-protein coding regions. Extrinsic content sensors exploit the similarity between a genomic sequence region and a DNA sequence present in a database to establish whether the region is transcribed or coding. Local alignment tool BLAST is used for detecting this similarity. To find biwords that are over-represented or under-represented rho values and z-scores are used. The count of exons, donor site and acceptor site are also examined as they are essential discriminators of a gene.

In summary, a total of 43 features including intrinsic and extrinsic properties of each gene sequence are defined. The training set for the multi-class classification problem includes 1000 feature vectors with a dimension 43. As ten genes namely SHANK3, TSC1, MECP2, FMR1, CACNA1C, CHD8, FOXP2, GABRB3, HOXA1, CNTNAP2 are considered for the study, the feature vectors are assigned class labels from 1 to 10.

The detailed description of the feature extraction and dataset creation is given in Section 4.1 of Chapter 4.

Mutation Dataset

In a clinical environment to enable the ASD patients for precise genetic tests, it is indispensable to identify the kind of genetic mutations that are the causal factors of the phenotype. Change or mutation in the gene sequence alters the structure of the sequence which implies the cause of disease. These structural changes are captured as characteristics from mirrored sequences to learn the prediction model. The gene code for making a protein is altered by the mutation which causes the protein to malfunction. When a mutation alters a protein that has a major role in the body, it can interrupt normal development or cause a disorder. It is rather a complex task of classifying mutations in complex syndromic ASD genes taking into account the genetic variations due to synonymous and non-synonymous nucleotide polymorphisms. To distinguish mutations in ASD gene sequences, gene specific features (GS), substitution matrix features (SM), amino acid residue changes (AARC) are vital. Hence a dataset that includes 15 attributes that describe a mutation on different aspects was designed. These attributes can be categorized into 5 gene specific, 6 features extracted from published substitution scoring matrices and 4 features related to amino acid residue changes. The attributes of this dataset are described below.

The features like mutation start position, mutation end position, length of mutation, length of cDNA sequence, the type of mutational variation i,e Nonsense, Missense, Frameshift, Silent Mutations are characteristics extracted from a gene. In sequence alignment, scoring matrices are used to decide the relative score made by matching two characters. They are computed as the log-odds of the probability of two characters that are derivatives of a common ancestral character. Many types of scoring matrices exist for nucleotide sequences, codon sequences and amino acid sequences derived by aligning the known homologous sequences [84]. These alignments are then used to determine the likelihood of one character being at the same position in the sequence as another character.

This work utilizes the values of 6 scoring matrices namely (i) WAC matrix constructed from amino acid comparative profiles, (ii) Log-odds scoring matrix collected in 6.4-8.7 PAM , (iii) BLOSUM80 substitution matrix, (iv) PAM-120 matrix, (v) Substitution matrix obtained by

maximum likelihood estimation and (vi) Mutation matrix for initially aligning which are collected from the AAIndex database.

The mutated sequences are translated to generate protein sequences which in turn provide the amino acid observed values whereas the amino acid expected value is extracted from SFARI autism database. The 2-gram encoding method extracts different patterns of two consecutive amino acid residues in a protein sequence and counts the number of occurrences of the extracted residue pairs. There are 20^2 combinations of 2-grams which is huge and hence the standard deviation and the mean z-score between the values of the 400 bigrams with respect to the protein sequence are calculated.

The dataset with a total of 15 features for distinguishing the mutations is developed with 1000 feature vectors with a dimension 15. There are four types of mutations considered for the study namely Nonsense, Missense, Frameshift, Silent Mutations and hence the feature vectors are assigned four different class labels.

The detailed description of the feature extraction and dataset creation is given in Section 4.2 of Chapter 4.

Pooled Mutation Dataset (PMDS)

In most of the candidate ASD genes, the actual mutations that increase the risk for autism have not been identified. There is a need for a more comprehensive learning approach that exploits the correlations between ASD genes and the mutations that underlie them. In order to identify the ASD genes and the co-occurring mutations which exhibit dependency among them, multi-dimensional approach is attempted. To address this need, a dataset reflecting both gene characteristics and mutation aspects is designed. The gene specific features like nucleotide composition, Rho values of biwords, Z scores of biwords, Alignment score, count of exons, donor sites, acceptor sites are pooled with the mutation features such as mutation features (GS), substitution matrix features (SM) and amino acid change residues (AARC) and 1000 instances of dimension 58 are created.

The dataset with features contributing to the gene – mutation identification is developed with 1000 feature vectors with a dimension 58. There are two class labels assigned for this

dataset namely the gene class and mutation class. The gene class has ten class labels as there are ten genes taken for the study. There are four types of mutations considered for the study and hence the dataset consists of four class labels for mutation class.

The detailed description of the feature extraction and dataset creation is given in Section 4.3 of Chapter 4.

Gene Susceptibility Dataset

The exploration for genetic factors underlying ASD has led to the identification of hundreds of genes containing mutations that differ in the mode of inheritance, frequency and function. Each of the mutations has its own associated risk to ASD and hence it is challenging to assess the collective substantiation for the gene's susceptibility to the disorder. It is necessary to perfrom a systematic evaluation of a gene's susceptibility to ASD by considering different types of genetic variants implicated in ASD. To address this need, a dataset with significant attributes representing the gene susceptibility is developed. The approach is based on an integrated assessment involving multiple attributes of gene, mutation, conserved protein domains, gene expression profiles and pathway interactions.

Various characteristics determine the gene's vulnerability to a disorder. Initially publication evidences collected from PubMed are considered along with gene properties like exon count, protein length, whether protein altered or not, conserved domains. Mutation specific properties like mutation type, start, end position, inheritance pattern, rare or common variant are studied. The association of a gene with biological processes and cellular components related to ASD are included. The presence of a gene in ASD linked pathways indicates its link to the disorder and so axon guidance, neuronal system, interaction of neurexin and neuroligin at synapses, developmental biology and synaptic transmission are investigated. A consolidated score by summing the various features are generated for each individual variant of an ASD-implicated gene leading to a clear understanding of their relevance to the disorder.

Finally one of the three class labels low (score < 0.5), medium (score >=0.5 and < 0.8), high(score >=0.8) is assigned to the gene depending on the range. Thus the Gene Susceptibility Dataset **(**GSDS) dataset is designed with 1000 instances of dimension 25. The detailed description of the feature extraction and dataset creation is given in Chapter 5.

Table VI summarizes the training dataset used in both conventional machine Learning and contemporary deep learning approaches.

Table VI Summary of Training Datasets

Measures	Codon Measures Dataset (CMDS)	Mutation Dataset (MDS)	Pooled Mutation Dataset (PMDS)	Gene Susceptibility Dataset (GSDS)
Gene sequences	1000	1000	1000	1000
Genes considered for study	10	10	10	10
Number of features	43	15	58	25
Class Labels	10	4	10	3
Dataset size	1000 x 43	1000 x 15	1000 x 58	1000 x 25

To facilitate deep learning through representation learning and to provide appropriate solution for gene type identification, mutation prediction and gene susceptibility identification, the three datasets namely CMDS, MDS, GSDS can be used.

For next level research, two types of encoding schemes are proposed and two different datasets are designed to utilize the self learning power of deep architectures.

Codon Encoded Dataset

Codons are good differentiators of genes and hence the synthetic gene sequences are converted into categorical values ranging from 1 to 64. The diseased gene sequences are converted into a one dimensional representation by encoding technique. Among the genes considered, CHD8 has the maximum number of 2582 codons . Hence the length of each record is taken as 2582 timesteps of a feature vector which is the maximum number codons in a gene sequence. The Codon Encoded Dataset (CEDS) consisting of the encoded feature vectors is created with 1000 instances of dimension 2582 where each instance is assigned with one hot encoded class label ranging from 1 to 10.

One Hot Encoded Dataset

Another encoding mechanism of one-hot encoding is attempted to give binary representation for the input sequences without losing positional information of each nucleotide. The simulated mutated sequences undergo the process of one hot encoding where each input sequence of length l is transformed into a 4 × l representation. The nucleotide bases adenine (A), cytosine (C), guanine (G), thymine (T) match the components from top to bottom respectively. If one of the nucleotide appears, the corresponding component is set to one and the others are set to 0. All sequences are not of the same length, but in order to make it uniform, 0 padding is done to make them equal in length. The One Hot Encoded Dataset (OHEDS) consisting of encoded feature vectors is created with 1000 instances of dimension 7746 x 4 where each instance is assigned with one hot encoded class labels ranging from 1 to 10 for the ten possible genes.

These two datasets are utilized for the development of deep learning models to predict the ASD causing genes. The detailed description of the above two datasets is given in Chapter 8.

3.3 TRAINING AND TESTING

The training datasets mentioned in the previous sections are used to train the classifiers for multi classification task as the problem objectives are to identify ASD causing genes, their susceptibility and mutations. The algorithm learns from the observations in the training set which forms the experience. Each observation in supervised learning problems consists of an observed output variable and one or more observed input variables. The machine learning algorithms like Decision trees, SVM and MLP are employed to construct the models employing CMDS, MDS, GSDS datasets. In contemporary method, Deep Neural Network (DNN), Recurrent Neural Network (RNN) variants namely Bidirectional Recurrent Neural Network (BRNN), Long Short Term Memory (LSTM) and Gated Recurrent Units (GRU) are employed to build the prediction models.

Various experiments have been carried out using Sci-kit learn and Keras with Tensorflow as backend for implementing the traditional machine learning and deep learning approaches respectively. Training the features of ASD for different types of genes helps to create the learned model. In the training phase, the set of class labeled tuples is presented and the model is trained

by pairing the input with expected output. The test set is used to estimate how well the model has been trained and to estimate classification errors for classifiers, recall and precision.

The common techniques to assess the accuracy of a classifier are hold-out, k - fold cross validation and leave one out cross validation. The hold-out method is the simplest kind of cross validation where the dataset is separated into training set and test set. The training set is used by function approximator to fit a function and then it predicts the output values for data in the test set. The test error is used to evaluate the model and the errors are accumulated to give the mean absolute test set error. Cross validation is mainly used in machine learning to estimate the skill of a machine learning model on unseen data. In this technique the entire data set is not used to train a learner and some of the data is removed before training begins. The data that was removed during training can be used to test the performance of the learned model on new data. To determine the accuracy of a learning algorithm in predicting untrained data leave one out cross validation is used. In this method the learning algorithm is trained multiple times using all but one of the training set data points.

K-fold cross validation evaluates the data across the entire training set, but it does so by dividing the training set into K folds and then training the model as many times as K. During each iteration a different fold of the training data is left out and is used as a validation set. At the end, the performance metric is averaged across all K tests. Lastly, as before, once the best parameter combination has been found, the model is retrained on the full data. As can be seen, every data point gets to be in a validation set exactly once, and gets to be in training set K-1 times. This significantly reduces bias as most of the data is used for fitting, and also significantly reduces variance as most of the data is also being used in validation set. To add to the effectiveness of this method, interchanging of the training and test sets can also be done. As a common rule and empirical evidence usually K = 5 or 10 is ideal.

In this work, the value of K is fixed as 10 and the entire dataset is divided into 10 folds out of which 9 folds are used for training and 1 for testing in each iteration. 10 fold cross validation is performed to test the performance of the prediction models.

Performance Evaluation Measures

The performances of the models are evaluated using 10 fold cross validation with various metrics like precision, recall, accuracy and F-measure. During performance evaluation the number of observations correctly identified are denoted by True positive (TP) whereas the number of observations correctly rejected are indicated by True Negative (TN). False Positive (FP) gives the number of observations incorrectly identified by the model and False Negative (FN) refers the number of observations which are incorrectly rejected by the model.

Accuracy

The proportionate number of times the predictive model is right when applied to data is measured by accuracy. Accuracy can be calculated from formula given in equation 3.1.

$$\text{Accuracy} = TP + TN / (TP + TN + FP + FN) \quad (3.1)$$

Precision

It is the proportion of the samples which truly have class z among all those which were classified as class z. Precision can be calculated from formula given in equation 3.2.

$$\text{Precision} = TP / (TP + FP) \quad (3.2)$$

Recall

It is the ratio of samples of a particular class z correctly classified as belonging to that class z. Recall can be calculated from formula given in equation 3.3.

$$\text{Recall} = TP / (TP+TN) \quad (3.3)$$

F-measure comparison

F-measure is used to decide the accurate classification of document labels within dissimilar classes. It measures the efficacy of the algorithm on a single class and higher values indicate better results. It is defined as given in equation 3.4.

$$F = 2 * (\text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (3.4)$$

SUMMARY

The core research component of problem modeling has been elucidated in detail in this chapter with various tasks such as corpus development, features and dataset preparation, training and testing. The process of corpus development was described and the design of different datasets was presented. A note on training and testing of the models was also given. The method of performance evaluation and the metrics used have been mentioned in this chapter. Various models built using the traditional machine learning algorithms trained by CMDS, MDS, and PMDS datasets will be described in Chapter 4. The gene susceptibility prediction model built by training pattern recognition algorithms using GSDS dataset will be discussed in Chapter 5. The models built using deep learning approach with CMDS, MDS, GSDS datasets for predicting the type of ASD causing gene, their susceptibility and mutations will be presented in Chapter 6 and 7. The deep models built using encoding schemes to predict the type of ASD causing gene will be described in Chapter 8.