International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-2, July 2019

Human Protein Sequence Classification using Machine Learning and Statistical Classification Techniques

ChhoteLal Prasad Gupta, AnandBihari, SudhakarTripathi

Abstract: ***In the field of computational biology, to gauge the meaningful and accurate feature for protein function predications, either the profile-based protein data or sequence-based data has been used. As we know that the prediction of enzyme class from an unknown protein is most interacted research in the current era. In this context, machine learning and statistical classification technique has been used. In this article, we have use six different machine learning and statistical classification technique such as CRT, QUEST, CHAID, C5.0, ANN and SVM for classification of 4314 number of human protein sequence data. These data are extracted form UniprotKB databank with the help of PROFEAT server. The extracted data are categorized in seven different classes. To manipulate the high dimensional protein sequence data with some missing value, the SPSS has been used for classification and estimation of the performance of classification technique. The experimental results highlight that the class C4, C5, C6 and C7 data are imbalanced that affect the overall performance of classification technique. This article provides an extensive comparative analysis of different classification technique on sequence-based protein data. The experimental analysis highlights that the SVM and C5.0 classification technique gives better result than others and can be used for protein classification and predictions.***

Keywords: ***Protein function prediction; enzyme classification; classification techniques; UniProtKB; FASTA; Protein Sequence; etc.***

I. INTRODUCTION

Protein is a sequence of amino acids binding with the peptide bond that plays a significant role in maintaining the life [1]. It helps in improving the function of organs and tissues of human body [2-5] and to determine the function and structure of the protein by using the sequence of the amino acids. Basically, protein has three structures primary, secondary and tertiary. The structure of the protein helps in determining the functional behavior of the proteins and function predictions.It can also be used for finding sequence similarity [6], to cluster the similar type of proteins [7], to find the interaction between proteins [8] and many other works [9]. In biological research basically the protein function prediction is done with the help of either sequence or structure similarity. But this type of prediction takes a lot of resources and computation time [10]. To improve the computation accuracy with reduction of resources and computation time, the machine learning classification technique is used [11-15]. Dobson and Doig (2005) [16] discussed a new methodology using machine learning for enzyme class prediction from protein with the overall accuracy of 35%. Further Luizet. al. (2006) [17] used Bayesian classification technique and claimed with higher accuracy than Dobson and Doig model; i.e. 45%. Lee et al. (2009) used random forest classification technique and claimed that total 484 features. Apart from these researches several other research have been conducted for protein function prediction with the help of machine learning technique such as SVM, ANN and Decision Tree and many research indicates that the SVM gives better results than other classification techniques [10]. Kumar et al. [18] used the SVM for protein function predictions. Lou et al. [19] used SVM for finding DNA-binding sites. Liu et al. [20] used the random forest to predict DNA-binding with the help of DNA-binding proteins and amino acid. Amidi et al. (2016) [21] used the SVM and Nearest Neighbor classification technique for function prediction and claimed that the accuracy has been reached to 93.4%. Gupta et. al [40] used machine learning approach and claimed that the C5.0 gives 86.49% accuracy. Generally, the SVM classification technique is frequently used for protein function predictions. But as we know that the SVM classification technique is more suited for non-linear high dimensional data. The nature of the protein data is non-linear and high dimension also. It also seems that many of the proteins data having missing features value. In this case, SVM did not perform well. In this article we made a comparative analysis of six different classification technique such as CRT, QUEST, CHAID, C5.0, NEURAL, and SVM and found that the accuracy is above of 94% for all models. The experimental results highlight that the C5.0 and SVM classification technique are more suited for protein classification and predictions. The SVM are capable to handle the non-linear data and C5.0 capable for handling non-linear high dimensional data with missing value. This article discusses the six different classification techniquesand implemented on protein sequence data that are classified into seven different classes. Section 1 discusses how the proteins are beneficial for human body and the uses of protein features for disease therapy, Section 2 discuss the protein data extraction process and the classification technique that are used for protein classification.

Revised Manuscript Received on 30 July 2019.
***** Correspondence Author
ChhoteLal Prasad Gupta, Computer Science & Engineering, Dr. APJ Abdul Kalam Technical University, Lucknow, India, clpgupta@gmail.com
AnandBihari, School of Information Technology & Engineering, VIT University, Vellore, Tamil Nadu, India.anand.bihari@vit.ac.in
SudhakarTripathi,Deportment of Information Technology, Rajkiya Engineering College, Ambedkarnagar, Uttar Pradesh, India, p.stripathi@gmail.com

Section 3 gives the experimental result and analysis of protein data, finally section 4 conclude the article.

II DATA AND METHODS

A. Data collection

To gauge the functional behavior of proteins and their importance in human body from a broad collection of proteins is a widely attracted assignment. In this context, researchers are using the UniProtKB (UniProtKnowledgebase) [22]. The UniProtKB is a protein data-bank contains a huge amount of reviewed and un-reviewed protein data. The reviewed protein dataset is the Swiss-Prot and the un-reviewed protein dataset is the TrEMBL. The reviewed data set contains 559228 and the un-reviewed dataset contains 146106279 protein data till now. The reviewed data are inspected and corrected by the different protein forums and communities [23] and widely used to gauge the functional behavior of proteins. The reviewed dataset contains different organism. This study considers only the human organism data. Out of 559228, the 20417 proteins data are available for human. For this study, we have considered only enzyme class of data to estimate the functional behavior of proteins. The dataset contains 4314 number of proteins for experimental analysis and gauge the functional behavior of proteins. The human enzyme class of data is classified into seven different class (shown in Table 1.).

Table 1: Class wise data description of human enzyme class.

Class	Class Name	No. Protein
EZC1	Oxidoreductases	556
EZC2	Transferases	1712
EZC3	Hydrolases	1604
EZC4	Lyases	149
EZC5	Isomerases	117
EZC6	Ligases	125
EZC7	Translocases	51

B. Feature extraction

Nowadays, the feature selection to gauge the functional behavior of protein function and prediction is one of the most important tasks, because every protein contains a massive amount of feature. In real world a protein has many features and some of the features having very less significance in function prediction. The uniport databank contains two types of features (i) Profile Feature and (ii) Sequence Feature. Our previous study is based on the profile feature [40]. In this research we have consider the sequence-based features that are present in all seven class of protein data. To do this, first we have extracted the protein sequence data form uniport databank. The extracted data is in FASTA (Fast Adaptive Shrinkage Threshold Algorithm) format. After that, we have extracted the feature form FASTA file of protein sequence data with the help of PROFEAT sever. The profile-based protein contains only 48 features; however, sequence-based protein contains 1436 features. More number of feature considerations in function prediction may give better results than the consideration of relatively lesser number of features [24-25].

C. Classification of protein in enzyme class

As discussed in previous section, the human protein for enzyme class is classified into seven different enzyme class but has not guarantee the all proteins are accurately classified. The wrongly classified proteins may affect the performance and accuracy of prediction of function. In machine learning, there are many classification techniques are defined to classify such type of data. In this study, we have considered the following well-known highly recommended classification techniques such as (i) CRT, (ii) QUEST, (iii) CHAID, (iv) C5.0, (v) ANN, and (vi) SVM.

(i) CRT: In this classification, data is classified with the help of classification tree and the predictions are based on the regression tree [26].

(ii) QUEST: It is a tree based binary classification technique. It reduces the computation time than the others tree-based classification. In this classification, the statistical test has been conducted to select an input field. It also separates the input selection and the splitting of trees [27].

(iii) CHAID: It is a tree-based model for classification and prediction of variables and also finds the interaction between variables. It builds a non-binary tree by using multiple regressions. The main objective of the CHAID technique is to find how one variable affect the performance of other variables [28].

(iv) C5.0: It is an extension of ID3 algorithm of decision tree. It produces a binary tree with multiple branches. It deals with all possible data including the missing values. It is discrete and continuous in nature [29].

(v) ANN: It is basically used to estimate the performance of biological networks. In this technique the learning process is based on adjustment of weight between connection of neurons and the output of the model is depends on the activation function [30].

(vi) SVM: It is one of the most influenced classification techniques based on statistical learning for classifications and prediction of data [31-35]. It deals with wide variety of classification problems including the non-linearly high dimensional problem.

D. Performance evaluation metrics

To estimate the performance of above mentioned seven classifiers, the following evaluation metrics: (a) Accuracy (AC), (b) Sensitivity (ST), (c) Specificity (SP), (d) F-measure and (e) MCC (Matthew's correlation coefficient) are used in this research [38-39], which are described as:

$$AC = \frac{\mathrm{TP} + \mathrm{TN}}{TP + FN + FP + TN} \quad (1)$$

$$ST = \frac{\mathrm{TP}}{TP + FN} \quad (2)$$

$$SP = \frac{\mathrm{TN}}{FP + TN} \quad (3)$$

$$F - measure = \frac{2 \text{ x PR x RC}}{PR + RC} (4)$$

$$MCC = \frac{(\text{TP x TN}) - (\text{FP x FN})}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} (5)$$

Where TP, FP, TN and FN are number of true positive, false positive, true negative and false negative of proteins respectively. Precision (PR) and Recall (RC) are equivalent to sensitivity and specificity respectively.

III. EXPERIMENTAL RESULT AND ANALYSIS

In this section, our primary goal is to find the minimal number of features to gauge the functional behavior of proteins and predict the appropriate class for the respective proteins. In this context, we have extracted 4314 number of proteins data and every protein contains all 1438 features. After study of protein sequence-based features data, we found that some of the features value is relatively too high for some of the proteins; however, some of the features value is 0 or very less. That may not affect much in computations, but we have considered all features and used six highly recommended classification techniques for classification of protein. All experiments are done with the help of SPSS Clementine tools. The experimental results highlight that the some of the classification technique considers all features to classify the protein data. Model wise number of selected features is shown in Table 2.

Table 2: Model-wise selected feature

Model	Total No. of selected features	Model	Total No. of selected features
CRT	275	**C5.0**	1347
QUEST	75	**ANN**	1438
CHAID	250	**SVM**	1438

From the Table 2, it can be clearly concluded that the QUEST classification technique selected only 75 numbers of features, however the ANN and SVM is selected maximum number of features i.e. 1438. C5.0 classification technique also selected near about the SVM and ANN. Our main objective is to consider maximum number of features in classification, as per our assumption the C5.0, ANN and SVM will perform well. To validate the above statements, we have used Accuracy, Specificity, Sensitivity, Precision, Recall and MCC. Now, we present the experimental result of all six models discussed above.

(i) CRT based classification

In this classification technique, the classification tree has been used to classify the data and the regression tree has been used for prediction. The CRT classification technique is implanted on total 4314 data of protein of all seven-enzyme class with 1438 features. After classification it produces total 275 numbers of features for predication of appropriate class of respective proteins. The performance of the CRT model on the given dataset is shown in Table 3.

Table 3: Result of the CRT model on Protein sequence data

CRT							
	C1	**C2**	**C3**	**C4**	**C5**	**C6**	**C7**
True positive (TP)	40	967	1073	0	0	0	0
True negative (TN)	3742	1756	1338	4165	4197	4189	4263
False positive (FP)	16	846	1372	0	0	0	0
False negative (FN)	516	745	531	149	117	125	51
	C1	**C2**	**C3**	**C4**	**C5**	**C6**	**C7**
Accuracy	87.66806	63.12007	55.88781	96.54613	97.2879	97.10246	**98.8178**
Sensitivity	0.071942	0.564836	**0.668953**	0	0	0	0
Specificity	0.995742	0.674865	0.493727	**1**	**1**	**1**	**1**
Precision	**0.714286**	0.53337	0.438855	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!
Recall	0.071942	0.564836	**0.668953**	0	0	0	0
F-measure	0.130719	**0.548652**	0.530007	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!
MCC	0.20036	**0.237586**	0.158663	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!

(ii) QUEST:

It is an extended version of the tree-based classification technique. It works on the binary based either the data is classified or not and the statistical test has been done by the splitting of tree and the input field. The QUEST has been implanted on total 4314 data of protein of all seven-enzyme class with 1438 features. After classification it produces only 75 numbers of features for function predication. The performance of the QUEST model on the given dataset is shown in Table 4.

Table 4: Result of the QUEST model on Protein sequence data

QUEST							
	C1	C2	C3	C4	C5	C6	C7
True positive (TP)	0	1701	24	0	0	0	0
True negative (TN)	3758	43	2680	4165	4197	4189	4263
False positive (FP)	0	2559	30	0	0	0	0
False negative (FN)	556	11	1580	149	117	125	51
	C1	C2	C3	C4	C5	C6	C7
Accuracy	87.11172925	40.42651831	62.67965	96.54613	97.2879	97.10246	**98.8178**
Sensitivity	0	**0.993574766**	0.014963	0	0	0	0
Specificity	**1**	0.016525749	0.98893	**1**	**1**	**1**	**1**
Precision	#DIV/0!	0.399295775	**0.444444**	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!
Recall	0	**0.993574766**	0.014963	0	0	0	0
F-measure	#DIV/0!	**0.569658406**	0.028951	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!
MCC	#DIV/0!	**0.044447506**	0.01692	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!

(iii) CHAID (Chi-square Automatic Interaction Detection)

This classification technique has been used to find how one variable affect the performance of other variables. The CHAID has been implanted on total 4314 data of protein with1438 features. After classification it produces only 250 numbers of features for function predication. The performance of the CHAID model on the given dataset is shown in Table 5.

Table 5: Result of the CHAID model on Protein sequence data

CHAID							
	C1	C2	C3	C4	C5	C6	C7
True positive (TP)	82	1233	785	0	0	0	0
True negative (TN)	3637	1266	1953	4165	4197	4189	4263
False positive (FP)	121	1336	757	0	0	0	0
False negative (FN)	474	479	819	149	117	125	51
	C1	C2	C3	C4	C5	C6	C7
Accuracy	86.2077	57.92768	63.46778	96.54613	97.2879	97.10246	**98.8178**
Sensitivity	0.147482	**0.72021**	0.489401	0	0	0	0
Specificity	0.967802	0.486549	0.720664	**1**	**1**	**1**	**1**
Precision	0.403941	0.479953	**0.509079**	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!
Recall	0.147482	**0.72021**	0.489401	0	0	0	0
F-measure	0.216074	**0.576034**	0.499046	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!
MCC	0.182416	0.206106	**0.211838**	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!

(iv) SVM:

It is one of the most influenced classification techniques based on statistical learning for classifications and prediction of data [7,8,9]. It deals with wide variety of classification problems including the non-linearly high dimensional problem. It provides more efficient predication than others classification problem. The SVM has been implanted on total 4314 data of protein with 1438 features. After classification it produces 1438 number of features for function predication. The performance of the SVM model on the given dataset is shown in Table 6.

Table 6: Result of the SVM model on Protein sequence data

SVM							
	C1	C2	C3	C4	C5	C6	C7
True positive (TP)	523	1634	1533	122	106	110	43
True negative (TN)	3727	2500	2614	4158	4196	4183	4263
False positive (FP)	31	102	96	7	1	6	0
False negative (FN)	33	78	71	27	11	15	8
	C1	C2	C3	C4	C5	C6	C7
Accuracy	98.51646	95.82754	96.12888	99.21187	99.72184	99.51321	**99.8146**

Sensitivity	0.940647	0.954439	**0.955736**	0.818792	0.905983	0.88	0.84314
Specificity	0.991751	0.960799	0.964576	0.998319	0.999762	0.998568	**1**
Precision	0.944043	0.941244	0.941068	0.945736	0.990654	0.948276	0.999987
Recall	0.940647	0.954439	**0.955736**	0.818792	0.905983	0.88	0.84314
F-measure	0.942342	0.947796	**0.948345**	0.877698	0.946429	0.912863	0.91489
MCC	0.933831	0.913111	0.917464	0.876071	**0.945998**	0.911035	0.91736

(V) C5.0

The one of the important features of the C5.0 is, it considers missing feature data for function prediction. In case of protein data, there are several data have missing feature value. This classification technique is more suited in protein classification. The C5.0 has been implanted on total 4314 data of protein of all seven-enzyme class with 1438 features. After classification it produces 1347 number of features for function predication. The performance of the C5.0 model on the given dataset is shown in Table 7.

Table 7: Result of the C5.0 model on Protein sequence data

C5.0							
	C1	C2	C3	C4	C5	C6	C7
True positive (TP)	507	1604	1519	94	78	78	33
True negative (TN)	3659	2475	2568	4148	4189	4182	4262
False positive (FP)	99	127	142	17	8	7	1
False negative (FN)	49	108	85	55	39	47	18
	C1	C2	C3	C4	C5	C6	C7
Accuracy	96.56931	94.55262	94.73806	98.33102	98.91052	98.74826	**99.55957**
Sensitivity	0.911871	0.936916	**0.947007**	0.630872	0.666667	0.624	0.647059
Specificity	0.973656	0.951191	0.947601	0.995918	0.998094	0.998329	**0.999765**
Precision	0.836634	0.926632	0.914509	0.846847	0.906977	0.917647	**0.970588**
Recall	0.911871	0.936916	**0.947007**	0.630872	0.666667	0.624	0.647059
F-measure	0.872633	**0.931746**	0.930475	0.723077	0.768473	0.742857	0.776471
MCC	0.853908	0.886462	**0.888519**	0.722906	0.772522	0.751107	0.79062

(Vi) ANN

The ANN has been implanted on total 4314 data of protein of all seven-enzyme class with 1438 features. After classification it produces 1347 number of features for function predication. The performance of the ANN model on the given dataset is shown in Table 8.

Table 8: Result of the ANN model on Protein sequence data

ANN							
	C1	C2	C3	C4	C5	C6	C7
True positive (TP)	125	1039	1173	0	0	0	0
True negative (TN)	3642	1863	1588	4165	4197	4189	4263
False positive (FP)	116	739	1122	0	0	0	0
False negative (FN)	431	673	431	149	117	125	51
	C1	C2	C3	C4	C5	C6	C7
Accuracy	87.32035	67.26936	64.00093	96.54613	97.2879	97.10246	**98.8178**
Sensitivity	0.22482	0.606893	**0.731297**	0	0	0	0
Specificity	0.969133	0.715988	0.585978	**1**	**1**	**1**	**1**
Precision	0.518672	**0.584364**	0.511111	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!
Recall	0.22482	0.606893	**0.731297**	0	0	0	0
F-measure	0.313676	0.595415	**0.601693**	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!
MCC	0.282974	**0.320927**	0.3073	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!

IV ANALYSIS OF RESULT

In this article, we have considered sequence-based feature data to classify and predict the appropriate class of protein. To find the appropriate model, we have made a comparative analysis of performance of different models. This comparative analysis helps in finding the appropriate model for the particular class of protein as well as whole proteins. The model wise comparative analysis highlights that the which model is good for the which class. The class wise comparative analysis of performance of CRT based model is shown in Table 9.

Table 9: A Comparative analysis of performance of CRT for all class

CRT							
	Accuracy	**Sensitivity**	**Specificity**	**Precision**	**Recall**	**F-measure**	**MCC**
C1	87.66805749	0.071942446	0.995742416	**0.714285714**	0.071942446	0.130718954	0.200359754
C2	63.12007418	0.564836449	0.674865488	0.533370105	0.564836449	**0.548652482**	**0.23758626**
C3	55.88780714	**0.668952618**	0.493726937	0.438854806	**0.668952618**	0.530007409	0.158663022
C4	96.54612888	0	**1**	#DIV/0!	0	#DIV/0!	#DIV/0!
C5	97.28789986	0	**1**	#DIV/0!	0	#DIV/0!	#DIV/0!
C6	97.10245712	0	**1**	#DIV/0!	0	#DIV/0!	#DIV/0!
C7	**98.8178025**	0	**1**	#DIV/0!	0	#DIV/0!	#DIV/0!

It can be clearly seen that the accuracy of class 7 is 98.8178025, the sensitivity of class 3 is 0.668952618, the specificity of the class 4, 5, 6, & 7 is 1, the Precision of class 1 is 0.714285714, the recall is 0.668952618 for class 3, the F-measures is 0.548652482 for class 2 and the MCC is 0.23758626 for class 2 is high as compare to other classes. If we consider the accuracy as a measure then this model works good for class 4, 5, 6 and class 7, because both of the classes' accuracy have very less difference, but their precision is "#DIV/0!" and recall ration is zero. Therefore, we can say that this model is under-fitting in case of class C4, C5, C6 and C7 and over-fitting in case of class C1, C2 and C3. If we consider class C1, C2 and C3, then it can be clearly seen that the class C1 accuracy is more than the other 2 classes, but their recall value is relatively lesser than the class C2 and C3, but its specificity and precision value is relatively higher than class C2 and C3. If we consider the higher F-measures and MCC, then the performance of this model is good for class C2. Finally, we can conclude that the CRT based model works efficiently for class C1. The class wise comparative analysis of performance of QUEST based model is shown in Table 10.

Table 10: A Comparative analysis of performance of QUEST for all class

QUEST							
	Accuracy	**Sensitivity**	**Specificity**	**Precision**	**Recall**	**F-measure**	**MCC**
C1	87.11172925	0	**1**	#DIV/0!	0	#DIV/0!	#DIV/0!
C2	40.42651831	**0.993574766**	0.016525749	0.399295775	**0.993574766**	**0.569658406**	**0.044447506**
C3	62.67964766	0.014962594	0.988929889	**0.444444444**	0.014962594	0.028950543	0.016920448
C4	96.54612888	0	**1**	#DIV/0!	0	#DIV/0!	#DIV/0!
C5	97.28789986	0	**1**	#DIV/0!	0	#DIV/0!	#DIV/0!
C6	97.10245712	0	**1**	#DIV/0!	0	#DIV/0!	#DIV/0!
C7	**98.8178025**	0	**1**	#DIV/0!	0	#DIV/0!	#DIV/0!

Here it can be clearly seen that the performance of the QUEST model on this data set is degraded as compare to the CRT model. Model is under-fitted in case of class C1, C4, C5, C6, and C7 and over-fitted in case of class C2 and C3, because the data are imbalanced. This model works efficiently for class C7, when we consider high accuracy, if we consider high sensitivity, recall, F-measure or MCC, it works well for class C2 and in case of high specificity, it works efficiently for C1, C4, C5, C6 and C7, if we consider high precision value, then model works efficiently for class C3. Finally, we can conclude that the QUEST model works efficiently on class C2 data and produce an average performance in every measure. The class wise comparative analysis of performance of CHAID based model is shown in Table 11.

Table 11: A Comparative analysis of performance of CHAID for all class

CHAID							
	Accuracy	**Sensitivity**	**Specificity**	**Precision**	**Recall**	**F-measure**	**MCC**
C1	86.20769587	0.147482014	0.967802022	0.403940887	0.147482014	0.216073781	0.182416115
C2	57.92767733	**0.72021028**	0.486548809	0.479953289	**0.72021028**	**0.576033637**	0.206105874
C3	63.46777932	0.489401496	0.720664207	**0.509079118**	0.489401496	0.499046408	**0.211837664**
C4	96.54612888	0	**1**	#DIV/0!	0	#DIV/0!	#DIV/0!
C5	97.28789986	0	**1**	#DIV/0!	0	#DIV/0!	#DIV/0!
C6	97.10245712	0	**1**	#DIV/0!	0	#DIV/0!	#DIV/0!
C7	**98.8178025**	0	**1**	#DIV/0!	0	#DIV/0!	#DIV/0!

In Table 11, it can be clearly seen that the performance of CHAID is much better than the CRT and QUEST. Here, the model is under-fitted in case of Class C4, C5, C6 and C7 and over-fitted in case of other classes. The CRT and QUEST model are under-fitted in case of class C4, c5, C6, and C7 and also here it is under-fitted. If we consider high accuracy, then this model works efficiently for, class C7, when consider sensitivity, Recall and F-measures then model fitted for class C2, in case of high specificity model works efficiently for class C4, C5, C6, and C7 when we consider precision,

then model works efficiently for class C3. Finally, we can conclude that this model works efficiently for class C3 and produce an average performance in every measure. The class wise comparative analysis of performance of C5.0 based model is shown in Table 12.

Table 12: A Comparative analysis of performance of C5.0 for all class

C5.0							
	Accuracy	**Sensitivity**	**Specificity**	**Precision**	**Recall**	**F-measure**	**MCC**
C1	96.56930923	0.911870504	0.9736562	0.836633663	0.911870504	0.872633391	0.853908371
C2	94.55261938	0.936915888	0.951191391	0.926632005	0.936915888	**0.931745571**	**0.886462195**
C3	94.73806212	**0.947007481**	0.947601476	0.914509332	**0.947007481**	0.930474732	0.888518814
C4	98.3310153	0.630872483	0.995918367	0.846846847	0.630872483	0.723076923	0.722906452
C5	98.91052388	0.666666667	0.998093877	0.906976744	0.666666667	0.768472906	0.772522139
C6	98.74826147	0.624	0.998328957	0.917647059	0.624	0.742857143	0.75110714
C7	**99.55957348**	0.647058824	**0.999765423**	**0.970588235**	0.647058824	0.776470588	0.79061983

Table 12 shows that the performance of C5.0 model is much better than the CRT, QUEST and CHAID. This model classifies data appropriately and produce balanced result and the accuracy in all classes in above of 94%. Similarly, if we consider the accuracy as a measure, then this model works efficiently for class C7, in case of sensitivity, it works efficiently for class C3, in case of specificity and precision, the model works efficiently for class C7. If we consider F-measures or MCC, it works efficiently on class C2 data. If we consider Recall as a measure, then it works efficiently for class C3. If we consider all measures, then it works efficiently for class C2 and C3, because all measures value is above of 88%. The class wise comparative analysis of performance of ANN based model is shown in Table 13.

Table 13: A Comparative analysis of performance of ANN for all class

ANN							
	Accuracy	**Sensitivity**	**Specificity**	**Precision**	**Recall**	**F-measure**	**MCC**
C1	87.32035234	0.224820144	0.969132517	0.518672199	0.224820144	0.313676286	0.282973522
C2	67.26935559	0.606892523	0.715987702	**0.584364454**	0.606892523	0.595415473	**0.320927158**
C3	64.00092721	**0.731296758**	0.58597786	0.511111111	**0.731296758**	**0.601692742**	0.307300065
C4	96.54612888	0	**1**	#DIV/0!	0	#DIV/0!	#DIV/0!
C5	97.28789986	0	**1**	#DIV/0!	0	#DIV/0!	#DIV/0!
C6	97.10245712	0	**1**	#DIV/0!	0	#DIV/0!	#DIV/0!
C7	**98.8178025**	0	**1**	#DIV/0!	0	#DIV/0!	#DIV/0!

This model performs as similar to CRT and under-fitted for class C4, C5, C6 and C7 and over-fitted for rest of the classes. As per accuracy, one can say that the model works efficiently for class C7 and produce similar specificity value for class C4, C5, C6 and C7. If we consider high sensitivity, recall or F-measure value, it works efficiently for class C3. When we consider Precession or MCC as an evaluation metrics then the model works efficiently for class C2. Overall, we can say that the ANN network model works efficiently for class C2. Table 14 shows the class wise comparative analysis of performance of SVM based model.

Table 14: A Comparative analysis of performance of SVM for all class

SVM							
	Accuracy	**Sensitivity**	**Specificity**	**Precision**	**Recall**	**F-measure**	**MCC**
C1	98.51645804	0.940647482	0.991750931	0.944043321	0.940647482	0.942342342	0.933831468
C2	95.82753825	0.954439252	0.960799385	0.94124424	0.954439252	0.947795824	0.913110959
C3	96.12888271	**0.955735661**	0.964575646	0.94106814	**0.955735661**	**0.94834519**	0.917463716
C4	99.21186834	0.818791946	0.998319328	0.945736434	0.818791946	0.877697842	0.876070897
C5	99.72183588	0.905982906	0.999761735	0.990654206	0.905982906	0.946428571	**0.945997619**
C6	99.5132128	0.88	0.998567677	0.948275862	0.88	0.912863071	0.911035282
C7	**99.81455726**	0.843137255	**1**	0.999987	0.843137255	0.914893617	0.917364691

It performs almost similar to C5.0 and can say that the successor model of CRT, QUEST, CHAID and Neural. As similar to previous model, it works efficiently for class C7 when we consider accuracy as an evaluation metrics. If we consider either sensitivity, Recall or F-measures, it works efficiently for class C3. If we consider either precision or specificity, it works efficiently for class C7; it works efficiently for class C5, when we consider MCC as measure.

When we find the overall performance of this model in every measure, it works efficiently for class C1. From the above discussion it can be clearly seen that the accuracy of all model is high for class C7 as compare to other classes and rest of the measures value is different for different class, however the precision and recall value for class C7 is relatively very low or negligible in every model except C5.0 and SVM based model. From the above analysis we can conclude that the almost all of the above classification technique work efficiently for class C2 and C3 except CRT. The CRT classification technique work efficient for class C1. The experimental results highlight that the Class C4, C5, C6 and C7 data are comparatively more imbalance than the class C1, C2 and C3. Finally, we can conclude that the C5.0 and SVM classification technique can be used for data classification and prediction of appropriate class, because the protein data is non-linear high dimensional sequence data and also have missing features value. As we know that the SVM can be used for non-liner high dimensional data, the C5.0 can be used for all possible data including missing value. As a result, we can conclude that the C5.0 and SVM classification technique is more suited in case of protein data classification and prediction. Above discussion is based on the individual class of data. Finally, we have computed the overall performance of the model in terms of accuracy and elapsed time. The overall performance of all models is shown in table 15.

Table 15: Overall performance table of all models on sequence-based protein data.

S.No.	Model	Data		Accuracy		Elapsed time for model build
		Correct	Wrong	Correct	Wrong	
1	CRT	2,080	2,234	48.22%	51.78%	0 hours, 2 mins, 23 secs
2	QUEST	1,725	2,589	39.99%	60.01%	0 hours, 0 mins, 9 secs
3	CHAID	2,100	2,214	48.68%	51.32%	0 hours, 1 mins, 4 secs
4	C5.0	3,913	401	90.70%	9.30%	0 hours, 0 mins, 10 secs
5	NEURAL	2,337	1,977	54.17%	45.83%	0 hours, 0 mins, 27 secs
6	SVM	4,071	243	94.37%	5.63%	0 hours, 9 mins, 38 secs

Table 15 gives the brief summary of the accuracy, total number of correct and incorrect data and elapsed time taken by the machine (SPSS tool) for classification and prediction of data. Form the above table; it can clearly see that the accuracy of the SVM and C5.0 is above of 90%;and also we can conclude that the accuracy of the above model is increased at commutative level, however the performance of the other classification technique is degraded. Finally, we can conclude that the sequence-based protein data with SVM and C5.0 classification technique can be used for protein classification and predictions. If we consider accuracy with elapse time, then C5.0 can be used for protein classification and predictions.

V. CONCLUSIONS

In the field of computational biology to gauge the meaningful and accurate feature for protein function predications, either the profile-based protein data or sequence-based data has been used. This article used total 4368 number of protein sequence-based data of human enzyme class for computation and predictions. To classify the protein data, the CRT, QUEST, CHAID, C5.0, NEURAL, and SVM classification technique has been used. The experimental analysis highlights that the C5.0 and SVM classification technique is more suited for protein classification and predictions. The features have been selected by the C5.0 and SVM classification technique can be used for protein classification and predictions. Further, it can also be seen that the class C4, C5, C6 and C7 data are more imbalanced than class C1, C2 and C3 in both of the data set. Due to imbalanced dataset, the precision and recall value of these classes are relatively very low than the other classes. These also affect the overall performance of the classification techniques.

ACKNOWLEDGMENT

The authors wish to express their gratitude to anonymous reviewers for their valuable comment.

REFERENCE

1. Karunapala, E. D. S. C. (2015). Protein Function Prediction Using Machine Learning (Doctoral dissertation).
2. Das S, Sillitoe I, Lee D, Lees JG, Dawson NL, Ward J, et al. CATH FunFHMMer web server: protein functional annotations using functional family assignments. Nucleic Acids Res. 2015; 43: W148–153.
3. Jackson SP, Bartek J. The DNA-damage response in human biology and disease. Nature. 2009; 461: 1071–1078.
4. Weinberg SE, Chandel NS. Targeting mitochondria metabolism for cancer therapy. Nat Chem Biol.2015; 11: 9–15
5. Yang H, Qin C, Li YH, Tao L, Zhou J, Yu CY, et al. Therapeutic target database update 2016: enriched resource for bench to clinical drug target and targeted pathway information. Nucleic Acids Res. 2016; 44: D1069–1074.
6. Piovesan D, Giollo M, Leonardi E, Ferrari C, Tosatto SC. INGA: protein function prediction combining interaction networks, domain assignments and sequence similarity. Nucleic Acids Res. 2015; 43: W134–140.
7. Rentzsch R, Orengo CA. Protein function prediction using domain families. BMC Bioinformatics. 2013; 14 Suppl 3: S5.
8. Kotlyar M, Pastrello C, Pivetta F, Lo Sardo A, Cumbaa C, Li H, et al. In silico prediction of physical protein interactions and characterization of interactome orphans. Nat Methods. 2015; 12: 79–84.
9. Li, Ying Hong, Jing Yu Xu, Lin Tao, Xiao Feng Li, Shuang Li, Xian Zeng, Shang Ying Chen et al. "SVM-prot 2016: A web-server for machine learning prediction of protein functional families from sequence irrespective of similarity." PloS one 11, no. 8 (2016): e0155290.
10. Singh, Upendra, and SudhakarTripathi. "Protein Classification Using Hybrid Feature Selection Technique." In International Conference on Smart Trends for Information Technology and Computer Communications, pp. 813-821. Springer, Singapore, 2016.

11. Lee, B.J., Lee, H.G., Ryu, K.H.: Design of a novel protein feature, enzyme function classification. In: CIT Workshops 2008. IEEE 8th International Conference on Computer and Information Technology Workshops, pp. 450–455. IEEE (2008)
12. Yadav, A., Jayaraman, V.K.: Structure based function prediction of proteins using fragment library frequency vectors. Bioinformation 8(19), 953–956 (2012)
13. Garg, A., Raghava, G.P.: A machine learning based method for the prediction of secretory proteins using amino acid composition, their order and similarity-search. Silico Biol. 8(2), 129–140 (2008)
14. Mer, A.S., Andrade-Navarro, M.A.: A novel approach for protein subcellular location prediction using amino acid exposure. BMC Bioinform. 14(1), 1 (2013)
15. Jensen, L.J., Skovgaard, M., Brunak, S.: Prediction of novel archaeal enzymes from sequence-derived features. Protein Sci. 11(12), 2894–2898 (2002)
16. Dobson, Paul D., and Andrew J. Doig. "Predicting enzyme class from protein structure without alignments." Journal of molecular biology 345, no. 1 (2005): 187-199.
17. Borro, Luiz C., Stanley RM Oliveira, Michel EB Yamagishi, Adaulto L. Mancini, José G. Jardine, Ivan Mazoni, EH Dos Santos, Roberto H. Higa, Paula R. Kuser, and GoranNeshich. "Predicting enzyme class from protein structure using Bayesian classification." Genet. Mol. Res 5, no. 1 (2006): 193-202.
18. Yadav, S.K., Bhola, A., Tiwari, A.K.: Classification of enzyme functional classes, subclasses using support vector machine. In: 2015 International Conference on Futuristic Trends on Computational Analysis and Knowledge Management (ABLAZE), pp. 411–417. IEEE (2015)
19. Lin, W.-Z., Fang, J.-A., Xiao, X., Chou, K.-C.: iDNA-Prot: identification of dna binding proteins using random forest with grey model. PLoS One 6(9), e24756 (2011)
20. Wu, J., Liu, H., Duan, X., Ding, Y., Wu, H., Bai, Y., Sun, X.: Prediction of DNAbinding residues in proteins from amino acid sequences using a random forest modelwith a hybrid feature. Bioinformatics 25(1), 30–35 (2009)
21. Amidi, Afshine, ShervineAmidi, DimitriosVlachakis, Nikos Paragios, and Evangelia I. Zacharaki. "A machine learning methodology for enzyme functional classification combining structural and protein sequence descriptors." In International Conference on Bioinformatics and Biomedical Engineering, pp. 728-738. Springer, Cham, 2016.
22. Poux S, Arighi, CN, Magrane M, Bateman A, Wei C-H, Lu Z, Boutet E, Bye-A-Jee H, Famiglietti ML, Roechert B. On expert curation and sustainability: UniProtKB/Swiss-Prot as a case study bioRxiv (2017)
23. Karp, Peter D. "What we do not know about sequence analysis and sequence databases." Bioinformatics (Oxford, England) 14, no. 9 (1998): 753-754.
24. Caruana R, de Sa VR: Benefitting from the variables that variableselection discards. J Mach Learn Res 2003, 3:1245-1264.
25. Hall MA, Holmes B: Benchmarking attribute selection techniquesfor discrete class data mining. IEEE Trans Knowl Data Eng2003, 15:1-16.
26. Schwartz CE, Sprangers MA, Oort FJ, Ahmed S, Bode R, Li Y, Vollmer T. Response shift in patients with multiple sclerosis: an application of three statistical techniques. Quality of life research : an international journal of quality of life aspects of treatment, care and rehabilitation. 2011;20(10):1561–1572.
27. Zwartjes, Ardjan, Paul JM Havinga, Gerard JM Smit, and Johann L. Hurink. "QUEST: Eliminating online supervised learning for efficient classification algorithms." Sensors 16, no. 10 (2016): 1629.
28. Milanović, Marina, and Milan Stamenković. "CHAID Decision Tree: Methodological Frame and Application." Economic Themes 54, no. 4 (2016): 563-586.
29. Pang, Su-lin, and Ji-zhang Gong. "C5. 0 classification algorithm and application on individual credit evaluation of banks." Systems Engineering-Theory & Practice29, no. 12 (2009): 94-104.
30. Bose, Nirmal K., and P. Liang. "Neural Network Fundamentals with Graphs, Algorithms, and Applications (McGraw-Hill Series in Electrical Computer Engineering)." (1996).
31. Vapnik VN. The Nature of Statistical Learning Theory, 1995.
32. Nizar A, Dong Z, Wang Y. Power utility nontechnical lossanalysis with extreme learning machine method. PowerSystems, IEEE Transactions on, 2008; 23: 946-955. https://doi.org/10.1109/TPWRS.2008.926431
33. Xiao H, Peng F, Wang L, Li H. Ad hoc-based featureselection and support vector machine classifier for intrusion detection, in 2007 IEEE International Conference on GreySystems and Intelligent Services, 2007; pp. 1117-1121.https://doi.org/10.1109/GSIS.2007.4443446
34. Samb, M.L., Camara, F., Ndiaye, S., Slimani, Y., Esseghir, M.A.: A novel RFESVM-based feature selection approach for classification. Int. J. Adv. Sci. Technol.43, 27–36 (2012)
35. Tiwari, A.K., Srivastava, R.: A survey of computational intelligence techniques in protein function prediction. Int. J. Proteomics (2014)
36. Cheeseman, Peter C., Matthew Self, James Kelly, Will Taylor, Don Freeman, and John C. Stutz. "Bayesian Classification." In AAAI, vol. 88, pp. 607-611. 1988.
37. Han, Jiawei, Jian Pei, and MichelineKamber. Data mining: concepts and techniques. Elsevier, 2011.
38. Baldi P, Brunak S, Chauvin Y, Andersen CAF, Nielsen H: Assessingthe accuracy of prediction algorithms for classification: anoverview. Bioinformatics 2000, 16:412-424.
39. Matthews BW: Comparison of the predicted and observed secondarystructure of T4 phage lysozyme. BiochimBiophys Acta1975, 405:442-451.
40. Gupta, C.L.P., Bihari, A. and Tripathi, S., 2019. Protein Classification using Machine Learning and Statistical Techniques: A Comparative Analysis. arXiv preprint arXiv:1901.06152.

AUTHORS PROFILE

ChhoteLal Prasad Gupta has done B. Tech. in Computer Science and engineering in 2005, M. Tech. in computer science and engineering in 2010. Currently pursuing Ph.D. in computer science and engineering from AKTU, Lucknow, India has 14 years of teaching experience and has hold different position in academic field.

Dr. AnandBihari currently working as an Assistant Professor, Senior Grade at School of Information Technology & Engineering, VIT University, Vellore, Tamil Nadu, India. He has completed his Ph. D. from NIT Patna. He has published many research article in reputed international journal and conferences. His research area is Scientometrics, Social Network Analysis, Machine Learning, Protein Classification and Brain Computation.

Dr. SudhakarTripathi, Associate Professor, Information Technology Rajkiya Engineering College Ambedkar Nagar, U. P. India. He has completed his Ph. D. from IIT(BHU) Varanasi. He has published many research article in Reputed International Journal. His research area is Artificial Intelligence,Artificial Neural Networks,Computational Biology,Brain Computational Modeling and Research,Machine Learning,Datamining