Improved Machine Learning using Adaptive Boosting algorithm in Membrane Protein Prediction

Anjna Jayant Deen, Manasi Gyanchandani

Abstract**:** ***Membrane protein are very important and play significantly in the field of biology and medicine. The main purpose is to find suitable features of a membrane protein. Various features extraction methods are use to find membrane protein and their types. PseAAC (Pseudo Amino Acid Composition) is a one of the feature extraction method which was used to find the localization of the protein, which helps in the detection of membrane types. Therefore, in this study, a novel feature extraction method which is an integration of the pseudo amino acid composition integer values mapped in discrete sequence numbers in a matrix. The proposed scheme avoids biasing among the different membrane proteins and their types. Decision making for predicting the identification of membrane protein types was performed using an algorithm framework to improve the learning accuracy, by putting the training samples weights in the learning process of AdaBoost. The performance of different ensemble classifiers such as Random Forest, AdaBoost, is analyzed. The best accuracy achieved is 91.50% for with the Matthews correlation coefficient is 83.0%, and Cohen's Kappa value is 82.7%.***

Keywords : Membrane Protein Types, Random Forest; AdaBoost; Decision tree ; PseAAC..

I. INTRODUCTION

EVERYliving organism's basic building block is cell. Each cell is surrounded by cell membrane which separated by outside environment. AdaBoost, a short form of Adaptive Boosting, in paper [7] suggested first to practical use of boosting algorithm. It focuses and aims to convert a set of weak classifiers into a strong one, main task to improve learning quality. An accuracy are key wards for successful learning methods, boosting is one of the methods to improves the learning quality and accuracy enhancement. Many informatics area has a wide range of classification problem during learning, due to the weak learner. In paper[1] proposed a new regression-based algorithm on a boosted support vector machine. As a transfer learning machine can learn one data task to another task but also learn on labeled and transfer them in different views randomly.

The transfer learning model is one of another important key factor for learning success.

In paper[2] study, an Integrated multi-dimension learning transfer with Adaboost , in this regard the source and target task as a collection of different views and each of these two tasks can be learned from every view at the same time. In extreme learning machine pay attentions in various field due to straight, fast and strong observation in pattern recognition, however, the extreme learning machine performance is still an open challenge due to high-dimensional protein data, Therefore in this[3] paper, introduce hyperspectral image classification with an ensemble extreme learning machine based on bagging and AdaBoost for the classification task SVM based learning. Same learning task needs more performance and accuracy, in this [4] paper describes SVM based AdaBoost framework to improve the performance and accuracy by using the training samples weights of the in the re-sampling process of AdaBoost. To base classifier performance improve the process by AdaBoost algorithm and the complexity of the whole ensemble learning is simplified, in [4] this paper presents an SVM ensemble method based on an improved iteration process of Adaboost algorithm.

Proteins are macromolecules which is made of amino acid residues. Based on their localization, Protein has different types. Some membrane protein encodes gene in most genomes among them, membrane proteins are found around different types of cell membranes and perform many crucial tasks in the cell. Due to their flexibility, absence of stability and partly hydrophobic surfaces membrane proteins have proven to be hard to study [9]. Moreover, Due to the problems in determining structure with experimental techniques, the amount of outer membrane proteins in the structure database is very limited. Therefore learning outer membrane protein from non-outer membrane protein with computational methods is drug and genome sequencing necessity [10]. Cell membrane protein carry out, many of the functions of the vital component of a cell, that are imperative to cells survival, Knowledge of the cell membrane protein type is essential in the determination of its functional types and behaviour of the cell ,they become an attractive target for drug design and new research [15].

Moreover, when membrane proteins become abnormal, major illnesses such as cancer, neurological diseases, cardiovascular diseases and Alzheimer's disease occur[11]. Membrane proteins make up about 50% of drugs [10],[21].

Revised Manuscript Received on October 30, 2019.

* Correspondence Author

Anjna Jayant Deen*, Department of Computer Science and Engineering Maulana Azad National Institute of Technology, Bhopal anjnadeen123@gmail.com

Manasi Gyanchandani, Department of Computer Science and Engineering Maulana Azad National Institute of Technology, Bhopal manasi_gyanchandani@yahoo.co.in

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an open access article under the CC-BY-NC-ND license http://creativecommons.org/licenses/by-nc-nd/4.0/

For the discovery of new drugs and genomes, knowledge of types of membrane proteins and their functions are essential [9], [10]. So predicting the membrane protein types is a crucial task in the field of bioinformatics. But the traditional biological experimental techniques used to predict the type of membrane protein are expensive and time-consuming. Therefore, a quick, efficient and automated technique must be developed to detect the type of uncharacterized protein[25].

There are classes of membrane proteins: single-pass type I,II,III and IV, multi-pass trans-membrane, lipid chain-anchored membrane, GPI anchored and peripheral membrane [20,24].

Recently various feature extractions and classification techniques have been used in order to predict membrane protein types. To predict membrane protein types Amino Acid Composition (AAC) [11],[19], [13], was first used [14a,b], but Amino acid composition is not able to store information of sequence order .Therefore Pseudo Amino Acid Composition resolved these problem (PseAAC) [6],[11], [18],[13],[19]. Different decision tree classifiers and ensemble techniques have been used for identification of membrane protein types and analyse their performance, As evidenced by latest publications[13],[18],[19] in accordance with PseAAC by Chou. This article presents a novel technique of machine learning to identify the types of membrane protein using their sequence of amino acids as the only input[27]. In our study, the dataset taken is imbalanced and therefore ensemble classifiers are used. In this study focuses on a decision about the presence of membrane types by applying ensemble of machine learning classifiers on features extracted from the output of eight different membrane types, like, a singlepass type I, II, III, IV, multipass transmembrane, lipid chain-anchored membrane, GPI anchored and peripheral membrane. Our primary objective was to enhance the performance and accuracy . First, each sequence of proteins has been mapped into a vector based on feature extraction. Second, to recognise class of membrane, the better performing classifier was chosen. Third, ensemble method based an adaptive boosting algorithm using for prediction. The proposed method uses PseAAC feature extraction methods which gives a feature vector of 50 D (dimensions) and helps in categorizing membrane proteins into 8 classes.

II. MATERIALS AND METHOD

A. Data Set

The protein dataset of 560459 manually annotated and reviewed proteins was collected from Swiss-Prot PDB in this study various paper [18],[14],[16],[17]. Dataset is filtered and only membrane protein is chosen. The 62029 membrane proteins is selected and categorized into 8 types, which are: single-pass type-1,II,III IV, multi-pass trans-membrane, lipid chain-anchored membrane, GPI-anchored and peripheral membrane. Further classification of the 62029 membrane proteins done into 43418 training samples and 18611 test samples datasets. Table I demonstrates the sample details.

TABLE I Total samples in the dataset

Membrane protein (types)	No. of instances
single-pass type-I	2948
single-pass type-II	2194
single-pass type-III	211
single-pass type-IV	194
Multi-pass trans membrane	35480
Lipid chain anchoredmembrane	3032
GPI-anchored membrane	651
Peripheral membrane	17319
Total	62029

B. Feature Extraction Methods

Feature extraction is important in the learning process [14a,b][15a]. Specific characteristics are useful in identifying membrane proteins types. Computational biology's amino acid is the primary sequence as the input data, these sequences are in alphabetic order can be transformed into machine-usable information, it is most important as well as most difficult problem is how to convert the biological sequences with discrete vector and in a matrix form Because all current machine learning algorithms, as elaborated in [28] can manage the only vector. A vector described in a discrete model, however, may lose all data about the sequence pattern. The PseAAC or pseudo amino acid composition was suggested in order to prevent totally losing the protein sequence pattern data. [29]. PseAAC has penetrated almost all fields of computational proteomics since the notion of it was proposed [20]. Structure and physicochemical descriptor extracted from protein sequence and other functional interaction properties of the protein. In this study we used iFeature, a python based versatile toolkit for generating numerical descriptor value for feature extraction. iFeature is a powerful toolkit by which each protein should be transformed into a numerical vector that represents enough biological data to meet this objective. The sequences of proteins comprise 20 distinctive amino acids. Twenty amino acids have a prevalent but distinct chemical properties due to functional behavior variations.

C. Pseudo Amino Acid Composition (PseAAC)

Sequence of protein is a mixture of 20 distinct amino acid residues. The amino acid chemical composition is comparable, but for each amino acid it has varying side chain. They have distinct chemical properties due to the different side chains of each amino acid. The composition of the sequence's 20 amino acids is calculated and displayed in the feature vector as a 20-dimensional feature [11].

$$P = [p_1, p_2 \ldots p_{20}, p_{20+1} \ldots p_{20+\lambda}]^T \qquad (1)$$

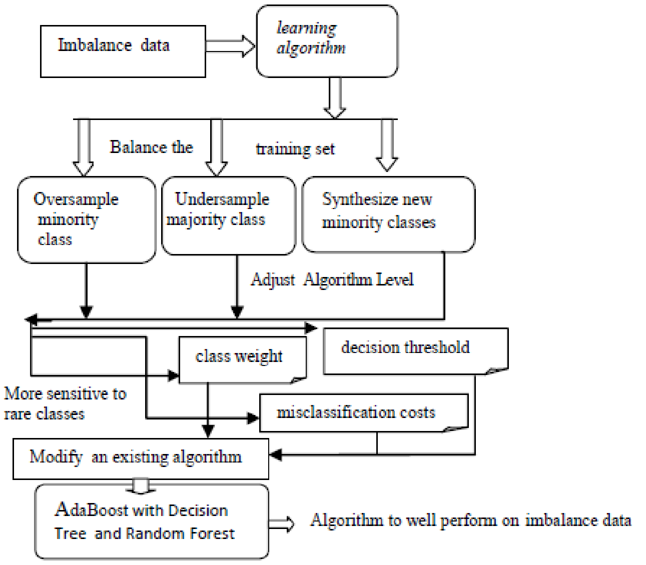
Where the elements are the amino acids composition that can be calculated using the hydrophilic value, hydrophilic value and mass of 20 amino acids.$p_1, p_2 \ldots p_{20+\lambda}$are calculated by,

$$p_u = \begin{cases} \frac{f_u}{\sum_{i=1}^{20} f_i + \omega \sum_{k=1}^{\lambda} \tau_k}, & (1 \le u \le 20) \\ \frac{\omega . \tau_{u-20}}{\sum_{i=1}^{20} f_i + \omega \sum_{k=1}^{\lambda} \tau_k}, & (20 + 1 \le u \le 20 + \lambda) \end{cases} \quad (2)$$

Where ω is the weight factor (set to 0.05) and τ_k is the k^th tier correlation factor that represents order of correlation between all the k^th-most continuous residues. Amino Acid Compositions is a 20-dimnesional vector that indicates 20 kinds of amino acids frequency of occurrence in the protein sequence. The length of PseAAC depends on value λ, which in our case is set to 30, thus giving 20+30=50D feature vector.Thus the feature dimension from PseAAC is 50 descriptor vector space.

D. Classification

In this study, an imbalanced dataset is taken from www.uniprot.org benchmark protein dataset. A predictor trained with a data set of uneven proteins would result in an incorrect prediction. It is therefore essential to discover an efficient strategy for optimizing the imbalanced protein dataset and minimizing the classifier's biased prediction[32]. AdaBoost Ensemble refers to the general technique of combining several weak models to achieve a stronger model. In boosting, AdaBoost classifier trains a series of weak classifiers and subsequently increase the penalty on misclassified points. The final prediction is the weighted vote of all weak classifiers. The various classifier used in this study is decision tree, random forest and AdaBoost, will help in achieving better accuracy therefore use of ensemble techniques based classifiers such as Random forest, Adaboost and Decision tree performances of classifiers are compared with each other with namely base classifiers. The main purpose is to make a model which predicts the target variable by using learning decision making rules from feature vector.


Imbalance data
learning algorithm
Balance the
training set
Oversample minority class
Undersample majority class
Synthesize new minority classes
Adjust Algorithm Level
class weight
decision threshold
More sensitive to rare classes
misclassification costs
Modify an existing algorithm
AdaBoost with Decision Tree and Random Forest
Algorithm to well perform on imbalance data

Work Flow Diagram of an ensemble class algorithm for predict Imbalance Membrane Protein Types ; Fig.2. Schematic diagram of the Proposed Model

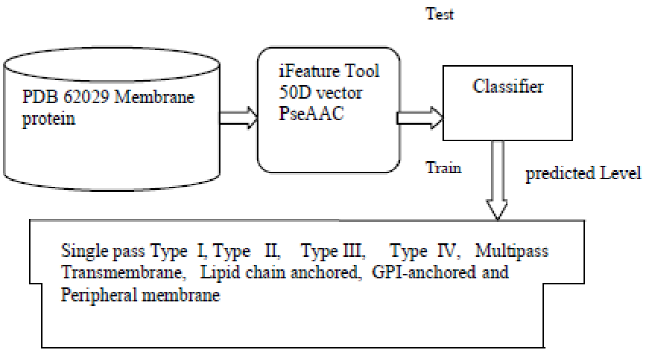

Test
PDB 62029 Membrane protein
iFeature Tool 50D vector PseAAC
Classifier
Train
predicted Level
Single pass Type I, Type II, Type III, Type IV, Multipass Transmembrane, Lipid chain anchored, GPI-anchored and Peripheral membrane

Fig 1. Prediction level of Membrane Protein Types

E. Performance Estimation of a Classifier

The aim to AdaBoost algorithm first resolved class imbalance problem, when the classes are balanced predictive accuracy works fine. Adaboost as the first successful boosting algorithm for binary classification problem. That is every class in data set are equally important. When the classifier classified a test data set with imbalanced class distribution, then predictive accuracy on its own is not a reliable indicator of classifiers effectiveness. Learning with AdaBoost model by weighting training instances and the weak learner selves. Predict with AdaBoost by weight predictions from weak learners.

The proposed of this study to performs more nearly with large multiclass data samples separated as majority class and minority class, the rest of the classifier for multiclass imbalance learning, this work n number of data samples creates balance training subset by random under sampling of the majority class samples, each subset learn process fixed weight update with instances, so that the training subsets depends on the degree of the class imbalance, resultant outcomes are computed based on majority voting.

In membrane protein learning tasks, the membrane types T often takes discrete value sets of classes: (T_{c1}, T_{c2}, T_{c3}.........T_{ck}). were membrane protein types might map a feature description of types of k possible membrane protein. The features PseAAC descriptors vector 50 dimensional can be map each one of k class. multiclass learning tasks can be generalized easily by using decision tree algorithms. In each leaf of decision tree can be labeled with a given number of classes, were k, and internal subtask of each instance can be selected to discriminate into their eight classes. The main aim is to learn k class functions form T tasks, were T task consisting of various sub-task T_1, T_2, T_3...T_k, one of each class. Instead of using bootstrap copies, and finding optimal cut for randomized features at each node, it randomly selects a cut-point. Another type of ensemble method known as AdaBoost attempts to enhance predictive outcomes by mixing weak learners to build a meta classifier *k.* Prediction of one learner is used to train another learner and so on, resulting in reduced errors in each steps.

Ensemble Algorithm

Let consider P represents the actual training data,
k represents the number of base classifiers,
and D represents the test data.

```
 for $i$ = 1 to $k$ do
Create  $P_i$ (Training set from P ).
Build T (Base classifier from P ).
end for
for each test record $x \in$ D do
$T_{C*}(x) = Vote\ (\mathbf{T}_{C1}(\mathbf{x}), T_{C2}(\mathbf{x}), \ldots, T_{Ck}(\mathbf{x}))$
end for
------------------------------------------------------
```

Decision tree is made by either using information gain and prunes it to reduce errors. In order to enhance the classification rate, Random forest utilizes many decision trees. In Random forest, different trees are made by taking randomly sampled vectors.

III. RESULT AND DISCUSSION

In this paper, feature extraction techniques, namely PseAAC are used which give feature vector of 50-dimension to train the model. For getting best accuracy, various type of classifiers such as Decision tree, AdaBoost , RF (Random Forest), AdaBoost with decision tress , AdaBoost with random forest classifiers are used and on the basis of result obtained from them, the model is built. It is visible in Table III that the ensemble classifiers AdaBoost, Random forest (RF) and Decision Tree classifier produces high accuracy i.e., approximately 88-91%. It gives the highest overall accuracy for single pass type-I, type-II, type-III, type-IV, multi-pass transmembrane, lipid and GPI anchored-membrane protein types.

A. Accuracy

Accuracy, Sensitivity and Specificity : Predictive accuracy € can be expressed in term of sensitivity and specificity

$$€ = \frac{TP+TN}{TP+FP+FN+TN} \quad (3)$$

$$= \frac{TP+TN}{P+N} \quad (4)$$

$$€ = \frac{TP}{P} \times \frac{P}{P+N} + \frac{TN}{N} \times \frac{N}{P+N} \quad (5)$$

$$€ = Sensitivity \times \frac{P}{P+N} + Specificity \times \frac{N}{P+N} \quad (6)$$

Based on the various performance metrics, resultant can be characterize a classifier. In the terms of F-1 score, Precision, Recall, TPR (true positive rate) ,FPR (false positive rate) and accuracy [19],[11],[6]. The overall accuracy of different classifiers are shown in Table II.

F-measure is a performance measure of the model. It uses both precision as well as recall to calculate the score.

$$Precision = \frac{TP}{TP+FP} \quad (7)$$

$$F_{measure} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (8)$$

Usually, F-measure is less for classes with less number of instances and high for classes with greater number of instances. F-measure of various classifiers are shown in table V. The specificity of classifiers is shown in Table III, the range of specificity is from 95% to 100%, because TNR (true negative rate) is good i.e., wrongly classified samples are very less.

Sensitivity of Classifiers is shown in Table IV. Shows that classes with more number of samples like multiclass -transmembrane and peripheral membrane are gives better sensitivity results, but for less sample classes like type-2, type-3, type-4, GPI, are less sensitive. All classifier has the highest sensitivity for the majority classes, single-pass type classes and minority classes among all the classifiers implemented does perform well.

TABLE II
OVERALL ACCURACY OF DIFFERENT CLASSIFIERS

Classifier	Accuracy
Decision Tree	83.26
Random Forest	89.38
AdaBoost	76.88
Adaboost+DT	90.01
Adaboost+RF	91.50

B. Cohen's Kappa

It measures the agreement between predictions and actual results, it can be considered as a good measure for imbalanced data set.

$$K = \frac{n_0 - n_e}{1 - n_e} \quad (9)$$

and $$n_0 = \frac{TP+TN}{TP+TN+FP+FN} \quad (10)$$

$$n_e = \left(\frac{TP+FP}{TP+TN+FP+FN}\right) \cdot \left(\frac{TP+FN}{TP+TN+FP+FN}\right) + \left(\frac{FN+TN}{TP+TN+FP+FN}\right) \cdot \left(\frac{FP+TN}{TP+TN+FP+FN}\right) \quad (11)$$

Where n is number of samples, $n_0$is the relative agreement among prediction and actual result (similar to accuracy), n_e is hypothetical agreements due to chance agreement between actual and predicted results. The value of Cohen's Kappa for different classifiers is shown in Table 6.The Cohen's Kappa value for the classifiers such as Decision tree is below 75%. But for the ensemble classifiers such as AdaBoost with DT is 81.8%, AdaBoost with Random Forest 82.70% value. As noted, kappa value for Random forest is 80.40%. Because protein sequences are patterns of different length, it is found that proposed PseAAC to integer encodings 50D vector features space with AdaBoost classifier provide better performance for membrane protein types.

C. Mathew's Correlation Coefficient (MCC)

It is popular performance evaluation parameter for any others prediction model. Mathew's Correlation Coefficient ranges from -1 to +1 where former representing total disagreement and latter representing total agreement with observation and prediction whereas 0 is equal to a random prediction. Therefore, the Mathew's Correlation Coefficient (MCC) (eq. 12) is considered to be the better performance for the classification problem of unbalanced data [14].

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (12)$$

From the study, it is found that all ensemble classifiers perform excellently compared to tree based classifiers on the various performance measures such as specificity, accuracy, , F-measure, Prediction result shows that the proposed method achieved good accuracy for the independent datasets, the various classifier prediction results shows bar graph as shown in Fig. 3.

Table- III: Performance of Classifier

Classifier	COHEN’S KAPPA VALUE	Mathew’s Correlation Coefficient
Decision Tree	79.0	73.1
Random Forest	80.4	80.8
AdaBoost	65.1	68.8
AdaBoost with Decision Tree	80.7	82.1
AdaBoost with Random Forest	82.6	83.2

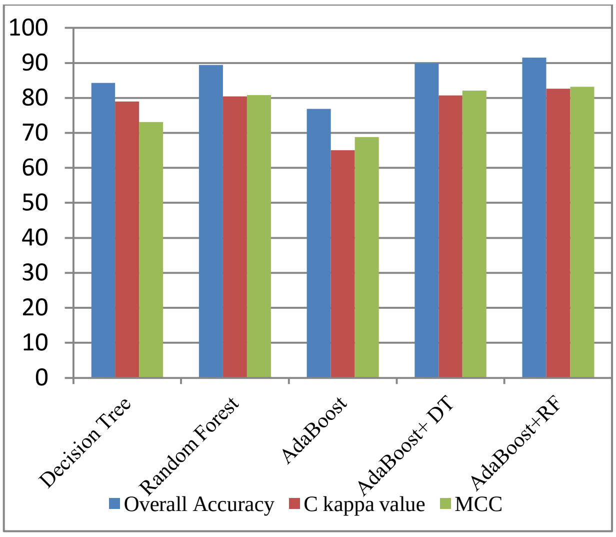


Fig.3. Comparison between performance measuring parameters of different classifiers

TABLE III SPECIFICITY OF DIFFERENT CLASSIFIERS

Types	Specificity				
	Decision tree	Random forest	Adaboost	Adaboost Decision Tress	Adaboost with Random forest
GPI	0.98	0.88	0.87	0.99	0.85
Lipid	0.97	0.93	0.85	0.98	0.89
Multi-pass	0.87	0.95	0.95	0.87	0.95
Peripheral	0.93	0.97	0.92	0.95	0.94
Type I	0.97	0.94	0.87	0.99	0.92
Type II	0.97	0.94	0.85	0.99	0.85
Type III	0.99	0.96	0.86	1.00	0.86
Type IV	0.99	0.93	0.88	1.00	0.90

TABLE IV SENSITIVITY OF DIFFERENT CLASSIFIERS

Types	Sensitivity				
	Decision Tree	AdaBoost	Adaboost Random Forest	Adaboost Decision Tress	Random Forest
GPI	0.12	0.16	0.06	0.03	0.06
Lipid	0.48	0.16	0.58	0.49	0.55
Multi-pass	0.92	0.89	0.97	0.98	0.97
Peripheral	0.87	0.86	0.93	0.95	0.93
Type I	0.56	0.19	0.61	0.55	0.58
Type II	0.53	0.04	0.45	0.46	0.45
Type III	0.38	0.19	0.49	0.35	0.49
Type IV	0.13	0.19	0.15	0.12	0.15

TABLE V F1-SCORE OF DIFFERENT CLASSIFIERS

TYPES	F1-SCORE				
	DECISION TREE	RANDOM FOREST	ADABOOST	ADABOOST DECISION TRESS	ADABOOST WITH RANDOM FOREST
GPI	0.10	0.07	0.14	0.04	0.07
LIPID	0.50	0.60	0.24	0.58	0.67
MULTI-PASS	0.92	0.95	0.87	0.95	0.96
PERIPHERAL	0.86	0.91	0.79	0.92	0.91
TYPE I	0.54	0.68	0.23	0.69	0.72
TYPE II	0.50	0.61	0.08	0.63	0.68
TYPE III	0.40	0.66	0.15	0.67	0.74
TYPE IV	0.20	0.57	0.24	0.60	0.57

IV. CONCLUSION

In this experiment results, PseAAC features are used from protein amino acid sequence datasets. The various kinds of classification methods are generate useful information to find diseases and relationship between other functions of protein. These 50D (Dimension) feature vectors are further used to learn various types of classifiers such as Random forest, Adaptive Boosting, AdaBoost with random forest and AdaBoost with Decision Tree. The performance of these classifier are measured on different performance measures and compared with each other which shows that ensemble methods perform well than other classifiers. In future, with help of different types of extraction methods or oversampling and under sampling techniques classifiers can give more better accuracy.

REFERENCES

1. Lin Gao, Feng Gao, Xiaohong Guan, DZhou, JLI "A regression Algorithm Based on Adaboost". 6th IEEE World Congress on Intelligent Control and Automation 2006 Volume: 1 pp1-12
2. Zhijie Xu, Shiliang Sun " multi view Transfer Learning with Adaboost". 2011 IEEE 23rd International Conference on Tools with Artificial Intelligence , pp 1-12
3. Alim Samat , Peijun Du , Sicong Liu , Jun Li , Liang Cheng, "E2LMs : Ensemble Extreme Learning Machines for Hyperspectral Image Classification". IEEE Journal of in Applied Earth Observations and Remote Sensing (Volume: 7 , Issue: 4 , April 2014 Page 1060 - 1069
4. Xiaolong Zhang , Fang Ren, "Improving Svm Learning Accuracy with Adaboost ". IEEE *Xplore* 2008 Fourth International Conference on Natural Computation ,DOI: 10.1109/ICNC.2008.841
5. Yiming Tian , Xitai Wang, " SVM ensemble method based on improved iteration process of Adaboost algorithm". 29th Chinese Control And Decision Conference (CCDC), 2017, pp 4026-4032
6. Cai, Y.D., Ricardo, P.W., Jen, C.H., Chou, K.C., "Application of SVM to predict membrane protein types". 2004, J. Theor. Biol. 226 (4), pp 373-376.
7. Yoav Freund and Robert E. Schapire. "A decision-theoretic generalization of on-line learning and an application to boosting". Journal of Computer and System Sciences, 55(1), August 1997, pp 119–139.
8. Elisabeth P carpenter, KonsatinosBeis, Alexander D , So Iwata., "Overcoming the challenges of membrane protein crystallography". CurrOpinStructBio , 2008, PMCID. 18(5), page 581-586 .
9. Chen Z, Zhao P, Li F, Leier A, Marquez LogoTT, Ang Y, Webb GI, Smith AI, Daly RJ, Chou KC, Song J. " iFeature:a python package and web server for feature extraction and selection from protein and peptide sequence". Bioinformatics,Volumw 34 issue 14,15 july 2018 page2499-2502 doi:10.1093/bioinformatics/bty140
10. Qiao Ning, Zhiqiang Ma, Xiaewi Zhao. "dformKNN -PseAAC detecting formylation site from protein sequence using K-nearest neighbor algorithm via chou's 5-step rule and pseudo component". 2019 Journal of Theoretical Biology.470,43-49.
11. Ali, F., Hayat, M."Classification of membrane protein types using voting feature interval in combination with Chou’s pseudo amino acid composition". 2015,J. Theor, Biol. 384 pp 78-83.
12. Marco Punta, Lucy R. Forrest, Henry Bigelow, Andrew Kernytsky, Jinfeng Liu, and BurkhardRost. "membrane protein prediction methods" 2007. NIH Public access PMC ; 41(4): pp 460–474.
13. Chen, W., Ding, H., Feng, P., "iACP: a sequence-based tool for identifying anti-cancer peptides". 2016, Oncotarget 7,pp 16895-16909.
14. Chou, K.C., "Insights from modeling the tertiary structure of BACE2". J. Proteome Res. 3, 2004a 1069-1072.
15. Chou, K.C., "Insights from modelling three-dimensional structures of the human potassium and sodium channels". 2004b, J. Proteome Res. 3,pp 856-861.
16. Chou, K.C., "An unprecedented revolution in medicinal chemistry driven by the progress of biological science".2017, Curr. Top. Med. Chem. Doi:10.2174/1568026617666170414145508.
17. Chou, K.C., Elrod, D.W., "Prediction of membrane protein types and sub cellular locations. Proteins" Struct. Funct. Bioinf. 1999. 34 (1), pp 137-153.
18. Chou, K.C., Shen, H.B., "MemType-2L: a web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM". Biochem. Biophys. Res. Commun. 2007,360 (2), pp 339-345.
19. Gao, Q.B., Ye, X.F., Jin, Z.C., He, J., " Improving discrimination of outer membrane proteins by fusing different forms of pseudo amino acid composition" . J. Anal. Biochem. 2010, 398, pp 52-59.
20. Chen, Y.K., Li, K.B.,"Predicting membrane protein types by incorporating protein topology, domains, signal peptides, and physiochemical properties into the general form of Chou’s pseudo amino acid composition". 2013, J. Theor, Biol. 318, pp1-12.
21. Golmohammadi, S.K., Kurgan, L., Crowley, B. Reformat, M., "Amino acid sequence based method for prediction of cell membrane protein types". 2008, Int. J. Hybrid Inf. Technol. 1 (1), pp 95-109.
22. Hayat M., Khan, A., "Mem-Phybrid: hybrid features-based prediction system for classifying membrane protein types". Anal. Biochem. 2012, 424, pp 35-44.
23. Hayat M., Khan, A., Yeasin, M., "Prediction of membrane proteins using split amino acid and ensemble classification". Amino Acids 42 (6), 2012, pp 2447-2460.
24. Howe, W.J., "Prediction of the tertiary structure of the beta-secretase zymogen". Biochem. 2002, Biophys. Res. Commun 292, pp 702-708.
25. Huang, C., Yuan, J.Q., "A multi label model based on Chou’s pseudo-amino acid composition for identifying membrane proteins with both single and multiple functional types". 2013, J. Membr. Biol. 246 (4), pp 327-334.

26. Hayat M., Khan, A, "Predicting membrane protein types by fusing composite protein sequence features into pseudo amino acid composition". 2011, J. Theor. Biol. 262, pp 10-17
27. Liu, B., Wu, H., "Pse-in-One 2.0: an improved package of web servers for generating various modes of pseudo components of DNA, RNA, and protein sequences". 2017,Nat. Sci. 9, pp 67-91.
28. Liu, H., Wang, J.M., Xue, L., Chou, K.C., "Using Fourier spectrum analysis and pseudo amino acid composition for prediction of membrane protein types". Protein J. 24 (6), 2005, pp 385-389.
29. Shen, H., Chou, K.C, "Using optimized evidence-theoretic K-nearest neighbour classifier and pseudo-amino acid composition to predict membrane protein types". 2005, Biochem. Biophys. Res. Commun. 334 (1), pp 288-292.
30. Mahdavi, A., Jahandideh, S., "Application of density similarities to predict membrane protein types based on pseudo-amino acid composition". J. Theor. Biol.2011, 276, pp 132-137.
31. Nanni, L., Lumini, A., "An ensemble of support vector machines for predicting the membrane protein type directly from the amino acid sequence". 2008, Amino Acids 35 (3), pp 573-580.
32. Shen, H.S., Chou, K.C., "Using ensemble classifier identify membrane protein types". 2007, Amino Acids 32, pp 483-488.
33. Wang, J., Li, Y., Wang, Q., You, X., Man, J., et al., "ProClusEnsem: predicting membrane protein types by fusing different modes of pseudo amino acid composition". 2012, Comput. Biol. Med. 42.pp 564-574.

AUTHORS PROFILE

Anjna Jayant Deen received her B.E Computer Technology from Govt. Engg College, Barkatullah University Bhopal and M.Tech. in CSE from MANIT Bhopal since 1993 and 2007 respectively. After graduated in Engineering she joined Govt. Engg. college Bhopal as a faculty . she was currently a PhD student in MANIT Bhopal. Her research interest is in Data science, Network security, Bioinformatics, Artificial Intelligence, Machine learning and Neural network. Her publications in more than 14 research papers in Scopus-journal, International Conference and International journal. She is a member of IEEE.

Manasi Gyanchandani received her B.E and M.E. in Computer science from College of Engg Badnera Amravati and Ph.D. in CSE from MANIT Bhopal since 1995,1997 and 2005. After Post graduated in Engineering she joined MANIT Bhopal as a faculty. she was currently an Ass.Professor in MANIT Bhopal. Her research interest is in Internet of Things(IoT), Network Security, Big-data, Artificial Intelligence, Machine learning, Neural network, Intrusion Detection & Information Retrieval. Dr. Manasi Gyanchandani, publications in more than 32 research papers in Sci-journal, Scopus-journal, International Conference, book chapter, International Journal and National Conference. She is a Life member of ISTE.