Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
(2017) 25: 3932 – 3946
© TÜBİTAK
doi:10.3906/elk-1608-173

Prognosis of muscular dystrophy with extrinsic and intrinsic descriptors through ensemble learning

Sathyavikasini KALIMUTHU[*], **Vijaya VIJAYAKUMAR**
Department of Computer Science, PSGR Krishnammal College for Women, Coimbatore, India

Received: 16.08.2016 • **Accepted/Published Online:** 13.02.2017 • **Final Version:** 05.10.2017

Abstract: Muscular dystrophy is a neuromuscular disorder that impairs the functioning of the locomotive muscles. Large deletion and duplication mutations in the gene sequences pave the way for these muscular dystrophies. Any heritable change can be used as input in computational studies such as pattern and classification models. Mutated gene sequences are generated by adopting the positional cloning approach on the reference cDNA sequence with mutational information from the Human Gene Mutational Database (HGMD). The extrinsic and intrinsic descriptors of the mutated gene sequence are indispensable to identifying the disease. This work describes a computational approach of building a disease classification model by extracting the exonic and intronic descriptors from the mutated gene sequences through a combined learning technique. An ensemble hybrid model is developed through LibD3C classifier. The hybrid learned model gained an accuracy of 98.3% in diagnosing the neuromuscular disorder, based on deletion and insertion/duplication mutations. Furthermore, this paper analyzes the implementation of ensemble-learning classifiers based on features related to synonymous and nonsynonymous mutations, in order to detect muscular dystrophy performed with the same data set. Experiments showed high accuracy for the models built using LibD3C classifier, which proves that ensemble learning is effective for predicting disease. To the best of our knowledge, for the first time the models established here explore a scheme of disease prediction through pattern recognition from the sequence of nucleic acid molecule and associated mutations.

Key words: Ensemble learning, extrinsic, intrinsic, LiBD3C, muscular dystrophy

1. Introduction

Muscular dystrophy is a kind of neuromuscular genetic disorder caused by a deformity in the genes that deteriorates the locomotive muscles [1]. At present, there are no effective therapeutic strategies to halt progression or to cure this type of disease. Certain types of muscular dystrophy are Duchenne, Becker, Emery–Dreifuss, limb–girdle muscular dystrophy, facioscapulohumeral, myotonic, and Charcot Marie tooth disease.

Variation in the genetic code bequeathed from parents to offspring can cause a perpetual change in the gene sequence, known as mutation. Single character change in a gene makes an impact on the gene, which, in turn, changes its function. Nonsynonymous mutations that show an impression on protein sequence include missense, nonsense, insertions, deletions, splicing, and frame-shift mutations. Replacement of a nucleotide by another that makes an impact on protein change is termed as missense mutation [2,3]. Nonsense mutations are those where the protein succeeds in stopping codon when a change occurs in the DNA sequence. A cluster

[*] Correspondence: mail2sathyavikashini@gmail.com

of bases are added or deleted during translation, which leads to the insertion or deletion mutation in the gene sequence that gradually roots in a change of protein sequence [4,5].

Altering the bases in a codon encodes for the same amino acid and the resulting protein does not reflect any change during silent mutations [6]. The information from the genes transfers the nucleic acid to proteins in the form of codons. During the process of translation, the synonymous codons have different frequencies [7], which are referred to as codon usage bias. The functionality of the gene depends on codon usage bias.

In medical applications, the capability of machine learning is well-suited to analyzing complex diseases such as diabetes [8], hepatitis [9], rheumatoid arthritis [10], and schizophrenia [11]. However, not many studies have been carried out on variation in muscular dystrophy using machine-learning algorithms. Furthermore, the classification of this complex disease is performed with either the protein data or microarray data as their inputs. Classification of facioscapulohumeral muscular dystrophy (FSHD) disease is performed by monitoring expression levels. Usually, microarray gene expression analysis is mainly focused on cancer diseases. In [12], the authors proposed a model using a support vector machine to classify the types of FSHD, using microarray gene expression data from the DUX4 gene.

The authors in [13] 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. RSCU values are calculated for the gene sequences by converting them into a vector of 59 elements. The support vector machine achieved an accuracy of 99.3%.

Nisha et al. proposed a new approach, based on codon usage pattern, to classify the type of hepatitis C virus (HCV), which is the primary reason for liver infection. To classify the subclass of its genotype, a model was created using codon usage bias as input to multiclass SVM [14]. Falk and Gilchrist [15] developed a model using neural networks to identify limb–girdle muscular dystrophy (LGMD). Using family details data, a classification of disease status was made using the neural network, achieving an accuracy of 98%.

The authors in [16] constructed a protein–protein interaction network to classify the subtypes of muscular dystrophy with a multiclass support vector machine. Microarray gene expression data sets are analyzed, the protein data and their interaction data are collected, and a network is constructed to classify the subtypes.

The authors in [17] employed a machine learning approach, based on ensemble classifier LibD3C, to predict the cytokines. The analysis was carried out on the physicochemical properties and the distribution of whole amino acids.

In [18], the authors identified large mutations, such as duplications and deletions, through a 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. In [19], the authors predicted the disease-causing mutations with the ensemble learning technique. The protein sequence data set from Swiss-Prot database was used for classification. A comparative analysis was made between the traditional and ensemble approaches, and it was found that the LogitBoost ensemble technique achieves the highest performance among all the methods compared.

From the background study, it is observed that the muscular dystrophy identification problem can be modeled as a pattern recognition task and solved using machine learning techniques. The difficulties involved in the disease identification system need to be analyzed and taken into account when 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 predicting muscular dystrophy susceptibility and mortality. It allows the clinician to make a diagnosis without needing a muscle biopsy, raises clinician response time, and helps to treat disorders.

In our previous work [20] predicting muscular dystrophy, this was performed by creating 150 cloned gene sequences. Missense and nonsense mutation-related features are extracted from the gene sequences to build the classifier and classify the type of disease. A model was developed based on pattern classification algorithms, and high accuracy was attained from the decision tree classifier. In other work, silent mutational features are captured by calculating the RSCU values from the disease gene sequences, and a model was build using standard classification techniques to identify the muscular dystrophy disease. An accuracy of 86% was attained using a support vector machine [21]. With the same set of RSCU features, 90% of accuracy is achieved from LibD3C classifier in predicting muscular dystrophy with silent mutations.

The primary focal point of this research is to classify the major forms of muscular dystrophy disease such as Duchenne muscular dystrophy (DMD), Becker's muscular dystrophy (BMD), Emery–Dreifuss muscular dystrophy (EMD), limb–girdle muscular dystrophy (LGMD), and Charcot Marie tooth disease (CMT). As the mutated sequence is not readily available for this disease, the corpus is developed on the reference cDNA sequence, with the mutational information obtained from the HGMD. The HGMD [22] is a core collection of data on germ-line mutations in genes coupled with the human inherited disease, which is grasped from various works in the literature. The study is performed by extracting the well-defined descriptors pertaining to insertions/duplications and deletions from 300 mutated gene sequences in the corpus. The experimentation is executed with a hybrid approach of the ensemble learning technique with the LibD3C classifier. Additionally, this study predicts muscular dystrophy related to features of synonymous and nonsynonymous mutations based on the ensemble LibD3C classifier with this corpus of data.

In machine learning, the hybrid approach has been an ongoing research area for achieving better performance for classification or prediction problems with a single learning approach. The motivation behind the hybrid model is that a hybrid classification model can be composed of one unsupervised learner to preprocess the training data and one supervised learner to learn the clustering result [23,24]. The familiar ensemble learning methods for classification problems are bagging, boosting, and random forests. To reduce information redundancy within multilabel learning, a model-shared subspace boosting algorithm was constructed [25], which automatically finds shared subspace models, where every model was made to learn from the random feature subspace and bootstrap data and combined a number of base models through multiple labels [26]. Ensemble classification is a technique that combines multiple basic classifiers with their own decision-making capacity. The prediction ability of an ensemble classifier is excellent compared to that of a single classifier, because the former can address the differences produced by the latter more efficiently when challenged with different problems. LibD3C is a type of ensemble classifier with a clustering and dynamic selection strategy. A method that blends two types of discriminating ensemble techniques is known as dynamic selection and circulating combination-based clustering (D3C). LibD3C employs two types of selective ensemble techniques, namely ensemble pruning based on k-means clustering and dynamic selection and circulating combination. LibD3C is a selective ensemble classifier, where various candidate classifiers are trained, and a set of classifiers that are accurate and diverse are selected to rectify the problem.

2. Materials and methods

The arrangement of the bases in the gene sequences differs in every human. The main objective of this research is to pinpoint discriminative descriptors and provide an efficient machine learning solution for predicting the type of muscular dystrophy with insertion, deletion, and duplication mutations. Multiclass classification is worked out through data modeling of gene sequences. The availability of diseased gene sequences is a challenge for this intricate disease, which stimulates the need for generation of synthetic mutational gene sequences. Identification of muscular dystrophy disease involves various phases that utilize extrinsic and intrinsic features. At first, the mutational gene sequences are generated using positional cloning. The discriminative descriptors are identified and then extracted. The selected indispensible features are utilized to train the model, and the processing flow of the approach is depicted in Figure.1.

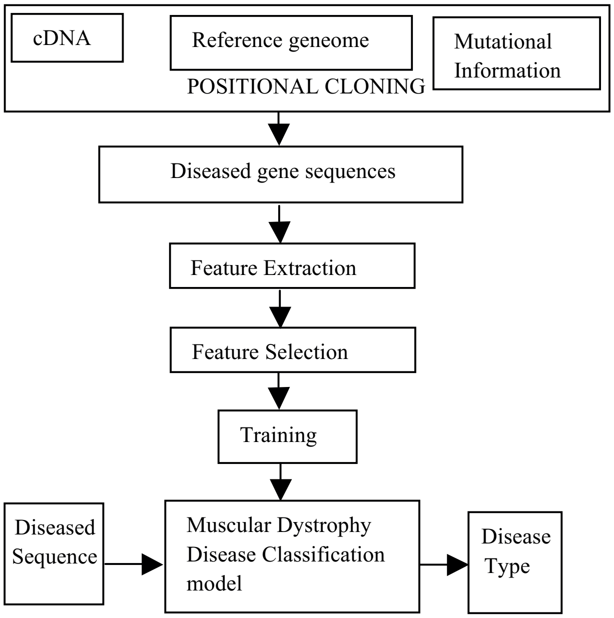


Figure 1. Disease identification model.

2.1. Corpus preparation through positional cloning

Various types of genes associated with the five types of neuromuscular disorder are studied. We analyze 55 genes that are associated with five types of muscular dystrophy, namely DMD, BMD, EMD, LGMD, and CMT. Several types of mutated sequences based on mutations such as missense, nonsense, synonymous, insertion/duplication, deletion mutations, and splicing mutations are collected. For the purpose of this research, in each category of muscular dystrophy, 60 synthetic mutated gene sequences are generated and a corpus comprising 300 sequences for all five categories of muscular dystrophy is developed.

The reference genes for the mutated genes are downloaded from NCBI. The raw sequence obtained from the HGMD is processed to form a cDNA sequence. Nucleotide base alteration is carried out based on mutational information through an R script and new sequences are generated. Using 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 the HGMD database. Using the traditional 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; for example, nucleotide change is 2 T $>$ C, which indicates that in position 2 the nucleotide changes from T to C and alters the protein from Met to thr.

For example, the cDNA sequence of EMD gene is

ATGGACAACTACGCAGATCTTTCGGATACCGA......

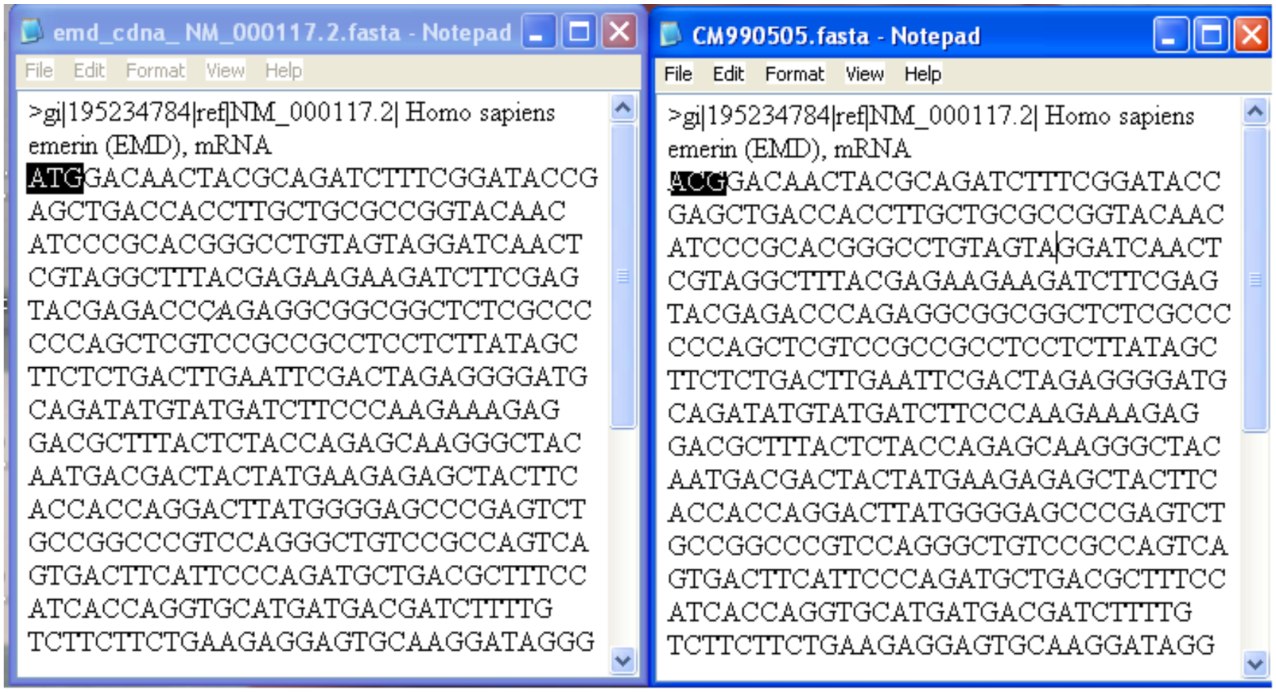
↑

After the nucleotide change in position 2

ACGGACAACTACGCAGATCTTTCGGATACCGA......

↑

A sample output of the cloning technique for gene sequence using positional information is shown in Figure 2.


emd_cdna_ NM_000117.2.fasta - Notepad
File Edit Format View Help
CM990505.fasta - Notepad
File Edit Format View Help

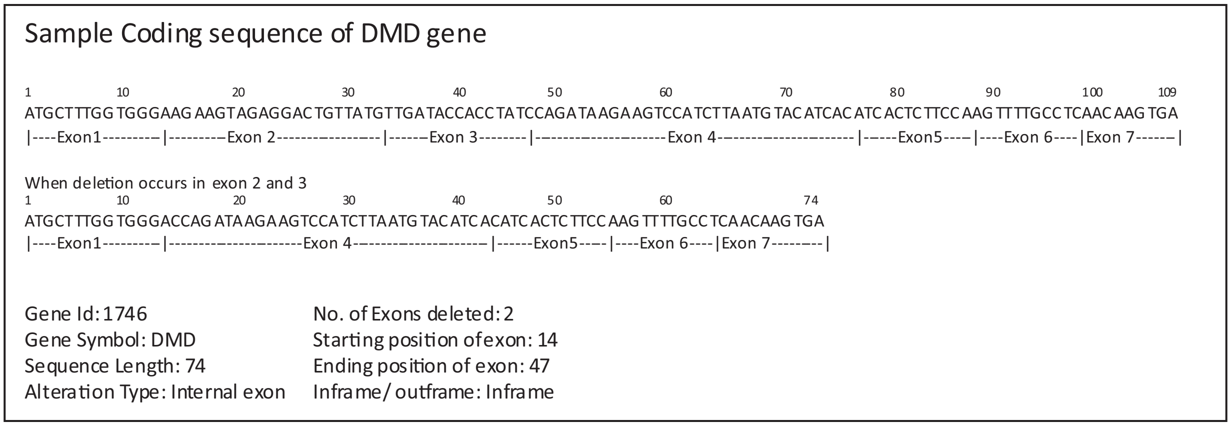
```
>gi|195234784|ref|NM_000117.2| Homo sapiens
emerin (EMD), mRNA
ATGGACAACTACGCAGATCTTTCGGATACCG
AGCTGACCACCTTGCTGCGCCGGTACAAC
ATCCCGCACGGGCCTGTAGTAGGATCAACT
CGTAGGCTTTACGAGAAGAAGATCTTCGAG
TACGAGACCCAGAGGCGGCGGCTCTCGCCC
CCCAGCTCGTCCGCCGCCTCCTCTTATAGC
TTCTCTGACTTGAATTCGACTAGAGGGGATG
CAGATATGTATGATCTTCCCAAGAAAGAG
GACGCTTTACTCTACCAGAGCAAGGGCTAC
AATGACGACTACTATGAAGAGAGCTACTTC
ACCACCAGGACTTATGGGGAGCCCGAGTCT
GCCGGCCCGTCCAGGGCTGTCCGCCAGTCA
GTGACTTCATTCCCAGATGCTGACGCTTTCC
ATCACCAGGTGCATGATGACGATCTTTTG
TCTTCTTCTGAAGAGGAGTGCAAGGATAGGG
```

```
>gi|195234784|ref|NM_000117.2| Homo sapiens
emerin (EMD), mRNA
ACGGACAACTACGCAGATCTTTCGGATACC
GAGCTGACCACCTTGCTGCGCCGGTACAAC
ATCCCGCACGGGCCTGTAGTAGGATCAACT
CGTAGGCTTTACGAGAAGAAGATCTTCGAG
TACGAGACCCAGAGGCGGCGGCTCTCGCCC
CCCAGCTCGTCCGCCGCCTCCTCTTATAGC
TTCTCTGACTTGAATTCGACTAGAGGGGATG
CAGATATGTATGATCTTCCCAAGAAAGAG
GACGCTTTACTCTACCAGAGCAAGGGCTAC
AATGACGACTACTATGAAGAGAGCTACTTC
ACCACCAGGACTTATGGGGAGCCCGAGTCT
GCCGGCCCGTCCAGGGCTGTCCGCCAGTCA
GTGACTTCATTCCCAGATGCTGACGCTTTCC
ATCACCAGGTGCATGATGACGATCTTTTG
TCTTCTTCTGAAGAGGAGTGCAAGGATAGG
```

Figure 2. Sample output of generated mutated gene sequence.

2.2. Feature extraction and training data set

Genetic disorders are caused by changes or mutations occurring in gene sequences. The descriptors are captured from structural changes in the gene sequence to learn the prediction model. So far, no attempt has been made in the literature to extract descriptors pertaining to insertion, deletion, and duplication mutations. Hence, it is significant to propose these features to build a disease identification model. The exonic and intronic features are considered from diverse gene families to extract the well-defined descriptors related to insertion, deletion, and duplication mutations in the mutated gene sequences. Twenty-three such evocative descriptors are extracted and feature vectors are created for learning the disease prediction model. These extrinsic and intrinsic features, extracted from 300 gene sequences, are depicted in Table 1. Code is written using R for extracting most descriptor values from the mutated gene sequences.



Sample Coding sequence of DMD gene

```
1          10           20          30          40        50           60           70          80         90         100     109
ATGCTTTGG TGGGAAG AAGTAGAGGACTGTTATGTTGATACCACCTATCCAGATAAGAAGTCCATCTTAATG TACATCACATCACTCTTCCAAG TTTTGCCTCAACAAG TGA
|----Exon1---------|---------Exon 2----------------|------Exon 3--------|--------------------Exon 4--------------------|-----Exon5----|----Exon 6----|Exon 7------|

When deletion occurs in exon 2 and 3
1          10           20          30          40        50           60           74
ATGCTTTGG TGGGACCAG ATAAGAAGTCCATCTTAATG TACATCACATCACTCTTCCAAG TTTTGCCTCAACAAG TGA
|----Exon1---------|--------------------Exon 4--------------------|------Exon5----|----Exon 6----|Exon 7--------|
```

Gene Id: 1746
Gene Symbol: DMD
Sequence Length: 74
Alteration Type: Internal exon

No. of Exons deleted: 2
Starting position ofexon: 14
Ending position of exon: 47
Inframe/ outframe: Inframe

Table 1. Extrinsic and intrinsic descriptors and their description.

Descriptors	Description
Gene ID	Gene identifier
Gene symbol	Gene name
Gene start position	Gene starting position in chromosome
Gene end position	Gene ending position in chromosome
Length	Sequence length
Alteration type	Type of alteration either deletion or duplication
Number of exons	Number of exons inserted or deleted
Starting position of exons	Exon starting position
Ending position of exon	Exon ending position
Inframe or outframe	Deletion is inframe or outframe
Type of exon	Initial, internal, terminal exon, single exon
Exon conservation score	Conservation score of all exons
Protein coding region score	Distinguishing protein coding region from noncoding region and calculates its score
Probability of exon	Exon's probability score
Alignment scores (3)	Blast alignment scores such as phred quality scores, edit distance scores, quality scores
Nucleotide composition values (6)	A,T,G,C,AT,GC nucleotide compositional values

Descriptor 1, 2 – Annotation features: Gene sequences are identified by attributes such as gene identifier and symbol. As many-to-one relationships occur between the gene and the disease, these descriptors are considered to differentiate the gene sequence in every disease type. The attributes of gene sequences, such as gene ID and gene symbol, are identified by using the biomart package in R, and are extracted using getgenes(id). The gene ID is the NCBI gene identifier for the affected phenotype. For example, GeneID 1746 and gene symbol DMD are for Dystrophin gene, GeneID 2010 and gene symbol EMD for Emerin gene, etc.

Descriptor 3 – Alteration type: The next descriptor alteration type denotes the type of mutation such as insertion, deletion, and duplications. This feature is captured by hardcoding the mutation type to its corresponding numeric values from 3 to 5, such as 3 for insertion, 4 for duplication, and 5 for deletion.

Descriptor 4, 5 – Gene starting position and gene ending position: A chromosome comprises several genes and every gene has a starting and ending position. If an insertion/duplication or deletion mutation occurs in a sequence, then there may be a change in the gene's starting or ending position. Hence, these features aid the classification of the disease type. Nucleotide blast is used to capture the starting and ending position of a gene by aligning the sequence with its reference gene sequence.

Descriptor 6 – Sequence length: The length of the sequence plays an important role in examining the difference in length of the sequence. When the insertion or deletion mutation occurs, the length of the sequence varies automatically. This feature is determined using length() function by converting the fasta file into a data frame.

Descriptor 7 – Number of exons inserted/deleted: Severe effect on the deletion of exons leads to DMD, and mild deletion of exons results in BMD. While gross insertions and gross deletions occur, the severity of the disease is determined by the number of exons inserted or deleted. This descriptor is calculated using the deletion region information column in the HGMD.

Descriptor 8 – Exon and intron boundary: Every gene sequence comprises coding (exonic) and non-coding (intronic) regions. The boundary of the exonic and intronic regions is altered when insertion/duplication or deletion of exons occurs. Therefore, these descriptors are captured to identify the differences in the boundary between the normal and the diseased sequences. By visualizing the sequences in Geneious Pro, these descriptors are captured.

Descriptor 9 – Deletion type: If the sequence can still be read after deletion mutation occurs, then it is considered as an inframe deletion. In the outframe deletion type, the sequence cannot be read after the deletion mutation occurs. Deletion type is a contributive feature in identifying the type of the disease, as in some diseases like BMD, where the sequence can still be read after deletion, and in some diseases like DMD, the sequence cannot be read after deletion as it is outframe. This feature is captured by translating the diseased sequences into its corresponding amino acid sequence. Splitseq (), tablecode () functions from biostrings, and seqinr packages are used to capture this descriptor.

Descriptor 10 – Exon type: Depending on the location of the exon, the type of exon may be initial, internal, terminal, or single exons. The mutation in each type of exon has its own severity, which aids in classifying the disease type. This discriminative feature is captured using Geneious Pro.

Descriptor 11 – Conservation score: The structure or the function of the sequence is identified by calculating the conservation score by aligning the sequence with all organisms. The University of California Santa Cruz (UCSC) genome browser is employed to calculate the conservation score.

Descriptor 12 – Protein-coding region score: The score of the protein-coding region is calculated with a coding potential calculator, based on the sequence features to distinguish protein coding from noncoding regions. When a deletion occurs in an exon, the protein-coding region score decides the severity of the deletion on the sequence.

Descriptor 13 to 19 – Nucleotide composition values: The base composition A, C, G, T are calculated to count the number of occurrences of the four different nucleotides ("A", "C", "G", and "T") in the sequence. GC content is the fraction of the sequence that consists of Gs and Cs, i.e. the GC content can be calculated as the percentage of the bases in the genome that are Gs or Cs. That is,

AT content = (number of As + number of Ts) × 100 / (genome length)

GC content = (number of Gs + number of Cs) × 100 / (genome length)

Therefore, six different descriptor values are calculated as the nucleotide composition values.

Descriptor 20 – Stop codon position: The position of the stop codon reveals the end of the coding part in the sequence. This position may be altered with the occurrence of mutation and, hence, it is noted. Match pattern () function is employed to identify and capture the position of the stop codon.

Descriptor 21 to 23 – Alignment scores: Alignment scores are considered as an important feature in disease prediction. The global pairwise alignment, based on edit distance, is performed with the mutated sequence against the reference cDNA sequence, and the three alignment scores are calculated using the edit distance scoring method. The PhredQuality measures are calculated with pattern quality and subject quality to examine the quality-based match and mismatch bit scores for DNA/RNA. The substitution scores are calculated by setting the error probability to 0.

The above 23 features are extracted from each diseased gene sequence and a data set with 300 feature vectors is created.

3. Experiment and results

In machine learning, the hybrid approach has been an ongoing research area for achieving better performance for classification or prediction problems over a single learning approach. LibD3C is a type of ensemble classifier with a clustering and dynamic selection strategy. In this experiment, a muscular dystrophy prediction model is built using the LibD3C algorithm, based on clustering and dynamic strategy in WEKA environment [27]. The benefit of the positional cloning approach in this experiment is that it supports the generation of synthetic mutated gene sequences. The diseases are identified by extracting the well-designed descriptors from the cloned gene sequences. A training data set with instances related to five categories of muscular dystrophy, i.e. Duchenne muscular dystrophy, Becker's muscular dystrophy, Emery–Dreifuss, limb–girdle muscular dystrophy, and Charcot Marie Tooth disease has been developed as described in Section 2.2.

A standard k-fold cross validation technique is used to evaluate the generalization power of the classifiers and estimate their predictive capabilities for unknown samples. As the data set comprises 300 instances, it is appropriate to use cross validation with K = 10. This 10-fold cross validation iterates the algorithm 10 times with different groupings of training and testing data sets. The comparison of cross validation results is made between standard supervised learning techniques and the LibD3C ensemble classifier. The hybrid approach of the LibD3C algorithm shows a high accuracy of 94.34% over traditional classification algorithms such as Naïve Bayes, decision tree, and artificial neural networks. The results of the experiments are summarized in Table 2 and the performances are depicted in Figure 3.

Table 2. Predictive performance of the classifiers.

Performance criteria	Decision tree classifier	Artificial neural network	Naïve Bayes classifier	LibD3C classifier
Kappa statistic	0.902	0.88	0.854	0.931
Mean absolute error	0.095	0.1	0.15	0.098
Root mean squared error	0.21	0.41	0.43	0.19
Relative absolute error	29.944	32.1	40.45	19.44
Root relative square error	50.95	52.95	55.95	45.5
Time taken to build the model (s)	6.09	6.7	6.1	5.47
Correctly classified instance	275	265	260	283
Incorrectly classified instance	25	35	40	17
Prediction accuracy	91.69%	88.3%	86.6%	**94.34%**

3.1. Feature selection

Feature selection or attribute subset selection look for the best descriptors for model construction. Here the information gain selection attribute method is used to select the subset of attributes. Information gain measures

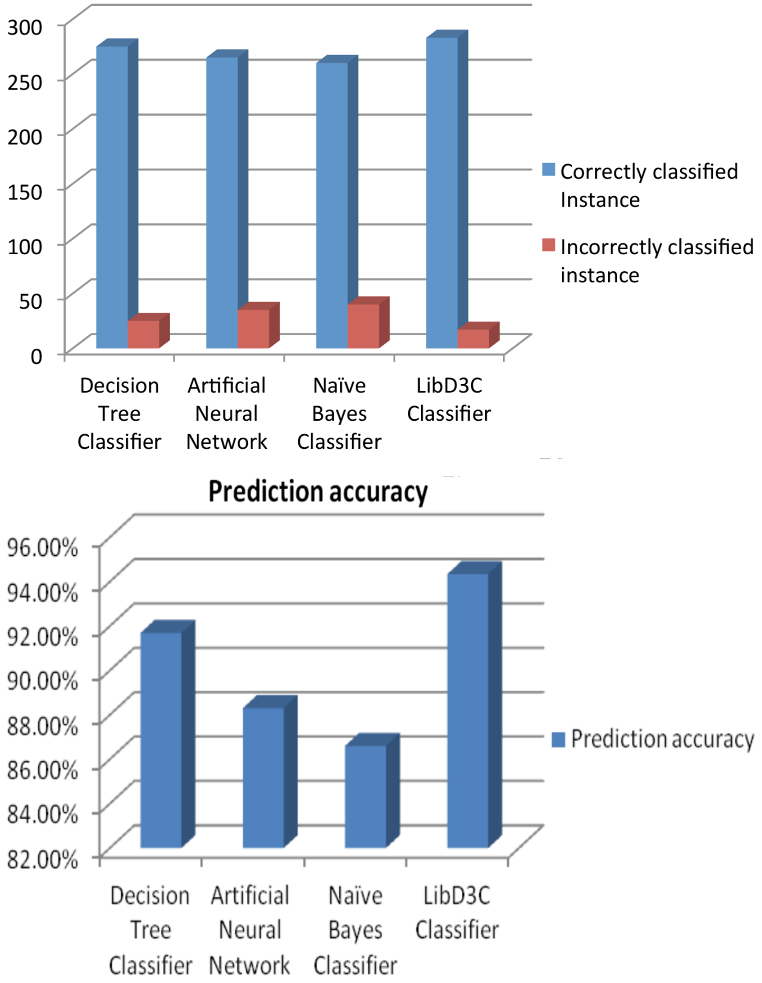


Figure 3. Prediction accuracy of LibD3C classifier.

the number of bits of information obtained for a category prediction of disease by knowing the type of mutation in the disease gene sequence. The difference observed and the expected uncertainty from attribute X is the information gain of attribute X. By ranking the two attributes X and Y, if the gain of X is greater than that of Y, then attribute X is preferred. Likewise, highly ranked attributes are identified using the information gain method and are listed in the following order: in-frame out-frame boundary, type of exon, number of exons, exon start position, exon conservation score, protein conservation score, nucleotide composition, sequence length, and alignment scores. The model is built using a LibD3C classifier and the performance is evaluated in the same manner. An accuracy of 98.33% is observed, and the results are shown in Tables 3 and 4 and are depicted in Figure 4.

3.2. LiBD3C classifiers based on synonymous and nonsynonymous descriptors

As shown in our previous work [20,21], standard classification algorithms, such as decision tree, Naïve Bayes, and artificial neural network, yielded an average of 92% for 150 mutated synthetic gene sequences. Section 3 demonstrated that the LiBD3C classifier earned promising results in predicting the type of muscular dystrophy based on extrinsic and intrinsic features. This stimulated the performance of two independent implementations

Table 3. Predictive performance of the LibD3C classifier after applying feature selection technique.

Performance criteria	Decision tree classifier	Artificial neural network	Naïve Bayes classifier	LibD3C classifier
Kappa statistic	0.932	0.91	0.88	0.97
Mean absolute error	0.098	0.091	0.13	0.065
Root mean squared error	0.17	0.27	0.31	0.14
Relative absolute error	18.4	23.1	38.15	23.44
Root relative square error	45.95	50.15	50.15	41.65
Time taken to build the model (s)	3.94	4.03	3.75	2.76
Correctly classified instance	284	274	267	295
Incorrectly classified instance	16	26	33	5
Prediction accuracy	94.6%	91.3%	89.6%	**98.33%**

Table 4. Assessment of overall accuracies before and after feature selection.

Method	Correctly classified instance	Incorrectly classified instance	Time taken to build the model (s)	Prediction accuracy
Classification without attribute selection	275	25	5.47	91.69
Classification with attribute selection	286	14	2.76	98.33

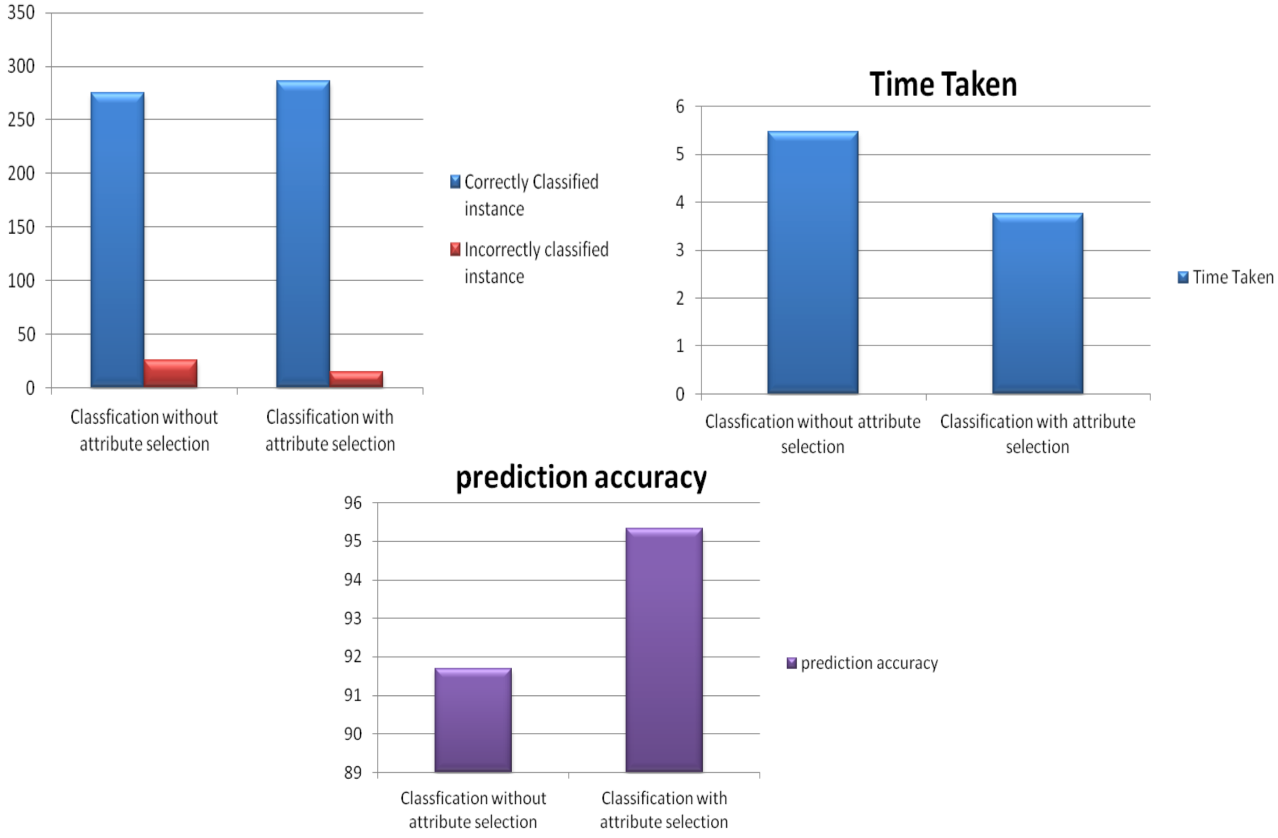


Figure 4. Prediction accuracy of LibD3C classifier (feature selection).

for disease classification through ensemble learning, based on nonsynonymous and synonymous mutational features to predict the type disease. The nonsynonymous mutational features are structural, annotation, and alignment features. Relative synonymous codon usage (RSCU) values for 59 codons form synonymous mutational features. The size of the data set is increased to 300 gene sequences and used for training with the LibD3C ensemble classifier after extracting the respective features. LibD3C-hybrid classifiers are built for classifying neuromuscular disorder based on nonsynonymous and synonymous features. The results of the performance evaluation are presented in Table 5 and drawn in Figure 5. Predictive performance of the LibD3C classifier for all three types of mutation is shown in Table 6.

Table 5. Predictive performance of the LibD3C classifier based on nonsynonymous and synonymous mutations.

Performance criteria	LibD3C classifier (nonsynonymous mutation)	LibD3C classifier (synonymous mutation)
Kappa statistic	0.95	0.85
Mean absolute error	0.015	0.09
Root mean squared error	0.15	0.202
Relative absolute error	20.74	29.844
Root relative square error	39.45	50.801
Time taken to build the model (s)	4.5	7.36
Correctly classified instance	291	202
Incorrectly classified instance	9	28
Prediction accuracy	97%	90%

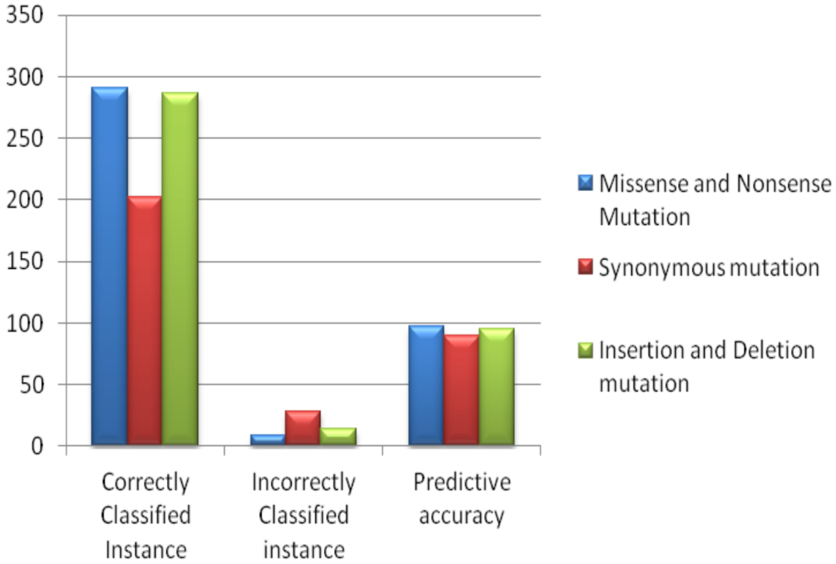


Figure 5. Prediction accuracy of LibD3C classifier (all mutations).

Prediction accuracy of about 97%, nonsense mutational descriptors with a kappa of 0.95, and learning time of 4.5 s are attained for missense. We attained 90% prediction accuracy for silent mutational descriptors with a kappa of about 0.85 and a minimum time of about 7.36 s for 300 cloned gene sequences. Various other measures, such as TP rate, FP rate, precision, recall, F-measure, and ROC area, are evaluated for all three LibD3C classifiers, and the results are depicted in Table 7. The precision recall curve and receiver

Table 6. Predictive performance of the LibD3C classifiers for all three types of mutations.

Problem	Correctly classified instance	Incorrectly classified instance	Predictive accuracy
Nonsynonymous mutation	291	9	97
Synonymous mutation	202	28	90
Insertion and deletion mutation	286	14	98.33

operating characteristic (ROC) metric are illustrated in Figures 6 and 7 for each class, respectively. The existing approaches classify either the gene or the disease with the micro array, protein, or family details data. The proposed approach can classify five types of muscular dystrophy from gene sequence data by extracting mutational features. The results are compared with existing studies and presented in Table 8.

Table 7. Evaluation measures of the classifiers.

Problem	Precision	Recall	F-measure	TP rate	FP rate	ROC area
Missense and nonsense mutation	0.967	0.23	96.8	96.8	0.014	0.97
Synonymous mutation	0.901	0.72	89.3	89.3	0.018	0.91
Insertion and deletion mutation	0.983	0.45	98.2	98.2	0.08	0.997

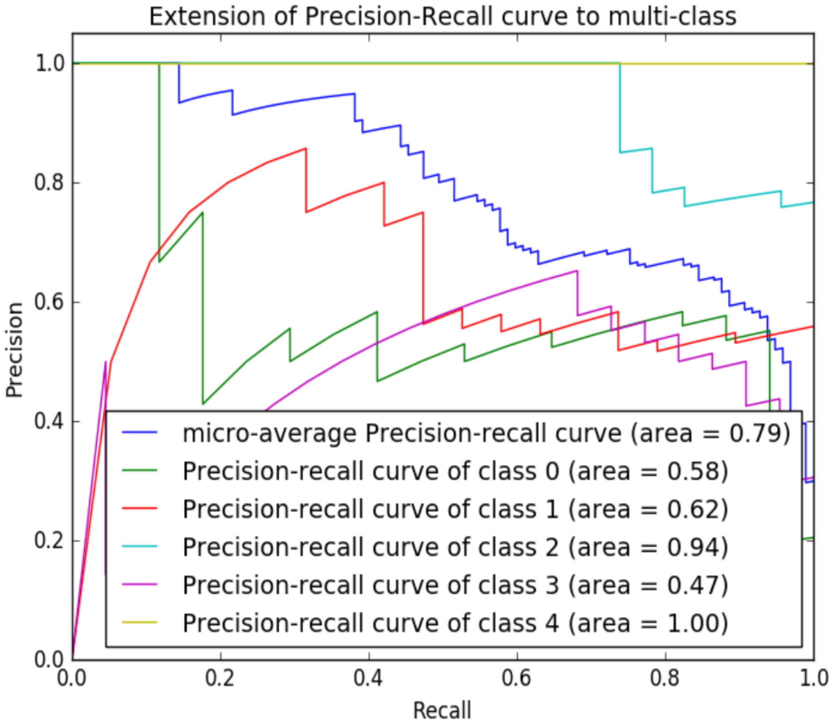


Figure 6. Precision recall curve for each class.

3.3. Discussion and findings

These experiments confirm that the hybrid approach of the LibD3C classifier yields better results than the standard pattern classification algorithms for predicting the disease from the mutated gene sequences, and high accuracy is attained. This work attains elevated kappa statistic and prediction accuracy through ensemble

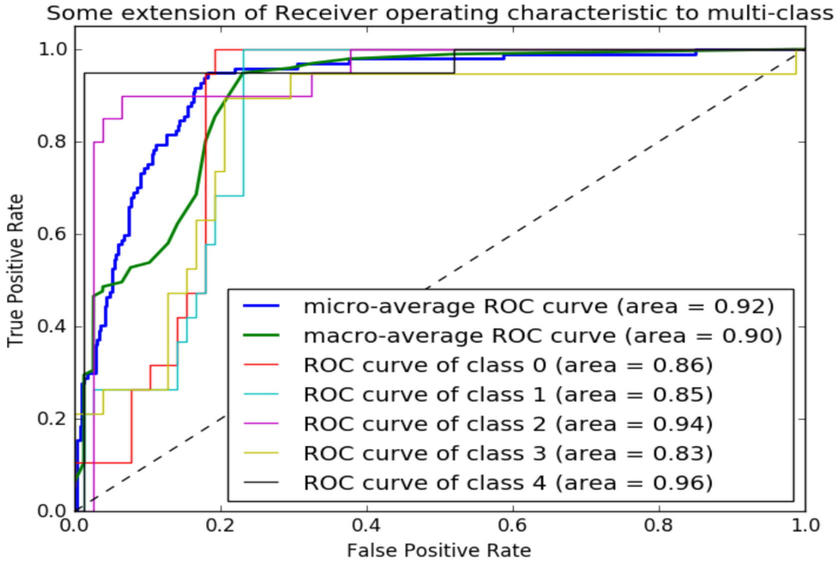


Figure 7. ROC curve for each class.

Table 8. Comparison of the results with existing studies.

	Classification	Data	Approach	Algorithm (method)	Accuracy (%)
Results of existing studies	LGMD	Family details	Machine learning	ANN	98
	6 types of MD	Microarray protein–protein interaction	Machine learning	MSVM	86
	FSHD	Microarray	Machine learning	SVM	84.65
	Gene type classification	HLA gene	Machine learning	SVM	99.3
	Virus type classification	HCV virus	Machine learning	SVM	100
	Cytokine classification	Protein	Ensemble learning	LibD3C	93.3
Results of proposed work	5 types of MD	Nonsynonymous mutations	Machine learning	LibD3C	97
	5 types of MD	Synonymous mutations	Machine learning	LibD3C	90
	5 types of MD	Insertion/deletion mutations	Machine learning	LibD3C	98.33

learning. The mean absolute error is minimized and the consistency of the system is improved. LibD3C is a hybrid approach that is a combination of unsupervised over supervised learning, a powerful approach for predicting the class label of the unseen instance. The time taken to build the model is a minimum of 3.75 s.

Reevaluation of features through feature selection facilitated the improvement of the outcome, and the prediction accuracy of the hybrid classifier built using high ranked features was elevated to 98.33%. It is well proven that the ensemble learning technique using LibD3C classifier is powerful and suitable for predicting muscular dystrophy when any type of mutational features are utilized for building the models. The work to date ascertains the highly efficient creation of muscular dystrophy prediction models through ensemble learning and discriminative mutational descriptors.

4. Conclusion

Muscular dystrophy disease identification is a multiclass classification problem that classifies the type of disease from the mutated gene sequences. Exonic and intronic descriptors pertaining to insertion, deletion, and duplication mutations are generated from the cloned gene sequences, and an eminent model is built by engaging the machine learning technique through a LibD3C ensemble classifier in hybrid learning method. Feature selection technique is introduced to improve prediction accuracy. The performance of the classifier was evaluated based on various metrics and the results indicate that the LibD3C algorithm is best suited for predicting the type of muscular dystrophy. Several types of mutations, such as missense, nonsense, and silent mutations, are also considered for building hybrid models through ensemble learning technique, and their results are analyzed. This work establishes that the ensemble learned model is highly efficient in the prognosis of neuromuscular genetic disorders.

References

[1] Fajkusova L, Lukas Z, Tvrdikova M, Kuhrova V, Hajek J, Fajkus J. Novel dystrophin mutations revealed by analysis of dystrophin mRNA alternative splicing suppresses the phenotypic effect of a nonsense mutation. Neuromuscular Disord 2001; 11: 133-138.

[2] Zubrzycka-Gaarn EE, Bulman DE, Karpati G, Burghes AH, Belfall B, Klamut HJ, Talbot J, Hodges RS, Ray PN, Worton RG. The Duchenne muscular dystrophy gene product is localized in sarcolemma of human skeletal muscle. Nature 1988; 333: 466-469.

[3] Kann MG. Advances in translational bioinformatics: computational approaches for the hunting of disease genes. Brief Bioinform 2009; 11: 96-110.

[4] Tranchevent LC, Capdevila FB, Nitsch D, De Moor B, De Causmaecker P, Moreau Y. A guide to web tools to prioritize candidate genes. Brief Bioinform 2010; 12: 22-32.

[5] Jones KJ, North KN. Recent advances in diagnosis of childhood muscular dystrophies. J Pediatr Child Health 1997; 33: 195-201.

[6] Koenig M, Hoffman EP, Bertelson CJ, Monaco AP, Feener C, Kunkel LM. Complete cloning of the Duchenne muscular dystrophy (DMD) cDNA and preliminary genomic organization of the DMD gene innormal and affected individuals. Cell 1987; 50: 509-517.

[7] Charif D, Thioulouse J, Lobry JR, Perriere G. Online synonymous codon usage analyses with the ade4 and seqinR packages. Bioinformat Oxford J 2005; 21: 545-547.

[8] Ban HJ, Heo JY, Oh KS, Park KJ. Identification of Type 2 diabetes-associated combination of SNPs using support vector machine. BMC Genet 2010; 23: 11-26.

[9] Uhmn S, Kim DH, Ko YW, Cho S, Cheong J, Kim J. A study on application of single nucleotide polymorphism and machine learning techniques to diagnosis of chronic hepatitis. Expert Syst 2009; 26: 60-69.

[10] Briggs FB, Ramsay PP, Madden E, Norris JM, Holers VM, Mikuls TR, Sokka T, Seldin MF, Gregersen PK, Criswell LA et al. Supervised machine learning and logistic regression identifies novel epistatic risk factors with PTPN22 for rheumatoid arthritis. Genes Immunol 2010; 11: 199-208.

[11] Nicodemus KK, Callicott JH, Higier RG, Luna A, Nixon DC, Lipska BK, Vakkalanka R, Giegling I, Rujescu D, Muglia P et al. Evidence of statistical epistasis between DISC1, CIT and NDEL1 impacting risk for schizophrenia: biological validation with functional neuroimaging. Human Genet 2010; 127: 441-452.

[12] González-Navarro FF, Belanche-Muñoz LA, Silva-Colón KA. Effective classification and gene expression profiling for the facioscapulohumeral muscular dystrophy. PLoS ONE 2013; 8(12).

[13] Ma J, Nguyen MN, Rajapakse JC. Gene classification using codon usage and support vector achines. IEEE-ACM T Comput Bi 2009: 1: 1545-5963.

[14] Nisha CM, Bhasker P, Pardasani KR. SVM model for classification of genotypes of HCV using relative synonymous codon usage. J Adv Bioinform Appl 2012; 3: 357-363.

[15] Falk CT, Gilchrist JM, Pericak Vance MA, Speer MC. Using neural networks as an aid in the determination of disease status: comparison of clinical diagnosis to neural-network predictions in a pedigree with autosomal dominant Limb-Girdle muscular dystrophy. Am J Hum Genet 1998; 62: 941-949.

[16] Wang C, Ha S, Xuan J, Wang Y, Hoffmann E. Computational analysis of muscular dystrophy sub-types using a novel integrative scheme. Neurocomputing 2012; 1: 9-17.

[17] Zou Q, Wang Z, Guan X, Liu B, Wu Y, Lin Z. An approach for identifying cytokines based on a novel Ensemble classifier. Hindawi Publishing Corporation 2013: 686090.

[18] Kalari KR. Computational approach to identify deletions or duplications within a gene. PhD, University of Iowa, Iowa City, IA, USA, 2006.

[19] Wu J, Zhang W, Jiang R. Comparative study of ensemble learning approaches in the identification of disease mutations. In: IEEE 2013 International Conference on Biomedical Engineering and Informatics; 16–18 October 2010; Yantai, China. New York, NY, USA: IEEE. pp. 2306-2310.

[20] Sathyavikasini K, Vijaya MS. Predicting muscular dystrophy with sequence based features for point mutations. In: 2015 IEEE Conference on Research in Computational Intelligence and Communication Networks; 20–22 November 2015; Kolkata, India. New York, NY, USA: IEEE. pp. 235-240.

[21] Sathyavikasini K, Vijaya MS. Muscular dystrophy disease classification using relative synonymous codon usage. Int J Mach Learn Comput 2016; 6: 139-144.

[22] Stenson PD, Mort M, Ball EV, Shaw K, Phillips A, Cooper DN. The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. Hum Genet 2013; 133: 1-9.

[23] Lin C, Chen W, Qiu C, Wu Y, Krishnan S, Zoua Q. LibD3C: Ensemble classifiers with a clustering and a dynamic strategy. Els Neurocomput 2014; 123: 424-435.

[24] Nasierding G, Kouzani AZ, Tsoumakas G. A triple-random ensemble classification method for mining multi-label data. In: IEEE 2010 International Conference on Data Mining Workshops; 13 December 2010; Sydney, NSW, Australia. New York, NY, USA: IEEE. pp. 667-685.

[25] Yan R, Tesic J, Smith JR. Model-shared subspace boosting for multi-label classification. In: ACM 2007 International Conference on Knowledge Discovery and Data Mining; 12–15 August 2017; San Jose, CA, USA. New York, NY, USA: ACM, pp. 834-843.

[26] Zhou ZH, Wu J, Tang W. Ensembling neural networks: many could be better than all. Artif Intell 2002; 137; 239-263.

[27] Witten IH, Frank E, Trigg L, Hall M, Holmes G, Cunningham SJ. Weka: Practical Machine Learning Tools and Techniques with Java Implementations. Working Paper No. 99/11, 1999.