Human Protein Features Extraction and Function Prediction using improved Machine Learning and critical analysis with WEKA

S. Sharma, A. Singh, P. Singh, G. Singh, R. Singh

Abstract—Molecular class Prediction of a protein is highly relevant for conducting research in domains of disease detection and drug discovery process. Numerous approaches incorporated to increase the accuracy of Human protein Function (HPF) prediction task but it is very challenging due to very nature of this domain which is wide and versatile. This research is focused on sequence derived attributes/features (SDF) approach for HPF prediction and critically analyzed with WEKA data analysis tool. More SDFs were identified and included, training data-set from Human protein reference database, enhanced as in number of sequences and the related features for deriving the relation with various protein classes. A range of Machine Learning approaches were analyzed for prediction effectiveness and a comprehensive comparison is carried out to achieve higher classification accuracy. Machine Learning approach is also analyzed for its limitation on application of broad spectrum data domain.

Index Terms— Bagging, BayesNet,, C5, Decision Tree, HPF, IBK , J48, Logistic Approach, PART, Random Forest, SDF Weka

1 Introduction

PROTEIN classification is a vast domain with enormous amount of data available for research and analysis yet the knowledge about its correct perception is very limited. On the other hand Machine learning (ML) provide promising answers to not-so-clearly defined areas of research.

Decision tree based prediction approach of machine learning is very clear and reliable for protein classification. Being a white-box approach it clearly illustrates the sequence of computations involved at each and every stage. This plus point enables its usage by computational experts even without much knowledge of the concerned domain. Likewise, it enables an expert from the concerned domain to critically examine the steps followed by a computational expert. So it bridges the gap between technical know-how and domain expertise. Decision tree comprises of nodes and edges depicting various functionalities at different levels of computations. A decision tree clearly illustrates the required results or outputs amongst various outcome possibilities. It clearly defines the problem structure and its interpretations in a hierarchical way which is much easier to comprehend. As the model has a unique ability of taking into account various input parameters and reaching a goal. [13] [20]

Recent studies indicate that protein function prediction is one of the area where ML faces serious challenges. [22]

2 Introduction to Weka and Other Mining Tools

WEKA [18] comes with enhanced capability of dealing with huge databases which other popular data analytics tool lack. WEKA is a workhorse containing blend of tools and calculus for information processing and its detailed representation with a user friendly GUI which provide ease of use for a range of computational capabilities. It is highly suitable for a wide range of Research and development activities.
Factors favoring use of WEKA are:

- Free access as it is an open source product.
- Highly versatile and equally portable
- Extensive coverage of information acquisition and handling and visual modeling capabilities.
- User friendly GUI. [19]

Alternate data mining and data analysis tools are See5 SIPINA etc. They have data mining as well as machine learning capabilities. SIPINA uses decision trees approach for classification, under supervised learning. SIPINA contains classification algorithms like ID3, ASSISTANT

- *S. Sharma, Assistant Professor, PG Department of computer Science, Hindu College, Amritsar. E-mail: sunny.dcse@gndu.ac.in.*
- *A. Singh, Assistant Professor, Department of Computer Science, Guru Nanak Dev University, Amritsar. E-mail: amritpal.dcse@gndu.ac.in*
- *P. Singh, Research Scholar, Department of Computer Science, Guru Nanak Dev University, Amritsar. E-mail: amritpal.dcse@gndu.ac.in*
- *G. Singh, Dean Faculty, Professor Department of Computer Science, Guru Nanak Dev University, Amritsar. E-mail: gsbawa@yahoo.com*
- *R. Singh, Professor& Head, Department of Computer Science, Guru Nanak Dev University, Amritsar. E-mail:tovirk@yahoo.com*

86,GID3, C4.5, One Vs All Decision Tree, Improved CHAID etc. it supports neural network, decision list, discriminant analysis, rule induction, etc. [20].

3 LITERATURE SURVEY

Many of the protein classification methods based on Sequences, Genomics context, Phylogenomics, Structure, Protein-Protein Interaction, Gene expression, Data Integration etc, since they use the features extraction methodology for protein function prediction or protein classification; Survey shows that feature extraction form protein sequence using vector space integration methods of data Integration for protein classification plays vital role in bioinformatics.

Jensen, L. et al. (2002) focused research on developing sequence based method which recognizes and combines important features for the purpose of assigning protein function to respective classes and enzyme classification and show the benefits for protein function prediction through linear sequence of amino acids over protein structure. They identified important attributes features related with post-translational-modifications; also include simple features like length, composition of the polypeptide chain as well as isoelectric point [1].

Cai, C.Z et al. (2003) use the five physicochemical features from linear amino acid sequence like normalized van dar Waals volume, surface tension, polarity, charge using the SVM-Prot method for function classification and showcase its importance by achieving accuracy of 71.4% on data set of 49 plants proteins [14].

Friedberg, I. (2006) expressed that as diversification and enhancement in the volume of pure sequence and structure related data is growing, which leads unequal enlargement in the number of un-characterized gene products. Recently well-known methods for gene as well as for protein annotation, like homology based transformation, they are annotating fewer data and in some cases they are amplifying existing erroneous-annotation. The author said Contextual and Subjective definition of protein function which is cumbersome in nature and expresses his views for quality of function predictions [6].

Lobley et al. (2007) predict the protein function with IDRs (intrinsically disordered regions) in human protein sequence on the basis of length and position dependencies. Sequence based features were used like length, molecular weight, charge, hydrophobicitys, transmembrane residue, pest region peptide and disordered related features for protein function prediction through machine learning approach [21].

Singh, M. et al. (2007)describes how the protein related data is increasing day by day and suggested to solve the problem related to human protein function prediction and said express the need of machine learning algorithms for drug discovery. The author use Decision tree induction approach through C4.5 algorithm for the selection of best attribute for protein function prediction and presents the accuracy of 72% for human protein function prediction in contrast to accuracy of 44% of existing prediction methodology [2].

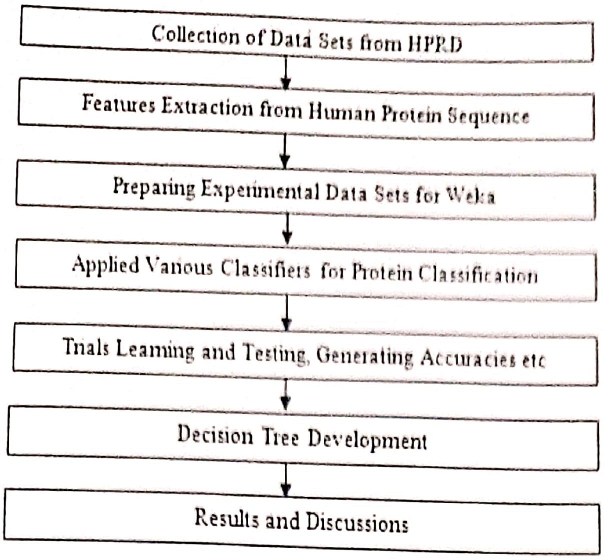
Singh, M. et al. (2011) describe unsupervised learning and cluster analysis approach prediction of human protein classes. The database extracted from human protein reference database (HPRD) was used and 5 amino acid sequences were taken for each molecular class then sequence derived features were grabbed for each protein sequence with the help of web based tools. Clustering technique predicts the class of the query sequence [9].

Wass, M.N et al. (2012) used sequence based features protein-protein-interaction features as well as gene expression based features and incorporate CombFunc method for protein function prediction. CombFunc was also evaluated for the predictions of gene ontology molecular function on the data set of 6686 proteins. The Uni-Prot GOA annotation's taken out for the proteins, only 5000 among were used for training, rest used for testing purpose. The CombFunc method obtains recall value of 0.64 and precision of 0.71[11].

Ofer et al. (2015) used the feature extracted technique and offer the community with hundreds of features of biological interpretability and predict localization-structure classes and its distinctive functional properties [15]. **Qingtian Gong et al. (2016)** describe sequence-based-function prediction method called GOFDR [16]. GOFDR takes input as query protein sequence, and predicts gene ontology terms or features for the query sequence from similar sequences to the query that are retrieved from a sequence-database, key point towards GOFDR is that it takes residues that are specific for a particular gene ontology term in contrast to gene ontology term to the query [16]. **Sayoni Das et al. (2016)** retrieve dalike sequences to a query from databases using two methods, the first one is BLAST and other is HMMER3 a hidden Markov Model-based-tool, and then extracted gene ontology terms from the retrieved sequences. The key point of their research is that it considers taxonomy information of sequences [12]. **Enrico Lavezzo et al. (2016)** reviews sequence and structure-based function prediction methods with a spotlight on their protein classification database (CATH-Gene3D, andFunFHMMer server). Where CATH-Gene3D does classification of query sequences to the CATH protein-structural-domain-classification-database and the FunFHMMer server matches a query sequence with sequence present in CATH-Gene3D with incorporation of hidden Markov model then predict function of the query sequence [17].

4 RESEARCH METHODOLOGY

The Methodology used for the research process from data set collection to results and analysis with trial learning and testing can be expressed as follow.


Collection of Data Sets from HPRD
Features Extraction from Human Protein Sequence
Preparing Experimental Data Sets for Weka
Applied Various Classifiers for Protein Classification
Trials Learning and Testing, Generating Accuracies etc
Decision Tree Development
Results and Discussions

5 Results and Discussions

The dataset used for analysis is extracted from HPRD[7] and the 25 features are extracted from 70 sequences which is considered as input for analysis through Weka as shown in Fig 1.

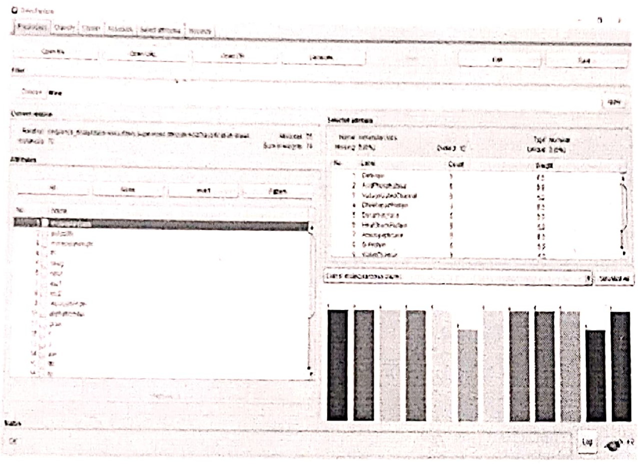


Fig. 1 Sequence derived Attributes or Features detail with Classes

The contribution of each sequence derived feature with its maximum and minimum range in protein class prediction is shown in Fig 2.

The various classification techniques based of Decision Tree, Rule Mining, Lazy, Bayes Network, Meta and Functions are implemented on the data set. In depth the various classification algorithms like Random Forest, J48, PART, BayesNet, Logistic Approach, IBK and Bagging are applied on the data set, among them Random Forest outperform all of them with achievement of overall accuracy 57.14% shown in Fig 3., and classification accuracy of 97.14% for protein classes as shown in Fig 4.

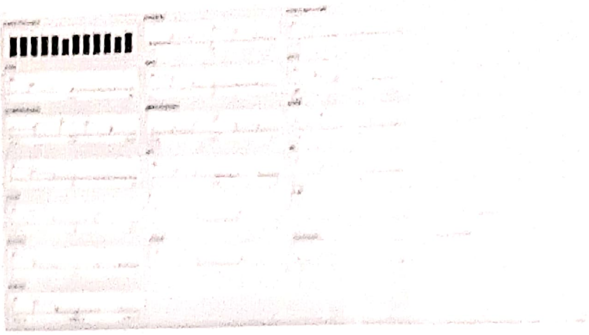


Fig. 2 Contribution sequence derived features with max and minimum ranges

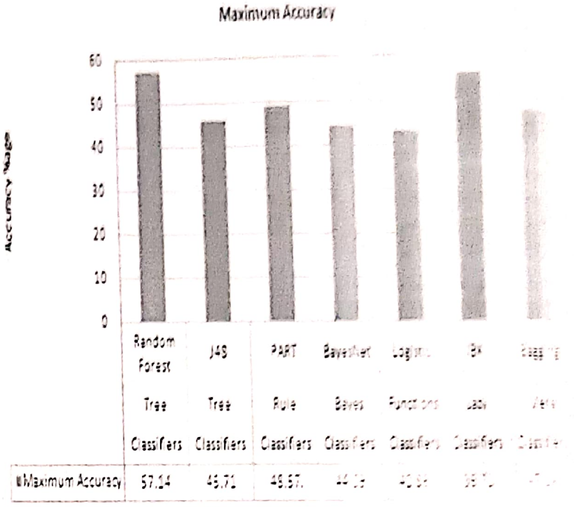


Fig. 3.Accuracy comparisons of classifiers

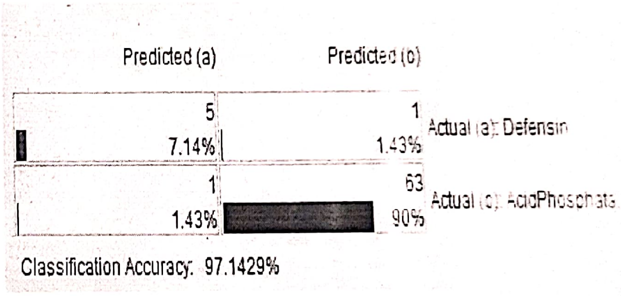


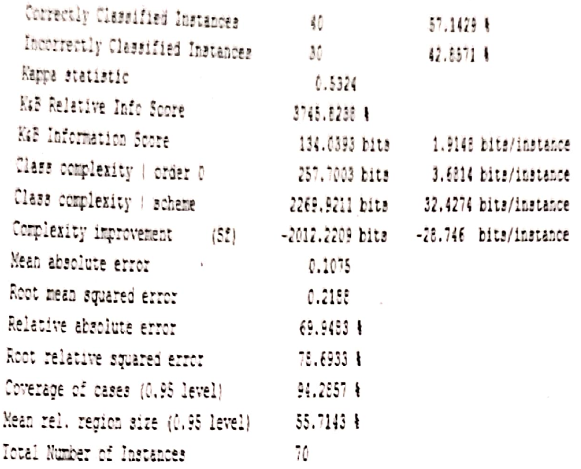
Fig. 4 Protein class classification accuracy

The 10fold cross validation is done for protein class prediction or we can say classification of instances on random forest method the detailed summary is shown in Fig 5, which shows the mean absolute error is 0.10 % and the Root mean squared error of 0.21% and describe the detail of correctly and incorrectly classified instances. It also showcase complexity improvement of -28.746 bits/instance.

The detailed accuracy achievement by protein classes with weighted average of true positive rate, false positive rate, precision, recall value, F-Measure, MCC and Area under ROC and PRC is shown in Fig 6, which expressed the true classification of protein classes.

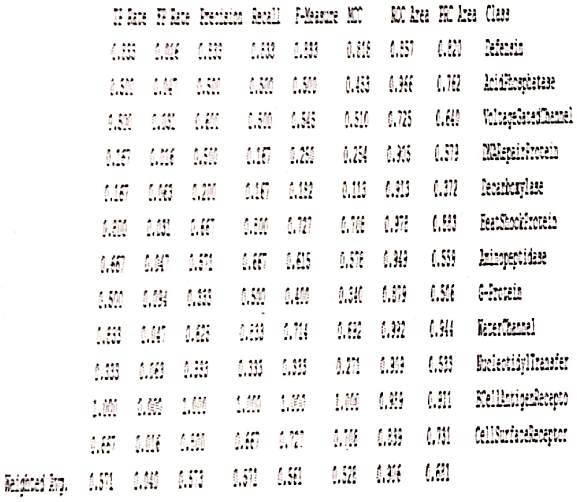
The cost matrix for minimizing cost describe a gain of 8.97 with random prediction of 10.9 at a minimum cost of 2.0 with good true positive rate for protein class prediction on 70 sequences as shown in Fig 7, it describe the classification accuracy is achieved for defensin class at random cost gain is 97.14% is depicted in Fig 8.

The Threshold curve for protein Defensin class for sample size and true positive rate is depicted in Fig 9; the X Axis is taken as for sample size and Y axis as fortrue positive rate. The defensin class having true positive rate of 0.833 along with the precision and recall value of 0.833 which is quite good.



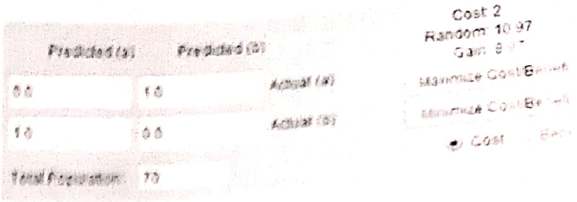
```
Correctly Classified Instances          40               57.1429 %
Incorrectly Classified Instances        30               42.8571 %
Kappa statistic                          0.5324
K&B Relative Info Score               3745.8238 %
K&B Information Score                  134.0393 bits      1.9148 bits/instance
Class complexity | order 0             257.7003 bits      3.6814 bits/instance
Class complexity | scheme             2269.9211 bits     32.4274 bits/instance
Complexity improvement     (Sf)      -2012.2209 bits    -28.746  bits/instance
Mean absolute error                      0.1075
Root mean squared error                  0.2158
Relative absolute error                 69.9483 %
Root relative squared error             78.6933 %
Coverage of cases (0.95 level)          94.2857 %
Mean rel. region size (0.95 level)      55.7143 %
Total Number of Instances               70
```

Fig. 5 Cross validation summary for classification of instances with error



	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.833	0.016	0.833	0.833	0.833	0.818	0.957	0.820	Defensin
	0.500	0.047	0.500	0.500	0.500	0.453	0.966	0.762	AcidPhosphatase
	0.500	0.031	0.600	0.500	0.545	0.510	0.725	0.640	VoltageGatedChannel
	0.167	0.016	0.500	0.167	0.250	0.254	0.905	0.579	DNARepairProtein
	0.167	0.063	0.200	0.167	0.182	0.113	0.813	0.372	Decarboxylase
	0.800	0.031	0.667	0.800	0.727	0.706	0.978	0.883	HeatShockProtein
	0.667	0.047	0.571	0.667	0.615	0.576	0.949	0.559	Aminopeptidase
	0.500	0.094	0.333	0.500	0.400	0.340	0.879	0.506	G-Protein
	0.833	0.047	0.625	0.833	0.714	0.692	0.992	0.944	WaterChannel
	0.333	0.063	0.333	0.333	0.333	0.271	0.919	0.533	NucleotidylTransfer
	1.000	0.000	1.000	1.000	1.000	1.000	0.929	0.911	BCellAntigenRecepto
	0.667	0.016	0.800	0.667	0.727	0.706	0.939	0.751	CellSurfaceReceptor
Weighted Avg.	0.571	0.040	0.573	0.571	0.561	0.528	0.906	0.681	

Fig. 6 Accuracy achievement summary for protein classes



	Predicted (a)	Predicted (b)	
	0.0	1.0	Actual (a)
	1.0	0.0	Actual (b)
Total Population	70		

Cost 2
Random 10.97
Gain 8.97

Fig. 7 Cost/benefit Analysis

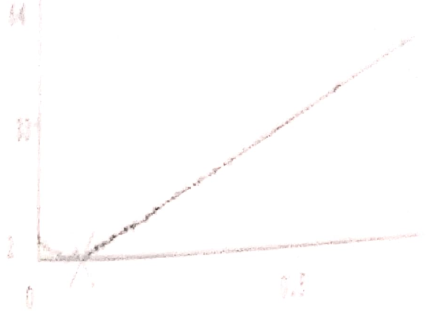


Fig. 8 Cost/Benefit Curve for Protein Class Defensin

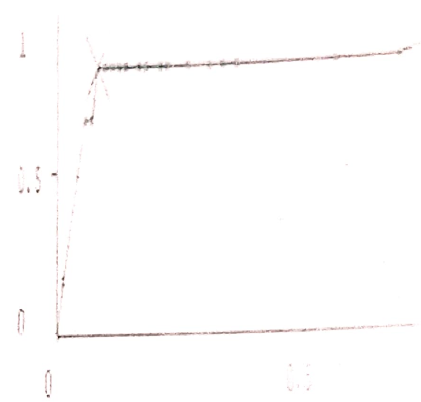


Fig. 9 Threshold curve for protein Defensin class

The 25 featured attributes taken for the experimental verification were solubility, molecular weight, PI, nneg, npos, excl, exc2, instability index, aliphatic index, gravy, t, s, ser, thr, tyr, mean, d, prob, expaa, predhel, ProbN, Absorbance, IsoelectricPoint and Volume. The Best First and CfsSubsetEval methods were applied on attribute evaluation. The PI, excl, mean expaa, predhel outperform all of the other attribute contribution. They contributed their max for class prediction as well as for decision tree formation. Their contribution for molecular class prediction is expressed in Fig 10, 11, 12, 13. The decision tree in accordance to the attribute contribution for protein classification prediction is shown in Fig 14. The confusion matrix for random forest classifier on experimental data set- is expressed in Fig 15. This describes the true positive/ negative and false positive/negative predictions for protein molecular classes. It demonstrate correctness of the classification process.

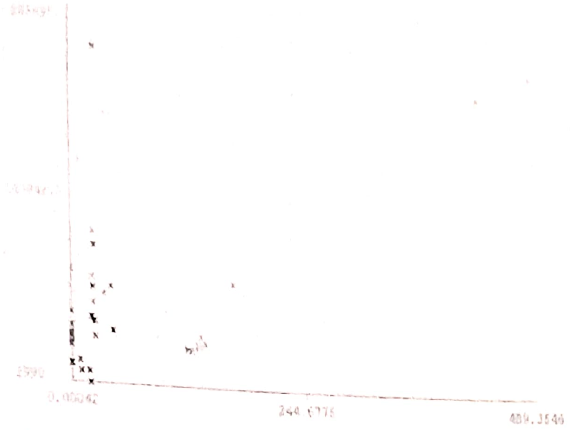


Fig. 10 Contribution Plot for Molecular classes detection using expaa along (X axis) and exc1along (Y axis)

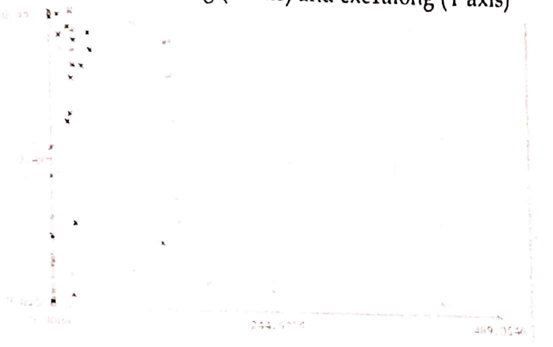


Fig. 11 contribution Plot for Molecular classes detection using expaa along (X axis) and mean along (Y axis)

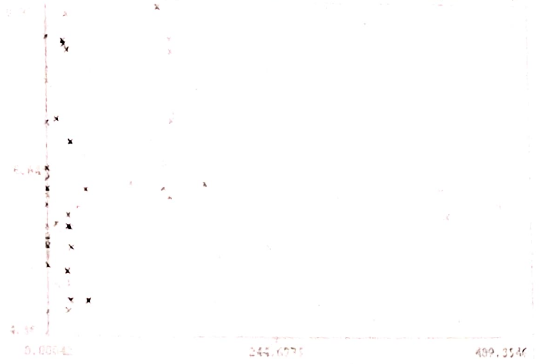


Fig. 12 contribution Plot for Molecular classes detection using expaa along(X axis) and PI along(Y axis)

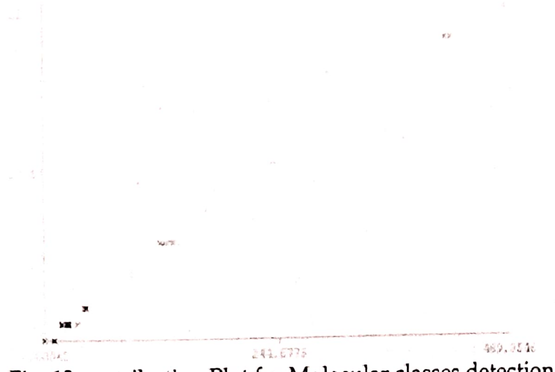


Fig. 13 contribution Plot for Molecular classes detection using expaa along (X axis) and Predhel along (Y axis)

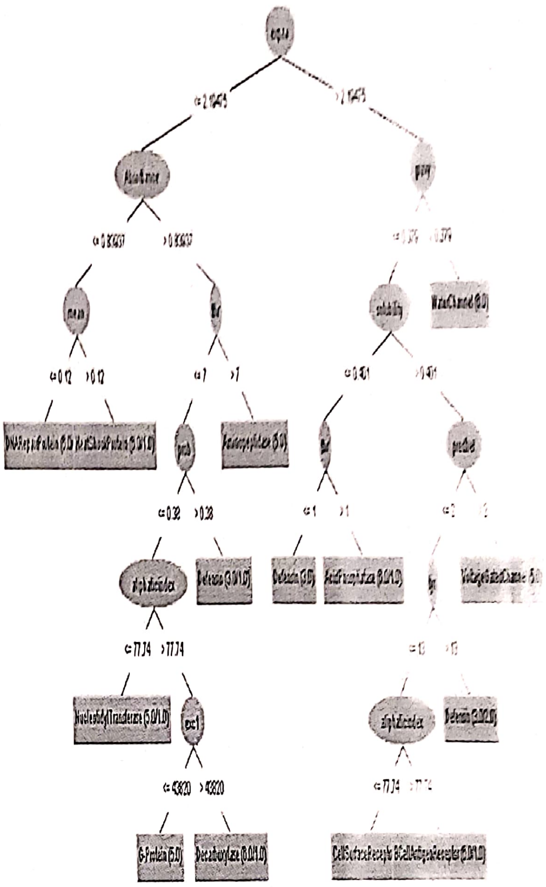
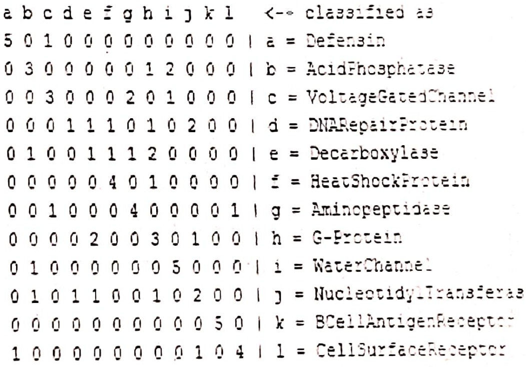


Fig. 14 Decision tree for Protein class prediction from sequence derived attributes or features



```
a b c d e f g h i j k l   <-- classified as
5 0 1 0 0 0 0 0 0 0 0 0 | a = Defensin
0 3 0 0 0 0 0 1 2 0 0 0 | b = AcidPhosphatase
0 0 3 0 0 0 2 0 1 0 0 0 | c = VoltageGatedChannel
0 0 0 1 1 1 0 1 0 2 0 0 | d = DNARepairProtein
0 1 0 0 1 1 1 2 0 0 0 0 | e = Decarboxylase
0 0 0 0 0 4 0 1 0 0 0 0 | f = HeatShockProtein
0 0 1 0 0 0 4 0 0 0 0 1 | g = Aminopeptidase
0 0 0 0 2 0 0 3 0 1 0 0 | h = G-Protein
0 1 0 0 0 0 0 0 5 0 0 0 | i = WaterChannel
0 1 0 1 1 0 0 1 0 2 0 0 | j = NucleotidylTransferase
0 0 0 0 0 0 0 0 0 0 5 0 | k = BCellAntigenReceptor
1 0 0 0 0 0 0 0 0 1 0 4 | l = CellSurfaceReceptor
```

Fig. 15 Confusion matrix for protein classes

6 Conclusion

Research gaps suggested the applicability of the machine learning approach for protein classification and also indicated its weakness in this domain due to very vast and versatile data set of the domain. So this critical analysis clearly indicate how formulation and incorporation of 5 new features enhanced the accuracy of machine learning algorithm's classification accuracy for 'defensin' class to a remarkable level of 97% with 90% true positive rate in the confusion matrix against the combined classification accuracy of 57% on the data set with random forest algorithm and also highlighted the importance of doing these steps at early stages of Machine Learning implementation, else the upcoming research results built with ML approach will be biased and the error will propagate to further investigations.

This is equally applicable in other research domains for scope of improvement in result obtained from ML by working on individual components of the classification problem rather than tackling it all at once.

Acknowledgment

The authors wish to thank Department of Computer Science, Guru Nanak Dev University, Amritsar for providing the research environment.

References

[1] L.J. Jensen, R. Gupta, N. Blom, D. Devos, J. Tamames C. Kesmir, H. Nielsen, H.H. Stærfeldt, K. Rapacki, C. Workman C.A.F. Andersen, S. Knudsen, A. Krogh, A.Valencia and S. Brunak , "Prediction of Human Protein Function from Post-Translational Modifications and Localization Features ", *Journal of Molecular Biology*, vol. 319, issue 5,pp 1257-1265, 2002.

[2] M. Singh, P. K. Wadhwa and P. S. Sandhu, " Human Protein Function Prediction using Decision Tree Induction ", *IJCSNS International Journal of Computer Science and Network Security*, vol. 7, no.4, pp. 92-98, 2007.

[3] H. Wei-Feng, G. Na, Y. Yan, L. Ji-Yang, Y. Ji-Hong, "Decision Trees Com-bined with Feature Selection for the Rational Synthesis of Aluminophos-phate AlPO4-5", *National Natural Science Foundation of China*, vol 27, no.9, pp 2111-2117, 2011.

[4] B. Bergeron, "Bioinformatics Computing", pp 257-270, 2002.

[5] J. Han and M. Kamber, "Data Mining Concepts and Techniques", *MorganKaufmann Publishers, USA* pp 279-322, 2003.

[6] I. Friedberg, "Automated Protein Function Prediction- the Genomic Chal-lenge", *Briefings in Bioinformatics*, vol 7, no.3, pp 225-242.

[7] https://www.hprd.org.

[8] D. Arditi and T. Pulket, "Predicting the outcome of construction litigation using boosted decision trees ", *Journal of Computing in Civil Engineering*, vol. 19, no. 4, pp 387-393, 2005.

[9] M. Singh, G. Singh, "Cluster Analysis Technique based on Bipartite Graph for Human Protein Class Prediction", *International Journal of Computer Applications (0975 – 8887)*, vol. 20, no.3, pp. 22-27, 2011.

[10] http://rulequest.com/see5-info.html.

[11] Wass, M.N., Barton, G., Sternberg, M.J.E.: Combfunc: predicting protein function using heterogeneous data sources. Nucleic Acids Res **40**(Web server issue), W466–W470 (2012)

[12] Sayoni Das, Christine A. Orengo, Protein function annotation using protein domain family, resources, Methods 93 (2016) 24–34.

[13] Sharma, Sunny, Amritpal Singh, and Rajinder Singh "Enhancing Usability of See5 (Incorporating C5 Algorithm) for Prediction of HPF from SDF." *International Journal of Computing and Technology* 3.4 (2016)

[14] Cai, C.Z., Han, L.Y., Ji, Z.L., Chen, X., Chen, Y.Z. SVM-Prot: Web-based support vector machine software for functional classification of a protein from its primary sequence. Nucleic Acids Res 31(13) ,2003

[15] Ofer, D., Linial, M.: ProFET: Feature engineering captures high-level protein functions. Bioinformatics **31**(21), 3429–3436. 2015.

[16] Qingtian Gong, Wei Ning, WeidongTian, GoFDR: a sequence alignment based method for predicting protein functions, Methods 93 (2016) 3–14

[17] Enrico Lavezzo, Marco Falda, Paolo Fontana, Luca Bianco, Stefano Toppo, Enhancing protein function prediction with taxonomic constraints – the Argot2.5 web server, Methods 93 (2016) 15–23.

[18] https://en.wikipedia.org/wiki/Weka_machine_learning

[19] http://www.cs.waikato.ac.nz/ml/weka/

[20] Sunny Sharma, " Harnessing the Power of Decision Tree approach in Machine Learning for Cervical Cancer Stage Prediction using See5 and SIPINA", International Journal of Scientific Research in Science, Engineering and Technology(IJSRSET), Print ISSN : 2395-1990, Online ISSN : 2394-4099, Volume 2 Issue 2, pp.1176-1182, March-April 2016.

[21] Lobley, A., Swindells, M.B., Orengo, C.A., Jones, D.T. Inferring function using patterns of native disorder in proteins. PLoSComputBiol3(8), e162 (2007)

[22] Shehu, Amarda, Daniel Barbará, and Kevin Molloy "A Survey of Computational Methods for Protein Function Prediction." *Big Data Analytics in Genomics* Springer International Publishing, 2016. 225-298.

Biography

S. Sharma is an Assistant Professor at PG Department of computer Science of Hindu College, Amritsar, Punjab, India. (sunny.dcse@gndu.ac.in) He received his MCA degree in Computer Science from Guru Nanak Dev University, Amritsar, Punjab and Cleared **UGC NET**, **GATE** & now **pursuing Ph.D.** in Computer Science from Guru Nanak Dev University, Amritsar Pb, (INDIA). He has published 25 International and 08 National research papers. His current research interest is Bio-Informatics, Machine Learning and Data mining. He works on the Prediction of Protein function & Structure, Rule Mining, Machine Learning.

A. Singh is an Assistant Professor at Department of computer Science of Guru Nanak Dev University, Amritsar, Punjab, India (amritpal.dcse@gndu.ac.in) He received his MCA degree in Computer Science from Guru Nanak Dev University, Amritsar, Punjab and now **pursuing Ph.D.** in Computer Science from Guru Nanak Dev University, Amritsar Pb. (INDIA). He has published 10 International and 02 National research papers. His current research interest is Bio-Informatics, Machine Learning and Data mining.

P. Singh is a Research Scholar at Department of computer Science of Guru Nanak Dev University, Amritsar, Punjab, India. (amritpal.dcse@gndu.ac.in) He received his M.Tech degree in Computer Science & Engineering from Guru Nanak Dev University, Amritsar, Punjab and now **pursuing Ph.D.** in Computer Science from Guru Nanak Dev University, Amritsar Pb. (INDIA). He has published 08 International and 03 National research papers. His current research interest is Bio-Informatics, Machine Learning and Data mining.