8. CONCLUSION

The thesis titled “Identification of Rare Genetic Muscular Dystrophy from Sequence based features through Shallow and Deep Learning” portrays the research work carried on genetic disease identification through shallow and deep learning approaches with gene sequences as input.

As it is significant to develop an accurate model for predicting the rare genetic disorder, this research work has been evolved. The goal of this research is to develop disease identification models for genetic disease such as muscular dystrophy based on mutated gene sequences using two learning approaches. Muscular dystrophy disease identification models are built using shallow learning approach by identifying and extracting the hand crafted features from gene sequences. Deep learning approach aids in building disease identification models through self-extraction of features. The proposed approaches significantly makes the genetic disease prediction clear-cut and suggests an expedient solution by extracting relevant gene based features.

A corpus of 1000 synthetic gene sequences containing diseased gene sequences that covers the entire mutation spectrum of all type of mutations with gene sets of major five kinds of muscular dystrophy disease has been developed using positional cloning approach.

The foremost task in shallow learning approach is to design the discriminative features from the mutated gene sequences and to build data driven models for identifying the type of the genetic disorder. Handcrafted mutational features were identified and extracted from the disease gene sequences based on the type of mutations such as missense, nonsense, silent, insertion, deletion, duplication, splicing and five independent datasets have been created. Five different disease identification models have been built using supervised learning algorithms such as Decision Tree Induction, Naïve Bayes, Artificial Neural Network, Support Vector Machine. Several experiments have been carried out under different environmental setup and the disease identification models are evaluated using 10-Fold cross validation. Performances of the classifiers are analyzed using various measures such as precision, recall, F-score, Kappa statistic and time taken to build the model. The comparative analysis is carried out and the interpretations are presented. The outcome of the experiments proves that, the disease identification model is effectual when the collective features are used in learning.

Hybridization approach is implemented though ensemble learning and disease identification models are built using the LiBD3C classifier for the same datasets. It is observed that the LibD3C classifier outperforms the standard classifiers in its predictive accuracy.

Deep learning is an added significant approach exploited for muscular dystrophy disease identification problem. Synthetic gene sequences are encoded into its decimal values nucleotide and codon mapping schemes. Self extraction of features is done in the deep learning approach and disease identification model is built using deep neural network classifier in the Tensor flow environment. Deep learning aids in identifying the muscular dystrophy disease by self-extraction of intelligent hints from the diseased gene sequences. Representations of gene sequences using two kinds of mapping schemes are also proposed. Since the deep neural network has an ability of extracting high-level abstraction through self-extraction of features, promising results were obtained in this disease classification. Comparative analysis between linear classifier and tensorflow deep neural network is made and the results are analyzed.

Comparison between shallow and deep learning approaches is done by measuring up ANN models against deep neural work. It is perceived that the predictive accuracy shown by the deep neural network is comparatively higher than that of other supervised learning approaches.

The work to date ascertains highly efficient creation of muscular dystrophy disease prediction models through supervised, ensemble and deep learning approaches. Exhaustive experimentation carried out shows that the classification modeling is feasible with the support of discriminative mutational descriptors for muscular dystrophy disease identification. The thesis has also generalized the disease identification problem as an automated practice, which can be applied to identify any kind of genetic disease. These approaches for disease identification exceedingly simplify the 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.

The observations and the interpretations acquired from this research work are summarized as follows:

- It is noticed that the synthetic gene sequences can be generated using positional cloning approach for disease identification through computational approaches
- Gene based features can be identified for any kind of mutations to predict any type of genetic disease

- Traditional machine learning approaches perceived an enviable accuracy that motivates to extent the research higher
- LibD3C classifier attains an elevated accuracy when compared with the standard classification algorithms
- Self-extraction of features through deep learning eliminates the feature extraction process
- It is found that the accuracy of the disease identification model based on deep learning approach based on nucleotide mapping and codon mapping schemes is higher than that of supervised learning algorithms in shallow learning approach
- Deep learning approach performs well when the size of the input is large

The research contributions made in this thesis are

- Rare genetic disorder – muscular dystrophy disease has been taken into consideration for this research
- Disease identification was done using gene sequences whereas gene expression data and protein data were used in the existing works
- Positional cloning approach was used to generate synthetic gene sequences
- Identified and captured discriminative descriptors from gene sequences related to all type of mutations
- Implementation of deep learning with two mapping schemes- nucleotide and codon mapping
- Tools used - R, Mat lab, Bio Edit, Geneious pro, Weka, Scikit-Learn, Jupyter notebook, TensorFlow

It is concluded that shallow and deep learning techniques are suitable in predicting muscular dystrophy disease when any type of mutational features are utilized for building the models. The promising results obtained from this research work have encouraged the use of novel approaches based on hand crafted and self extracted features from gene sequences in global genetic disease prediction. In future, the work can be extended with different contributing aspects like protein misfolding, gene to gene interaction, protein protein interaction and using hybrid approaches with emerging computational techniques.