1. INTRODUCTION

Data mining has caused revolutionary changes in research and has been used extensively to solve problems in many sectors. Data mining is the process of discovering new correlations, patterns and trends that are significant by passing a large number of data stored in a vault. The non-trivial extraction of implicit, previously unknown and potentially useful information from data paves way for great prospects to solve problems. Data mining has quickly become pervasive and has contributed immensely towards providing new insights and increased efficiency in various fields like bioinformatics, telecommunications, healthcare, finance, retail and intrusion detection. It is one of the major frontiers that facilitates in solving societal issues like improving healthcare and reducing costs, finding alternative energy sources, predicting the impact of climate change and reducing poverty by increasing agriculture production. Data mining facilitates scientists to perform automated analysis of huge datasets. Data mining techniques and application tools are used extensively in healthcare sector as they reduce the complexity of the data transactions in this domain. It is used in the health sector for various tasks like diagnosing the disease, determining the treatment for a disease, classifying the patients and determining the high risk factors in surgeries.

Bio-informatics is a domain in which data mining is strongly embraced as it is rich in data. Genome sequencing has contributed to an exponential increase in complete and partial sequence databases. Progress in technologies such as microarrays, resulted in the beginning of the sub domains of genomics and proteomics. These technologies examine the genes, proteins and the circuit inside the cell that regulates the gene-expression. Lots of data are being generated and needs to be mined by the mankind to interpret the mysteries of cells. During the last few years, enormous progress has been achieved, but there remain a number of fundamental problems in bio-informatics, such as finding genes and predicting protein structures. Data mining will play an essential role in the understanding of gene expression, the development of medicines and other problems in genomics and proteomics. Furthermore, text mining will be important to filter knowledge from the escalating literature concerning bioinformatics. Applications of data mining in bioinformatics include gene finding, protein function inference, disease diagnosis,

disease prognosis, disease treatment optimization, protein and gene interaction network reconstruction, data cleansing and protein sub-cellular location prediction.

This research focuses on the application of data mining and machine learning to build predictive models for the identification of ASD causative genes, their susceptibility and the triggering mutations. A brief introduction about data mining, machine learning and deep learning is presented in this chapter. An overview of ASD including the types of ASD, genes associated and the mutations are also elaborated. Subsequently the detailed report of literature survey and motivation behind this study is highlighted. The objectives of this research work are also stated clearly in this chapter.

1.1 DATA MINING

Data mining is a promising technology which draws ideas from machine learning, artificial intelligence, pattern recognition, statistics and database systems. It is a process of analyzing large amounts of data, stored in databases or data warehouse to find hidden knowledge [1]. The data mining tasks can be classified generally into predictive and descriptive as depicted in Fig.1.1 based on what a specific task tries to achieve.

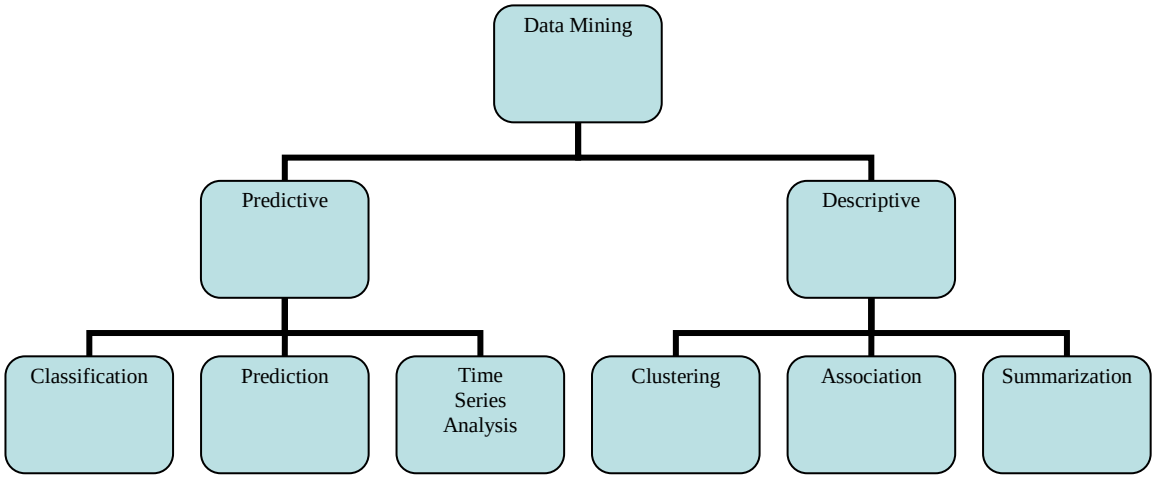

Fig. 1.1 Data Mining Tasks

Predictive Tasks

Predictive data mining tasks perform inference on the presented data set to predict the behavior of a new data set. Predictive data mining tasks come up with a model from the existing data set that is useful in predicting unknown or future values of another data set of interest [2]. A

sample predictive data mining task is a medical practitioner trying to diagnose a disease based on the medical test results of a patient. The three types of predictive tasks include classification, prediction and time-series analysis.

Classification

Classification is a standard data mining technique based on machine learning. Fundamentally, classification is used to classify each item in a set of data into a predefined set of classes or groups. Classification method makes use of mathematical techniques like decision trees, linear programming, neural network and statistics [3]. It is the process of building a model that describes the data classes or concepts. The idea is to use this model to predict the class of objects whose class label is unknown. This model is developed by analyzing the training data. There are different ways of presenting the derived model namely

- Classification (IF-THEN) Rules
- Decision Trees
- Mathematical Formulae
- Neural Networks

A model to categorize loan applicants is a typical classification example. The model to predict credit risk is built using the observed data of the loan applicants for a period of time. The data will also include personal details like employment status, own house details, investments details and so on. Classification aims to accurately predict the target class for each instance present in the data. In the above example credit rating is the target, the other attributes are the predictors, and the data for each customer constitutes an instance. The simplest type of classification problem is binary classification in which the target attribute has only two possible values like high credit rating or low credit rating. Multiclass targets typically have more than two values like low, medium, high or unidentified credit rating.

Classification is applied in many areas like business modeling, biomedical, credit analysis, customer segmentation etc. Examples of classification task include categorizing news stories as finance, weather, entertainment, sports etc., classifying land covers using satellite data, identifying intruders in the cyberspace, predicting tumor cells as benign or malignant and classifying secondary structures of protein as alpha-helix, beta-sheet, or random coil. Some of

the classification techniques are Decision tree, K-nearest neighbor, Support Vector Machines, Naive Bayesian classifiers and Neural Networks [4].

Prediction

Prediction is one of the most valuable data mining techniques, since it is used to forecast the types of data seen in the future. In predictive tasks, historical data is fed into a mathematical model that considers key patterns in the data. The model is then applied to current data to predict what will happen in the future. Prediction is one of the data mining techniques that detects the relationship between dependent and independent variables. Most prediction techniques are based on mathematical models such as regression, neural networks, Radial Basis Function (RBF). All are centered on fitting a curve through the data to find the relationship from the predictors to the predicted variable. For instance, the prediction analysis technique can be used in sales to predict the profit. In this case, if sales is considered as an independent variable, profit can be a dependent variable. The historical sales and profit data can be used to draw a fitted regression curve which in turn can predict profit. Predicting fuel efficiency based on a linear regression model of engine speed versus load is another example. Other real life examples include predicting level of sales that will result from a price change and predicting the rainfall based on current humidity.

Time - Series Analysis

Time series is a chain of events in which the next event is decided by one or more of the previous events [5]. Time series analysis comprises methods to analyze time-series data to mine valuable patterns, trends, rules and statistics. Predicting stock market prices is done using time-series analysis. The goal of time series analysis is to find the patterns in correlated data. The widespread patterns are trends and seasonality. Trend is the overall movement or general direction of the data, ignoring any short term effects like cyclical or seasonal variations. For example, the enrollment trend at a particular university may be steadily increasing on an average over the past 100 years.

Trends are typically linear and moving averages or regression analysis is used to find trends. Seasonality is a trend that repeats itself systematically over time. Temperatures usually show seasonal variation, dropping in winter and rising again in summer. If the time series exhibits seasonality, there should be four to five cycles of observations in order to fit a seasonal

model to the data. Other examples of time series analysis include analyzing patient's heartbeat, hourly readings of air temperature, predicting airline traffic volume, rainfall, yearly sales etc. Prediction or forecasting is widely used in economics and business.

Descriptive Tasks

Descriptive data mining tasks usually find data describing patterns and comes up with novel, significant information from the available data set. The descriptive function handles the common properties of data in the database. It can be defined to discover interesting regularities in the data, to expose patterns and find interesting subgroups in the bulk of data. Descriptive analytics focuses on the summarization and conversion of the data into significant information for reporting and monitoring [6]. The descriptive analysis is used to extract data and offer the most recent information on past or current events. The various operations performed in the descriptive approach are standard reporting, query or drill down and ad-hoc reporting which are capable of finding the source of problem. Descriptive mining employs unsupervised learning functions while predictive uses supervised learning techniques. An example of a descriptive data mining task is identifying products that are purchased together. The list of descriptive functions includes clustering, mining of associations and summarization.

Clustering

Clustering is an unsupervised data mining technique that makes a meaningful or useful cluster of objects which have similar characteristics. In clustering, a set of data items is organized into groups such that there is high intra cluster similarity and low inter cluster similarity. Clustering defines the classes and puts objects in each class, while classification assigns objects into predefined classes [7]. Customer segmentation is a real time application of clustering. Clustering can be divided into two subgroups namely hard and soft clustering. Every instance in hard clustering belongs to a cluster completely. For example, each customer can be assigned into only one cluster out of the k clusters. Soft clustering assigns a probability of a data point belonging to all k clusters instead of assigning it to a separate cluster [8]. For example, from the above scenario each customer is assigned a probability of belonging to k clusters of the retail store.

k-means is a well accepted machine learning algorithm for cluster analysis. It operates on a given set of data iteratively and the number of clusters to be formed is fixed. The data given as input to k - means algorithm is partitioned into k clusters. In case of globular clusters, k-means produces denser clusters than hierarchical clustering [9]. Given a smaller value of k, the algorithm computes faster than hierarchical clustering for large number of variables. Clustering helps marketers improve their customer base and work on the target areas. A clustering problem is where the inherent groupings in the data is discovered such as grouping customers by purchasing behavior. Some of the applications of clustering include customer profiling for targeted marketing, grouping related documents that are similar to each other, grouping genes and proteins that have similar functionality, grouping stocks with similar price fluctuations.

Association Rule Mining

Association rule mining is one of the well-known data mining techniques. In association analysis, a pattern is revealed based on a relationship between items in the same transaction. Given a set of records each of which contain number of items from a given collection, association analysis produces dependency rules that predicts the occurrence of an item based on occurrence of other items [10]. Association rule mining problem finds rules that describe massive parts of information, like people who purchase X additionally tend to shop for Y.

Association is linked to finding patterns, but is more precise to dependently related variables. The attributes that are highly correlated with another attribute are identified. This technique is used in market basket analysis for sales promotion, shelf management and inventory management. An example of an association discovered is that 64% of the shoppers who bought milk also purchased bread. Association rule mining is a popular technique for market basket analysis because all possible combinations of potentially interesting product groupings can be explored. In medical informatics association rules are used to find combination of patient symptoms and test results associated with certain diseases.

Apriori algorithm is a popular algorithm that generates association rules from a given data set. Association rule denotes that if an item A occurs, then item B also occurs with a definite probability. The fundamental principle of Apriori algorithm is that if an item set occurs frequently then all the subsets of the item set also occur frequently. On the other hand an item set

that occurs infrequently will have all its supersets with infrequent occurrence. Apriori algorithm is easy to implement and can be parallelized easily.

Summarization

Summarization is the generalization of data. A set of related data is summarized yielding a reduced set that presents aggregated information of the data. Data is mapped into subsets with associated plain descriptions in summarization. Basic statistics such as mean, standard deviation, variance, mode and median can be used as summarization approaches. For example, the shopping done by a customer can be summarized into total products, total spending, offers used, etc. Such high level summarized information can be useful for sales or customer relationship team for detailed customer and purchase behavior analysis. Data can be summarized in different abstraction levels and from different angles.

Data Mining Types

Data mining is categorized into various types like web mining, spatial data mining, text mining, sequence mining, graph mining etc. Some of them are mentioned below.

Web Mining

The web poses great challenges for efficient resource and knowledge discovery. The sources are web pages, web data repositories, web logs, click-streams, web traffic and web links. Web mining is defined as the discovery and analysis of useful information from the World Wide Web to generate extraordinary results [11]. Web structure mining, web content mining and web usage mining are the sub categories of web mining. All three categories focus on the process of knowledge discovery of implicit, previously unknown and potentially useful information from the web. Web content mining targets the knowledge discovery from multimedia documents such as images, video, and audio, which are linked to the web pages. Web structure mining is centered on the analysis of the web link structure and generates structural summary about the web site and web pages. It tries to find out the link structure of the hyperlinks at the inter-document level. The main focus of web usage mining is on techniques that predict behavior of web users. It discovers user navigation patterns from web data and tries to discover useful information from the interactions of the users while surfing on the web. Web usage mining finds user access patterns of web pages by gathering the data from web log records. Opinion mining is an example of web

content mining and is of great significance for marketing intelligence and product benchmarking. Automatically segmenting web pages to extract the main content of the pages is another interesting example.

Spatial Data Mining

The immense explosion in geographically referenced data due to advancements in information technology highlight the significance of developing novel approaches to geographical data analysis. Spatial data mining is the extraction of implicit knowledge, spatial relations, or patterns not explicitly stored in spatial databases [12]. Spatial objects characterized by spatial data types and spatial relationships are stored in a spatial database. The complexity of spatial data types, spatial relationships and spatial autocorrelation limits the effectiveness of conventional data mining techniques for extracting spatial patterns. Classifying a spatial entity into a particular class like school or park is an example of a spatial classification task whereas finding outliers like crime location is a spatial clustering task. In spatial association non spatial attributes can be associated to spatial attributes. A spatial association rule for example states that 80% of schools that are close to sports centers are also close to parks, and 0.5 % of the data belongs to such a case.

Text Mining

The main purpose of text mining is to identify facts and relationships from huge textual data. It involves data mining, machine learning, statistics, and natural language processing to extract high quality, useful information from unstructured formats. It is used to optimize regular operational competency and to develop long-standing strategic decisions in domains like automobile, healthcare, and financial sector. In order to identify patterns and trends in huge volumes of unstructured data, methods like categorization, entity extraction, and sentiment analysis are used. The primary step in text mining is to organize the data into a more structured form by using natural language processing technology [13]. In most of the customer care applications text mining and natural language processing are used. Text analysis is also used for faster and automated customer response, thus reducing dependency on call center operations. It helps businesses and organizations to get valuable insights useful for their business. Text mining finds its application in sentiment analysis. The social media which is a potential source of

unstructured data is considered as a valuable source of information for market and customer intelligence. Many companies are using text mining to predict customer needs and evaluate the awareness of their brand. Text mining techniques are implemented to improve the effectiveness of spam filtering methods.

Sequence Mining

The enormous amount of sequence data available today has increased the real need for accurate and fast techniques to analyze these sequences. Understanding sequence data, and the ability to utilize this hidden knowledge, creates a significant impact on many aspects of the society. Some instances of sequence data are gene sequences, protein, user buying history and web user records. Sequence mining involves methods to discover statistically connected patterns among data samples in which the values are distributed in a sequence [14]. The discovered patterns can be used to implement competent systems that can aid in making predictions, improve usability of systems, identify events and facilitate in making strategic decisions. Sequences of ordered elements or events stored irrespective of time constitute a sequence database. Sequential pattern mining methods have been applied in various domains. It has plenty of applications as data is naturally encoded as sequence of symbols in many fields such as bioinformatics, weather prediction, market basket analysis, webpage click-stream analysis, e-learning, production processes and network intrusion detection.

There is immense amount of biological and clinical data from genomic and proteomic sequences as biologists are trying to comprehend the biological processes that lie behind disease pathways. Biological sequence analysis attempts to exploit these data for discovering new knowledge that can be translated into clinical applications. Biological sequence analysis compares, aligns, indexes and analyzes biological sequences and thus plays a critical role in bioinformatics.

Graph Mining

Graph mining has gained much attention in the last few decades as it is one of the novel approaches for mining the graph structures. Graph structures are found in biological pathways or networks, chemical compounds, protein structures, traffic flow, XML databases, Web and social networks. Graph mining aims to discover patterns from graphs that portray the original data

which can be used subsequently for classification or clustering [15]. Various data mining approaches are used to mine the graph-based data and to further carry out constructive analysis on it. It is used to find strongly connected groups in social networks and in several scientific domains like finding frequent molecular structures. A variety of social networks of unparalleled scales have emerged due to Web 2.0 and social media which present new problems for more effective graph mining techniques. Most networks demonstrate strong community structures and hence community detection which uncovers the group membership of actors in a network is a basic task in social network analysis using graph mining. Graph mining is applied in various problem domains including bioinformatics, program flow structures, computer networks, social networks etc.

1.2 MACHINE LEARNING

Machine learning attempts to make computers operate like human beings and to make them learn autonomously by providing data and information in the form of observations [16]. It is the practice of using algorithms to parse data, learn from it, and then make a prediction. The acquired knowledge from available data is in the form of structural description that can be represented in different ways allowing computers to generalize to new settings. It encompasses a wide variety of techniques used for the discovery of rules, patterns and relationships in sets of data. Machines that learn are useful to humans because, with all of their processing power, they are able to find patterns quickly in unseen data. Machine learning is a tool that can be used to enhance human ability to solve problems and make informed inferences on a wide range of problems, from helping diagnose diseases to coming up with solutions for global climate change.

The applications for machine learning include machine perception, computer vision, natural language processing, syntactic pattern recognition, search engines, medical diagnosis, bioinformatics, brain-machine interfaces, cheminformatics, fraud detection, stock market analysis, speech and handwriting recognition, object recognition in computer vision, gaming, software engineering and robot locomotion. Different types of machine learning are listed below.

Supervised learning: It generates a function that maps inputs to desired outputs.

*Unsupervised learning***:** It models a set of inputs and labeled examples are not available.

Semi-supervised learning: This method combines both examples to generate an appropriate function or classifier.

Reinforcement learning: It learns to act when provided with an observation of the world. To every action there is an influence in the environment, which provides response in the form of incentives that directs the learning algorithm.

Transduction: This type of learning attempts to predict outputs based on inputs and outputs from training and test inputs.

1.2.1 Supervised Learning

Supervised learning is an automatic learning which focuses on modeling input, output relationships [17]. The goal of supervised learning is to identify an optimal functional mapping between the input data X, to the output variable i.e., a class label Y such that Y = f(X). This is performed based on a sample of observations of the input variables X, which are the characteristics of the examples for a given problem.

Supervised learning aims to construct a classifier provided a set of classified training examples [18]. A pair consisting of an object and its associated class label is called as a labeled example. The training set consists of labeled examples given to the learning algorithm. The classifier is constructed based on training data supplied to the classification algorithm. The key challenge of the supervised learning algorithm is generalization i.e., the ability to predict the correct label on unseen data.

The performance of the classifier is evaluated by employing a different set of labeled examples called the test set. The reason for using a separate test set for evaluating the classifier is that the most learned classifiers can accurately predict the class label of the training examples, not the new examples. Hence it is more appropriate to use a different data set for testing the ability of the learned model to generalize new data points. The percentage of correctly classified test examples is called as the classification rate or prediction accuracy. The classification rate estimate is more accurate when the test dataset is larger.

Some of the standard supervised classification algorithms namely Support Vector Machines, Decision Tree Induction, Multi-layer Perceptron used in this research are presented below.

Decision Tree Induction

Decision tree is the most important representative of the family of symbolic machine learning techniques. Decision tree learning represents the procedure of learning decision trees from the set of labeled training examples. The output of a decision tree classification algorithm is a binary tree like structure in which each internal leaf node signify a test on an attribute [19]. The branch in the tree indicates an outcome of the test and each leaf node contains a class label. The uppermost node in a decision tree is the root node. The feature values of the new instance are tested against the decision tree and its class label is predicted. The class label of that instance is obtained by traversing a path from the root to a leaf node. Decision trees can be effortlessly transformed into classification rules. Decision tree induction algorithm is given below.

- The tree starts with a single node N, representing the training instances in D
- If all the instances in D belong to the same class, then node N becomes a leaf and is labelled with that class and the procedure is terminated. Otherwise an attribute 'A' is selected using attribute selection measure based on the splitting criterion and the node N is labelled with the splitting condition. A branch is grown from the node N for each of the decisions of the test condition
- The instances in D are partitioned accordingly
- Apply the algorithm recursively to each of the subsets D_i of D to form a decision tree such that the partitions are as pure as possible.
- The recursive partitioning stops when one of the following conditions is satisfied.
    i. All the instances in a partition D_i belong to the same class
    ii. There are no remaining attributes on which the node N may be further partitioned. Convert the node N into a leaf node and label it with most common class of the instances in D_i.
    iii. A partition D_i is empty. A leaf node is created with the majority class in D_i.

An attribute selection measure is used for selecting the splitting criterion that splits a given data set D, of labeled training instances into individual classes. If D is divided into smaller partitions based on the outcomes of the splitting condition, then normally each partition would be pure i.e., all the instances in a given partition would belong to the same class. Attribute selection measures are also known as splitting rules because it determines the manner in which a given node is split.

The attribute selection measure is used to give a ranking for every attribute. The attribute having the highest rank is chosen as the splitting attribute for the given training set. If the splitting attribute is continuous-valued, then the split point must be determined as a part of the splitting process. The node N created for a partition D is labeled with the splitting criterion, branches fan out for each outcome of the condition and the instances are partitioned accordingly. There are three possible cases. Let A be the splitting attribute and A has v distinct values a_1, a_2, …a_v.

- If A is discrete valued, branches are created for each value a_j of A at node N and labeled with that value. Partition D_j corresponds to the subset of labeled instances in D having value a_j for A. If all the instances in a given partition have the same value for A, then the attribute A need not be considered again for partitioning of the instances.
- If A is continuous-valued, the test at node N has two possible outcomes corresponding to the condition A < = split-point and A > split-point. Normally, the split-point, 'a', is taken as the midpoint of two known adjacent values of A and therefore it may not be a pre-existing value of A from the training data. Two branches are grown from N one for each condition A <= split-point, A> split-point and the resultant nodes are labeled accordingly. The instances are partitioned such that D1 holds the partition of instances in D for which A<=split-point, whereas D2 holds the remaining.
- If A is discrete and binary valued, two branches are grown from N. The left branch of N is labeled as 'yes' such that D1 corresponds to the partition of instances in D that satisfy the test. The right branch of N is labeled 'no' so that the partition D2 contains instances of D that does not satisfy the test.

Decision tree algorithm does not require making any assumption on the linearity in the data and hence can be used in circumstances where the parameters are non-linearly related. Decision trees unreservedly execute feature selection which is very essential in predictive analytics. During the process of fitting a decision tree to a training dataset, the decision tree is split based on the nodes at the top which are considered as important variables in a given dataset. Decision trees are not affected by missing values and outliers and hence help in saving data preparation time.

Support Vector Machine (SVM)

Support Vector Machine is a supervised machine learning algorithm that has achieved recognition than any other machine learning methods. It works by finding a hyperplane that separates the training data set into classes. Support vector machines are a set supervised learning methods used for classification and regression. It can handle noisy and large datasets. SVMs are not influenced by local minima and do not experience the problem of dimensionality [20]. The position of the dividing hyper plane is decided by support vectors that are the decisive elements of the training set. Different linear hyperplanes can be formed but SVM algorithm attempts to maximize the distance between the various classes. The optimal line that can divide the two classes is the line with the largest margin and is known as the Maximal-Margin hyperplane. The margin is considered as the perpendicular distance from the line to the nearest points.

SVMs are classified into two categories, linear SVM in which the training data are separated by a hyperplane and non-Linear SVM in which it is not possible to separate the training data using a single hyperplane. Given some data points each belonging to one of two classes, the goal is to decide which class a new data point will be in. In the case of support vector machines, a data point is viewed as a p-dimensional vector of a list of p numbers, and one wants to know whether such points can be separated with a p - 1 dimensional hyper plane. This is called a linear classifier. Even though there are many hyper planes that classify the data, the maximum separation of margin between the two classes is typically preferred. So the hyper plane is chosen in such a way that the distance from it to the nearest data point on each side is maximized. If such a hyper plane exists, it is clearly of interest and is known as the maximum-margin hyper plane and such a linear classifier is known as a maximum margin classifier.

Mathematical Formulation of Support Vector Machine

Learning machine algorithms are implementations of statistical inference principles. Typically, the machine is presented with a set of training examples, (x_i, y_i) where the x_i is the real world data instances and the y_i are the labels indicating which class the instance belongs to. For the two class pattern recognition problem,the class labels are either $y_i = +1$ or $y_i = -1$. A training example (x_i, y_i) is named positive when $y_i = +1$ and negative if $y_i = -1$.

From the geometric point of view, the support vector machine constructs an optimal hyper plane given by $w^T x - \gamma = 0$ between two classes of examples as shown in Fig.1.2. The problem of selecting optimal hyper plane from the multiple possible hyper planes is an ill-posed one. The free parameters are a vector of weights w which is orthogonal to the hyper plane W and a threshold value γ.

Each training example is related to a separating hyper plane by a quantity called margin. The functional margin is given by $y_i (w^T x - \gamma)$. If the functional margin is greater than zero, then the data point is correctly classified by the hyper plane. When the weight vector w of the functional margin is normalized ($w=w/||w||$), the geometric margin is obtained, which measures the distance of data points from the hyper plane in Euclidean space. The expression margin of a training set is used to refer to the maximum geometric margin over all possible hyper planes. The hyper plane defined by the maximum geometric margin is unique and it is known as a maximal margin hyper plane. The expected generalization error is minimized when the classes are separated with a large margin. One of the simplest models of SVM based maximal margin is Maximal Margin classifier.

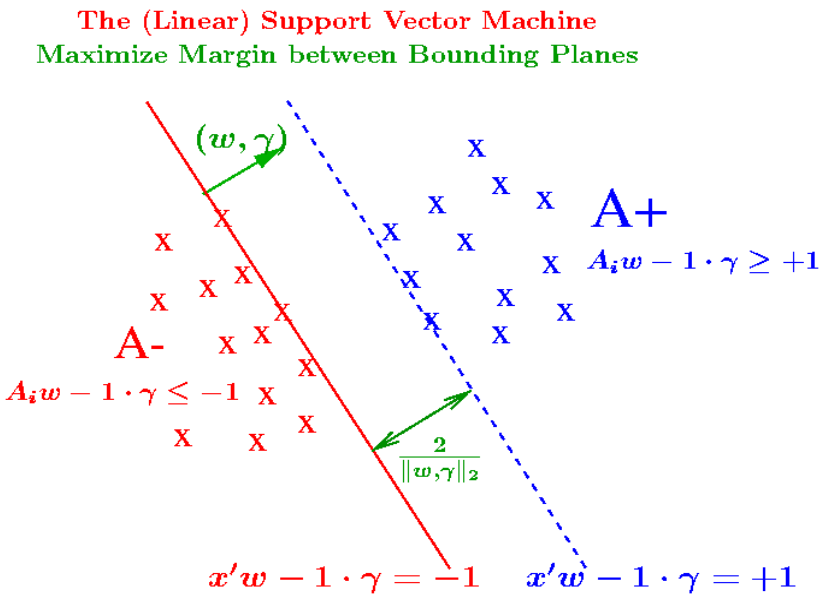

Fig. 1.2 Linear Support Vector Machine

The two planes parallel to the hyper plane which passes through one or more points called bounding hyper planes are given as in equation 1.1.

$w^T x - \gamma = 1$ (1.1)

where w is weight vector realizing functional margin 1 on the positive point X+ and on the negative point X-.

The margin between the optimal hyper plane and the bounding plane is 1/||w||, and so the distance between the bounding hyper planes is 2/||w||.

Distance of the bounding plane $w^T x - \gamma = 1$ from the origin is $|-\gamma + 1|/||w||$ and the distance of the bounding plane $w^T x - \gamma = -1$ from the origin is $|- \gamma - 1|/||w||$. Support vectors are the points falling on the bounding planes and they play decisive role in the theory. The data points x belonging to two classes A+ and A- are classified based on the conditions given in equations 1.2 and 1.3.

$w^T x_i - \gamma \geq 1$ for all $x_i \in A^+$ (1.2)

$w^T x_i - \gamma \leq 1$ for all $x_i \in A^-$ (1.3)

These inequality constraints can be combined to give $D_{ii}(w^T x_i - \gamma) \geq 1$ for all x_i where

$D_{ii} = 1$ for A+ and $D_{ii} = -1$ for A-.

The learning problem is hence to find an optimal hyper plane which separates A+ from A- by maximizing the distance between the bounding hyper planes. A unique property of SVMs is that it minimizes the empirical classification error and also reduces the geometric margin and hence it is called as maximum margin classifier [21]. SVM constructs a separating hyper plane to maximize the margin between the two data sets in an n-dimensional space. To calculate the margin, two parallel hyper planes are constructed one on each side of the separating one, which are pushed up against the two data sets. The hyper plane that has the largest distance to the neighboring data points of both classes achieves a good separation. The larger the margin or distance between these parallel hyper planes, the better the generalization error of the classifier will be. SVM offers best classification performance on the training data and renders more efficiency for correct classification of the future data. SVM does not make any strong assumptions on data and does not over-fit the data.

Naive Bayes Classifier

The Naive Bayes classifier (NB) is a simple but effective classifier, which has been used in numerous applications of information processing including, natural language processing, information retrieval, etc. This principle behind this technique is the Bayesian theorem and is more suitable when the dimensionality of the inputs is high. Naive Bayes classifier assumes that the effect of a variable value on a given class is independent of the values of other variable. The Naive Bayes inducers compute conditional probabilities of the classes given the instance and pick the class with the highest posterior. Depending on the specific nature of the probability model, Naive Bayes classifiers can be trained very efficiently in a supervised learning setting.

A Naive Bayes classifier is based on Bayes theorem with strong independence assumptions. Given classes w_j and dataset x, general formulation is given by equations1.4 and 1.5.

$P(w_j \mid x) = p(x \mid w_j) P(w_j) / p(x)$ where (1.4)

$P(x) = \sum p(x \mid w_j) P(w_j)$ (1.5)

Bayesian classification model is given in equation 1.6.

Posterior = likelihood * prior / evidence (1.6)

where prior probability reflects knowledge of the relative frequency of instances of a class, likelihood is a measure of the probability that a measurement value occurs in a class and evidence is a scaling term. The classification of unseen data x is performed by calculating P (C_i / x) for each class and assigning x to class i, if P (C_i / x) > P (C_j / x) for all i not equal to j.

Naive Bayes Classifier algorithm achieves fine with categorical input variables. It converges faster and needs comparatively less training data than other discriminative models [22]. It is easier to predict class of the test data set with Naive Bayes Classifier algorithm. It is a good bet for multi class predictions as well. Though it requires conditional independence assumption, Naive Bayes Classifier has presented good performance in various application domains like document classification, disease prediction.

Artificial Neural Network

An Artificial Neural Network (ANN) is a mathematical model motivated by neural networks in brain. A neural network includes an interconnected group of artificial neurons, which process information using network connections [23]. Neural networks are used to model intricate relationships between inputs and outputs and to recognize patterns in data. Multi Layer Perceptron (MLP) consists of one or more layers between input and output layer and is a feed forward neural network. Feed forward signifies that data flows in one direction from input to output layer. This network is trained through back propagation learning algorithm. MLPs are used for pattern classification, prediction and approximation as it is capable of solving problems which are non-linearly separable. Similarly an N - layer neural networks is trained with the same ideas of single layer networks.

An MLP includes a network of neurons called perceptrons that computes a single output from multiple real-valued inputs by forming a linear combination according to its input weights and then provides the output through a nonlinear activation function. MLP networks are usually used in supervised learning problems with the back-propagation algorithm. The algorithm consists of two phases namely forward pass and backward pass. During the forward pass, the predicted outputs matching the given inputs are evaluated. The partial derivatives of the cost function with respect to the various parameters are propagated back through the network in the backward pass. The architecture of an ANN is depicted in Fig.1.3.

A multi-layer perceptron is especially useful for approximating a classification function that maps input vector (x_1, x_2 ... x_n) to one or more classes c_1,c_2,...c_m. The output cost function is minimized by adjusting the network weights [24]. It is only the output of the final layer that emerges in the output error function. The earlier layers of weight decide the final layer output and the learning algorithm will adjust all of them. The learning algorithm repeatedly adjusts the outputs of the earlier layers so that they form appropriate intermediary depictions.

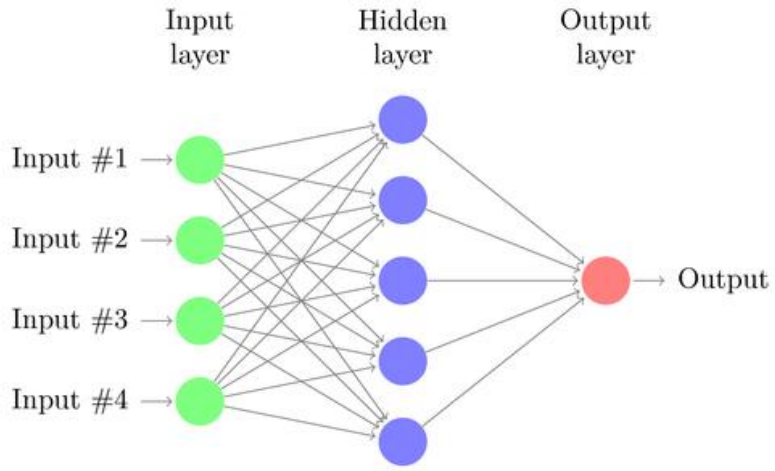

Fig. 1.3 A Typical ANN Architecture

The network represents a broad variety of classification functions by optimizing weights and thresholds for all nodes. Supervised learning helps to optimize the weights and the network gains knowledge from a large number of examples. The network computes the actual vector for every instance and compares it with the original output. Subsequently adjustment of weights and thresholds is done, relative to their role in the error made at the corresponding output. The popular method used is the backpropagation in which the errors are broadcasted into the lower layers in a continuous manner, and are utilized for the alteration of weights.

1.2.2 Deep Learning

Deep learning is a latest development in machine learning that achieves great power and flexibility by learning to symbolize the world as nested hierarchy of concepts. Each concept is defined relative to simple concepts whereas more conceptual depictions are computed using less abstract ones. Deep architectures are compositions of many layers of adaptive non-linear components, in other words, they are cascades of parameterized non-linear modules that contain

trainable parameters at all levels. Deep architectures allow the representation of wide families of functions in a more compact form than shallow architectures, because they can trade space for time. The lower layers produce features symbolizing low level abstractions that are merged in the subsequent layers to form high level features that correspond to higher level abstractions [25].

Over the last few years, deep learning has turned out to be successful in discovering intricate structures in high-dimensional data and has obtained remarkable performances for object detection in images, speech recognition, natural language understanding and translation. They enable the discovery of high-level features, improving performances over traditional models, increasing interpretability and providing additional understanding about the structure of the data. Record-setting results on many important clinical applications have been demonstrated thus initiating the way toward a potential new generation of intelligent tool based deep learning for real-world medical care.

Deep learning techniques are a form of representation learning that uses multiple transformation steps to generate very complex features. Fig.1.4 compares the architecture of conventional ANN and Deep Neural Networks (DNN). ANNs are limited to three layers and are trained to acquire supervised representations that are optimized for the precise task and are not comprehensive. But every layer of a DNN produces a representation of the observed patterns based on the data it receives as inputs from the layer below, by optimizing a local unsupervised criterion. Traditional techniques consist of a solitary linear transformation of the input space and are restricted in their ability to process natural data in their unprocessed form.

Deep learning is different from traditional machine learning in the manner representations are learned from the raw data. Deep learning is a technique that allows neural network based models with several processing layers to study representations of data with multiple levels of abstraction. Deep learning in high-throughput biology is used to capture the internal structure of increasingly larger and high-dimensional data sets like DNA sequencing and RNA measurements. Deep neural networks process the inputs in a layer-wise nonlinear manner to pre-train the nodes in subsequent hidden layers to learn deep structures and representations that are generalizable [26]. A supervised layer is provided with these representations as input and the entire network is adjusted using the backpropagation algorithm for representations optimized

for the specific task. Deep learning has shown spectacular success in computer vision and speech recognition. A detailed description about deep neural network is presented in Section 2.1.

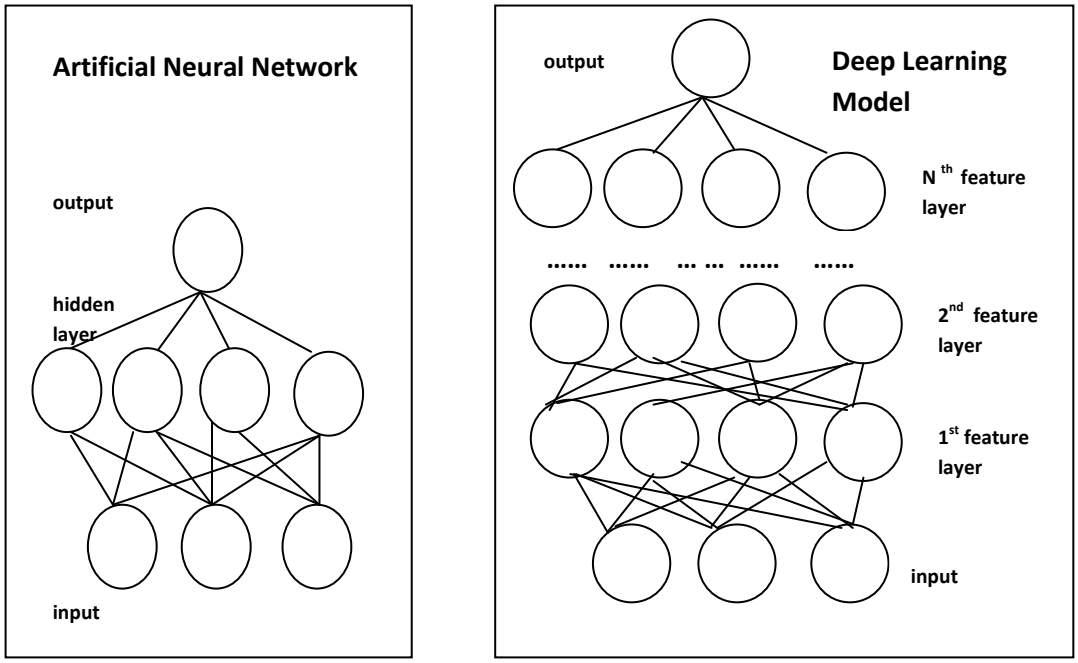

Fig. 1.4 Comparison Between Artificial Neural Network and Deep Learning Model

Advantages of Deep Learning

Deep learning has excellent performance on problems that significantly surpass other results in various fields, including speech, language, vision, gaming etc. It is an architecture that can be adapted to new problems relatively easily. Some of the important benefits of deep learning over traditional methods are stated below.

Feature Engineering Not Required: Feature engineering is the process of extracting features from raw data to better describe the underlying problem. It is a primary task in machine learning as it enhances the accuracy of the model. This procedure necessitates knowledge of the domain in which the problem is formulated. One of prime advantages of deep learning over other machine learning algorithms is its capacity to execute feature engineering on its own. Careful engineering and domain expertise is required to build a machine learning system that alter the raw data into an appropriate internal representation which enables the classifier to detect patterns in the data set. A deep learning algorithm will examine the data to look for features that associate and combine to enable faster learning without being explicitly programmed to do so. This ability

means that data scientists can save a great deal of time. The neural networks present in a deep learning algorithm can uncover new, complex features that humans can miss. The advantage of deep learning algorithms is that they try to learn high-level features from data in an incremental manner.

Best Results with Unstructured Data: Deep learning models perform well with unstructured data like images, audio, text and derive insights that are relevant to the purpose of its training. It is not possible for humans to capture the essence of a multi dimensional large dataset like images. The ability to process considerably more number of features makes deep learning very dominant when dealing with unstructured data. For example, a deep learning algorithm can unearth existing relation between pictures, social media chatter, industry analysis, weather forecast and can be used to predict future stock prices of a given company.

Data Labeling Not Needed: One of the major difficulties in machine learning is acquiring fair quality of training data because data labeling can be a complex and costly job which consumes more time. It requires the judgement of highly skilled domain experts and so getting good quality training data can be very expensive for particular industries. Deep learning stands out at learning without teaching and hence it supersedes the need for well-labeled data. In a deep network, every node gains knowledge about the features by repeatedly attempting to rebuild the input from which it takes its samples during training. This minimizes the gap between the forecasted and the probability distribution of the input data itself. In this way, these neural networks discover the relationship between significant features and optimal results. The correlation between feature signals and feature representations are represented well by them.

High-Quality Results: Deep learning methods include multi layer processing with less time and better accuracy performance. Deep learning does perform better because it mimics the brain functions with multiple layers of neural networks stacked one after another like the classical brain model. As the network grows deeper, abstract representation of data are created by deep learning techniques and thus the model automatically extracts features and gives higher accuracy results. The quality of its work never reduces, unless the training data comprises raw data that does not symbolize the problem.

Applications of Deep Learning

Over the last few years deep learning has been applied to varied problems in natural language processing and has shown remarkable success. The mounting research and development activities expands the application of deep learning systems in defense, aerospace, healthcare, telecommunication, information technology, retail, banking and financial sector, automotive, etc. Some of the deep learning applications are listed below.

Visual Recognition: Neural networks have significantly improved computer vision applications. Image processing is being used for object recognition and video processing is being used to automate scene classification or people recognition. Deep Learning enables to sort images based on locations detected in photographs, faces, a combination of people, or according to events, dates, etc. Searching for a particular photo from a library requires state-of-the-art visual recognition systems consisting several layers to recognize elements. Large-scale image visual recognition through deep neural networks is boosting growth in this segment of digital media management by using convolutional neural networks, Tensorflow, and Python extensively.

Self-Driving Cars: Deep Learning is the power that is bringing autonomous driving to life. The system is fed with enormous data in order to build a model, trained and then tested in a safe environment. The Uber AI Labs is working on making driverless cars by integrating several smart features such as food delivery options with the use of driverless cars. Classy models are created to find the way through traffic, identify paths, signage, pedestrian routes, traffic volume and road blockages with data from cameras, sensors, geographical mapping.

Natural Language Processing (NLP): One of the hardest tasks for humans to learn is to understand the complexities associated with language whether it is syntax, semantics, tonal nuances, expressions, or even sarcasm. NLP through deep learning is trying to achieve this by training machines to catch linguistic nuances and frame appropriate responses. Document summarization is extensively being used in the legal domain. Deep learning is applied in NLP tasks like answering questions, language modeling, text classification, social media analysis and sentiment analysis. Previously logistic regression or SVM were used to build time-consuming complex models but now distributed representations, convolutional neural networks, recurrent

and recursive neural networks, reinforcement learning, and memory augmenting strategies are helping attain better maturity in NLP.

News Aggregation and Fraud News Detection: Extensive use of deep learning in news aggregation is strengthening efforts to customize news as per readers. Newer levels of sophistication to define reader personality are being met to filter out news as per geographical, social, economical parameters along with the individual preferences of a reader. Fraud news detection is an important asset in the current scenario where the internet has become the prime source of all genuine and fake information. It becomes tremendously hard to distinguish fake news as bots replicate it across channels automatically. Deep learning helps develop classifiers that can detect fake or biased news and remove it from the feed and warn possible privacy breaches. Training and validating a deep learning neural network for news detection is really tough as the data is overwhelmed with opinions and nobody can decide if the news is biased or unbiased.

Healthcare: Some of the deep learning projects picking up speed in the healthcare domain are helping early, accurate and speedy diagnosis of life-threatening diseases, augmented clinicians addressing the shortage of quality physicians and healthcare providers, pathology results and treatment course standardization, understanding genetics to predict future risk of diseases and negative health episodes. Readmission is a huge problem for the healthcare sector as it costs tens of millions of dollars in cost. Health risks associated with readmissions are mitigated by healthcare giants by using deep learning while bringing down the costs. It is also being used in clinical researches to treat fatal diseases but physician's uncertainty and lack of a huge dataset are the challenges to the use of deep learning in medicine.

Virtual *Assistants:* The trendiest application of deep learning is virtual assistants ranging from Alexa to Siri to Google Assistant. These assistants learn more about user voice and accent thus providing the user a resultant human interaction experience. Virtual assistants apply deep learning to know more about user preferences like visited spots or favorite songs. They evaluate natural human language to learn and understand commands to execute them. They are capable of translating speech to text, making notes for user and booking appointments. Virtual assistants can aid in creating email copy with deep learning applications such as text generation and document summarization.

Entertainment: Deep learning is used to analyse player emotions and expressions through hundreds of hours of footage to auto-generate highlights for telecast thus saving time and cost. Amazon provides a personalized experience to its viewers by learning their show preferences, time of access, history by enhancing their deep learning capabilities to recommend shows that are liked by a particular viewer. Deep video analysis can save time of manual effort required for audio or video synchronization and its testing and tagging. Content editing and auto-content creation are now realism due to deep learning and its contribution in face and pattern recognition. Deep learning is revolutionizing the filmmaking process as cameras learn to study human body language to imbibe in virtual characters.

Fraud Detection: Another domain gaining from deep learning is the banking and financial sector that is inundated with the task of detecting fraud as money transactions are going digital. In order to detect credit card frauds and to save huge cost in financial institutions autoencoders in Keras and Tensorflow are being developed. This is done on the basis of recognizing patterns in customer transactions, discovering anomalous behavior and outliers. Machine learning is mostly used for emphasizing cases of fraud and entails human intervention, whereas deep learning tries to minimize these efforts.

Personalizations: Chatbots are used in all platforms to offer visitors with personalized experiences with a human touch. Deep Learning is playing a major role in providing flawless personalized experiences in providing recommendations, offers and identifying revenue opportunities. Customer experiences are personalized by robots that are specialized in precise tasks and contribute the most suitable services.

A bigger impact of deep learning is found in the business world. Deep learning algorithms are applied to customer data in Customer Relationship Management systems, social media and other online data to better segment clients, predict churn and detect fraud. The financial industry is now dependable on deep learning to deliver stock price predictions at the right time. In the healthcare industry, exploration of the possibility of using known and tested drugs to treat new diseases is done to shorten the time before the drugs are made available to the general public. Governmental institutions are relying on deep learning to get immediate insights into metric like food production and energy infrastructure.

Several deep learning architectures like Convolutional Neural Network (CNN), Restricted Boltzmann Machines (RBM), Deep Belief Networks (DBN), Recurrent Neural Networks (RNN) have been developed and are proven to be powerful. CNN is comprised of one or more convolutional layers and subsampling layers followed by one or more fully connected layers. CNNs are easier to train and have fewer parameters than fully connected networks with the same number of hidden units. RBM is an undirected graphical model with only one layer of hidden units with no connections between them. RBMs essentially perform a binary version of factor analysis. It is an algorithm which is useful for dimensionality reduction, classification, regression, collaborative filtering, feature learning, and topic modeling. DBNs are composed of layers of RBMs for the pre-train phase and then a feed-forward network for the fine-tune phase. The primary step in training DBN is to study a layer of features from the observable units. Then, the next step is to treat the activations of previously trained features as visible units and learn features in a second hidden layer. They are used for video sequence recognition and motion capture. RNNs are connectionist models that explore the dynamics of sequences via cycles in the network of nodes. The success of RNNs can be attributed to their ability to deal with sequential data, as opposed to ANNs which are known for not having any notion of time. A detailed description of the various deep learning architectures is presented in Chapter 2. In this research work, deep learning based predictive models for ASD is developed using RNN.

1.3 OVERVIEW OF AUTISM SPECTRUM DISORDER

Bioinformatics research is driven by the vast potential for new understanding that can lead to new treatments, drugs and general expansion of knowledge. It has yielded a variety of experimental data like DNA sequences, gene expression patterns, three-dimensional models of macromolecular structure, protein data etc. This has exceedingly motivated computational studies to identify the cause of genetic disorders, to evaluate potential treatments or prevention strategies based on those findings. In this research Autism Spectrum Disorder (ASD) has been taken for study as it is a complex disorder and the genetic ground of this comprehensive developmental disability is very difficult to research. ASD occurs in the first three years of life and is discovered to be a lifetime neurological condition. Autism affected individual exhibits differences in communication, social interactions and imagination which are demonstrated by

repetitive and restricted play activities. The genetics associated with ASD, the mutations and genes involved in ASD are discussed below.

Genetic Disorder

Each cell in the human body depends on thousands of proteins to do their jobs in the right places at the right times. Genes are sequences of DNA that code for proteins that are made up of many amino acids. Human genetic variations primarily result from nucleotide polymorphisms that occur in nucleotide bases in the overall human population. Nucleotides are the repeating units of a DNA sequence. There are four nucleotides, each with a different nitrogenous base: thymine (T), adenine (A), guanine (G), and cytosine (C). The order of these nitrogenous bases determines the sequences of amino acids in a protein. A DNA sequences is made up of codons which are trinucleotide combinations of A(Adenine), C(Cytosine), G(Guanine), T(Thymine). Some features of codons are listed below

- Most codons specify an amino acid
- Three stop codons TAA, TAG, TGA mark the end of protein coding region
- One start codon ATG, marks the beginning of a protein and also encodes the amino acid methionine

The first step in making a protein, called transcription, is to use the DNA sequence of a gene to make an RNA molecule. RNA is made up of nucleotides and nitrogenous bases just like DNA, except that thymine is replaced with uracil (U). This RNA molecule is used to assemble a chain of amino acids, or a protein, in a process called translation. In order to translate into proteins, genes are first transcribed and the ends modified, with attachment of a 5′ cap and 3′ polyadenylation sequences. Exons are then identified and joined together, with introns removed, giving rise to mature mRNAs. Mature mRNAs are then transported out of nuclei and are translated into proteins. Translation is a process in which the codons in an mRNA are read from a start codon till a stop codon [27]. The order of amino acids in a protein is given by mRNA codons which are read from 5' to 3'. A gene mutation can cause a protein to malfunction by changing a gene's instructions for making the protein. Protein plays a critical role in the body and when it is altered by a mutation normal development is disrupted. These mutations are responsible for causing illnesses and if the gene mutations are present in the egg or sperm cell,

children will inherit the defective gene from their parents. Defect in a single gene or a set of genes can cause diseases. Diseases are categorised into the following types as per the degree of gene mutation.

- Chromosomal diseases which are caused when the complete chromosome, or bulk segments of a chromosome, is missing, duplicated or distorted. An example of chromosomal abnormality is Down Syndrome.
- Single-gene disorders that arise when a modification occurs in a gene making one gene to stop working. Sickle-cell anaemia is an example of a single gene disorder.
- Multifactorial disorders caused by mutations in multiple genes, coupled with environmental causes. An example of a multifactorial disorder is diabetes.
- Mitochondrial disorders which are rare disorders caused due to mutations in non-chromosomal DNA located within the mitochondria. Any part of the body including the brain and the muscles can be found to be affected by this disorder.

Gene Mutation

A permanent alteration in the DNA sequence such that the sequence differs from what is found in most people is called a gene mutation. A mutation can affect a single DNA building block to a large segment of a chromosome that includes multiple genes. Gene mutations can be categorized in two major groups:

- Mutations that are inherited from a parent and are present throughout a person's life in almost every cell in the body are called hereditary. These mutations also are known as germline mutations as they are present within the parent's egg or spermatozoan cells, which are also called germ cells. When associate egg and a spermatozoan unite, the resulting fertilized egg cell receives DNA from both parents. If this DNA encompasses a mutation, the child that grows from the fertilized egg will have the mutation in each of his or her cells.
- Acquired or somatic mutations arise at any time during a person's lifetime and are present only in some cells, not in every cell in the body. This change is caused by environmental factors like UV from the sun, or can occur if an error is made as DNA copies itself during cell division. Acquired mutations in bodily cells cannot be passed to consequent generation.

Genetic changes that are known as de novo mutations can be either hereditary or somatic [28]. In certain cases, the mutation occurs in a person's egg or sperm cell but is not present in any of the person's other cells. In different cases, the mutation occurs in the fertilized egg shortly after the egg and sperm cells unite. When the fertilized egg divides, every ensuing cell within the growing embryo can have the mutation. When an affected child has a mutation in every cell of the body and there is no family history of the disorder, de novo mutations may be the cause for such disorders.

Point Mutations

Point mutations are single base changes in the DNA sequence. They can be further categorized into the following three types

- Mutations that cause a single amino acid change within the protein are missense mutations.
- Mutations that create a premature stop codon, causing the protein to be shortened are nonsense mutations.
- Mutations that do not cause amino acid changes are called silent.

Missense Mutation: It occurs when base substitution results in the generation of a codon that specifies a different amino acid and hence leads to a different polypeptide sequence. Depending on the type of amino acid substitution the missense mutation is either conservative or nonconservative. For example if the structure and properties of the substituted amino acid are very similar to the original amino acid the mutation is said to be conservative and will most likely have little effect on the resultant proteins structure or function. If the substitution leads to an amino acid with very different structure and properties the mutation is nonconservative and will probably be deleterious for the resultant proteins structure or function.

For example consider the DNA sequence

Normal Sequence	:	AUG	GCC	**T**GC	AAA	CGC	TGG
Amino acid	:	met	ala	cys	lys	arg	trp
Missense mutation	:	AUG	GCC	**G**GC	AAA	CGC	TGG
Amino acid	:	met	ala	arg	lys	arg	trp

In the above sequence, a single nucleotide change is encountered from T to G and a change has occurred in the amino acid sequence.

Nonsense Mutation: When a base substitution results in a stop codon ultimately truncating translation and most likely leading to a nonfunctional protein, nonsense mutation occurs.

For example consider the DNA sequence

Normal Sequence	:	AUG	GCC	TG**C**	AAA	CGC	TGG
Amino acid	:	met	ala	cys	lys	arg	trp
Nonsense mutation	:	AUG	GCC	TG**A**	AAA	CGC	TGG
Amino acid	:	met	ala	---	---	---	---

In the above sequence, a single nucleotide change is encountered from C to A and introduced a stop codon thus ending the translation.

Silent Mutation: If abase substitution occurs in the third position of the codon there is a good chance that a synonymous codon will be generated. Thus the amino acid sequence encoded by the gene is not changed and the mutation is said to be silent.

For example consider the DNA sequence

Normal Sequence	:	AUG	GC**C**	TGC	AAA	CGC	TGG
Amino acid	:	met	ala	cys	lys	arg	trp
Silent mutation	:	AUG	GC**T**	TGC	AAA	CGC	TGG
Amino acid	:	met	ala	cys	lys	arg	trp

In the above sequence, a single nucleotide change is encountered from C to T but no change has occurred in the amino acid sequence.

Frameshift Mutations

Addition or removal of one or more DNA bases lead to insertion mutations and deletion mutations. When they happen in multiples of three it causes shift in the reading frame of a gene, changing the grouping of bases into codons. These changes can significantly change a protein's amino acid sequence. The insertion of additional base pairs may lead to frameshifts depending on whether or not multiples of three base pairs are inserted. If one or two bases are deleted the translational frame is altered resulting in a garbled message and nonfunctional product. A deletion of three or more bases leaves the reading frame intact. A deletion of one or more codons results in a protein missing one or more amino acids.

For example consider the DNA sequence

Normal Sequence	:	AUG	GCC	TGC	AAA	CGC	TGG
Amino acid	:	met	ala	cys	lys	arg	trp
Insertion	:	AUG	GCT	**C** TGC	AAA	CGC	TGG
Sequence	:	AUG	GCT	CTG	CAA	ACG	CTG
Amino acid	:	met	ala	leu	gln	thr	leu
Deletion	:	AUG	GC-	TGC	AAA	CGC	TGG
Sequence	:	AUG	GCT	GCA	AAC	GCT	
Amino acid	:	met	ala	glu	asn	ala	

In the above sequence, insertion and deletion of a single nucleotide has altered the reading frames, changed the groupings of codons and hence altered the amino acid sequence.

Autism Spectrum Disorder

ASD is characterized by atypical social behavior, weak communication and typecast behavior. The signs of ASD usually emerge during early days and person suffering from ASD is unable to communicate and interact with others. ASD is defined by a range of conditions and behaviors that affect persons to varied degrees. Some of the indications associated with autism include deferred learning of language, complicatedness in making eye contact or making a conversation, monotonous behaviors and limited interests or activities, difficulty with executive functioning, deprived motor skills and sensory sensitivities [29].

People affected by ASD are found to have irregularities in several regions of the brain with abnormal levels of serotonin or other neurotransmitters. ASD typically results from the interruption of normal brain development in premature fetal development caused by flaws in genes that manage brain growth and neuron communication. Each person with ASD has a distinct set of strengths and challenges. The ways in which ASD affected person learn, think and solve problem can range from highly skilled to severely challenged. People with ASD may require varied levels of support from the society depending on the severity of the disorder. The prevalence of ASD had risen to 1 in every 59 births in the United States and almost 1 in 54 boys. The probability of boys affected by ASD is about four times more than girls. Studies in Asia, Europe, and North America have identified individuals affected by ASD with an average prevalence between1% to 2%.

Symptoms of ASD

Autism is identified as a spectrum disorder because there is wide variation in the type and severity of symptoms people experience. ASD indications are usually seen during the first two to three years of life whereas some children show signs from birth. Others seem to develop normally at first, but suddenly show symptoms when they are 18 to 36 months old. Symptoms of a communication disorder may not appear in certain individuals until demands of the environment go beyond their capabilities.

Children with ASD have difficulty in communicating. They have problem in understanding thoughts and feelings of others. Hence it is not easy for them to express themselves through words or gestures, facial expressions and touch. A child with ASD may be disturbed or affected by sounds, touches, smells, or sights which are normal to others. These children may have monotonous, stereotyped body movements like rocking, pacing or hand flapping [30]. They exhibit abnormal responses to people, affection to objects, confrontation to change in their routines or aggressive behavior. They may not notice people, objects, or actions in their surroundings and a few children with autism may also develop seizures. On the other side they may have extraordinarily developed skills in additional areas like drawing, music, crack problems or remembering facts.

Some of the diagnosing criterion for ASD from the Diagnostic and Statistical Manual of Mental Disorders version 5 is given below [31].

A. Constant deficits in social communication and interaction across multiple circumstances, as evident by the following

- Discrepancy in social-emotional reciprocity, nonverbal communicative behaviors used for social interaction, ranging from poorly integrated verbal and nonverbal communication to abnormalities in eye contact and body language or problems in understanding and use of gestures; deficient in facial expressions and nonverbal communication.
- Issues in developing, maintaining and appreciating relationships, ranging from complexities in adjusting behavior to match different social contexts; to problems in sharing imaginative play or in making friends; to lack of interest in peers.

B. Constrained, recurring patterns of behavior, interests, or activities, as manifested by at least two of the following

- Typecast or monotonous motor movements, speech, persistence on similarity, ritualized prototypes, rigid adherence to routines, or e.g., extreme sorrow at small changes, issues with transitions, inflexible thinking patterns, greeting rituals, wanting to take same route or eat food every day.
- Highly constrained, obsessed interests that are unusual in intensity or focus. For example strong attachment to or preoccupation with strange objects, extremely limited interest
- Overexcited or very less reactivity to sensory input or unusual interests in sensory facets of the environment, unresponsiveness to pain / temperature, unpleasant response to specific sounds or textures, undue smelling or touching of objects, visual attraction to lights or movement

Possible signs of ASD in babies and toddlers

- Lack of smiling while socially interacting with people, lack of happy expressions, lack of eye contact, unable to begin non verbal communications, no babbling or attempts to form baby-talk and words for communicating with others, not responding when their name is called out, no attempts to begin verbal communication with actual words or phrases are some of the signs of ASD in babies upto 2 years.

Risk Factors

Genetic and environmental factors are the primary reasons for ASD [32]. High age of the mother or the father augments the likelihood of an autistic child. The child of a pregnant woman who is exposed to certain drugs is more probable to be autistic. In some cases, autism has been linked to untreated phenylketonuria (called PKU, a disorder caused by the absence of an enzyme) and rubella. Some of the risk factors of ASD are listed below.

- Studies have shown that among identical twins, if one child has ASD, then the other will be affected about 36-95% of the time. In non-identical twins, if one child has ASD, then the other is affected about 0-31% of the time.
- The likelihood of parents with an ASD affected child having a second child also affected with the same disorder is 2%–18%

- ASD is likely to occur more frequently in people who are affected by certain genetic or chromosomal conditions. About 10% of kids with ASD are also recognized to have Down syndrome**,** tuberous sclerosis*,* fragile X syndrome or other genetic disorders.
- Almost 44% of children recognized to have ASD possess average to above average intellectual ability.
- Children born prematurely or with low birth weight are at risk for having ASD.
- ASD is more likely to occur together with other neurologic, developmental, psychiatric, chromosomal and genetic diagnoses. There is 83% co-occurrence of one or more non-ASD developmental diagnoses and 10% for psychiatric diagnoses.

Types of ASD

Autism is categorized into idiopathic or asyndromic caused by unknown factors which occurs in majority of cases and syndromic in which a chromosome irregularity, single-gene disorder or environmental agent can be recognized. About 85 percent people have idiopathic ASD and 15 percent of individuals with autism can be diagnosed with syndromic ASD.

Idiopathic or Asyndromic ASD

Among the causes of idiopathic ASD are very rare genetic disorders or prenatal exposures. In a majority of cases, the causes are

- A child is born to parents who are not autistic
- The familial history did not have autism
- The child was not premature
- The parents were under 35 years old
- Genetic anomalies like Rett syndrome that might originate autism in the child were not uncovered by tests
- When the mother was pregnant she was not exposed to any of the drugs known to add to the risk of autism

Some of the genes and gene functions associated with idiopathic ASD are listed in Table I.

Syndromic ASD

A disorder with a clinically defined pattern of somatic abnormalities and a neurobehavioral phenotype that may comprise ASD is called as syndromic ASD [33]. Syndromic ASD represents a group of childhood neurological conditions, typically associated with chromosomal abnormalities or mutations in a single gene. Identification of syndromic autism is done using targeted genetic testing. Table II shows the common genetic syndromes and genes associated.

Table I Genes and Gene Functions Associated with Idiopathic ASD

S.No	Gene	Protein	Gene function
1	NLGN3	Neuroligin-3 precursor	Neuroligins function as ligands for the neurexin family of cell-surface receptors
2	NLGN4	Neuroligin-4, X-linked precursor	Involved in the formation and remodeling of central nervous system synapses
3	CHD8	Chromodomain Helicase DNA Binding Protein 8	This gene functions in processes like transcriptional regulation, epigenetic remodeling and regulation of RNA synthesis.
4	NRXN1	Neurexin-1α precursor	Neurexins functionin the vertebrate nervous system as cell adhesion molecules and receptors
5	FOXP2	Forkhead Box P2	This gene is required for proper development of speech and language regions of the brain and is involved in a variety of biological pathways that control language development
6	HOXA1	Homeobox protein Hox-A1	Regulates multiple developmental processes including brainstem, cardiovascular development and morphogenesis as well as cognition and behavior
7	PTEN	Phosphatidylinositol 3,4,5-trisphosphate 3-phosphatase	A tumour-suppressor gene influencing G1 cell cycle arrest and apoptosis
8	CNTNAP2	Contactin Associated Protein Like 2	This gene is a member of the neurexin family acting in the vertebrate nervous system as cell bonding receptor
10	GABRB3	Gamma-Aminobutyric Acid Type A Receptor Beta3 Subunit	This gene serves as the receptor for gamma-aminobutyric acid, a major inhibitory neurotransmitter of the mammalian nervous system.

Table II Common Genetic Syndromes and Genes Associated

S. No	Syndrome (inheritance)	Gene symbol	Protein name	Gene function
1	Fragile X syndrome	FMR1	FMRP	FMRP is a nucleocytoplasmic shuttling protein that binds several mRNAs and associates with translating ribosomes
2	Rett syndrome	MECP2	Methyl-CpG-binding protein 2 (MeCP2)	Mediates transcriptional silencing and epigenetic regulation of methylated DNA
3	Angelman syndrome	UBE3A	Ubiquitin–protein ligase E3A	Involved in the ubiquitination pathway, which targets selected proteins for degradation
4	Neurofibromatosis syndrome type 1	NF1	Neurofibromin	It appears to activate ras GTPase, thereby controlling cellular proliferation and acting as a tumour suppressor
5	Sotos syndrome	NSD1	Histone-lysine N-methyltransferase, H3 lysine-36 and H4 lysine-20	This protein may act as a nucleus-localized, transcriptional factor and also as a bifunctional transcriptional regulator
6	Timothy syndrome	CACNA1C	Voltage-dependent L-type calcium channel subunit alpha-1C	Mediate the entry of calcium ions into excitable cells and are also involved in a variety of calcium-dependent processes, including muscle contraction, hormone release, gene expression, cell division, and cell death
7	Tuberous sclerosis complex	TSC1	Hamartin	Implicated as a tumour suppressor
8	Williams syndrome	ELN	Elastin	This gene encodes a protein that is one of the two components of elastic fibres
9	Cohen syndrome	VPS13B (COH1)	Vacuolar protein sorting 13B	This gene encodes a potential transmembrane protein that may function in vesicle-mediated transport and sorting of proteins within the cell

10	Phelan-McDermid syndrome	SHANK3	SH3 and Multiple Ankyrin Repeat Domains 3	Connects neurotransmitter receptors, ion channels, and other membrane proteins to the actin cytoskeleton signaling pathways. Shank proteins also play a role in synapse formation and dendritic spine maturation.
11	Joubert syndrome	AHI1	Jouberin	Involved in signal transduction, RNA processing and vesicular trafficking
12	Smith–Magenis syndrome (microdeletion syndrome)	RAI1	Retinoic acid-induced protein 1	Function as a transcriptional regulator

Most cases are known to involve chromosomal abnormalities, submicroscopic copy number variations and mutations in a single gene, such as in Fragile X syndrome (FXS), Rett syndrome (RTT), MECP2 duplication syndrome (MDS), tuberous sclerosis complex (TSC) and PTEN macrocephaly syndrome [34]. Tuberous sclerosis complex (TSC) is an autosomal dominant genetic disorder characterized by benign tumors in the brain and high penetrance of ASD. It is caused by mutations in the TSC1 or TSC2 genes producing hamartin and tuberin, respectively. Fragile X syndrome (FXS) is the most common monogenic cause of syndromic ASD. FXS accounts for about 2% of all ASDs and is caused by mutations in FMR1 gene, leading to hypermethylation of its promoter. Rett syndrome (RTT) is caused by mutations in the methyl-CpG-binding protein 2 (MECP2) gene. Timothy syndrome is caused by mutations in CACNA1C and is characterized by congenital heart malformations, cardiac arrhythmia, weakened immune system and premature death. Mutations in SHANK3 gene lead to Phelan-McDermid syndrome and is characterized by Epilepsy, kidney dysfunction and cardiac anomalies. The clinical features of the syndromes and their prevalence in ASD are depicted in Table III.

Table III Clinical Features of the Syndromes and their Prevalence in ASD

S.No	Syndrome	Prevalence in ASD	Clinical features
1	Fragile X syndrome	2%	Characteristic facial appearance, macroorchidism. Females are generally less affected than males.
2	Rett syndrome	1%in females, rare in males	Speech impairment, loss of purposeful hand use, ataxia, hyperventilation
3	Angelman syndrome	Rare	Lack of speech, inappropriate laughter, seizures, microcephaly, ataxia
4	Neurofibromatosis syndrome type 1	Rare	Skeletal dysplasia, growth of both benign and malignant nervous system tumors
5	Sotos syndrome	Very rare	Macrocephaly, advanced bone age, characteristic facial features and learning disabilities
6	Timothy syndrome	0.5%	Cardiac abnormalities, facial dysmorphism, seizure
7	Tuberous sclerosis complex	1%	Non-malignant tumors in the brain, kidneys, heart, eyes, lungs, and skin, seizures
8	Williams syndrome	Rare	Characteristic neurobehavioral profile, distinctive facial features, connective tissue abnormalities, endocrine abnormalities
9	Cohen syndrome	Very rare	Typical facial dysmorphism, retinal dystrophy, neutropenia, obesity, microcephal
10	Phelan--McDermid syndrome	0.5%	Absent or severely delayed speech, autistic behavior, seizures, hypotonia, decreased sensitivity to pain,
11	Joubert syndrome	Very rare	Distinctive cerebellar and brainstem malformation, ataxia, breathing abnormalities
12	Smith–Magenis syndrome	Rare	Hyperactivity, sleep disorder, seizures, self--mutilation, hoarse voice

Diagnosis

An early and accurate diagnosis of ASD markers before age 4 i.e., before 12 – 48 months is associated with significant gains in cognition, language and adaptive behavior. A child's overall development can be improved through early intervention. A diagnosis done late causes increased parental stress and hinders early intervention, which is vital to positive outcomes over time [35]. This suggests that early diagnosis and intervention are imperative in the long-term trajectories and quality of life for children with ASD as they are more likely to gain essential social skills and react better in society. Hence there is a critical need for inventive approaches to portray the genetic basis of ASD which will enable early detection.

Diagnosing ASD can be complex, since there is no medical test, like a blood test, to diagnose the disorders. Doctors observe the child's behavior and development to make a diagnosis. Diagnosing an ASD includes developmental screening and a comprehensive diagnostic evaluation. Developmental screening is a short test to identify whether the children are learning basic skills when they should, or if they might have delays. Comprehensive diagnostic assessment comprises hearing and vision screening, neurological testing, genetic testing and other medical testing. Genetic screening is a powerful tool to deal with monogenic Mendelian disorders, characterized by direct genotype - phenotype correlations. In the case of complex disorders like ASD, widespread genetic testing would be expensive, time consuming and inappropriate due to their complexity.

When ASD occurs along with a clinically defined syndrome recognizing these disorders manually depends on the familiarity of the clinician with the features of the syndrome, and the diagnosis is typically validated by targeted genetic testing for example, mutation screening of FMR1. In some cases, genome-wide testing using microarry or whole exome sequencing are carried out to identify ASD-associated variants. These ASD groups cannot be easily clinically defined as patients with a given variant have variable somatic abnormalities. These methods are also costlier and time consuming.

Hence this work is focused upon building models that will enable the automation process and the accurate diagnosis of ASD. Ten genes namely FMR1, MECP2, TSC1, CACNA1C, SHANK3, CHD8, FOXP2, CNTNAP2, GABRB3, HOXA1 and four types of mutations namely

missense, nonsense, silent and frameshift are taken into account for study. Three levels of susceptibility to ASD namely low, medium and high are considered for the research.

1.4 REVIEW OF LITERATURE

In the past decade, researchers have added significantly to the understanding of ASD genetics. New frontiers of genetic contributions to identify new genes that are implicated in increased risk for ASD or to better understand previously identified genetic risk factors were attempted. It is revealed that mutations play a major role in ASD and also that research has been done to identify integrated gene networks for ASD [36].

As ASD is thought to be among the most heritable of all developmental neuropsychiatric conditions, the identification of susceptibility genes would seem to hold tremendous promise for elucidating the underlying cellular and molecular mechanisms of disease and to pave the way for improvements in diagnosis and the development of novel therapeutic strategies. Past research has demonstrated that more detailed genetic analyses can identify mutations that are targets of pharmacological therapy. Though ASD research has shown right directions, much work remains to be done for a complete picture of the risk factors that play the most significant role.

Computational methods based on micro array gene expression and imaging data are applied in ASD related research works. Jamal et al. [37] investigated the occurrence of Autism using efficient brain connectivity measures resulting from EEG of children through face perception tasks. Samples were obtained from EEG signals for typical children and children affected by ASD. In each class, 12 subjects were used for the extraction of connectivity features from joyful, sad and scared faces. The discriminant analysis and support vector machine with polynomial kernels were investigated for the classification task. 94.7% accuracy was reported in the leave one out cross-validation of the SVM with sensitivity and specificity values as 85.7% and 100% respectively.

Latkowski et al.[38] explored the most important genes which are strictly related to autism using gene expression microarrays. The work applied different methods of gene selection, to select the most representative input attributes for an ensemble of classifiers. The authors developed a two stage ensemble system of automatic recognition of autism on the basis of gene

microarray. The results of selection combined with a genetic algorithm and SVM classifier showed increased accuracy of autism recognition.

Li Liu et al.[39] developed a system, Detecting Association with Networks (DAWN) that is efficient in predicting ASD genes and subnetworks by using genetic and gene expression data. The ensemble data is passed on as a hidden Markov random field in which the graph structure is established by gene co-expression. Along with these interrelationships node-specific observations like genetic data, gene identity, expression and the estimated effect on risk are combined. The findings prove that neurite extension and neuronal arborization are threats for ASD. The system recognized 127 genes which possibly affect risk and a set of ASD subnetworks.

Arjun Krishnan et al. [40] used a machine-learning approach to present a genome-wide prediction of autism-associated genes, based on a human brain-specific functional gene interaction network. The work was validated in an independent case-control sequencing study of approximately 2,500 families. A statistical model was built that captured the connectivity patterns of known autism genes in the brain-specific network. This model was used to predict whether each of the other unlabeled genes in the network resembled an autism gene based on its connectivity in the network. It also recognized probable pathogenic genes with most common autism-associated copy-number-variants (CNVs) and reported genes and pathways that are possible mediators of autism across multiple CNVs. The work involved an evidence-weighted linear support vector machine (SVM) classifier.

Yan Kou et al., [41] applied supervised learning techniques to prioritize ASD disease gene candidates depending on curated lists of known ASD and ID disease genes. Two network-based classifiers and one attribute-based classifier were employed to rank and predict genes for ASD. It is proved that neuronal phenotypes features in mouse knockouts can aid in classifying neurodevelopmental genes.

Yun Jiao et al., [42] constructed diagnostic models for ASD based on regional thickness measurements extracted from Surface-based morphometry. The study included 22 subjects with ASD and 16 volunteer controls and regional cortical thicknesses for 66 brain structures for each subject were collected. Four machine-learning techniques namely support vector machines

(SVMs), multilayer perceptrons (MLPs), functional trees (FTs), and logistic model trees (LMTs) were employed to generate diagnostic models. It was found that thickness-based diagnostic model LMT achieved the best classification performance, with accuracy of 87%.

Deep Neural Networks have been used in varied genomic and proteomic research works focusing on protein structure prediction, gene expression regulation, protein classification and anomaly classification. First systematic applications of deep learning methods in computational biology were focused on the prediction of splice sites and coding regions [43, 44]. Current modern applications of deep learning in genomics are focusing on the analysis of actual DNA or RNA sequences and the inference of functional properties and phenotypic consequences associated with mutations. Fakoor et al. [45] applied deep learning methods to extract key features from gene microarray data in predicting cancers. The work first applied PCA to eliminate the effects of redundant and noisy dimensions, then applied three auto-encoders methods. The stacked auto-encoder with fine-tuning achieved the best accuracy in six datasets with accuracy ranging from 76.67% to 95.15%, while the single-layer sparse auto-encoder performed the best in 5 datasets with ACC ranging from 46.76% to 91.50%.

Tan et al., [46] attempted denoising autoencoders (DAs) to identify and extract complex patterns from genomic data. Their performance was measured by implementing them to a collection of breast cancer gene expression data. Results showed that DAs effectively constructed features with both clinical and molecular information. Danaee et al.[47] used SDAE to transform high dimensional, noisy RNA-seq gene expression data to lower dimensional, meaningful representations, based on which they applied different machine learning methods to classify breast cancer samples from the healthy control.

Singh et al., [48] explored a deep learning strategy for tumor classification with gene expression data combined with layer-wise feature selection using stacked sparse auto-encoders (SSAE). The algorithm was tested on 36 datasets from the GEMLeR repository and the model performance with respect to area under ROC curve was found to outperform the benchmark results.

Jesse M.Zhang et al. [49] explored how deep Recurrent Neural Network (RNN) architectures can be used to capture the structure within a genetic sequence. The work leveraged

a bidirectional character-level RNN to predict the interaction of a given genomic sequence with transcription factors. It was empirically shown that the deep network surpassed a baseline model on a noteworthy majority of 919 binary labeling tasks.

Very few contributions have been done with Long Short Term Memory units (LSTM) in the field of genomics and specifically for gene sequence classification. Authors in [50] proposed DanQ, a novel hybrid convolutional and bi-directional long short-term memory recurrent neural network framework for predicting non-coding function de novo from sequence. In this model, the regulatory motifs were identified by convolution layer, while the long-term associations between the motifs were recognized by recurrent layer that improved predictions. More than 50% relative improvement was achieved by the system for few regulatory markers in the area under the precision-recall curve metric when compared to related models.

Researchers in [51] proposed a deep-learning-based hybrid architecture that used Gated Recurrent Units (GRU) to predict enhancers using the DNA sequence alone. The results demonstrated that common enhancer patterns were learnt by the system from the DNA sequence with high accuracy. It was more generalizable in enhancer prediction compared to other standard enhancer predictors based on sequence characteristics. Shen et al.[52], proposed a model, named KEGRU, to identify TF binding sites by combining Bidirectional Gated Recurrent Unit network with k-mer embedding. At first DNA sequences are divided into k-mer sequences with a specified length and stride window. Each k-mer was treated as a word and pre-trained word representation model was built through word2vec algorithm. Thirdly, a deep bidirectional GRU model was constructed for feature learning and classification. Experimental results have shown that the method has better performance compared with some state-of-the-art methods. The summary of literature survey is presented in Table IV.

Table IV Summary of Literature Survey

Authors	Data	Objective	Algorithm	Results
Wasifa Jamal, 2014	Phase synchronized patterns from 128-channel EEG signals obtained for typical children and children with autism spectrum disorder	To identify the presence of ASD using functional brain connectivity measures derived from electro-encephalogram (EEG) of children	SVM classifier	94.7% accuracy

Latkowski T, 2015	Gene expression microarrays	To recognize a case of autism using gene expression microarrays	SVM classifier	The results of selection combined with a genetic algorithm and SVM classifier have shown increased accuracy of autism recognition
Li Liu, 2014	Gene expression data	To predict ASD genes and subnetworks	Hidden Markov random field	127 genes which possibly affect risk and a set of ASD subnetworks were recognized
Arjun Krishnan, 2016	Human brain-specific functional interaction network	Genome-wide prediction of autism risk genes	Weighted linear SVM classifier	It recognized probable pathogenic genes with most common autism-associated copy-number-variants
Yan Kou, 2012	Curated list of 114 genes implicated in ASD	To predict and rank ASD genes	SVM	80%–98% accuracy
Yun Jiao, 2010	Brain images of 22 subjects with ASD and 16 volunteer controls	To construct diagnostic models for ASD, based on regional thickness measurements extracted from Surface-based morphometry	Logistic model trees	87% accuracy
Fakoor, 2013	13 array-type datasets with sample sizes ranging from 20 - 1,047	To predict cancers by extracting key features from gene microarray data	Stacked auto-encoder	97.5% accuracy
Tan, 2014	Breast Cancer dataset with 1,424 training 712 testing samples	To demonstrate that Denoising Autoencoders effectively extract key features from gene expression data	Denoising Autoencoders	ACC: 75%-99.6%

Danaee, 2016	TCGA RNAseq with 1,210 samples, including 1,097 breast cancer samples and 113 healthy samples	To detect cancer using deep learning approach	Stacked Denoising Autoencoder	Accuracy-98.26% sensitivity - 97.61% specificity 99.11% precision 99.17%
Singh, 2016	36 datasets from the Gene Expression Machine Learning Repository	To perform tumor classification using deep learning strategy	Stacked sparse auto-encoder	ACC-83.7%
Jesse M.Zhang, 2016	8000 sequence -label pairs	To perform character-Level Genome Prediction	Recurrent neural network	Deep network surpassed a baseline model on a noteworthy majority of 919 binary labeling tasks
Daniel Quang, 2016	Human GRCh37 reference genome was segmented into non-overlapping 200-bp bins with 919 labels	To build a predictive model for the function of non-coding DNA	CNN and RNN	DanQ achieved over 50% relative improvement in the area under the precision-recall curve metric compared to related models.
Bite Yang, 2017	1747 and 567 experimentally validated human and mouse noncoding elements	To predict enhancers using the DNA sequence alone	CNN with GRU-BRNN	AUC – 0.831
Shen, 2018	125 transcription factor binding sites from A549, MCF-7, H1-HESC and HUVEC datasets	To identify transcription factor binding sites from DNA sequence	Bidirectional GRU model	Average Precision - 0.9620

From the literature study, it is observed that researchers have not attempted deep learning for predicting ASD causing gene sequences, their susceptibility and the driving mutations. In the initial phases researchers had used data from images, controls and microarray gene expression extensively to identify the occurrence of ASD, associated genes, RNA biomarkers and its subnetworks. Traditional methods like discriminant analysis, hidden Markov random fields were applied but the potential of gene sequence data was not exploited for these tasks. Later machine learning approaches like SVM, ensemble of classifiers were used to recognize ASD genes and the interacting networks using microarray data. Recent advancements in deep learning has motivated geneticists to use this emerging area in their research works. They have explored deep learning to extract key features from microarray gene expression data and to identify diseases like cancer. Soon gene sequences were levaraged in various studies involving deep learning to find the similarity between gene sequences, to predict enhancers, splice sites, coding regions and to find the interaction between genes and transcription factors. While gene structural studies were carried out using gene sequence data, identification of ASD genes using this data is still not sought after. Though deep learning approaches are used significantly to improve the accuracy of the prediction task, so far researchers have not ventured deep learning together with gene sequence data for the recognition of any disease. Studies involving recognition of mutations which are the key drivers of ASD have also not been endeavored. Hence this research study is attempted to study the correlations between ASD genes and the mutations that underlie them and the problem is modeled as pattern recognition task using gene sequence data and deep learning techniques.

1.5 MOTIVATION

The genetic markers for ASD are hard to match from patient to patient and as there is no single indicator across every case, it is harder to treat this disorder. Each person needs a personalized treatment plan indicating that genetic diagnosis could play a major role in determining patient care and appropriate therapies. To provide guidance on personalized care to people with ASD, it is essential to determine and understand the factors that cause different variations of ASD so that tailored treatments can be developed. A reliable and accurate system to automate the process of genetic diagnosis smoothly will be of immense use to the clinicians. Gene sequences help in diagnosing genetic disorders and mutational information enable to reveal

the exact reason for the disorder. There is a need for more comprehensive learning approaches that exploit the association between ASD genes and the mutations that trigger them. Currently identification of ASD using gene sequences and mutations have not been attempted. Hence in this work, observations distinguishing the diseased gene are taken into account by also considering the appropriate mutational features from sequence data.

Machine learning models are more reliable and efficient providing accurate and interpretable results. Deep learning methods enable the discovery of high-level features using representation learning and demonstrate improved performances over traditional models. The major benefit of deep learning is its capacity to execute feature engineering on its own and hence does not require any domain expert. RNN architectures which are competent in handling varied length sequences have not been explored so far to identify ASD. Hence this work aims to explore traditional supervised machine learning algorithms and contemporary deep learning techniques specifically RNN architectures to identify genes, mutations triggering ASD and gene susceptibility to the disorder. The proposed model will be crucial in a clinical environment as it provides a better understanding of the affected gene sequences aiding focused treatment. The results of the work could be useful to guide clinicians in effective genetic diagnoses and pursue targeted genetic testing of individuals with ASD whose clinical phenotype matches a specific genetic or genomic etiology.

1.6 OBJECTIVES OF THE RESEARCH

The main focus of this research work is to develop models for predicting the causative ASD genes, their susceptibility and the underlying mutations through conventional machine learning and contemporary deep learning methods. The core objectives of this research work are as follows.

- To create a synthetic gene sequence database that mimics the causative ASD gene sequences
- To identify and capture distinctive features from the diseased gene sequences that contribute to the classification of genes, their susceptibility to the disorder and the underlying mutations
- To develop a framework based on conventional machine learning techniques for prediction of causative ASD genes and the underlying mutations

- To design a model to predict the susceptibility of the ASD genes to the disorder using supervised machine learning
- To build a deep learning framework for predicting the causative ASD genes, their vulnerability to the disorder and the driving mutations through user defined features
- To employ Recurrent Neural Network (RNN) variants namely Bidirectional Recurrent Neural Network (BRNN), Long Short Term Memory (LSTM) and Gated Recurrent Units (GRU) with user defined features for creating computational models to predict ASD causing genes, their susceptibility and the driving mutations
- To develop two kinds of encoding schemes namely codon encoding and one hot encoding to build DNN, BRNN, LSTM and GRU models through self-learnt features for predicting ASD causative genes

The thesis explains a novel and an unprecedented approach wherein the problem of predicting ASD genes, their vulnerability to the disorder and the driving mutations are formulated as pattern classification tasks and different classifiers are built. This task is carried out based on both user defined and self learned features through conventional supervised learning and contemporary deep learning methods. These approaches simplify the prediction problem generating reliable solution based on intelligent hints collected from mimicked gene sequences.

1.7 ORGANIZATION OF THE THESIS

The thesis is organized into nine chapters. The outline of the thesis is as follows:

Chapter 2 presents an overview of the deep learning architectures. A detailed description of the various deep learning architectures such as Deep Neural Networks, Recurrent Neural Networks, its variants Long Short Term Memory Units and Gated Recurrent Networks is presented. The basic definitions and concepts used in deep learning are also included.

Gene sequences required for the experiments are created by simulation using gene and mutational information collected from various sources like NCBI, OMIM and SFARI databases. Chapter 3 elucidates the process of data preparation and problem modeling. The chapter also gives a note on data collection, corpus development and creation of datasets that are used in this research.

Chapter 4 presents the implementation of supervised machine learning models for identifying the ASD causative mutations and genes. Experiments carried out using Decision Trees, Support Vector Machines, Multilayer Perceptrons on various datasets are discussed in detail. The chapter also deals with a multi dimensional model for concurrent prediction of ASD genes and mutations. The performance analysis of the results and findings of the experiments are reported in this chapter.

In Chapter 5, a brief introduction to gene susceptibility identification is presented. This chapter elucidates the supervised model to predict the susceptibility of candidate ASD gene based on the cumulative strength of evidence for each ASD gene sequence. It also explains the different models built with supervised classification algorithms such as Decision tree, Multilayer Perceptron, Support Vector Machines. Experiments carried out using various datasets using these algorithms are illustrated and the results are reported in this chapter.

Chapter 6 describes in detail about the deep learning approach used in this research work. The process of developing a deep neural network model through user defined features for identifying ASD causing genes, their susceptibility and mutations is elaborated and an analysis of the experimental results is presented with tables and charts.

In Chapter 7, the Recurrent Neural Network variants BRNN, LSTM and GRU attempted for the task of identifying the causative genes, their susceptibility and mutations are elaborated. These RNN algorithms are experimented with user defined features and the effectiveness of these methods are reported in this chapter.

In Chapter 8, two encoding schemes namely codon encoding and one hot encoding proposed to exploit the self-learning power of deep learning models are presented. Deep models built using DNN, BRNN, LSTM and GRU with encoded datasets are described. The results obtained by employing these two encoding schemes in deep networks are reported and illustrated in this chapter.

Chapter 9 summarizes the entire research work with various findings of the traditional machine learning and contemporary deep learning approaches for the efficient prediction of causative mutations, genes and their susceptibility to ASD. It also presents the research achievements of the proposed model and discusses the scope for future research.