1. INTRODUCTION

It is interesting to note that the biological databases are a gold mine of information for the researchers. In fact, analyzing the biological data using the Data mining (DM) techniques, with the invent of advanced computing technologies is virtually current research frontier which needs extraordinary attention. Mining biological data and making classifications, associations and predictions highlight many concealed facts from the historical data or existing information. Therefore, application of data mining techniques for modeling biological data is essential in many situations and hence focused in the present investigation.

1.1 Data Mining and Bioinformatics

Extraction of hidden predictive information from large databases is termed as Data Mining. Forthcoming trends and behaviors, permitting businesses to make proactive, knowledge-driven decisions are the predictions made using the data mining tools. It is an iterative process within which progress is defined by discovery, through either automatic or manual methods. Searching for new, valuable, and nontrivial information in huge volumes of data is done through data mining. Balancing the knowledge of human experts in labeling problems and setting goals with the exploration capabilities of computers achieves the best results.

Data mining is one of the blooming field in the computer arena. Data warehousing, data-mart, decision-support community and covering professionals from industries such as retail, manufacturing, telecommunications, healthcare, insurance, and transportation are the sources of data for mining the data. The development and applications of algorithms for discovery of a priori unknown relationships - associations, groupings, classifiers from data is done through data mining [1].

Data Mining is a process of finding patterns that are valid, novel, useful, understandable by analyzing large databases is also known to be Knowledge Discovery in Databases (KDD). The capabilities provided through data mining technologies are as follows:

- Predicting trends and behaviors automatically:

Data mining automates the process of discovering predictive information from huge volume of data. Data mining uses the past mailing data to identify the targets to amplify the return on investment. Forecasting bankruptcy and identifying similar segments of a population

that are respond to given events are some of the other problems which can be identified through prediction.

- Automated discovery of previously unknown patterns:

Analysing the retail sales data to identify the products that are often purchased together that are not related to each other is an example of discovering the patterns. Detecting frauds in credit card transactions and identifying anomalous data are some of pattern discovery problems. The efficiency of the approach is restricted by the creativity of the user to develop various hypotheses, in addition to the structure of the software being used. A discovery approach is utilized in data mining to examine multidimensional data relationships [2].

Data Mining Techniques

The most commonly used techniques in data mining are:

Rule induction: Based on statistical significance useful if-then rules can be extracted from data.

Classification – With the usage of a predictive learning function, a data item is classified into one of several predefined classes.

Regression – A data item is mapped into to a real-value prediction variable using a predictive learning function.

Clustering - A finite set of clusters is identified using clustering algorithm.

Summarization - Compact description for a set or subset of data.

Dependency Modeling – Based on the significant dependencies between variables or between the values of a feature in a dataset or in a part of a data set a model is identified.

Change and Deviation Detection - Changes in the data set are discovered.

Trends in Data Mining

Data mining tasks done with the various types of data. Therefore, data mining approaches holds numerous challenging research issues in data mining. The important tasks for data mining researches and application developers are the design of a traditional data mining languages, the development of efficient data mining methods, the building of interactive and integrated data mining environments and the applications of data mining to solve large applications problems. Some of the trends in data mining that reveal the pursuit of these challenges are as follows:

Standardization of data mining language

Commercially available data mining languages are Microsoft's SQL server 2005, IBM Intelligent Miner, SAS Enterprise Miner, SGI Mineset, Clementine, DBMiner. A standard data mining language provides the development of data mining solutions, improved interpretability among multiple data mining systems and functions.

Visual data mining

A picture is worth a thousand words is a saying. If the result of the mined data is shown in the visual form, it improvise the worth of the data which are processed data. Visual data mining is an effective way to discover knowledge from huge amounts of data. The systematic study and development of visual data mining techniques will stimulate the use for data mining analysis.

Web mining

The World Wide Web is massive pool of news, advertisements, consumer records, financial, education, government, e-commerce and many other services. Along with the above described data, the WWW also comprises of vast and dynamic collection hyper linked information, that provides a vital source for mining the data. Based on the above facts, the web poses great challenges for efficient resource and knowledge discovery. Multimedia information are numerous and extremely large is essential in many applications, and repositories of multimedia.

Spatial data mining

Applying data mining techniques to spatial data is known as Spatial data mining. Objective of the spatial data mining is to find patterns in geography. Two separate technologies such as data mining and Geographic Information Systems (GIS) visualize and analyses the data with its own methods, traditions and approaches. With the developments in the areas of Information Technology, digital mapping, remote sensing, and the global diffusion of GIS highlights the importance of developing data driven approaches to geographical analysis and modeling GIS-based decision-making enjoys the benefits of data mining, by searching the hidden patterns in large databases.

Pattern mining

Finding existing patterns in data is done using a data mining technique called Pattern mining. The association rules are to be termed as patterns in this context. Analyzing supermarket transaction data, to examine customer behavior in terms of the purchased products is one of the

pattern mining technique motivated with the association rules. Music Information Retrieval (MIR) is a new area in pattern mining where patterns are seen both in the temporal and non-temporal domains, which are introduced to classical knowledge discovery search techniques.

Subject-based data mining

Searching for associations between individuals in data is done with the data mining technique named as Subject-based data mining. In the context of combating terrorism, the National Research Council provides the following definition: "Subject-based data mining uses an initiating individual or other datum that is considered, based on other information, to be of high interest, and the goal is to determine what other persons or financial transactions or movements, etc., are related to that initiating datum.".

Sequence mining

Finding statistically relevant patterns between data examples is concerned with Sequence mining where the values are furnished in a sequence. It is usually believed that the values are discrete, where in Time series mining the values are closely related and considered as a different activity. Sequence mining is one of the product in structured data mining. String mining and Itemset mining are two different kinds of sequence mining.

In the usage of biological data, to examine gene and protein sequences, String mining is widely used sequences with a single member at each position. Alignment of the sequences is done with a variety of eminent algorithms. BLAST- is a kind of alignment that involves in matching a query with one subject, whereas Clustalw matches multiple query sets with each other.

Itemset mining one of the approach to text mining which are used in marketing and CRM applications. Key problems in this area includes building efficient databases and indexes for sequence information, extracting the frequently occurring patterns, comparing sequences for similarity, and recovering missing sequence members. The influential apriori algorithm and the more-recent FP-Growth technique are the two common techniques that are applied to sequence databases for frequent item set mining.

Data Mining Applications

Data mining helps to manage many organizations in all phases of the customer life cycle such as, identifying new customers, increasing returns from existing customers, and holding

good customers. By determining characteristics of good customers, a company can target prospects with similar characteristics. To detect fraudulent in the telecommunications and credit card companies mainly uses data mining techniques are applied to use of their services. Now a days, insurance companies and stock exchanges are also interested in applying this technology to reduce fraud. To predict the effectiveness of surgical procedures, medical tests or medications Data mining is used in medical domain. Companies active in the financial markets use data mining to determine market and industry characteristics as well as to predict individual company and stock performance. Decision on products to stock in particular stores, arrangement of the products in a store, as well as to assess the effectiveness of promotions and coupons, retailers make use of data mining techniques. Pharmaceutical firms are mining large databases of chemical compounds and of genetic material to discover substances that might be candidates for development as agents for the treatments of disease.

Data Mining in Bioinformatics

Bioinformatics is the science of storing, analyzing, and utilizing information from biological data such as sequences, molecules, gene expressions, and pathways. Development of novel data mining methods will play a fundamental role in understanding these rapidly expanding sources of biological data. It is motivating to note that the biological databases store a huge amount of information about the gene sequences, protein sequences and gene expression data are a gold mine of information for researchers in biological field. In fact, the study with biological data by using the data Mining (DM) techniques is nearly a new frontier that needs special attention. Bioinformatics and data mining are developing as interdisciplinary sciences.

Bioinformatics area which is data-rich, but is deficient in a comprehensive theory of life's organization at the molecular level is well suited to apply data mining approaches. The extensive databases of biological information create both challenges and opportunities for development of novel KDD methods [3]. Mining biological data aids in extracting useful knowledge from massive datasets, which are gathered in biology, and in other related life sciences areas such as medicine and neuroscience.

Data mining in bioinformatics is hindered by many facets of biological databases such as, their size, number, diversity and the lack of a standard ontology. The range of levels of domains of expertise present amongst potential users is an another problem and it can be tedious for the database curators to provide appropriate access mechanism.

Applications of Data Mining in Bioinformatics

Applications of data mining in bioinformatics include gene finding, protein function domain detection, function motif detection, protein function inference, disease prognosis, disease treatment optimization, protein and gene interaction network reconstruction, data cleansing, and protein sub-cellular location prediction.

Gene finding: Gene finding aids in identifying the coding regions in a gene that encode for protein. Extrinsic and intrinsic methods are available in finding the gene.

Protein function prediction: Bioinformatics methods that aim to detect the function of the protein. The protein sequences and structures are the inputs to the tools that provides testable predictions of function.

Protein and gene interaction network reconstruction: Large data sets of reliable protein–protein interactions are now available and thus Protein networks have many applications in genomics. Network modularity is an important implication for identifying the human genetic diseases, where protein databases collect the data and adhere to community standards..

Protein sub-cellular location prediction: This application involves in predicting the location of the protein where the protein lies in the subcellular location. Tools that accurately predict the outcome of protein targeting in cells need to be developed.

1.2 Genomics

Deoxyribonucleic acid (DNA) is a complex, long-chained molecule that encodes the genetic characteristics of a living organism. The growth and functioning of all organisms are programmed with the genetic instructions encoded by Deoxyribonucleic acid (DNA). DNA is the chemical information database that carries the complete set of instructions like the nature of the proteins produced by the cell, its life span, maturity, function and death. DNA is located in the nucleus of each cell which is the hereditary material in all organisms. The DNA is a double stranded that forms a double helix like structure and each strand is made up of millions of chemical building blocks called bases. The four types of bases that make up the DNA are Adenine (A), Thymine (T), Cytosine (C) and Guanine (G). The order of these bases is changed with permutation and combination in a sequence and unique sequences code for proteins.

Human DNA comprises of 3 billion bases, where 99 percent of the bases are identical in all humans. The sequence arrangement of the 4 bases will determine the information available for

building and maintaining an organism. Base pairs are formed by pairing up the DNA bases with each other, A with T and C with G. A sugar molecule and a phosphate molecule is attached with each base. Together, a base, sugar, and phosphates are called a nucleotide. A double helix is a spiral where the four bases are arranged in two long strands. The structure of the double helix is somewhat like a ladder, with the base pairs forming the ladder's rungs and the sugar and phosphate molecules forming the vertical sidepieces of the ladder.

Replicating or making copies of itself is an important property of DNA. The double helix can serve as a pattern for duplicating the sequence of bases in each strand of DNA. Similar to the way the order of letters in the alphabet can be used to form a word, the order of nitrogen bases in a DNA sequence forms genes, which in the language of the cell, forms the proteins. Another type of nucleic acid, ribonucleic acid, or RNA, translates genetic information from DNA into proteins. Double helix structure of a DNA is shown in Fig.1.1.

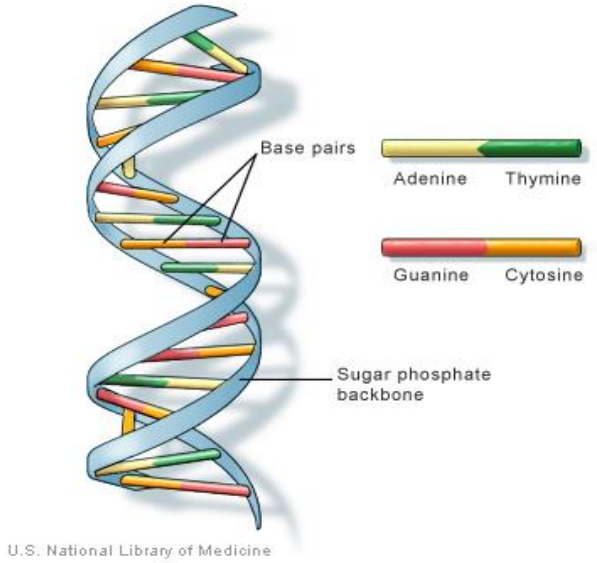

Fig.1.1 Double Helix Structure of a DNA

Cells may be Eukaryotic or Prokaryotic where animals, plants, humans, fungi are eukaryotes and bacteria, archaea are prokaryotes. In the eukaryotic cell's DNA is organized into long structures called chromosomes. These chromosomes are replicated during cell division by providing its own complete set of instructions to each of its cells. Eukaryotic organisms store most of their DNA inside the cell nucleus and some of their DNA in organelles, such as mitochondria or chloroplasts. Cytoplasm is the only place where the prokaryotes store their DNA. Chromatin proteins like histones compact and organize DNA within the eukaryotic

chromosomes. DNA is coiled tightly to fit inside cells and to form chromosomal structures and each chromosome comprises of a single DNA molecule. 23 pairs of chromosomes are present inside the cell's nucleus.

Genes are the basic unit of genetics which are the sections of DNA that code for proteins, that provides the structure and function to human bodies. Genes are hereditary material that is present within the nucleus of the cell. Generally two copies of gene are present in every human, one inherited from each parent. A genome is all genetic data of a single cell that includes the genes in the nucleus but also that of mitochondrial DNA. Human beings have 20,000 to 25,000 genes. The chromosome 1 is the largest chromosome that contains about 8000 genes and chromosome 21 is the smallest which contains about 300 genes. Mutations occur in the genes and the sequence can be altered by producing different alleles. Fig.1.2 shows the structure of a gene.

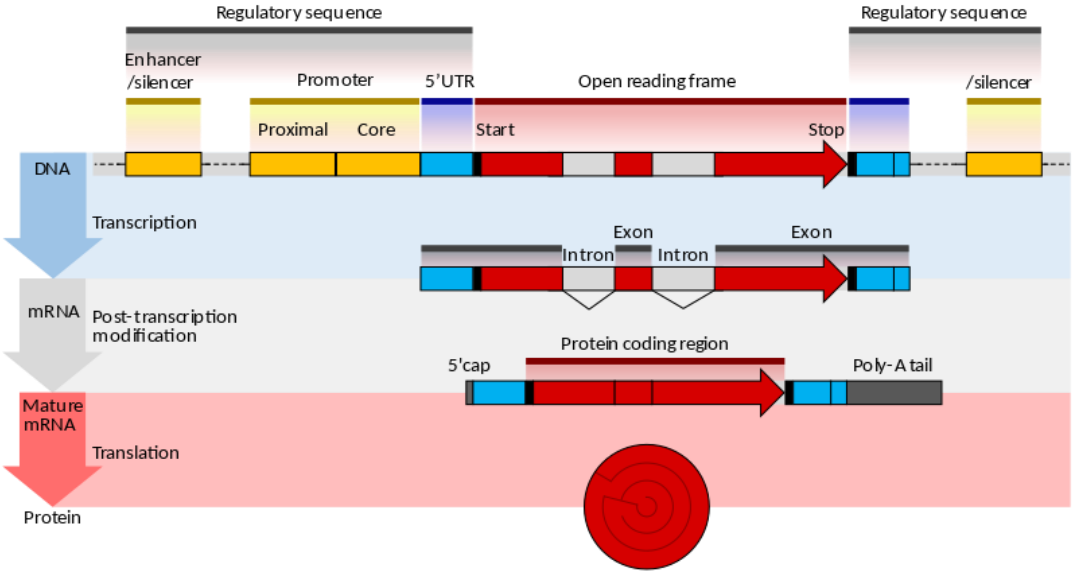

Fig.1.2 Structure of a Gene

Numerous elements composite the structure of a gene but only small part of the gene is used for coding the protein. The other elements include DNA regions that are not transcribed as well as untranslated regions of the RNA. Flanking the open reading frame, all genes contain a regulatory sequence that is required for their expression. Promoter sequence is required to express the genes which is recognized and bound by transcription factors and RNA polymerase to initiate transcription. 3' end and 5'end are the starting and ending position of the reading

frame of the gene sequence. A ribosome binding site, terminator and start and stop codons are present at the both ends of untranslated regions in the transcribed pre-mRNA. In addition, introns are present in the most eukaryotic open reading frames which are sliced out before the exons are translated. The exons are the coding regions of the gene sequence that are translated into protein. The sequences at the ends of the introns, dictate the splice sites to generate the final mature mRNA which encodes the protein or RNA product.

In the coding sequence of a gene, codon is a sequence of three adjacent nucleotides on a strand of a nucleic acid that constitutes the genetic code for a specific amino acid that is to be added to a polypeptide chain during protein synthesis. A codon is the triplet of nucleotides that code for a specific amino acid. Many to one relationship occur between the codon and amino acid. The portion of gene's DNA is the coding region of a gene, also known as the coding sequence or CDS (from Coding DNA Sequence) that are composed of exons whichcodes for protein. The start codon is bounded near the 5' end and the stop codon near the 3'end.

A pathologic condition impairs the normal function or structure of an organ in human beings. In the current genomic era, the identification of the disease is paramount. A disease is a specific aberrant condition, which distress the function of an organism. Some specific symptoms and signs are construed as a medical condition while a disease occurred. The external or internal factors cause these dysfunctions. The specific infective agents and inherent defects of the organism also lead to disease. Diseases can affect people not only physically, but also emotionally, as contracting and living with a disease can alter the affected person's perspective on life. Infectious diseases, deficiency diseases, genetic diseases (both hereditary and non-hereditary), and physiological diseases are the four major types of diseases. Genetic diseases are disorders that are inherited from parents or are related to some type of spontaneous genetic change. Mutations in the gene sequences are the vital reason for the genetic diseases.

Diagnostic Methods

Medical diagnosis is the process of determining condition that explains a person's abnormal symptoms and signs. The history and physical examination of the person seeking medical care are the information required for identification of the disease. One or more diagnostic procedures are done during the process to detect a medical indication to perform diagnostic procedure. Detection of any deviation from normal structure of the human body,

complaint or difficulties expressed from the patient are the procedures to diagnose a patient from a disease.

Types of genetic disease identification procedure:

A physical examination: A genetic disorder is suggested with the diagnosis of certain physical characteristics like distinctive facial features. A complete physical examination includes measurements such as head circumference, the distance between the eyes, and the length of the arms and legs. X-rays, computerized tomography (CT) scans, or magnetic resonance imaging (MRI) are the imaging studies that are employed to view the structures inside the body.

Personal medical history: A personal medical history includes information about an individual's health from birth. The past health implications that has already been that constitutes of hospitalizations, surgeries, allergies, medications and the results of any medical or genetic testing.

Family medical history: The health information about the family members is a critical tool for diagnosing the genetic disorders. Genetic conditions often run in families and hence the health conditions in an individual's parents, siblings, children, distant relatives are important source for diagnosing disease. This information provides hints about the diagnosis and inheritance pattern of a genetic condition in a family.

Laboratory tests: Genetic disorders are diagnosed with Molecular, chromosomal, and biochemical genetic testing. Other laboratory tests that measure the levels of certain substances in blood and urine can also help in suggesting a diagnosis.

The major genetic options include:

- Single gene sequencing

  The diseases that occur through genetic variation are identified using the gene sequencing methods. It is done by extracting the DNA from the blood samples of the family. The extracted blood samples are analysed through next generation sequencing technologies and Polymerize chain reaction (PCR). Mutations are identified using this procedure.

- Microarray Technique

  Analyzing the expression of large number of genes simultaneously is done with the microarray gene expression analysis technique. The hybridization of an mRNA molecule to the DNA template is involved in the microarray experiment. The expression level of the various genes is indicated with the amount of mRNA bounded to each site. Microarray

Expression Analysis, Microarray for Mutation Analysis, Comparative Genomic Hybridization are the types involved. Gene discovery, disease diagnosis, drug discovery, toxicological researches are the applications of using this technique.

- Whole exome sequencing

  Exome sequencing is one of the techniques for sequencing all of the expressed genes in a genome. The goal of this approach is to predict genetic variants that modify protein sequences.

DNA Mutation

Amend in the genetic code that causes a permanent change in the DNA sequence is termed as mutation. Normally mutations are found while a fault occurs during translation, problem with DNA, mistake during transcription. Mutations may be inherited from parents to offspring and also through spontaneous events such as environmental damage or an error occurs while a DNA replication takes place. Errors occur both during replication and distribution of the genetic material giving rise to sudden heritable changes in the characters of organisms and individuals showing these changes are known as mutants. Mutations arise at the DNA level but they show their effects at the protein level. Mutations are classified based on their effects on protein.

Single character change in a gene makes an impact on the gene which in turn changes the function of the gene. A mutation in DNA may do no harm in protein sequences in some of the mutations. Consider the codon table that constitutes 20 different amino acids that code for 64 different codons. In addition, some trinucleotides code for start and stop codon, where the stop codon serve as a signal of termination of the chain during the protein synthesis. TAA, TAG and TGA are the stop codons. The codon that initiates the start of a protein chain is ATG which is the start codon that codes for Methionine (Met). Fig.1.3 depicts the codon table.

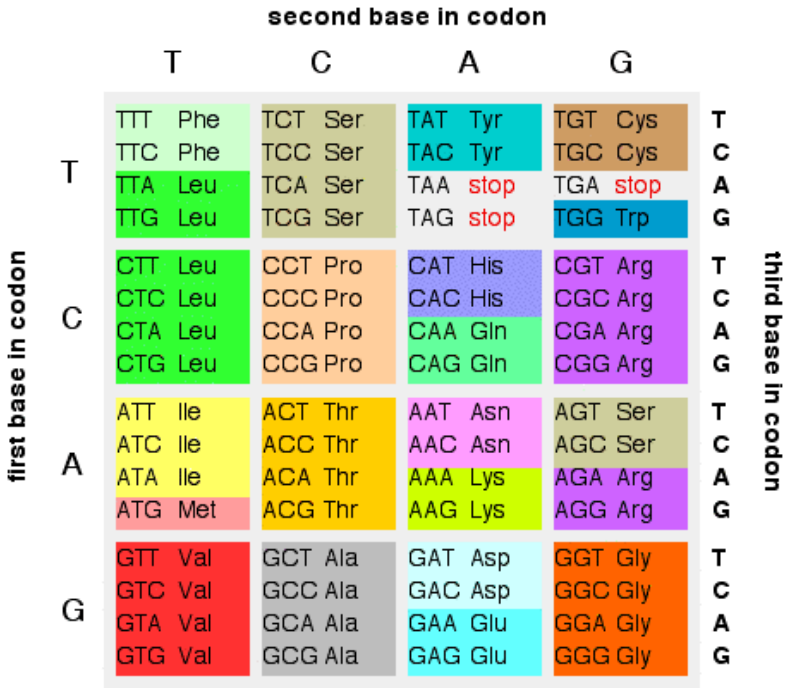

Fig.1.3 Codon Table

Types of Mutations

DNA mutations perceptibly root to genetic diseases. Mutation can result in many different types of change in sequences. Mutations in genes can have no effect, alter the product of a gene, or prevent the gene from functioning properly or completely. Mutations can also occur in nongenic regions but it is very harmful while occurred in the genetic regions. Mutations may arise due to a change in the base sequence of a gene. Such mutations are called as gene mutations or point mutations. Point mutation occurs at one point, which is a SNP (single-nucleotide polymorphism) mutation in the deoxyribonucleic acid (DNA) that occurs at one point. DNA replication occurs when one double-stranded DNA molecule creates two single strands of DNA. There are three types of point mutations, namely a missense mutation, nonsense mutation and silent mutation. Frameshift mutations change the reading frame of a sequence. Inserting or deleting one or more nucleotides in a sequence cause this kind of mutation. Changes in chromosomal number and structure also produce heritable changes in phenotype and these kinds are termed as chromosomal mutations.

Missense Mutation

Missense mutation is a kind of point mutation in which the substitution of single base in a codon encodes a different amino acid. Substitution is an exchange of one base to another, such as

swapping a base from A to G. Missense mutation is a type of non synonymous mutation. This type of mutation may occur due to copying errors, chemicals or any type of viruses. Therefore when a missense mutation occurs there will be an alteration in the amino acid sequence that causes defects in building a protein product.

DNA code for an amino acid sequence

For example consider a DNA sequence

Sequence:	ACT	CCT	GAG	GAG	GAG	ACT
Amino acid:	Thr	Pro	Glu	Glu	Glu	Thr

Sequence

Replacement of nucleotide when missense mutation occurs

Sequence:	ACT	CCT	GAG	GAG	GAG	ACT
Sequence:	ACT	CCT	GTG	GAG	GAG	ACT
Amino acid:	Thr	Pro	Val	Glu	Glu	Thr

sequence

In the above noted sequence, a single nucleotide change from A to T is occurred and thus it codes for Val instead of the amino acid Glu and thus incorrect amino acid in the sequence may do malfunctioning of protein.For example, a mutation entry in a HGMD database such as 347T>C indicates that codon changes CTC-CCC in the dystrophin gene results in DMD, where the protein Leu is altered to Pro [4].

Nonsense Mutation

Nonsense mutation is a kind of point mutation where the substitution in a codon that results in premature termination of protein or premature stop codon. A stop codon signals the end of the translation process and terminates protein production. A nonsense codon in the mRNA is truncated, incomplete and achieves a nonfunctional protein product with the effect of nonsense mutation. TAG ("amber"), TAA ("ochre"), TGA ("opal" or "umber") are the three stop codons.

DNA code for an amino acid sequence

For example consider a DNA sequence

Sequence:	ACT	CCT	GAG	GAG	GAG	ACT
Amino acid:	Thr	Pro	Glu	Glu	Glu	Thr

Sequence

Replacement of nucleotide when nonsense mutation occurs

Sequence: ACT CCT GAG GAG GAG ACT

Sequence: ACT CCT GAG TAG GAG ACT

Amino acid: Thr Pro Glu Stop Glu Thr

Sequence

In the above noted sequence, a single nucleotide change is encountered from G to T and thus it attains a stop codon in the middle of the sequence and therefore premature stop codon achieves nonfunctional protein product.

For example, a nonsense mutation in the dystrophin gene 433C>T, point out that the codon change CGA-TGA and the protein arg is terminated with amber stop codon and results in BMD [5].

Silent Mutation

In some cases, a DNA mutation may do no harm in protein sequences. It depends on the sort of DNA mutation and where it is located. A change in codon encodes the same amino acid and causes no change in the protein is called silent mutations [6]. Silent mutations are changes in codon that encodes for the same amino acid and therefore the protein is not altered. The silent mutation is a kind of point mutation which changes the codon usage pattern. The translated protein in the amino acid sequence is not modified with the synonymous codon changes.

Modern investigations show that the silent mutation changes can affect protein folding and function. More than 50 diseases are correlated with this kind of point mutation. Silent mutation alters the secondary structure of mRNA and hence the stability of the mRNA will be reconstructed. Even though several codons encode for the same amino acid the frequency will differ and this is referred as codon bias.

DNA code for an amino acid sequence

For example consider a DNA sequence

Sequence: ACT CCT GAG GAG GAG ACT

Amino acid: Thr Pro Glu Glu Glu Thr

Sequence

Replacement of nucleotide when silent mutation occurs

Sequence: ACT CCT GAG GAG GAG ACT

Sequence: ACT CCA GAG GAG GAG ACT

Amino acid: Thr Pro Glu Glu Glu Thr

Sequence

In the above noted sequence, a single nucleotide change is encountered from T to A and no change occurred in the amino acid sequence.

Consider an example, in CAPN3 gene 246G>A specifies CCG-CCA and the protein pro is not misrepresented, but it routes to the LGMD type 2 disease.

Insertion/Duplication Mutation

During insertions, new base pairs are added into the sequence that alters the function of a gene. The number of insertions may change from to 1 to 1000. Small number of base pairs inserted is termed as small insertions and if large number of base pairs or whole exons is inserted is termed as gross insertions. An increase in the number of the same nucleotides in a location is termed as duplications. When duplication occurs the protein can take on different functions.

DNA code for an amino acid sequence

For example consider a DNA sequence

Sequence:	ACT	CCT	GAG	GAG	GAG	ACT
Amino acid:	Thr	Pro	Glu	Glu	Glu	Thr

Sequence

Inserting a single nucleotide

Sequence: ACT CCT GAGAGAG GAG ACT

Sequence: ACT CCT GAG AGA GGA GAC T

Amino acid:	Thr	Pro	Glu	Arg	Gly	Asp

Sequence

In the above noted sequence, the entire amino acid sequence is altered when a single nucleotide is inserted into the frame of a sequence. Small or gross Insertions and deletions may be possible in this kind of mutation.

For example, EMD disease is caused by the duplications in the emerin gene for the nucleotide change 650_654dupTGGGC [7].

Deletion Mutation

Deletions occur in the genes when base pairs are deleted from a sequence that truncates the function of genes. Small number base pairs are deleted in small deletions where the whole number of exons is deleted while large deletion occurs. The total number of base pairs is alters when this kind of mutation occurs.

DNA code for an amino acid sequence

For example consider a DNA sequence

Sequence: ACT CCT GAG GAG GAG ACT

Amino acid: Thr Pro Glu Glu Glu Thr

Sequence

Deleting a single nucleotide g in the 6th position

Sequence: ACT CCT AG GAG GAG ACT

Sequence: ACT CCT AGG AGG AGA CT

Amino acid: Thr Pro Arg Arg Arg

Sequence

In the above noted sequence, the entire amino acid sequence is altered when a single nucleotide is deleted from the frame of a sequence.

For example, 253delG deletes G in 253 position in the SH3TC2 gene that directs for Charcot-Marie-Tooth disease 4C. Gross insertions and gross deletions occur when the whole number of exons is involved in the insertions are deletions.

Splicing mutation

In the Eukaryotic genes, the spliceosome catalyses the intervening sequences or introns which are spliced by the process of RNA splicing. Any change that occurs while splicing will lead to splicing mutation. Most of the genetic disorders caused by the mutations in the gene sequences show an impact on splicing. The splicing mutation that occurs in the splice sites comprises of donor site (5'end of the intron), branch site, Acceptor site (3'end of the intron).

Defects in the splice site lead to the loss of function of the site, premature stop codon, loss of exons, an inclusion of intron, variation in splice site location, insertion or duplication of amino acid and disruption in reading frame. The exons of a primary transcript may be spliced in various ways during pre-mRNA splicing which leads to Alternative splicing (AS). It facilitates the same gene to result into various splicing isoforms that contain diverse combinations of exons that result in different protein products. Splicing mutation leads to exon skipping, intron retention and alternative 3' splice site and 5' splice site. During the splice site mutation, the boundaries between exons and introns are affected.

1.3 Genetic Disorders and Muscular Dystrophy

A genetic disorder is caused by an abnormality in an individual's DNA that is predictable by the mutations in the gene sequences. Genetic disorders may be hereditary, passed down from the parents' genes or the defects may be caused by new mutations or changes to the DNA. Single gene disorders were caused by defects in one particular gene, often with simple and predictable Mendelian inheritance patterns such as Autosomal Dominant, Autosomal Recessive, X-linked Recessive, X-linked Dominant and Y-linked.

The autosomal dominant pattern involves, only one copy of the genetic defect to cause the disease. Anyone in the family with the gene mutation can pass the disorder to children. Both male and female are affected in this type of disorder. The disease is observed in multiple generations. Transmission of the disease occurred through this type of inheritance occurred by both male and female. Huntington diseases, marfan disease, achondroplasia, muscular dystrophy are some of the autosomal dominant disorders.

The recessive pattern of a disease requires two copies of inherited defective genes, one from each parent where both will be carriers of the disease but usually not affected by the disease. Both male and female are affected in this kind of recessive disorder. This kind of disease is observed in only single generation. Both gene alleles need to be affected when the disease is expressed. Two germline mutations one from each parent is acquired in this type of disease. Cystic fibrosis, galactosemia, muscular dystrophy are some kinds of autosomal recessive pattern.

In the case of X-linked dominant mutations occur only in X chromosome. Both male and female are affected and affected male will transmit the disease to their daughters while affected female transmits the disease to both sons and daughters.

In X-linked recessive inheritance also mutations occur only in the X chromosome. Here only the males are affected. The disease is passed only from mother to their children. Duchenne Muscular dystrophy is a kind of X-linked recessive inheritance. Hereditary diseases include hereditary hemochromatosis, down syndrome, muscular dystrophy, achondroplasia, usher syndrome, spherocytosis, hemophilia, sickle cell anemia, porphyria, turner syndrome, xeroderma pigmentosum, neurofibromatosis, galactosemia, , myotonic syndrome, albinism, polycystic kidney disease, retinoblastoma, klinefelter syndrome, tay-Sachs disease and phenylketonuria.

Chromosome disorders results in change in the number or structure of the chromosomes. Multifactorial disorders were caused by mutations in multiple genes, in a complex interaction

with environmental and lifestyle factors such as diet or cigarette smoke. Multifactorial disorders include heart disease and diabetes.

Generally diseases are classified as monogenic and polygenic diseases. Monogenic diseases are genetic diseases caused by a single mutation in a specific gene like Mendelian diseases [8]. This single mutation has various quantifiable phenotypes across patients [9]. Cystic fibrosis, sickle cell anemia, muscular dystrophies are a few monogenic diseases. In contrast, polygenic diseases are mostly non-genetic diseases caused by mutations in multiple genes and their phenotypic expression is often cumulative or cooperative. Polygenic diseases are complex diseases such as diabetes, cancer and heart disease.

Muscular Dystrophy

Muscular dystrophy (MD) is a cluster of successive muscle disorders stimulated by mutations in genes that encode for proteins that are vital for regular muscle function [10]. It is a monogenic disease that is caused by mutations in the genes. The change of protein in the muscles leads to alteration or malfunction of the muscles reveals muscular dystrophy. Progressive muscle weakness that affects limb, axial and facial muscles is the foremost cause of muscular dystrophy. The other muscles that function in respiratory, cardiac and swallowing are affected in some specific types of muscular dystrophy. In a rare variant, the brain, inner ear, eyes, or skin is impaired by muscular dystrophy disorder. Muscular dystrophy is believed as a genetic ailment flow in a family, even if only one blood relation in the ancestor is affected. Disorders may be X-linked recessive, autosomal recessive or autosomal dominant. In X linked case the males are mostly affected than female children because the male comprises of only one X chromosome and gets flawed by muscular dystrophy and hence in most cases the trait is identified in male children. In females, two pairs of X chromosomes are present and therefore the daughters turn out into carriers, and generally not affected by the disease.

There is no cure for muscular dystrophy. Diagnosis often involves blood tests and genetic testing. The majority of hereditary disorders place a significant burden on the families immortalizing the condition for the lack of effective treatment [11]. There are nine main categories of muscular dystrophy that contain more than thirty specific types. Duchenne muscular dystrophy, Becker muscular dystrophy, Emery-Dreifuss muscular dystrophy, Limb-girdle muscular dystrophy, Facioscapulohumeral muscular dystrophy, Myotonic muscular

dystrophy, Spinal muscular dystrophy, distal muscular dystrophy and Charcot Marie tooth disease are the few rare forms of muscular dystrophy [12].

Duchenne muscular dystrophy (DMD) is the most common form of muscular dystrophy which is X-Linked type and caused is by the mutations in the dystrophin gene located on the X chromosome. Dystrophin is the massive human gene that is 2.5MB long and encompasses of 79 exons. The absence of dystrophin gene occurs when a large number of exons are deleted, which is the major cause of DMD [13]. DMD causes out frame deletions that happen in the piece of the codon and the sequence read cannot be done post occurrence of deletion mutation. The patients affected by DMD are diagnosed around children in five years of age when the physical ability deviates obviously from their companion. When untreated, the strength of the muscle strength gets worse, and boys are wheelchair dependent at their early stages of the life. The other complications like respiratory, orthopedic, and cardiac emerge, that shortens the life of the patients [14].

Becker muscular dystrophy (BMD) is also one of the X-Linked and is as well caused by the mutations in the dystrophin gene located on the X chromosome. It upholds muscle fiber strength, reduces muscle rigidity and increases sarcolemmal deformability. Less defective mutations in the dystrophin gene result display a much milder dystrophic phenotype in affected patients, known as Becker's muscular dystrophy [15]. BMD causes in frame deletions that take place beyond the codons and the sequence still can be read after deletions.

Emery-Dreifuss muscular dystrophy (EMD) can be affected in patients, typically in their childhood and in the early adolescent years with muscle contractures. The symptoms include cardiac conduction defects, muscle weakness and arrhythmias. If the patients left untreated in the early stage, it leads to increasing the risk of stroke and sudden death. The mutations in the Emerin (EMD) and Lamin A/C (LMNA) genes cause Emery- Dreifuss muscular dystrophy. Mutations like point mutations, insertions and deletions in the genes direct to EMD. X-Linked, autosomal dominant and autosomal recessive are three subtypes of EMD muscular dystrophy disease. Each type varies in their prevalence and symptoms.

Limb-girdle muscular dystrophy (LGMD) can be seen in both boys and girls. Nearly mutations in 18 genes are the reason of LGMD. The defects in LGMD show a related distribution of muscle weakness that has an effect on both upper arms and legs. The different

patterns of inheritance in LGMD are autosomal and recessive. Missense, insertions and deletion mutations in the genes route to LGMD.

Charcot Marie tooth disease (CMT) includes a number of disorders with an assortment of symptoms grounds damages in peripheral nerves. The disorder affects the peroneal muscle in the lower leg and hence the disease also is known as hereditary motor and sensory neuropathy (HMSN) and peroneal muscular atrophy [16]. CMT causes mild and also severe muscle degeneration, which is dependent on its mutation. There may be mild problems limited to skeletal muscle and also a severe problem like muscle degeneration corresponding with upshot on the brain. More than 30 forms of CMT are noticed and 30 genes are concerned, some may show severe brain malformations, such as lissencephaly and hydrocephalus and hearing loss.

The Facioscapulohumeral Muscular Dystrophy (FSHD) is an autosomal dominant neuromuscular disorder. The deletions of D4ZA microsatellite repeats in DUX4 gene on chromosome 4q cause Type 1 FSHD. Mutations such as missense, splice site and small deletions in SMCHD1 gene reflects in Type2 FSHD. The weakness of muscles in the shoulder, upper arm muscles, shoulder girdle, stomach and lower limbs results in FSHD [17].

The Myotonic dystrophy is also known as Steinert's disease. The expansion of an unstable CTG trinucleotide repeat in the DMPK gene on chromosome 19 is the base for this disease. The normal individual has the repeats ranging between 5 and 37, if the repeats exceed 50 then it causes myotonic dystrophy. CTG repeats sizes ranges from 50 to 4000 [18].

Distal muscular dystrophy (DD) also known as Distal myopathy is a group of disorders that mainly affect distal muscles. The distal muscles that are located in the hands, feet, lower arms or lower legs are get flawed in this type of muscular dystrophy. There are about eight forms of distal myopathy caused by the defects in various genes.

Spinal muscular atrophy (SMA) is an autosomal recessive neuromuscular disease that results in progressive proximal muscle weakness and paralysis exemplify by degeneration of alpha motor neurons in the spinal cord. On the basis of the age of onset the patients SMA is classified into four types based on the age of the patients. The SMN1 gene is responsible for this genetic alteration that results in a reduction of survival motor neuron (SMN) protein [19, 20]. Different forms of muscular dystrophy disease types are tabulated in Table 1.1.

Table.1.1 Muscular Dystrophy Disease Types

Disease Type	Age at onset	Symptoms, rate of progression and life expectancy
DMD	2 - 6 years	General muscle weakness and wasting that affects pelvis, upper arms and upper legs.
BMD	Youth to early adulthood	Symptoms almost same as Duchenne, but are less defective and progresses slowly than Duchenne.
Emery-Drefiuss	Childhood to early teens	Weakness and wasting of muscles in shoulder, upper arm and shin muscles. Sudden death may occur from cardiac problems.
Limb griddle	Late childhood to middle age	Weakness and wasting of muscles in shoulder griddle and pelvic griddle. Progression is slow.
Fascioscapulohumeral	Childhood to early adults	Weakness and atrophy of the muscles around the eyes and mouth, shoulders, upper arms and lower legs. Weakness can spread to abdominal muscles and to hip muscles.
Myotonic	20 to 40 years	Muscle wasting and weakness. Clouding of the lens of the eye and abnormalities of the electrical signals that control the heartbeat. infertility in men
Distal muscular dystrophy (DD)	40 to 60 years	Weakness and wasting of distal muscles in forearms, hands, lower legs and feet.
Spinal muscular atrophy	From birth to early ages	The muscles of the shoulders, hips, thighs and upper back. Special complications occur in the muscles, which are used for breathing and swallowing.
Charcot marie tooth disease	Early childhood to 30 years or 40 years	Mild and also severe muscle degeneration. Loss of touch sensation in the feet, ankles, legs, hands, wrists and arms occur with various types of the disease.

Diagnostic Methods based on Mutations

The disease can be diagnosed using the laboratory approaches with the results of muscle biopsy, electromyography, electrocardiography and DNA analysis.

Serum creatinine kinase is a straightforward and economical indicative test for severe forms of dystrophy. The analysis is done by measuring the serum concentration of creatinine kinase. The higher concentrations of serum creatinine kinase than normal values suggest a disorder. In the diagnosis of DMD, serum creatinine kinase concentrations are extracted from the newborns, and the early diagnosis is done by testing in neonates which helps in reduction of disease further in the family. This method does not diagnose all forms of dystrophy [21].

Electromyography testing (EMG) is done in two phases. In the first phase a small needle is gently inserted into the electrical patterns of the muscles in the arm or thigh. The second phase determines the speed of the messages is being sent from the brain to nerves by stimulating the nerves of either arm or leg through a small electrical pulse being sent from the brain to the nerves. EMG test is uncomfortable, painful, lengthy procedure. EMG testing is less favored for children and it is mostly performed only on adults for disease identification. EMG tests are done mainly for the investigation in myotonic dystrophy. The performance of EMG is not satisfied for the patients having less creatinine kinase.

Muscle biopsy and DNA testing are widely used tests to predict muscular dystrophies. A muscle biopsy is a surgical practice where a tiny sample of a muscle is extracted and analyzed. The removal of muscle tissue is done using a biopsy needle and microscopic analysis is done to examine the level of the genes that cause muscular dystrophy. Performing muscle biopsy is costly, it is invasive, and at most care should be taken after the surgery. A muscle biopsy might be considered if speedy and trustworthy genetic testing is unavailable.

The clinical diagnosis of DMD is done through the laboratory analysis of dystrophin. The methodologies engage in recent dystrophin diagnostic experiment includes multiplex PCR, Multiplex ligation-dependent probe amplification (MLPA), Southern blot analysis, Detection of virtually all mutations-SSCP (DOVAM-S). The exact mutation in the DMD gene can be analyzed using gene therapy and missense, nonsense, insertions, deletions and splicing mutations are identified through direct sequencing [22, 23]. Molecular diagnostic methods at nucleotide level are required. The direct sequencing analysis is considered to be laborious, expensive and time-consuming. In some cases, the MLPA reports will be negative and point mutation detected

by the Sanger method requires direct full gene sequencing, and hence the role of direct sequencing in diagnosis of DMD is increased [24].

Polymerase chain reaction (PCR) is now common and often indispensable technique used in medical and biological research labs in the diagnosis of hereditary diseases. PCR has the benefit of being minimally invasive, efficient and very specific for the detection of large gene deletions. The major drawbacks in this approach are a lack of antimicrobial sensitivity data, complexity of the assay, and the price of PCR equipment and kits [25]. The demerits of these technologies are lengthy, painstaking procedure, and not able to detect duplication mutations precisely [26].

Genetic testing is an initial step tested on a blood sample to spot the alteration in the genes so as to help in the diagnosis of muscular dystrophy without performing a muscle biopsy. The risk involved in DNA analysis or genetic testing is minimal and the traits can be identified effectively as the disease-causing genes are explicitly known. Carrier mothers, those who may be at risk of passing this disease on to their children are identified by genetic testing and preventive measures can be provided [27]. To find the mutations in the genes for the patients identified through muscle biopsy, the genetic testing is again performed to confirm the diagnosis. However, the muscle biopsy is optional for the patients diagnosed by genetic testing, to distinguish from other phenotypes [28].

Genetic testing is also an option to confirm a Muscular dystrophy disease. As disease has several subtypes and there are different genes responsible for each subtype, it is important to narrow the possible type of disease as much as possible using the previously mentioned tests. If the gene change can be found and confirmed, this information can then be used to help in testing other family members to determine whether they are carriers of the disease [29].

Most of the genetic disorders caused by the mutations in the gene sequences show an impact on splicing. Bioinformatic tools that are designed to assess the impact of genetic variation on splicing are NNSplice [30] MaxEntScan [31], ESEFinder [32], Spliceman [33], Skippy [34] and Human Splice Finder [35]. Skippy is a web-based tool that defines exonic variants using the genomic features that modulate splicing. Single nucleotide variants relevant to splice-modulating genomic features variants are assessed and scored. Point mutations lay in the coding region show severe effects on gene function through disruption of splicing.

1.4 Review of Literature

Machine learning techniques have been successfully applied to identify disease-associated genes. The problem is formulated as a supervised learning problem, where the task is to make the classifiers to learn from training data and the prediction is made from the learned classifier. Several models were built to predict various diseases using machine learning techniques by extracting essential features. Various types of gene or protein annotation data used to solve the disease gene classification problem.

Schizophrenia is a genetic disease and also a heterogeneous syndrome characterized by perturbations in language, perception, thinking and social relationships. There is no set of symptoms finalized to categorize this disease other than the genetic factors. Disease gene association studies focused on SNP (Single Nucleotide polymorphism) aids in predicting the disease [36]. Seven datasets have been used containing 48 SNPs at the DRD3 and HTR2A genes associated with schizophrenia from the Galician population.252 classification models have been obtained using SNPs at two schizophrenia-related genes. Twelve machine learning techniques such as Linear Neural Networks, Multilayer Perceptron, Radial Base Functions, Bayesian Networks, Naïve Bayes, Support Machine Vectors, Decision Tables, Decision Table Naïve Bayes Hybrid Classifier, Best-First decision Tree classifier, Adaptive Boosting, Evolutionary Computation and Multifactor Dimensionality Reduction and seven datasets. The best relationships between the DNA molecule sequence and schizophrenia evaluated 78.3–93.8% of the DNA sequence from schizophrenia patients, for datasets with extra simulated negative subjects.

In [37] the authors, build a model to classify the types of spinal muscular atrophy. The biomarker studies are captured using Quantitative Muscle Ultrasound (QMU) and Electrical Impedance Myography(EIM). Features are extracted from ultrasound and EMI. Support vector machine is used to build the models to classify SMA2 muscles from SMA3 muscles and obtained an accuracy of 92.8%.

FSHD (Facioscapulohumeral Muscular Dystrophy) is an autosomal dominant neuromuscular disorder. The authors in [38] diagnosed FSHD through machine learning techniques with gene expression profiling data. This paper mainly aims in finding genes to discern between healthy cases and FSHD affected cases. Gene expression data from two databases are downloaded for this purpose. The Linear Discriminant analysis (LDA) and Linear

SVM is employed for classifying the healthy and non-healthy cases. To improve the accuracy feature selection algorithm is used to rank the features. 85.2% of accuracy is obtained while classifying the data.

Muscular dystrophy and its subtypes are classified by integrating protein-protein interaction (PPI) network, using interpretable gene set information and mRNA profiling data. Identification of gene sub-networks is done using a distance metric approach named affinity propagation clustering (APC) approach. The biomarkers are identified the functional gene set information is combined. Classification of muscular dystrophy is done with multi-class support vector machines (MSVMs) with the gene set features and sub networks. The proposed scheme is applied to gene expression data set to classify six different MD sub-types for their improved diagnostics [39]. Performing various analysis on the PPI data and combining the features accuracy of the classifier ranges from 72% to 90% of accuracy.

Only very few research has been carried out on gene sequences based on RSCU (Relative Synonymous Codon Usage) to predict or classify either type of gene or virus or diseases. The authors in [40] proposed a model to classify the types of Human Leukocyte Antigen (HLA) gene into different functional groups by choosing the codon usage bias as input. In their work they converted the gene sequence into 59 vector elements by calculating the RSCU values for the gene sequence. A model was created using Support vector machine and achieved an accuracy rate of 99.3 percent.

The authors Nisha C M, Bhasker Pant, and Pardasani K R proposed a new approach based on codon usage pattern to classify the type of Hepatitis C virus (HCV) that is the primary reason for the liver infection. To classify the subclass of its genotype a model was created using codon usage bias as input to multi class SVM [41]. 100% of accuracy is attained while classifying the 5 kinds of genotypes. The authors in [42] employed a machine learning approach based on ensemble classifier LibD3C to predict the cytokines. The analysis was made on the physicochemical properties and the distribution of whole amino acids. The cytokines are classified using the protein data and 93.3% of accuracy is attained.

The authors in [43] identified large mutations such as duplications and deletions through computational approach. A system SPeeDD was developed by utilizing the Logical Model Tree method based on machine learning technique for the gene BRCA1. High specificity was achieved with this technique. The authors in [44] predicted the disease-causing mutations

through ensemble learning technique. The protein sequence dataset from Swiss-Prot database was used for classification. A comparative analysis was made between the traditional approach and ensemble approach and LogitBoost ensemble technique achieves high performance among all the methods compared.

Mutpred splice is a machine learning approach for identifying the coding region substitutions that disrupt pre-mRNA splicing. Disease-causing splice altering variants, disease-causing splice neutral variants and polymorphic splice neutral variants are considered and discriminative descriptors are extracted from gene sequences. Supervised classification techniques such as Random Forest and SVM are employed for building models [45].

A deep learning model is made up of numerous computational layers that handle data in a hierarchical fashion. Every layer gains an input and generates an output, as a non-linear function of a weighted linear mixture of the input values. The end product of one layer becomes an input to the next processing layer, creating a deep architecture [46]. In each consecutive layer, the data have been indicated in a very more abstract way. In contrast to the shallow architectures which solely hold a limited feature layer and a weight-combination layer, deep architectures refer to the multi-layer network whereas every two neighboring layers are linked to each other in some manner. The sheer volume of data makes it usually unattainable to train a deep learning algorithm with a central processor and storage. Distributed environment with parallelized machines is chosen [47]. The creative models utilize clusters of CPUs or GPUs in increasing the training speed without devastating accuracy of deep-learning algorithms. Despite the fact that deep learning algorithms are not parallel, data and models are divided into blocks of in-memory data, and the forward and backwards propagations could be carried out in a parallel fashion [48, 49].

In the recent years, there has also been an updated attraction to the field of deep learning and the innovative research in the area of medical imaging using deep learning reveals a guaranteeing outcome. It has got acquired great successes in a broad area of applications such as speech recognition, computer vision, and natural language processing. Many challenging and complex problems for deep learning are there in the field of computational biology. It is appealing to observe that these techniques have not been utilized for solving these problems and also thought that these methods can be utilized to perform effective analysis of biological data

and they can provide biological insights by extracting highly effective features from the data [50]. Nearly all the papers identify the diseases from the image data.

An overview of the current state of the art deep learning architectures and optimization techniques is proposed in [51]. The ADNI hippocampus MRI dataset is used to diagnose the Alzheimer's disease using 3- dimensional hippocampal segmentation is proposed in this work. The author investigated the use of three different convolutional network architectures for patch-based segmentation of the hippocampi region in MRI images. The authors in [52] proposed a CAD system is proposed to classify lung nodules as either malignant or benign. This system uses deep features extracted from an autoencoder. 4303 instances containing 4323 nodules from the National Cancer Institute (NCI) Lung Image Database Consortium (LIDC) dataset is used as dataset. An overall accuracy of 75.01% with a sensitivity of 83.35% and false positive of 0.39/patient is attained over a 10 fold cross validation.

The authors in [53] proposed a deep CNN to classify lung CT image patches into 7 classes, that includes 6 different ILD patterns and a healthy tissue. A dataset of 14696 image patches, derived by 120 CT scans is used. The training was performed by minimizing the categorical cross entropy with the Adam optimizer. The classification performance 85% demonstrated the potential of CNNs in analyzing lung patterns. The feasibility of constructing a universal skin disease diagnosis system using deep convolutional neural network (CNN) is proposed in [54]. The CNN architecture is trained using the 23,000 skin disease images from the Dermnet dataset and its performance is tested with both the Dermnet and OLE images. The system achieved 73.1% Top-1 accuracy and 91.0% Top-5 accuracy when testing on the Dermnet dataset. For the test on the OLE dataset, Top-1 and Top-5 accuracies are 31.1% and 69.5%.

The Convolution neural networks used by DeepBind and DeepSEA are available deep learning approaches in computer vision. Various deep learning algorithms such as convolution neural networks, belief decision trees, Restricted Boltzmann Machines (RBM), Recurrent neural network (RNN) are used to obtain the results.

Deep Neural Networks (DNN) is adaptable stacked layered systems connected and communicating artificial neurons that perform different data transformations.DL4 techniques model high-level representations of data utilizing deep neural networks (DNNs). They possess a number of hidden layers of neurons, which number variation enables adjusting the degree of data abstraction. DNN is a set of machine learning algorithms supervised and unsupervised, strongly

dependent on the selection of representing the data that can be utilized on several layers of nonlinear output [55]. Such type of studies of deep neural network in supervised learning are to predicting noncoding RNA[56], predicting pharmacological characteristics of drugs and drug repurposing[57], protein-protein information extraction[58], Annotating the pathogenicity of genetic variants DANN[59].

The authors in [60] developed a DeepGDashboard for visualizing gene sequences using deep neural network. The models are trained and tested using 108 K562 cell ENCODE ChIP-Seq TF datasets. Several variations of each DNN architecture is implemented by varying the hyper parameters.The visualization is applied on the best performing models of each of the three DNN architectures.

The authors JiaMing Liu, et.al in [61] introduced a pretrained deep neural network (DNN), to the cough classification problem. Pretrain and fine-tuning are the two steps involved in building the deep neural network models, which is followed by a Hidden Markov Model (HMM) decoder to capture tamporal information of the audio signals. The experiments were conducted on a dataset that was collected from 22 patients with respiratory diseases. The results indicates that HMM-DNN framework performs better than the conventional GMM-HMM model.

The above pioneered literatures were the applications of deep neural network with biological background. None of the models involves gene sequences as input to predict the type of disease. Gene sequence can be considered as a sequence of categorical values. Machine learning based methods typically employ an encoding scheme to convert a DNA sequence into its numerical representation for downstream processing. Four types of representation for mapping DNA sequence into numerical values are proposed in [62]. Fixed mapping techniques, physicochemical property based mapping, Integer and real mapping techniques are employed to identify the donor and acceptor site from the DNA sequences using the Artificial neural network.

The authors in [63] use one-hot vector encoding to represent the gene sequences in numerical format. The DNA sequences are translated into words of size 3, and each word is represented by a one-hot vector of size 64. DNA sequence classification is achieved by using the convolution neural network. Computation prediction of splice junctions were achieved with deep belief neural network by encoding the DNA sequences. Nc bit 1-hot encoding is employed for 1-hot encoding, For nc = 4, A, C, G, and T are encoded by 1000, 0100, 0010, and 0001, respectively [64].

The classification of muscular dystrophy continues to evolve with the advances in understanding of their molecular genetics. A huge number of muscular dystrophy related defective genes and proteins are identified, but no effective treatments are known for many of its subtypes. At present, there is no effective method to identify and classify all types of muscular dystrophy. The proportion of mutations in deletions, duplications and point mutations differs in each type of disease and to date, no genetic testing has been developed to cover this whole mutational spectrum in a single platform. Large size and number of genes for all types of muscular dystrophy requires considerable effort, cost and time for direct sequencing. The direct sequence analysis of this spectrum involved in all kind muscular dystrophy requires a high level of the laboratory. However, it is more important to know the exact mutation site and type to predict prognosis and, therefore, all the mutation sites should be analyzed effectively. The review of literature is summarized in Table 1.2.

Table 1.2 Summary of the Existing Work

Disease	**Data**	**Approach**	**Algorithm (method)**	**Accuracy (%)**
DMD & BMD	Gene Sequences	MLPA – Laboratory	mPCR	75
DMD	Gene Sequences	MLPA – Laboratory	DHPLC	86
LGMD	Family Details	Machine Learning	ANN	98
Schizophrenia	SNP	Machine Learning	12 Supervised learning techniques	78.3–93.8
Spinal muscular atrophy	(QMU) and (EIM)	Machine Learning	SVM	92.8
6 types of MD	Micro array – Protein protein Interaction	Machine Learning	MSVM	72 -90
FSHD	Microarray	Machine Learning	SVM	85.2
Gene type Classification	HLA Gene	Machine Learning	SVM	99.3
Virus Type Classification	HCV Virus	Machine Learning	SVM	100
Cytokines	Physicochemical properties	Ensemble Learning	LibD3C	93.3

SPeeDD	BRCA gene	Machine Learning	Logical Model Tree	High specificity
Disease causing mutations	Protein sequence	Ensemble Learning	LogitBoost	99
Mutpred splice	Splicing variants	Machine Learning	Random Forest and SVM	98
Alzheimer's disease	ADNI hippocampus MRI dataset	Deep Learning	CNN	High performance on deep learning
Lung nodules classification	Lung Image Database Consortium	Deep Learning	DNN	75.01
Lung disease classification	Lung CT image	Deep Learning	CNN	85
Skin disease classification	Dermnet dataset	Deep Learning	CNN	91
Cough classification	Patient's Audio signals	Deep Learning	HMM-DNN	90
DNA sequence classification	DNA sequences	Deep Learning	CNN	96
Prediction of Splice junctions	DNA sequences	Deep Learning	DBN	95

Motivation

From the background study, it was observed that the laboratory approaches involve more cost for classifying the disease sequences to predict the type of muscular dystrophy. The disease identification is mainly based on direct sequencing, which is a tedious process and also accuracy may not be achieved. Few forms of muscular dystrophy are identified through computational methods based on full direct sequencing and micro array data. Usage of microarray gene expression data is convincing when multiple genes involved in a disease and also hereditary traits cannot be detected efficiently. Tests for only a few types of muscular dystrophy disease are already in clinical use. The apparent benefit of hereditary testing aids in identifying and understanding of risk for a certain type of disease. Predictive hereditary tests for all types of muscular dystrophy need to be done.

The current advancements in gene testing helps in identifying people at a risk of getting a disease in advance in ahead of any symptom appears. An accurate gene test results in finding the disease-related gene mutation. Identification of genetic factors in complex diseases like muscular dystrophy is a far more difficult task with the standard methods as it is difficult to analyze the data. The complex diseases provide a lot of challenges to standard data analysis techniques. Therefore, it is essential to model and represent this knowledge in a computational form with minimal loss of biological context through a gene sequences based approach. Disease-gene associations need to be designed and a suitable classification algorithm should be employed to handle this type of data.

Hence, it is proposed to model the muscular dystrophy disease identification problem as pattern recognition task and to provide solution using machine learning techniques. The intricacies involved in disease identification need to be analyzed and taken into account while modeling the disease identification task by considering the appropriate mutational features from sequence data. It is clear that machine learning methods can be used to significantly improve the accuracy of muscular dystrophy prediction model. It allows the clinician to diagnosis without needing a muscle biopsy and raises the clinician response time and helps to treat disorders.

An integrated approach based on computational intelligence technique should be demonstrated to detect major five forms like Duchenne muscular dystrophy (DMD), Becker's muscular dystrophy (BMD), Emery drefius muscular dystrophy (EMD), Limb Griddle muscular dystrophy (LGMD), Charcot Marie tooth disease (CMT) of muscular dystrophy with cloned gene sequences as input. Two extreme approaches can be employed for identifying muscular dystrophy disease through hand crafted and self extracted features. Data driven models with hand crafted mutational features pertaining to all kinds of mutations can be developed using supervised learning techniques to predict the disease precisely. Also a deep neural network approach which can learn complex and abstract features automatically from unlabelled data can also be employed to eliminate the extraction of features from this kind of multidimensional data.

1.5 Objectives of the Research

The main intention of this research work is to propose models for predicting the major five forms of muscular dystrophy disease such as Duchenne muscular dystrophy, Beckers muscular dystrophy, Emery derfiuss muscular dystrophy, Limb griddle muscular dystrophy and charcot marie tooth disease automatically from cloned gene sequences through shallow and deep learning. The core objectives of this research work are as follows

- To generate synthetic gene sequences by adopting positional cloning approach
- To identify and capture discriminative descriptors from the diseased gene sequences related to the mutations like missense, nonsense, silent, insertion, deletion, duplication and splicing.
- To build autonomous disease identification models based on all kind of mutational features using supervised learning techniques
- To build muscular dystrophy disease identification model using ensemble learning approach
- To define two mapping schemes namely nucleotide mapping and codon mapping schemes to encode the diseased gene sequences to encode gene sequences for deep learning
- To build Deep Neural Network classifier for predicting the category of diseased gene sequences into five types of muscular dystrophy in tensorflow environment

The thesis explains a novel and never before tried methodology in disease identification model wherein the muscular dystrophy disease identification problem is modeled as a classification problem. Muscular dystrophy disease classification is done based on both hand crafted mutational features and self extracted features through shallow and deep learning. These approaches for disease identification exceedingly simplify the process of traditional disease identification problem and the prediction model is more effective, reliable since it is generated based on intelligent hints collected from mutated gene sequences. In this work, the mutation spectrum covers all types of mutation for modeling and therefore the task of full sequencing is eliminated. This approach generalizes the disease identification task as an automated practice, which can be applied to identify any kind of genetic disease.

1.6 Organization of the Thesis

The rest of the thesis is structured as follows:

Chapter 2 presents some of the supervised learning techniques adopted in this work for pattern classification. Ensemble learning and LibD3C classifier is also explained briefly. Chapter 3 presents with a background study on deep learning.

The modeling approach used to design a muscular dystrophy disease identification problem is addressed in Chapter 4. The formulation of muscular dystrophy disease identification problem as pattern recognition problem is explained in detail in this chapter. Acquisition of raw gene sequences from various databases, development of corpus by applying positional cloning approach and framework of the disease identification model are also discussed.

In Chapter 5, the overview of the design and development of the disease identification model through the shallow learning approaches is elucidated. The discriminative mutational features from the mutated gene sequences are identified and extracted to build data driven models for predicting the type of the genetic disorder. The implementation of autonomous muscular dystrophy disease identification models using traditional supervised learning algorithms are detailed in this chapter. The performances of the models have been described with the results and findings.

In chapter 6, muscular dystrophy disease identification model built through ensemble approach with LibD3C classifier is described. Ensemble learning technique is adopted to analyse the performance of the classifier through combined learning technique based on D3C strategy. Various exhibits of the findings about muscular dystrophy disease identification models using LibD3C classifier is compared with the performance of the models built using supervised learning algorithms and the results are presented in this chapter.

Muscular dystrophy disease identification model using deep learning approach is described in chapter 7. Two mapping schemes namely nucleotide mapping and codon mapping schemes have been employed to encode the diseased gene sequences into numerical format are explained in detail. Results of shallow net using hand crafted features is compared against the results obtained from deep neural network through self-taught learning and the same is discussed in this chapter.

Finally, in chapter 8 the research work is concluded by giving an outline of whole process with various findings. This chapter summarizes the research contributions, discusses the results

and findings, sum up the achievements of the proposed muscular dystrophy disease identification models, and presents recommendations for future research.