

4. PROBLEM MODELING

The principal focus of this thesis is to propose an efficient machine learning solution for predicting the type of muscular dystrophy disease. This chapter deals with problem modeling wherein the muscular dystrophy disease identification task is modeled as a classification problem. This methodology simplifies the disease identification problem to a great extent and provides a suitable solution for muscular dystrophy disease identification problem by using shallow and deep learning approaches.

The arrangement of the bases in the gene sequences differs in every human. Occurrences of mutations alter the pattern of a gene sequence. Genetic diseases are identified by capturing the alteration in this pattern. The availability of diseased gene sequences is a challenge for this intricate muscular dystrophy disease and that stimulates the need for the generation of synthetic mutational gene sequences. The process of corpus development is also described in this chapter. Multi-class classification for disease identification is worked out through data modeling of gene sequences.

4.1 Data Acquisition Through Positional Cloning

Positional cloning is a traditional approach to view the alteration in the gene sequences. Positional cloning aids in disease identification even when minute information is known about the molecular basis of the trait. The first gene cloned by positional cloning methodology is dystrophin gene to identify of Duchenne Muscular Dystrophy (DMD) through laboratory methods such as Polymerase Chain Reaction (PCR)[117]. This approach mainly focuses on the structural features, i.e. annotation features of the gene. The steps involved in positional cloning are (i) identifying a candidate gene based on the chromosomal location (ii) retrieving genomic clones in the mapped region (iii) observing and analyzing the exons (iv) isolating the cDNA (v) characterization and mutation analysis. The method used to determine the candidate region is termed as linkage analysis.

Positional cloning is a fundamental step in diagnosing the genetic disorders such as Ducheane muscular dystrophy, Huntington's disease and Cystic fibrosis as the diseased gene sequences are not explicitly available [118]. The missense mutations in the genes were identified by positional cloning methods for the diseases such as cardiac arrhythmia. A

combined strategy of linkage analysis and next-generation sequencing (NGS) technology is applied to diagnose the disease acute myeloid leukemia (AML) with missense mutations. The mutations have been identified using the positional cloning approach and hence this method is applied in this research work to generate synthetic diseased gene sequences by encoding all the required genes.

Normally, this approach is done through laboratory methods like PCR and NGS but in this work, R scripts have been written for sequence generation by aligning the cDNA and reference gene sequences and store it as fasta files. The information on the position of mutations in the gene sequences is available in HGMD database, which are grasped from various literatures. The positional change is done in cDNA sequence against the reference gene sequence and the new mutated gene sequences for muscular dystrophy are generated through R script.

4.2 Corpus Development

The muscular dystrophy disease model is built by extracting the discriminative features from the mutated gene sequences. The mutated gene sequences are generated with the information retrieved from the HGMD -Human Gene Mutation Database with its corresponding reference gene sequence from NCBI - National Center for Biotechnology Information and OMIM - Online Mendelian Inheritance in Man databases.

Online Mendelian Inheritance in Man (OMIM) is a catalog of human genes and genetic disorders and traits. This database mainly focuses on the gene-phenotype relationship [119]. Each OMIM entry has a full-text summary of a genetically determined phenotype or gene and has numerous links to other genetic databases such as DNA and protein sequence, PubMed references, general and locus-specific mutation databases, HUGO nomenclature, MapViewer, GeneTests. Information in OMIM can be retrieved by queries on MIM number, disorder, gene name and gene symbol. Each OMIM entry is assigned a unique six-digit number whose first digit indicates whether its inheritance is autosomal, X-linked, Y-linked or mitochondrial. The six digit number assignment is as follows.

- 1- Autosomal loci or phenotypes (Entries before May 1994)
- 2- Autosomal loci or phenotypes (Entries before May 1994)
- 3- X-linked loci or phenotypes

- 4- Y-linked loci or phenotypes
- 5- Mitochondrial loci or phenotypes
- 6- Autosomal loci or phenotypes (Entries after May 1994)

1 and 2 indicates that the inheritance occurred is autosomal loci, where the entries are entered before the year 1994. X-linked inheritance phenotypes are noted by 3 and Y-linked inheritance loci are denoted by 4. 5 indicates that the inheritance is mitochondrial phenotype. 6 indicates that it is also autosomal loci, but the entries are entered after the year 1994.

For example, one of the OMIM entry for Emerin gene affected by Emery dreifuss Muscular dystrophy is 300384. Here, it indicates that it is an X-linked loci as the first digit is 3. The gene symbol is EMD and it gives the information about the diseases, which are likely to be caused by this gene. Association of various genes for the same disease is noted from the entry.

As many genes are responsible for the same phenotype, the genes associated with the disease are carefully examined using the OMIM database. In this research work, five types of muscular dystrophy diseases are taken into account and its corresponding genes are found out using this catalog. Various genes associated with different types of muscular dystrophy diseases are tabulated in Table 4.1.

Table 4.1 Genes Associated with Different Types of Muscular Dystrophy

Muscular dystrophy disease	Genes
Duchenne muscular dystrophy	Dystrophin
Becker's muscular dystrophy	Dystrophin
Emery-dreifuss muscular dystrophy	Emerin LMNA/C
Limb girdle muscular dystrophy	ANO5, CAPN3, CAV3, DYSF, FKRP, FKTN, LMNA, MYOT, POMGNT1, POMT1, POMT2, SGCA, SGCB, SGCD, SGCG, TCAP, TRIM32, TTN
Charcot marie tooth disease	AARS, AIFM1, BSCL2, DHTKD1, DNM2, DYNC1H1, EGR2, FGD4, FIG4, GARS, GDAP1, GJB1, HSPB1, HSPB8, INF2, KARS, KIF1B, LITAF, LMNA, LRSAM1, MED25, MFN2, MPZ, MTMR2, NDRG1, NEFL, PMP22, PRPS1, PRX, RAB7A, SBF2, SH3TC2, TRPV4, YARS

HGMD is a core collection of data on germ-line mutations in the genes coupled with the human inherited disease. It is a catalogue of all types of mutation. The positional information is retrieved from public version of HGMD database. The public version of HGMD (<http://www.hgmd.org>) is freely available to registered users from academic institutions/non-profit organizations [120]. The catalogue includes the information available in the database are accession id, codon change, amino acid change, nucleotide change, protein change, variant class, reported phenotype, journal reference and its corresponding cDNA sequence. This database consists of seven different variant classifications, which are summarized, in Table 4.2. The policy of this database is to upload any variation into the database that has been associated with disease, even if the functional aptitude is unclear. All such variations are clearly indicated.

Table 4.2 Definitions of HGMD Classifications

HGMD Classification	Definition
DM	Disease Causing Mutation
DM?	Probable/possibly disease-causing mutation; author expresses uncertainty
CNV	Copy number variations
FTV	Frameshift or truncating variant
FP	In vitro/laboratory or in vivo functional polymorphism
DFP	Disease-associated polymorphism with additional supporting functional evidence
DP	Disease-associated polymorphism

In this work, seven kinds of mutations such as missense, nonsense, silent, insertions, deletions, duplications and splicing are taken into account for building the disease identification models. Therefore, the corresponding mutation records are identified and captured from the database by specifying the required information. Fig.4.1. shows a sample of HGMD information that is retrieved for the DMD gene by providing the necessary details to HGMD database.

The screenshot shows the HGMD website interface. At the top, there's a search bar with 'Gene symbol: DMD' and a dropdown menu for 'Missense/nonsense'. Below the search bar, there's a summary of mutation types: Missense/nonsense (498), Splicing (37), Regulatory (3), Small deletions (14), Small insertions (12), Small indels (14), Gross deletions (184), Gross insertions (49), Complex (48), and Repeats (36). A table of mutations is displayed below, with columns for Accession Number, Codon change, Amino acid change, Codon number, Genomic coordinates, Phenotype, Reference, and Comments. The table lists several mutations, including TGG-TGG (Trp-Ter) at codon 3, GAG-AGT (Lys-His) at codon 18, and CAG-TGG (Gln-Ter) at codon 45.

Accession Number	Codon change	Amino acid change	Codon number	Genomic coordinates & RefSeq coordinates	Phenotype	Reference	Comments
CM031161	TGG-TGG	Trp-Ter	3	100000000-100000000	Muscular dystrophy, Becker	Phinney (2001) <i>Am J Hum Genet</i> 74: 931	Additional report available via PubMed
CM092281	TGG-TGG	Trp-Ter	4	100000000-100000000	Muscular dystrophy, Becker	Gomoch (2009) <i>Hum Mutat</i> 30: 433	
CM024512	GAG-AGT	Lys-His	18	100000000-100000000	Cardiomyopathy, dilated	Feng (2003) <i>Mol Genet Metab</i> 77: 119	Additional report available via PubMed Additional report available via PubMed
HM080103	CAG-TGG	Gln-Ter	35	100000000-100000000	Muscular dystrophy, Duchenne	Gomez-Panichi (2009) <i>Hum Genet</i> 128: 318	
CM050377	CAG-TGG	Gln-Ter	45	100000000-100000000	Muscular dystrophy, Duchenne	Burns (2005) <i>Hum Mutat</i> 25: 177	
CM081570	GAT-GTT	Asp-Val	46	100000000-100000000	Muscular dystrophy, Becker	Keenan (2009) <i>Hum Mutat</i> 29: 728	
CM930191	CTC-CCG	Leu-Arg	54	100000000-100000000	Muscular dystrophy, Duchenne	Phar (1991) <i>Nat Genet</i> 4: 337	Additional report available via PubMed Additional report available via PubMed
CM0910024	GAG-TGG	Glu-Ter	55	100000000-100000000	Muscular dystrophy, Becker	Sedlitz (2009) <i>Neurogenet Disord</i> 19: 748	
CM940334	CAG-TGG	Gln-Ter	60	100000000-100000000	Muscular dystrophy, Duchenne	Roberts (1994) <i>Hum Mutat</i> 4: 1	
CM022943	GAG-TGG	Glu-Ter	65	100000000-100000000	Muscular dystrophy, Duchenne	Dobson (2001) <i>Neurogenet Disord</i> 17: 845	

Fig.4.1 HGMD Information for DMD Gene

Various types of genes associated with the five types of neuromuscular disorder are studied. An analysis is made of fifty-five genes that are associated with five types of muscular dystrophy like (i) DMD, (ii) BMD, (iii) EMD, (iv) LGMD, (v) CMT. Several types of mutated sequences based on mutations like Missense, Nonsense, synonymous, Insertion/duplication, deletion mutations and splicing mutations are collected. The mutational information is extracted from the HGMD database using the gene information for the required phenotype. For example, Emery-Dreifuss disease (EMD) is affected by the mutations in the Emerin and LMNA gene. To view the mutational information for EMD, the search is made on the gene name such as Emerin and the details are captured. The same procedure is followed for LMNA gene and mutational information for EMD is extracted.

The raw cDNA sequence is downloaded from HGMD. The reference gene sequences are downloaded from NCBI database [121]. This sequence is required to generate the mutated gene sequence, which is done by blasting the raw sequence against the reference sequence. To calculate the similarity scores in the gene sequences and protein sequences BLAST algorithm is used. It is a powerful tool which searches for the sequences which are related to the query sequence within the same organism or in different organisms. The query sequence is searched on NCBI databases and the results are posted in the browser. Results of NCBI-BLAST includes

with all the hits found, sequence identifiers for the hits having scored related data in the table format, alignments for the sequence and the BLAST scores based on the hits received are presented in graphical format.

The raw sequence obtained from HGMD is processed to form cDNA sequence, the nucleotide base alteration is done based on the mutational information through R script, and new synthetic sequences are generated. Using the built-in functions a set of programs are executed from the R library to identify the required position to be altered and is replaced with the nucleotide specified in the nucleotide change column of HGMD database. Using the positional cloning approach the mutated sequences are generated and stored as fasta files.

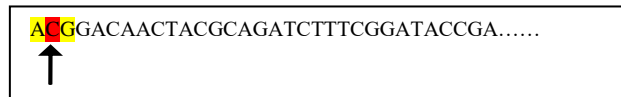
Consider the missense mutational information for the EMD phenotype from the Emerin gene such as nucleotide change is 2 T>C, which indicates the position 2 the nucleotide changes from T to C alters the protein from Met to thr.

For example, the cDNA sequence of EMD gene is



A rectangular box containing the text "ATGGACA...". The first two characters, "AT", are highlighted in yellow. A black arrow points upwards from the second character, "T", to the text below.

After the nucleotide change done in the position 2



A rectangular box containing the text "ACGGACA...". The first two characters, "AC", are highlighted in yellow. A black arrow points upwards from the second character, "C", to the text below.

The sample output of cloning technique for generating mutated gene sequences using positional information is shown below in Fig.4.2.



Fig.4.2 Sample Output of Generated Mutated Gene Sequence

The performance of the disease identification models hugely depends on the data with which it is trained. Hence, it is essential to have a large corpus with many examples that includes all five types of diseases that are affected by seven types of mutations. This will aid the muscular dystrophy disease identification model to learn the features of the diseased gene sequences to predict the category of disease accurately. For this purpose, in each category of muscular dystrophy disease, 200 synthetic mutated gene sequences are generated and a corpus comprising of 1000 sequences for all five categories of muscular dystrophy is developed. Few gene sequences generated through positional cloning approach are shown in Appendix - A.

4.3 Framework of Disease Identification Model

In this research work, two distinct learning approaches such as shallow learning and deep learning have been adopted to build models. The modeling procedures are illustrated in the following sections 4.3.1 and 4.3.2 respectively.

4.3.1 Muscular Dystrophy Disease Identification through Shallow Learning Approach

The principal focus of this work is to propose an efficient data driven model for predicting the type of muscular dystrophy disease. The general muscular dystrophy disease identification model comprises of phases such as generation of mutational gene sequences, feature extraction, feature selection, building the model and classification. The process flow of this approach is depicted in Fig.4.3.

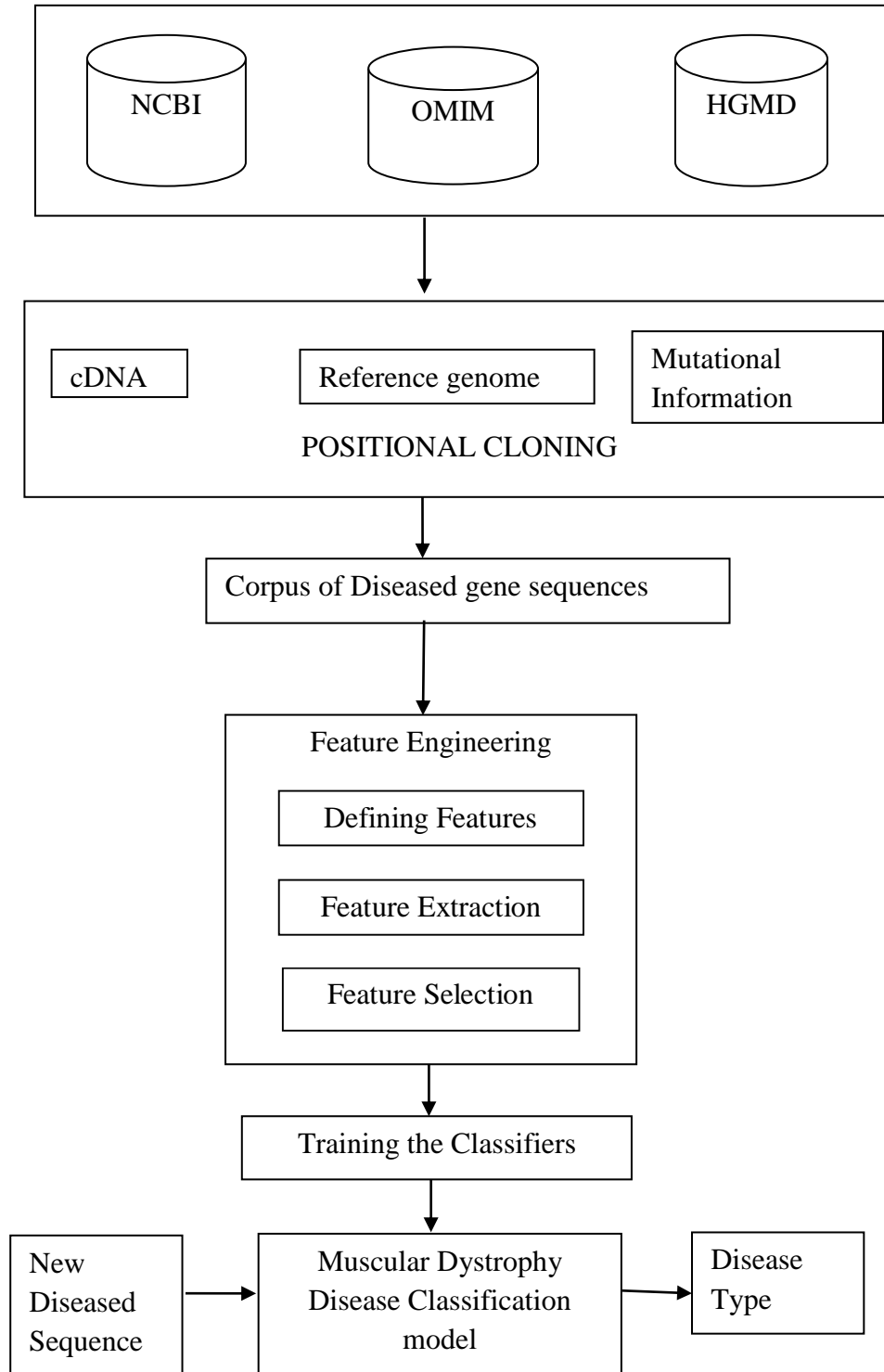


Fig.4.3 Disease Identification Model

The key point of this research is to pinpoint discriminative descriptors and to provide model for efficiently predicting the type of muscular dystrophy disease. Change or mutation in

the gene sequence alters the structure of the sequence, which implies the cause of disease. These structural changes are captured as features from mutated sequences to learn the prediction model. It is pioneered from the literatures that the disease identification problem can be modeled as a pattern recognition task to identify the disease. As machine-learning technique can automatically learn the model by taking intelligent hints from the data and predicts the output more accurately, it has been influenced in this work to extract various discriminative features from the diseased gene sequences for building disease prediction models.

Seven kinds of mutations are taken into account and four exclusive datasets have been form based on various mutational features. In the first case, the features related to missense and nonsense (Non-synonymous) mutations are considered. Annotation, structure and alignment features have been extracted from the corpus of sequences and the dataset NSM with dimension 26 is generated.

In the second case, the silent mutational features are taken into account, as it is required to identify the disease that is caused due to synonymous mutations. The codon usage patterns are considered as the contributing features for representing the mutated gene sequences. Since codon usage patterns are diverse in different gene families, this feature input is a well-chosen descriptors for specifying different gene families for all types of diseases. Codon usage bias helps in identifying Silent mutations and hence 59 RSCU (Relative Synonymous Codon Usage Bias) values have been determined from the same corpus and the dataset SYM with 1000 feature vectors is created.

Prediction of muscular dystrophy disease using features related to insertion and deletion mutations was done in the next experiment. The extrinsic and intrinsic features more solely depend on the exons and introns that enable to identify the disease affected by large insertions and deletions. Twenty-three such exonic and intronic descriptors related to Insertion/Duplication, deletion was extracted from the gene sequences and dataset IDM is prepared.

In the next case, the mutations occurred while splicing is considered to know the alteration after the splicing process as the exons are formed by splicing out the introns during transcription. Exon, Single Nucleotide Polymorphism (SNP) and gene features are taken into account. Position of the spliced introns and exons are carefully examined and twenty four such features are identified and extracted to capture the variations due to splicing in the mutated gene sequences. The dataset SPM of size 1000 instances with dimension 24 is generated.

Normally, the type of mutation caused in the gene sequence may not be known explicitly and hence all the mutational features are accumulated by eliminating the repetitive features without losing information to facilitate efficient learning for predicting the disease caused by any mutation. Information gain feature selection method is employed to select high ranked effective features and the dataset AGM is generated.

In all above cases, for each feature vector, the class label is assigned a sequence number 1 to 5 designating the category of muscular dystrophy diseases as DMD – 1, BMD – 2, EMD – 3, LGMD – 4, CMT - 5. The feature extraction process is explained in depth in Chapter 5.

Non-Synonymous (NSM), Synonymous (SYM), Insertion duplication & deletion (IDM), Splicing mutation (SPM), AGM (Aggregated mutational features) are the five independent datasets generated with different dimensions and the profile of these datasets are depicted in Table.4.3.

Table 4.3 Profile of Training Datasets

Type of Mutation	No. of Features	Dataset	Size of dataset
Non- Synonymous	26	NSM	1000*26
Synonymous	59	SYM	1000*59
Insertion/duplication & Deletion	23	IDM	1000*25
Splicing	24	SPM	1000*24
Aggregated	106	AGM	1000*106

Autonomous data driven models have been built based on the above mutational feature sets using supervised classification algorithms such as Decision tree, Artificial neural network, Naïve bayes and Support Vector Machines. Five independent experiments were carried out using the above datasets by implementing the above stated algorithms. The predictive performance of the disease classification models is evaluated using 10-Fold cross validation and analyzed using various metrics like predictive accuracy, precision, recall, F-measure and time taken to build the model. The experimental setup and the implementation results are demonstrated comprehensively in Chapter 5.

In machine learning, the hybrid approach has been an ongoing research area for gaining better performance for classification or prediction problems over a single learning approach. Ensemble models have been developed using LibD3C classifiers by learning the above five independent data sets. The performance of these ensemble models is evaluated in the similar fashion and the results are compared with standard pattern recognition algorithms. The process of building models using ensemble learning approach is presented in Chapter 6.

4.3.2 Muscular Dystrophy Disease Identification Model through Deep Learning Approach

Deep models were built with the self-extraction of features using deep neural network coded through a Jupyter notebook in TensorFlow environment. Self-taught learning attempts to automatically learn good features or representations based on training data. A deep neural network approach is also employed in this research for identifying the genetic disease, which can learn complex and abstract features automatically from unlabeled data. Tensorflow Deep neural network classifier was implemented based on scikit flow with 1000 gene sequences. The process flow of the proposed deep learning based muscular dystrophy disease prediction model is shown in Fig.4.4.

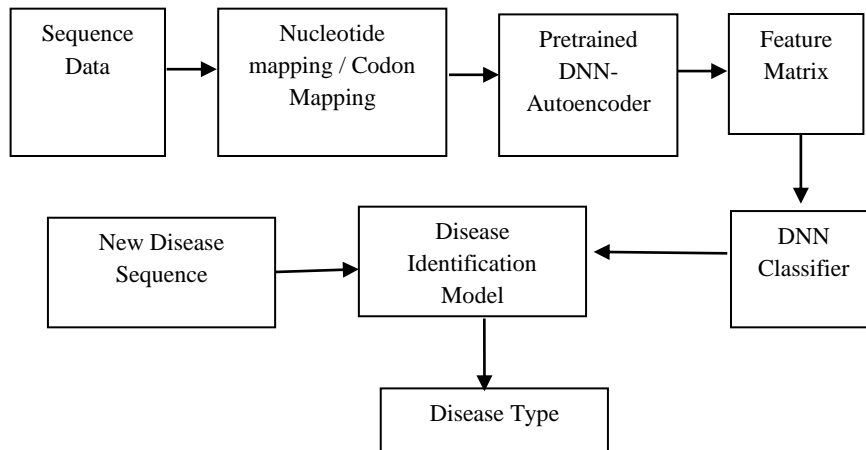


Fig.4.4 Process Flow of Proposed Muscular Dystrophy Disease Identification Model

The first phase in constructing a deep learning model to encode the gene sequence data. The DNA sequence is a string of categorical values and there is a need to hardcode these values into a numerical array as the deep neural network classifier accepts the input as an array of

numerical values. Hence, two mapping schemes namely nucleotide and codon mapping have been proposed to encode A, T, G, C nucleotide values. The trick in deep learning approach to represent biological data is converting disease gene sequence into 1-D representation through encoding schemes. The nucleotide and codon mapping schemes are explained in detail in chapter 7 under sections 7.1 and 7.2.

The next step is to train each auto encoder to encode and decode their input. After pre training is done the weights of DNN can be set with the weights of all encoder.

A feature matrix is generated as an array of numerical values and the output of the encoder is fed into the deep neural network classifier. TensorFlow Deep neural network classifiers are employed to build disease identification models using a python library in a Tensorflow environment.

4.4 Summary

This chapter portrayed the modeling of the muscular dystrophy disease identification problem as a pattern recognition problem. Corpus development through synthetic gene sequences, an important activity carried out in this research work, is also described. The framework for the muscular dystrophy disease identification model through shallow and deep learning approaches are presented in this chapter. The training and implementation of the muscular dystrophy disease identification model using shallow and deep learning approaches are discussed in the subsequent chapters.