6. MUSCULAR DYSTROPHY DISEASE IDENTIFICATION MODELS THROUGH ENSEMBLE LEARNING

The focus of this chapter is to give a detailed explanation on the implementation of muscular dystrophy disease identification models using ensemble learning. In machine learning, the hybrid approach has been an ongoing research area for pulling better performance for classification or prediction problems over a single learning approach. LibD3C is a kind of ensemble classifier with a clustering and dynamic selection strategy. Two types of selective ensemble techniques namely, combination of the ensemble pruning based on k-means clustering and dynamic selection and circulating combination have been employed using LibD3C Classifier.

In this experiment, muscular dystrophy disease prediction models are built using LibD3C algorithm by implementing LibD3C jar files in matlab environment. The performance of the LibD3C ensemble models is compared against the performance of standard supervised disease identification models and the results are analyzed.

6.1 Building Disease Identification Models through LibD3C Classifier

Based on various mutational features and the respective datasets, five independent experiments were carried out such as, (i) Predicting muscular dystrophy disease using features related to missense and nonsense (Non-synonymous) mutations (ii) Predicting muscular dystrophy disease using features related to synonymous mutations (iii) Predicting muscular dystrophy disease using features related to synonymous mutations insertion/deletion an duplication mutations (iv) Predicting muscular dystrophy disease using features related to splicing mutations (v) Predicting muscular dystrophy disease using pooled features using LibD3C classifier.

(i) Predicting Muscular Dystrophy Disease using Features related to Missense and Nonsense (Non-Synonymous) Mutations

In this work, ensemble disease identification model is built with Non synonymous mutational features and NSM dataset (section 5.2) using LibD3C classifier. Evaluating the generalization power of the classifiers and to estimate their predictive capabilities for unknown samples, a standard 10- fold cross-validation technique is used to split the data randomly and repeatedly into training and test sets. The results of performance evaluation for various metrics are tabulated in the Table 6.1and shown in Fig.6.1.

Table 6.1 Predictive Performance of the LibD3C Classifier (Non Synonymous mutation)

Performance criteria	**LibD3C Classifier**
Kappa Statistic	0.95
Mean Absolute Error	0.015
Root Mean Squared Error	0.15
Relative absolute error	20.74
Root relative square error	39.45
Time taken to build the model (in sec)	4.5
Correctly classified Instance	970
Incorrectly classified instance	30
Prediction accuracy	97%

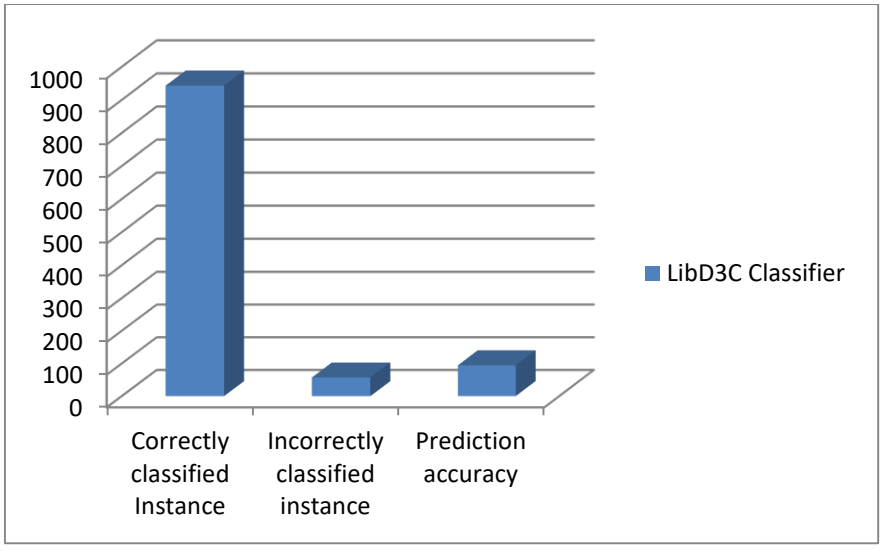

Fig.6.1 Prediction Accuracy of LibD3C Classifier (Non Synonymous Mutation)

Findings

Prediction accuracy of about 97% is attained for missense, nonsense mutational descriptors with a kappa of 0.95 and learning time of 4.5 sec. The mean absolute error and mean squared error scores very low values of 0.015 and 0.15.

(ii) Predicting Muscular Dystrophy Disease using Features related to Synonymous Mutations

Relative Synonymous Codon Usage (RSCU) values for 59 codons contributing synonymous mutational features and the corresponding SYM dataset (refer section 5.3) is used here to built ensemble based disease identification models through LibD3C classifier. The performance of the trained models is evaluated using the same validation technique and measured in terms of various metrics as in previous case. The results are tabulated in Table 6.2and shown in Fig.6.2.

Table 6.2 Predictive Performance of the LibD3C Classifier (Synonymous Mutation)

Performance criteria	**LibD3C Classifier**
Kappa Statistic	0.85
Mean Absolute Error	0.09
Root Mean Squared Error	0.202
Relative absolute error	29.844
Root relative square error	50.801
Time taken to build the model (in sec)	7.36
Correctly classified Instance	900
Incorrectly classified instance	100
Prediction accuracy	90%

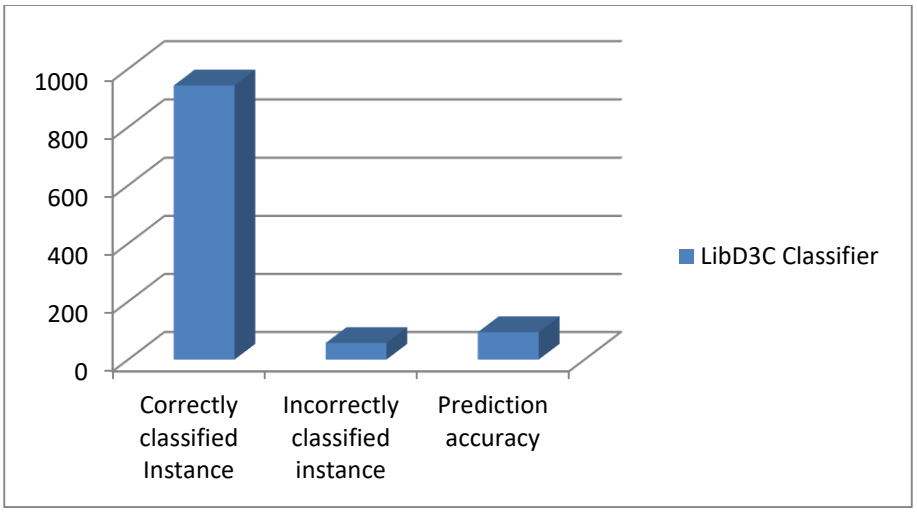

Fig.6.2 Prediction Accuracy of LibD3C Classifier (Synonymous Mutation)

Findings

With the 10-fold cross validation testing the prediction accuracy of about 90% is attained for synonymous mutational descriptors with a kappa of 0.85 and learning time of 7.36 sec. The mean absolute error and mean squared error scores achieved are 0.09 and 0.202.

(iii) Predicting Muscular Dystrophy Disease using Features related to Insertion/Deletion and Duplication mutations

The third experiment aims at building the disease identification model with IDM dataset having exonic and intronic features (refer section 5.4) through ensemble learning.LibD3C classifier is employed in performing training and classification and the results of classifier is evaluated with same performance metrics. The values obtained on various metrics are summarized in the Table 6.3and shown in Fig.6.3.

Table 6.3 Predictive Performance of the LibD3C Classifier (Insertion, Deletion, Duplication mutation)

Performance criteria	**LibD3C Classifier**
Kappa Statistic	0.97
Mean Absolute Error	0.065
Root Mean Squared Error	0.14
Relative absolute error	23.44
Root relative square error	41.65
Time taken to build the model (in sec)	2.76
Correctly classified Instance	983
Incorrectly classified instance	17
Prediction accuracy	**98.3%**

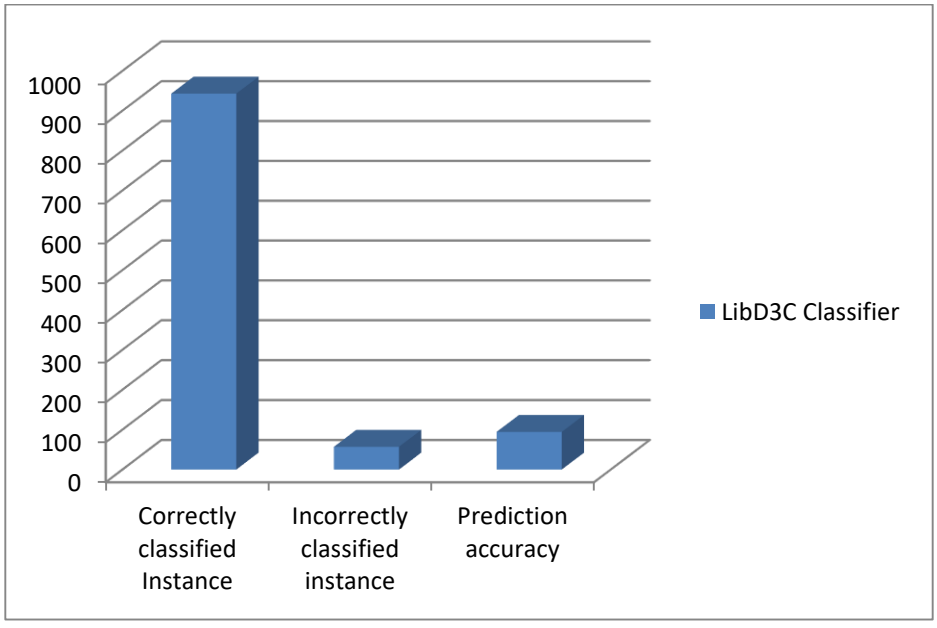

Fig.6.3 Prediction Accuracy of LibD3C Classifier (Insertion, Deletion, Duplication mutation)

Findings

The prediction accuracy of about 98.3% is attained in case of exonic and intronic descriptors for identifying the muscular dystrophy disease. Kappa statistic of 0.97 and learning time of 2.76 sec is attained. Minimal of 0.065 and 0.14 is achieved as the mean absolute error and mean squared error scores.

(iv) Predicting muscular dystrophy disease using features related to Splicing mutations

This experiment aims in building muscular dystrophy disease identification model with SPM dataset containing features related to splicing mutations (refer section 5.4) using LibD3C classifier. Predictive capabilities for unknown samples are estimated using a standard k- fold cross-validation technique with K=10 and the results of the experiment are summarized in the Table 6.4and shown in Fig.6.4.

Table 6.4 Predictive Performance of the LibD3C Classifier (Splicing Mutation)

Performance criteria	**LibD3C Classifier**
Kappa Statistic	0.931
Mean Absolute Error	0.098
Root Mean Squared Error	0.19
Relative absolute error	19.44
Root relative square error	45.5
Time taken to build the model (in sec)	3.47
Correctly classified Instance	943
Incorrectly classified instance	57
Prediction accuracy	**94.3%**

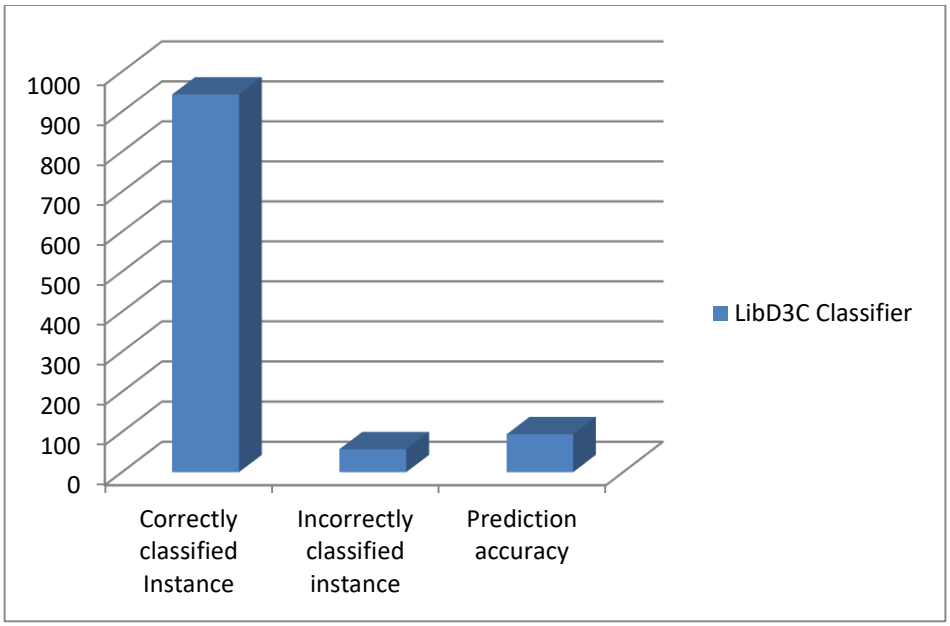

Fig.6.4 Prediction Accuracy of LibD3C Classifier (Splicing mutation)

Findings

94.3% of prediction accuracy is attained while learning the model with exonic and intronic features that modulate splicing for identifying the muscular dystrophy disease. Kappa statistic of 0.931 and learning time of 3.47 sec is attained. The mean absolute error and mean squared error scores are achieved with minimal of 0.098 and 0.19.

(v) Predicting Muscular Dystrophy Disease using Pooled Features

In this work, a generalized muscular dystrophy disease identification model is built using LibD3C classifier by utilizing the aggregated features and corresponding AGM dataset (refer section 5.5). As the result analysis of the experiment described in section 5.5, proves that feature selection method produce better accuracy, in this experiment only high ranked feature set is used to train LibD3C classifier. The cross validation results of the classifier are summarized in the Table 6.5 and shown in Fig.6.5.

Table 6.5 Predictive Performance of the LibD3C Classifier (AGM Dataset)

Performance criteria	**LibD3C Classifier**
Kappa Statistic	0.97
Mean Absolute Error	0.065
Root Mean Squared Error	0.14
Relative absolute error	23.44
Root relative square error	41.65
Time taken to build the model (in sec)	2.76
Correctly classified Instance	985
Incorrectly classified instance	15
Prediction accuracy	**98.5%**

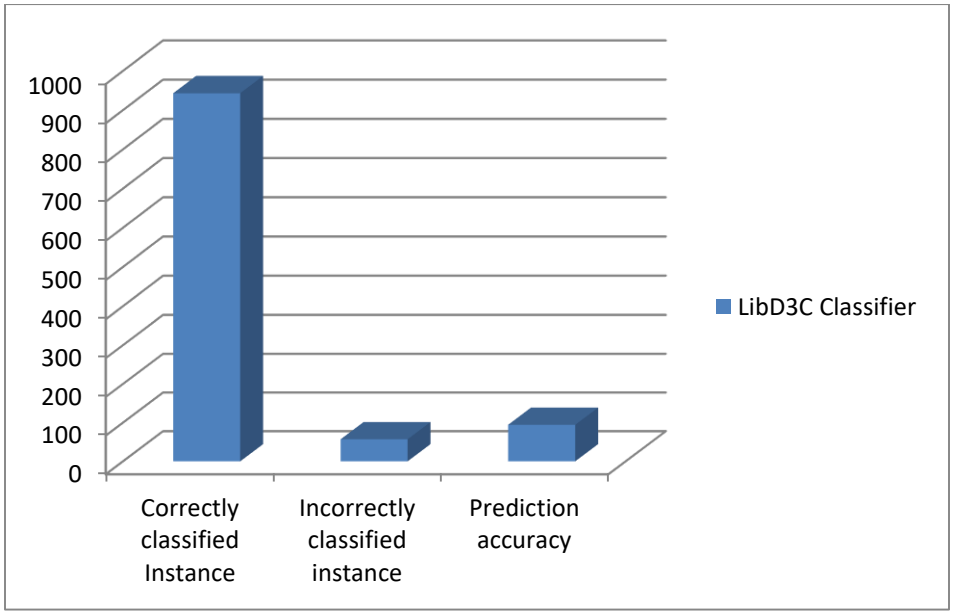

Fig.6.5 Prediction Accuracy of LibD3C Classifier (AGM Dataset)

Findings

With the 10-fold cross validation testing the high prediction accuracy of about 98.5% is achieved by aggregating all the mutational descriptors with a kappa of 0.97 and learning time of 2.76 sec. The AGM model is a generalized model which can identify any kind of disease effectively by aggregating all type of mutational features. The mean absolute error and mean squared error scores achieved are 0.065 and 0.14.

Comparison of all five LibD3C based disease identification models

The performance of the above LibD3C based disease identification models are compared with respect to various measures such as Precision, Recall, F-measure, TP rate, FP rate, ROC area and tabulated in Table 6.6. The results indicates that high performance is achieved with the ensemble model built on AGM dataset with feature selection. Thus enables a generalized disease identification model, which can predict the disease when any type of mutation incurs.

Table 6.6 Comparison of LibD3C Classifiers

Datasets	Precision	Recall	F- Measure	TP Rate	FP Rate	ROC Area
NSM	0.967	0.23	96.8	96.8	0.014	0.97
SYM	0.901	0.72	89.3	89.3	0.018	0.91
IDM	0.983	0.45	98.2	98.2	0.08	0.997
SPM	0.94	0.52	93.2	96.5	0.091	0.967
AGM	0.985	0.49	98.4	98.4	0.09	0.997

6.2 Performance Comparison of LibD3C Classifiers with Supervised Learning Models

The performances of the five ensemble models built using LibD3C classifier is compared with the results of supervised disease identification models (refer Chapter 5) and the performances are analysed with respect to accuracy. The below depicted Table 6.7 and Fig.6.6 elucidates the performance comparison of the disease identification models on various datasets.

Table 6.7 Predictive Accuracy Comparison of LibD3C Classifier with Supervised Learning Algorithms

Dataset	Decision Tree classifier	Naïve Bayes classifier	Artificial Neural Network	SVM	LiBD3C Classifier
NSM	82.5	82.3	71.8	86.8	97
SYM	85.3	85.6	83.1	86.3	90
IDM	87.5	87.8	85.3	88.5	98.3
SPM	86.9	85.7	83.5	88.9	94.3
AGM	86.9	85.7	83.5	88.9	98.5

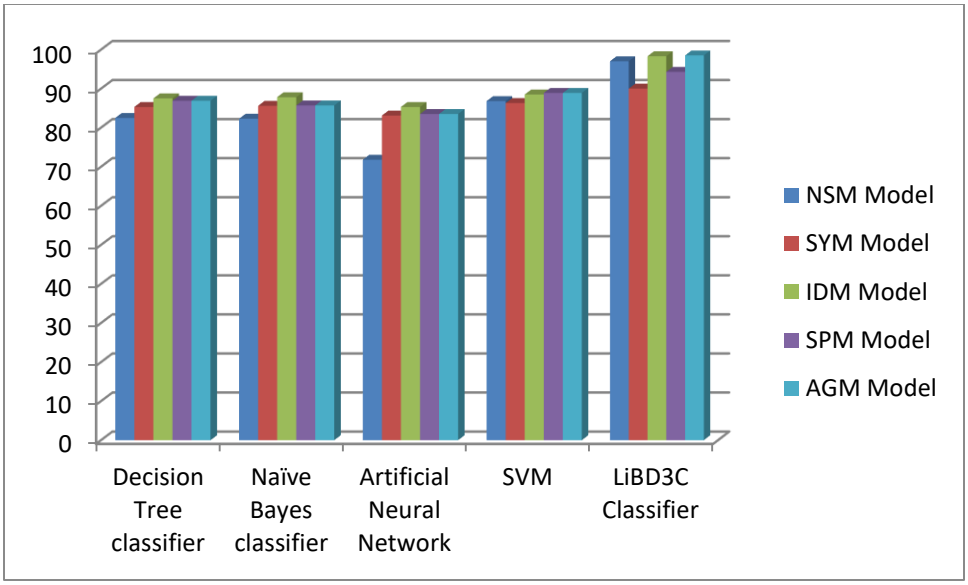

Fig.6.6 Performance Comparison of LibD3C with Supervised Models

Findings

From the above analysis, it is perceived that LibD3C is suitable for predicting the disease from the mutated gene sequences and their discriminative mutational descriptors as elevated prediction accuracy is attained from the ensemble learning methodology. It is confirmed that the hybrid approach of LibD3C classifier yields better results than the standard pattern classification

algorithms for predicting the disease from the mutated gene sequences as enviable accuracy is attained. The time taken to build the models is very minimal. Also, it is proved that ensemble learning technique using LibD3C classifier is suitable to predict muscular dystrophy disease when any type of mutational features are utilised for building the models.

6.3 Summary

Muscular dystrophy disease identification models have been built using the LibD3C classifier. The performance of the models are analysed and the observations are discussed. In addition, the predictive performances of LibD3C classifiers are compared against the performances of the models built using supervised learning algorithms. It was observed that LibD3C classifier attains a better performance than the standard pattern classification algorithms discussed in chapter 5. The upcoming chapter demonstrates the development of muscular dystrophy disease identification model through deep learning approach.

Remarks

1. Paper titled “Ensemble Learning for identifying Muscular dystrophy diseases using codon bias pattern”, has been published in the Proceedings of the 5th International Conference on Frontiers in Intelligent Computing: Theory and Applications, Advances in Intelligent Systems and Computing 515, ISBN 978-981-10-3152-6, Vol 1, pp: 21-29, Springer (AISC) series (Scopus indexed).
2. Paper titled "Prognosis of Muscular dystrophy disorder with Extrinsic and intrinsic descriptors through ensemble learning", has been accepted for publication in the Turkish Journal of Electrical Engineering and Computer Sciences (SCIE Indexed).