Journal of Theoretical and Applied Information Technology
30th April 2018. Vol.96. No 8

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

MACHINE LEARNING CONFIGURATIONS FOR HUMAN PROTEIN CLASSIFICATION USING SDFES

SUNNY SHARMA [1], AMRITPAL SINGH [2], GURVINDER SINGH [3], RAJINDER SINGH[4]

[1, 2] Research Scholar, DCS, Guru Nanak Dev University, Amritsar-143001, Punjab, India

[3]Professor and Dean, DCS, Guru Nanak Dev University, Amritsar-143001, Punjab, India

[4]Professor, DCS, Guru Nanak Dev University, Amritsar-143001, Punjab, India
E-mail: [1]sunnysharma05@yahoo.co.in

ABSTRACT

The identification of target proteins for diseased condition yields the development of the disease detection recommender system and drug discovery processes whose reticence can demolish the pathogen. The testing of this drug discovery is done through clinical and in addition through pre-clinical observations first on the creatures then on people. Thereafter the discovered drug is ready for public use. But if the drug discovery testing phase does not show the suitable consequences, then the entire task must be repeated. This repetitive clinical as well as the preclinical experimentation task is very cumbersome. But keeping in view the importance of the disease detection and drug discovery phase in protein identification as well as in the protein classification process this task must be done by researchers. The advancements in computational biology reveal the importance of computational prediction of protein function or to identify the target on the basis of protein sequence extracted features. To accurately predict the human protein functionalities, lots of approaches are incorporated but this is a very cumbersome task due to the large and versatile nature of the domain. The present work will help to do this job through computational prediction. This paper involves the development of a model which use associative rule mining to extract the sequence derived features at a single platform (SDFES-Sequence derived feature extraction server) from the given human protein sequence and then critically analyzed with machine learning (ML) approaches under the aegis of data analysis tool WEKA. The new sequence derived features are identified and incorporated in the data set, and the scopes of ML approaches were examined for effective prediction. The important configuration incorporation and their configured comparison of approaches are completed to accomplish higher accuracy. In addition to comparative analysis, the limitation of ML approach is discussed along with its remedies by changing the configurations. The proposed work will assist to derive the sequence extracted feature together at a single place and further predict the class or function of the protein which leads to the innovation in drug discovery and disease detection recommender systems.

Keywords: *Protein, Machine Learning, WEKA, Random Forest, Decision Tree.*

1. INTRODUCTION

Protein function prediction, protein classification, disease detection and drug discovery as well as their recommendation systems, are immense areas with the colossal amount of information. Lots of research activities are happening for protein classification and protein function prediction, but the learning about its right perception is quite low, so there is a need to explore this vast domain. The machine learning (ML) approaches gives promising results to vast as well as to not so clearly characterized zones of research, this encourages using the power of ML to explore such a vast domain and boost the present understanding of protein. The survey of 65 papers on ML approach shows that ML approaches are extensively used for human protein function predictions and it is the prominent area where ML got some challenges to show its supremacy [1]. The association of rule mining in data mining is a frequently used technique, which provides significant valuable rules or patterns. This association of rule mining extracts the possibility of the co-occurrence of common features in a data collection, this also encourages to use the power of rule mining to explore such a vast and not so clearly characterized zones of research and find the occurrence of common features while boosting up the present knowledge of protein. On the other

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

hand, there is very much reliable, vibrant and the white box decision tree [2] [3] technique of machine learning that helps to predict protein class, and do the classification of protein with the help of rule mining. Its step by step approach encourages computational experts to use this technique even without much information of the concerned area; with having nodes and edges this technique portrays different functionalities at various levels of the tree [4]. It clearly characterizes the issue structure and its elucidations progressively in the hierarchical way which is considerably less demanding to fathom. This hierarchical approach guides the input parameters to achieve its goals [1] [5]. So distinguishing challenges and the conceivable answers for protein classification problems is the key concentration of the investigation.

Human protein classification is an important research area because of its implication on various key research areas like disease detection, drug discovery, crop hybridization, etc.

The motivation for present research is to facilitate the sequence derived feature extraction process which in turn improves the accuracy of classification and prediction of human protein.

Second motivation is the use of machine learning approach for the process which is already applied in existing literature. But this research showcase that machine learning approach too have some pitfalls.

So to overcome those issues various configuration were tried and varying accuracy results of individual classes were obtained. This shows that for tasks like drug discovery, when the interest in identifying a particular class is more, these configuration are very useful.

1.1 Associative Rule Mining

In data mining association rule mining has been extensively studied & it is a frequently used technique, which produces significant valuable rules or patterns. Association is a rule mining which extracts the possibility of the co-occurrence of items or features in a collection. Association rules or the production rules convey the relationships between co-occurring features or items. Among the association the strong association rules favors the strong relation between data items & the weak association rules are considered least related items. To measure the interest, the support and confidence are two listed measures. To extract associated variables, the values for support and confidence are extracted to the production system. This grabbed the important relationship between variables only if satisfied minimum support and confidence. To find out all the frequently occurring items in the repository is a cumbersome because to find out the entire item from the repository with all combinations is tough job. The set of probable item sets is said to be the power set over all the items & has size 2n− 1. As the data items increases, the size of combination increases exponentially. The possible efficient search can be performed using anti- monotonicity [35].

1.2 Protein and Protein Function Information

Proteins are the complex foundations for all the living creatures on earth. Under the aegis of protein, the body's tissues and organs perform functionalities, structuring and controlling the body's parts. The building blocks of proteins are amino acids, they assist or anchored to each other and perform functionalities like store information, send information to body organs as well as structuring in the living cell. It can be portrayed as a string of 20 diverse amino-acids (AA) and they are anchored to make a protein of the living cell. These20 distinct kinds of amino acids are recorded as Serine, Alanine, Proline, Arginine, Valine. Asparagine,Threonine, Glumatic, Aspartic, Glycine, Tryptophan, Histidine, Cysteine, Isoleucine, Phenylalanine, Tyrosine, Glutamine, Leucine, and Methionine, Lysine.

Based on functionalities of proteins, they are categorized as Transport, storage, motor, structural, receptor, signaling, hormones, Antibodies, enzymes etc. Fundamentally the broad chain of amino acids is framed with peptide bond which is the bridge bond of amino corrosive structures with another amino acid corrosive structure. The progression of such peptide bonds is known as polypeptide bonds which is actually responsible for the correct functionality of every protein and depicting the interesting 3D structure [5].

1.2 Protein Features Extraction from Sequence

Protein classes having different protein sequences and these sequences constitute with various features which are actually responsible for functionalities of body organs and tissues. The sequence-based features are useful to anticipate protein class and also very helpful to build automatic protein class predictor classifier. The features can be extracted from protein sequence by utilizing different online bio-informatics servers or tools, which are freely available online such as. TMHMM Sererv.2.0 [6], SignalP4.1 Server [7], NetNGlyc1.0 Server [8], ProtParam [9], PSORT

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[10], PROFEAT [11], but extraction of various unique features from different online tools is not an easy task because for this purpose user has to visit online websites and upload the sequence and then get the features one by one of these tools. To extract unique features from the single sequence at the solitary stage is a very cumbersome task. To get rid of such task i.e. to visit these web tools the SDFES (Sequence derived features extraction server) is developed in MATLAB [12] [13], which extract lots of features at a solitary stage. SDFES is tested for extraction of accurate results by comparing its outcomes with the outcome generated from online web-based tools which are mentioned above. The supremacy of associative rule mining [14] helps to the development of SDFES server, which guides how different rules can be integrated together to develop a particular feature extraction platform. Using this very important approach the SDFES is developed to extract some unique features from the unique protein sequence. It shows the importance of rule mining and depicts how it can be helpful in machine learning for forthcoming research. The proposed way is reliable and clear to utilize, the only manual push to extract features is to just provide the protein sequence to the server. SDFES incorporate 5 new features for protein class prediction which further helpful in disease detection and drug discovery and its utilization strategies, which is a noteworthy commitment of this server. The data source considered for analysis is the human protein from the database HPRD [15], which contains thousands of AA sequences. SDFES extract the features from the protein sequence at solitary stage can be listed as: total number AA, the molecular presence of individual AA, positive as well as negative-charged residues, extinction-coefficient, Isoelectric-point(IP), molecular weight, Aliphatic-index(AI), absorbance, instability-index(II), gravy, computation of isoelectric point by molecular weight, volume, total atoms, C,H,N,O,S detail, tiny, polar, small, non-polar, charged, aromatic, basic, acidic, aliphatic residues, detail of codons, nucleotide density, nominal and mono-isotopic mass plots.

The importance of various sequence-based features with its detailed information is listed as:

1.2.1 Extinction coefficient (EC) of protein

Extinction coefficient or molar absorption coefficient is an important protein parameter. In the laboratory using UV spectrophotometry protein concentration in a solution is frequently extracted. How much light is absorbed by the protein depend upon size, composition & the wavelength of the light-beam in water. This feature can be extracted from the sequence using the composition of W, Y, C amino acid [16] [9].

1.2.2 Number of negatively & positively charged residues

It is be extracted by the total composition of Aspartic(D), Glutamic(E) acids for negative charges and by the total composition of R as well as K amino acids to grab positively charged residues [35].

1.2.3 Molecular weight (MW)

Atomic mass or sub-atomic weight is the mass of a particle. The molecular weight of the protein sequence can be ascertained from the protein sequence as the sum of the atomic weights of every constituent component multiplication by the quantity of particles of that component in the sub-atomic equation [35] [9].

1.2.4 Absorbance or optical density

Absorbance characterizes the material light blocking ability. This feature can be grabbed by the fraction of the protein's EC with the MW [17].

1.2.5 Isoelectric point (IP)

It describes the pH point about which the amino acid is neutral. It is the value at which the AA or the molecule does not carry an electrical charge or molecule does not transfer in an electric field [35].

1.2.6 Computation of IP/molecular weight

It is extracted by the fraction of the IP of individual AA with the MW of the given amino acid in a protein sequence [35] [9].

1.2.7 Aliphatic index

This feature is extracted by the composition of relative-volume (aliphatic-side-chains) which are basically A, V, I, L amino acids. This feature is the positive feature to increasing thermo-stability [18].

1.2.8 GRAVY

Gravy is Grand Average of Hydropathicity and it is extracted as the addition of hydropathy values of the whole AA of the protein sequence fraction by all the residues [19].

1.2.9 Instability index

It describes the estimated stability life of the protein. Guru parsad et al. [20] did analysis of 32 stable and 12 unstable proteins as well as the occurrence of di-peptides in the stable as compared to unstable proteins. They assigned an instability dipeptide value to each of the 400 dissimilar dipeptides. A protein whose instability index is >40 is expected as unstable & whose value <40 expected as the stable [20]

1.2.10 Volume

Journal of Theoretical and Applied Information Technology
30th April 2018. Vol.96. No 8

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Volume is extracted from the molecular weight of the peptide & an average protein fractional specific volume [21].

1.2.11 Residues properties

There are few residues properties like Polar, Non-Polar, Tiny, Small, Aliphatic, Acidic, Aromatic, basic, charged which also play an important role in protein function. SDFES extracts properties from the protein sequence with their molecular percentage and total composition of various acids present in the protein sequence.

1.2.12 CHNOS

Protein residues always contain C (Carbon), (H) Hydrogen, (N) Nitrogen, and (O) Oxygen. Sometimes, the protein contains (S)Sulphur also. The SDFES derive detail information of CHNOS can be calculated which describes the development of Atoms given in the protein sequence.

The proposed way to derive the sequence based features from the human protein sequence is reliable and efficient. It follows the modular approach to derive the features from the sequence. To extract the particular feature corresponding to its module will be called which return the required feature which is helpful in prediction.

The pseudo code for the Human Protein SDFES (Sequence Derived Features Extraction Server) is given below:

Human Protein Sequence Derived Features Extraction Server (HP-SDFES)
TAA: The total AA in Sequence
//Input Data
Read Protein _Sequence
Read Various variables to Calculate-Residues Properties
// Calculate-Protein _Sequence Properties
Call Calculate-Total-Amino-Acids(Protein _Sequence)
Call Calculate-Individual-Residues-M-Percentage(Protein _Sequence, TAA)
Call Calculate-Basic-Properties(Protein _Sequence, TAA)
Call Calculate-Extinction-Coefficient(Protein _Sequence)
Call Calculate-NP-Charge(Protein _Sequence)
Call Calculate-Molecular-Weight(Protein _Sequence)
Call Calculate-Absorbance(Protein _Sequence)
Call Calculate-Isoelectric-Point(Protein _Sequence)
Call Calculate-IP/MW(Protein _Sequence)
Call Calculate-Aliphatic-Index(Protein _Sequence, TAA)
Call Calculate-Gravy(Protein _Sequence, TAA)
Call Calculate-Instability-Index(Protein _Sequence, TAA)
Call Calculate-Volume(Protein _Sequence)
Call Calculate-CHNOS(Protein _Sequence)
// Termination
Display the Extracted Features and get the Next Protein-Sequence

The detailed information for each pseudo code is listed below. It describes how these features can be extracted from the sequence with the collaboration of each other.

The pseudo code for calculation of total AA is listed as [35]

A: The number of Alanine Amino Acid in Protein
R: The number of Arginine "
D: The number of Aspartic "
N: The number of Asparagine "
C: The number of Cysteine "
E: The number of Glumatic "
Q: The number of Glutamine "
G: The number of Glycine "
H: The number ofHistidine "
I: The number of Isoleucine "
L: The number of Leucine "
K: The number of Lysine "
M: The number of Methionine "
F: The number of Phenylalanine "
P: The number of Proline "
S: The number of Serine "
T: The number of Threonine "
W: The number of Tryptophan "
Y: The number of Tyrosine "
V: The number of Valine "
// Input Data is Protein _Sequence
// Calculate Total Amino Acids Present in Protein _Sequence
Read various diminutive units called amino acids
Calculate the TAA from the sequence
// Display
Display TAA;

The pseudo code to calculate individual residues with its Molecular Percentage is listed as:

MPA: Molecular Percentage of A
MPR: Molecular Percentage of R
// Input Data is Protein _Sequence &TAA
//Calculate Individual Residues Molecular Percentage
Read various 20 diminutive units called amino acids
Then
MPA=A*100/ TAA;
MPR=R*100/ TAA;
Similarly for other Amino Acid's present in Protein Sequence
//Display
Display Individual Residues Molecular Percentage

The pseudo code to for basic features is listed as:

// Input Data is Protein _Sequence
// Calculate Basic-Properties Present in Protein _Sequence
Read various 20 diminutive units called amino acids
Then
Calculate Polar, Non-Polar, Tiny, Small, Aliphatic, Acidic, Aromatic, Basic, and Charged residues as well as their Molecular percentage
//Display
Display Basic-Properties Polar, Non-Polar, Tiny, Small, Aliphatic, Acidic, Aromatic, Basic, Charged

The pseudo code for EC is listed as: [16] [9]

MPP: Molecular Percentage for Polar
EC: Extinction-Coefficient
// *Input Data* is Protein _Sequence
//Read variables e_{TYR}=1490; e_{TRP}=5500; e_{CYS}=125;
// calculate EC feature Present in Protein _Sequence
Read various 20 diminutive units called amino acids
$EC=Y* e_{TYR}+ W* e_{TRP}+C* e_{CYS}$
//Display

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

```
Display EC
```

The pseudo code to calculate the negative and the positive charge is listed as:

```
NC:          Negative Charge
PC:          Positive Charge
// Input Data is Protein _Sequence
//calculate NP charge feature Present in Protein
_Sequence
      Read various 20 diminutive units called amino acids

          NC=D+E;
          PC=R+K;
//Display
Display negative-positive charge NC, PC
```

The pseudo code to calculate MW is listed as: [9]

```
MW:          Molecular Weight
// Input Data is Protein _Sequence
//Read variables
  Read the individual Molecular Weight of  20 diminutive
units
// calculate the MW feature Present in Protein _Sequence
      Read various 20 diminutive units called amino acids

Then
Calculate  the Molecular Weight as MW
//Display
          Display MW
```

The pseudo code to calculate absorbance is listed as: [17]

```
// Input Data is Protein _Sequence
// Calculate Absorbance feature Present in Protein
_Sequence
      Call Calculate-Extinction-Coefficient(Protein
_Sequence)
      Call Calculate-Molecular-Weight(Protein _Sequence)
Absorbance= Extinction-Coefficient/ MW;
//Display
            Display Absorbance;
```

The pseudo code to calculate ISP is listed as:

```
ISP:       Isoelectric-Point
// Input Data is Protein _Sequence
// Calculate ISP feature Present in Protein _Sequence
ISP = 2143soelectric(Protein _Sequence); //MATLAB [26]
inbuilt function is used
//Display
Display ISP
```

The pseudo code for IP/MW computation is listed as:

```
// Input Data is Protein _Sequence
// Calculate IP/MW feature Present in Protein _Sequence
          Call Calculate- Isoelectric-Point (Protein
_Sequence)
          Call Calculate-Molecular-Weight(Protein
_Sequence)
IPMW = ISP / MW;
//Display
Display IP/MW (IPMW)
```

The pseudo code to calculate AI is listed as: [18]

```
MW:          Molecular Weight
AI:          Aliphatic Index
// Input Data is Protein _Sequence &TAA
//Read
Read the coefficient a=2.9, b=3.9 volume of the amino acid
(V & L) side-chains to the side-chains of A
// Calculate AI feature Present in Protein _Sequence
      Read various 20 diminutive units called amino acids

Then
      calculate the molecular percentage of A, I, L, V of
the
      diminutive units
Calculate AI
//Display
        Display AI
```

The pseudo code to calculate gravy is listed as [19]

```
SM:          Sum of AA
// Input Data is Protein _Sequence & TAA
//Read Gravy variables GV
Read the gravy information from each 20 diminutive units
//calculate Gravy feature Present in Protein _Sequence
    Repeat for i=1 to TAA
      ch= Protein _Sequence(i)
        Switch ch
          case match for each 20 amino acid and read
          an individual gravy information corresponding to
that
                    Otherwise
                    Unexpected Input
              end
                    SM=SM+GV(i);
          end
    Then
                    Gravy=SM/ TAA;
//Display
        Display Gravy feature Present in Protein
_Sequence
```

The pseudo code to calculate INI is listed as [20]

```
SM:          Sum of AA
INI:         Instability-Index
// Input Data is Protein _Sequence & TAA
//Read Instability Matrix or instability to each 400 different
dipeptides  (DIWV) as
Read instability matrix
//Calculate the INI  feature present in the protein
_sequence
          Temp='WCMHYFQNIRDPTKEVSGAL';
          Repeat for i=1 to TAA
            ch1= Protein _Sequence(i);
            ch2= Protein _Sequence(i+1);
                  Repeat for j=1 to 20
                        If( temp(j)==ch1)
                              Then m=j;
                        end
                        If( temp(j)==ch2)
                              Then n=j;
                        end
                  end
             SM=SM+ instability-matrix(m,n);
          end
INI=(10/TAA)*SM;
//Display
          Display INI
```

The pseudo code to calculate Volume is listed as: [21]

```
// Input Data is Protein _Sequence
//Read variable U=1.21
//Calculate the feature volume present in the protein _seq.
          Call Calculate-Molecular-Weight(Protein
_Sequence)
```

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

```
        Volume=U*MW;
//Display
        Display Volume
```

The pseudo code for Atoms and C, H, N, O, S is listed as:

```
// Input Data is Protein _Sequence
// Calculate CHNOS feature Present in Protein _Sequence
CHNOS=atomiccomp(Protein _Sequence);        //
MATLAB[26] Inbuilt function is used
Atoms= C+H+N+O+S;
//Display
Display C,H,N,O,S & Atoms
```

2. LITERATURE SURVEY

A wide range of techniques are used in bioinformatics for the protein characterization like data integration, genomics, sequence as well as the structure based, protein-interactions and so forth, a survey demonstrates that the vector-space incorporation method of data integration technique put many efforts for prediction protein functionalities, it utilizes the feature extraction philosophy for the prediction of protein functionality as well as for its classification. The ML and the rule mining (RM) are the two areas which deliver promising results to vast and to not so clearly defined areas of research, this encourages to use the power of ML and RM methodology along-with to explore such a vast domain. The literature shows how these techniques can be more beneficial when they assimilate with each other and provide promising results.

In 2002 **L. Jensen et al.** revealed the functions predictions of human protein through PTM and LF which are protein modifications & the local features of protein linear sequence for a particular class. The PTM is the modifications that happen to the protein at lateral stages. Different servers were used like Expasy [9], PSORT [10] to extract the features [22]. In 2003 **C.Z. Cai, et al.** did the classification of the protein plants through protein linear sequence features by the SVM method for a particular protein class and achieve the accuracy of 71% for classification. The important features they consider for classification like polarity, surface tension etc. [23]. In 2006 I. **Friedberg,** communicated his very valuable views on the quality of protein function prediction, according to his views there is remarkable growth in the sequence as well as the structure related data which leads unequal growth in un-portrayed genes products. The contextual, as well as in the subjective way the protein function is cumbersome in nature, and technique like protein homology transformation for protein-annotation are strengthening existing erroneous-annotation [24]. In 2007 **A.Al-Shahib et al.** revealed prediction on the protein using SVM-ML approach and computed linear sequence features. They also computed the percentage of known and unknown proteins, average GC contents. Authors subdivided dataset in to 4 identical size and performed experiments repeatedly for all datasets. They use PROFEAT for secondary-structure prediction, helices through TMHMM, and variant regions by DisEMBL [25]. In 2007 **Lobley et al.** anticipated the protein work with IDR in human protein through length and position conditional dependencies. The protein functions were predicted using ML approach with the help of protein sequence based features like weight, charge and their interdependencies etc. [26]. In 2007 **Kanakubo et al.** described the importance of association rule mining in data mining. The performance of Apriori methods degraded when the dataset was large. The author proposed an AI data mining technique which considered a solitary operation as a single association & interaction of randomly selected individuals was considered as association rules [14]. In 2010 **M. Nofal et al.** proposed survey and empirical comparative evaluation for classification based on an association rule mining techniques. They expressed the importance of prediction rate & runtime factors in association rule and classification tasks in data mining. They pointed Naïve Bayes & oneR is the fastest method over the RIPPER method to do classification [27]. In 2011 **M. Singh, et al.** portrayed the importance of clustering and unsupervised learning for the prediction of the protein classes. They extracted the data from the HPRD [15] and among that database; they extracted the sequence based features through web tools [36]. Similar protein classification from Sequence based features is applied in current research and multiple ML approaches were used to enhance the accuracy further. In 2011 **Jaiswal et al.** revealed the process to identify the protein for diseased conditions with the help of various bio-computational tools. They considered the physicochemical functional analysis and the secondary structure of human MMP family. They evaluated the importance of cysteine for diseased conditions & formulated the significance of the sequence extracted features for the physicochemical process [28]. In 2011 **M. Singh** proposed the server to extract the properties using associative rule mining [35]. Enhanced version of this server was build (in terms of number of sequences, various properties/features and protein classes) in this research effort. In 2014 **S. Saha et al.** revealed the prediction of the protein functionalities from the network-based protein interconnectivities with the help of

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

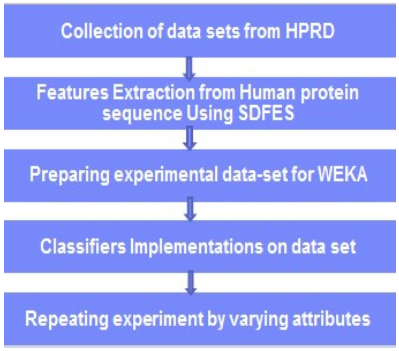
physicochemical properties and enlighten their importance [29] so more feature identification was done. In 2014 **A. Bhola et al.** revealed the ML-based approaches contributions for the prediction of the protein functionalities with the help of sequence-based features by the random forest-based method with reliefF [30] instead of relief parameter various combinations of features was tried. In 2015 **Ofer et al.** predicted the protein localization structure and the distinctive functional properties through features based extraction techniques [31] stress upon a specific class prediction. In 2015 **R. Singh et al.** proposed the improved protein function classification using SVM a machine learning non-linear classification using the sequence extracted properties through Protparam at Expasy, They used the Gaussian kernel non-linear process for the protein classification & express their importance towards sequence extracted features [32] other ML approaches were also used in our research and Random forest approach proves to be better. In 2016 **S. Das et al.** revealed their prediction by depicting gene ontology properties from the sequence grabbed from their database through two methods BLAST and HMMER3 [33] consolidated the use of features and different sequences for prediction task.

This research explores ML configuration settings on previous study conducted by M. Singh et al. and improved the accuracy for individual class prediction from 70 to 97% without much improvement on overall prediction accuracy. In 2016 A. Shehu et al. explored all the computational methods for protein function prediction and highlighted that ML is best suited for the prediction task. It has some problem areas like biasness of existing erroneous data propagating to further investigation. This limitation of ML was explored and verified in this research and the remedy for fixing this issue, in the form of ML configuration was identified and implemented.

3. RESEARCH METHODOLOGY

The methodology used for this work starting from the data collection from HPRD [15] to results and discussion is easy to understand. The proposed methodology obtained the results which are justified with the help of 10-fold cross-validation. The approach applied does not have any applicability issues and it is data-centric rather than algorithmic. The methodology phases are elucidated as:

1. The Human protein data collection from the HPRD [15] database which is readily available and can be downloaded very easily.
2. Extraction of Features after processing from the human protein sequence with the help of SDFES (sequence derived feature extraction server) which are helpful in protein classification. The features extracted by the server are already mentioned.
3. The analysis is done with the help of WEKA [23] and the features were extracted with the help of SDFES. The WEKA [34] is feed with the input in the excel format or in .CSV format. The SDFES is capable to derive the output in excel format, which is easily conceived by WEKA



4. The WEKA is equipped with power of GUI and ML approaches and these ML approaches or classifiers are used in this study for protein classification like Bayes Network, Rule Mining, Meta, Decision Tree, Functions and Lazy with the help of their significant algorithms like Bayes-Net, IBK, J48, Bagging, Logistic Approach, Random Forest and PART.
5. The experiment is done by applying above mentioned classifiers with changing classes, features as well as sequences configurations in order to achieve better results in terms of true positive classification and accuracy of individual classes as well as combined classification accuracy.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

4. RESULTS AND DISCUSSIONS

In this study, the human protein sequences is supplied as input to the SDFES, which predicts or extract all of the amino acids with their very important features of predictable properties. These features have their vital contributions in the protein classification and prediction of protein functionalities using ML approaches. The results are obtained after processing of input on the SDFES server (coded in MATLAB) [12]. The input is provided of the human protein amino acid sequence in capital letter format & the results are displayed as follow:

Protein-Sequence:- e.g.
MPSLLVLTFSPCVLLGWALLAGGTGGGGVG
GGGGGAGIGGGRQEREALPPQKIEVLVLLPQ
DDSYLFSLTRVRPAIEYALRSVEGNGTGRRLL
PPGTRFQVAYEDSDCGNRALFSLVDRVAAAR
GAKPDLILGPVCEYAAAPVARLASHWDLPML
SAGALAAGFQHKDSEYSHLTRVAPAYAKMG
EMMLALFRHHHWSRAALVYSDDKLERNCYF
TLEGVHEVFQEEGLHTSIYSFDETKDLDLEDI
VRNIQASERVVIMCASSDTIRSIMLVAHRHGM
TSGDYAFFNIELFNSSSYGDGSWKRGDKHDF
EAKQAYSSLQTVTLLRTVKPEFEKFSMEVKSS
VEKQGLNMEDYVNMFVEGFHDAILLYVLAL
HEVLRAGYSKKDGGKIIQQTWNRTFEGIAGQ
VSIDANGDRYGDFSVIAMTDVEAGTQEVIGD
YFGKEGRFEMRPNVKYPWGPLKLRIDENRIV
EHTNSSPCKSCGLEESAVTGIVVGALLGAGLL
MAFYFFRKKYRITIERRTQQEESNLGKHRELR
EDSIRSHFSVA

Total Number of Amino Acids in Sequence=540

Residues	Number	Individual Mole Percentage
A = Ala	45	8.333333
B = Asx	0	0.000000
C = Cys	7	1.296296
D = Asp	29	5.370370
E = Glu	40	7.407407
F = Phe	25	4.629630
G = Gly	54	10.00000
H = His	16	2.962963
I = Ile	24	4.444444
K = Lys	23	4.259259
L = Leu	53	9.814815
M = Met	14	2.592593
N = Asn	14	2.592593
P = Pro	18	3.333333
Q = Gln	16	2.962963
R = Arg	36	6.666667
S = Ser	40	7.407407
T = Thr	22	4.074074
V = Val	38	7.037037
W = Trp	6	1.111111
X = Xaa	0	0.000000
Y = Tyr	20	3.703704
Z = Glx	0	0.000000

Sequence extracted Features		Presence	Mole Percentage
Polar		236	43.7037
Non-Polar		304	56.2963
Tiny		168	31.11111
Small		267	49.44444
Aliphatic		115	21.2963
Acidic		69	12.77778
Aromatic		67	12.40741
Basic		75	13.88889
Charged		144	26.66667
Extinction coefficient		63675	-
Negatively Charged Residues		69	-
Positively Charged Residues		59	-
Molecular weight		59735.18	-
Absorbance		1.065955	-
Isoelectric point		6.312881	-
The computation of IP/MW		0.000106	-
Aliphatic Index		84.35185	-
GRAVY		-0.21204	-
Instability Index		40.91944	-
Volume Approximate		72279.57	-
Total ATOMS		8349	-
	C	2657	-
	H	4141	-
	N	739	-
	O	791	-
	S	21	-

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

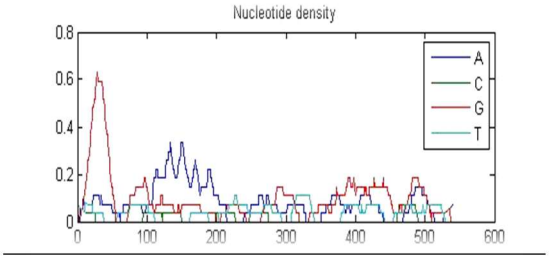


Figure 1: *Nucleotide Density*

The results obtained by SDFES from the sequence are mentioned above. The Figure 1 shows the nucleotide density obtained from sequence

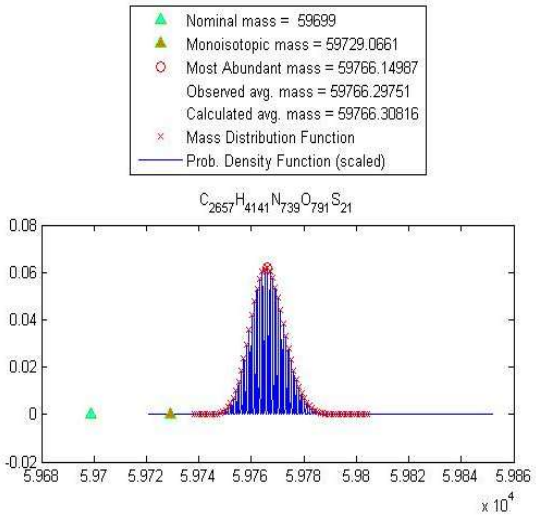


Figure 2: *Nominal & Mono-isotopic Mass*

The Figure 2 depicts nominal & mono-isotopic mass grabbed from the protein sequence along with the C, H, N, O, S and presence of atoms. Further, these features or impactful properties extracted by SDFES from the protein sequences which are taken from the database of HPRD are examined under the aegis of the WEKA controlled panel. The various classification ML approaches are used for protein classification like Bayes Network, Rule Mining, Meta, Decision Tree, Functions and Lazy with the help of their established algorithms like Bayes-Net, IBK, J48, Bagging,
Logistic Approach, Random Forest and PART, among all of these very significant algorithms the random forest surpassed all the other approaches with its individual protein class classification accuracy of 97.1429% and an overall the protein class classification of 70.7% shown in Figure 3.

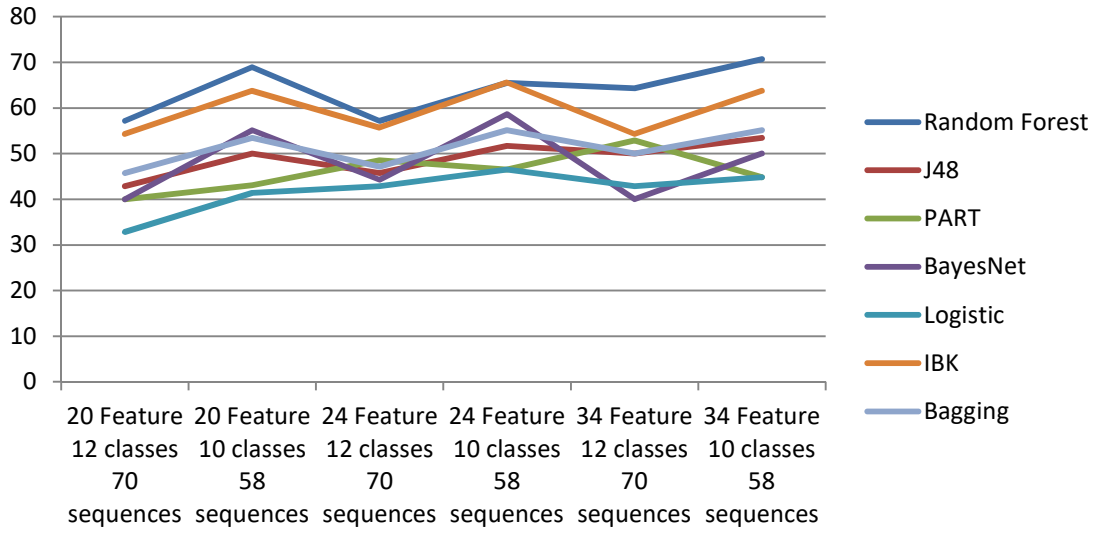


Figure 3: *Accuracy Comparison of ML approaches with the feature and the classes variations*

These experiments were conducted on 10 and 12 classes of protein with varying the features strength from 20 to 24 and then from 24 to 34 on the 58 & 70 protein sequences respectively. The resulting protein accuracies with its cost/benefit analysis are shown in Figure 4 and Figure 5.

The accuracy of classes can be examined from the cost/benefit matrix analysis.

Journal of Theoretical and Applied Information Technology
30th April 2018. Vol.96. No 8

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

The Figure 4 shows how the 'Defensin' class achieved the gain of 8.97 at the cost of 2.0 on 12 protein classes to grab the true positive rate of classification as 97.1429%, with the help of random forest approach. The 'Defensin' class achieved significant results for the human protein by getting the recall, precision as well as the true positive rate of 0.833.

The 10fold cross-validation is performed for protein classification or protein class prediction on random-forest with the instance ratio of 70:30. The resultant obtained the mean squared error of 0.21%, root-mean-squared error of 0.10% and the false positive rate of 0.019% with case complexity of -28.746% bits/instance.

Figure 3 shows how the random forest ML approach outshines the other ML approaches. The varying attribute configuration depicts that enhancement in the features reflects enhancement in overall accuracies but with dip in certain individual cases.

Table-1 shows the accuracies achieved by different classifiers with 20, 24, 34 features with 12 classes and 70 sequences. It depicts how accuracies are increasing in the case of random forest algorithm and J48. The Figure 6 shows graphical representation of the accuracies achieved by classifiers on 12 classes.

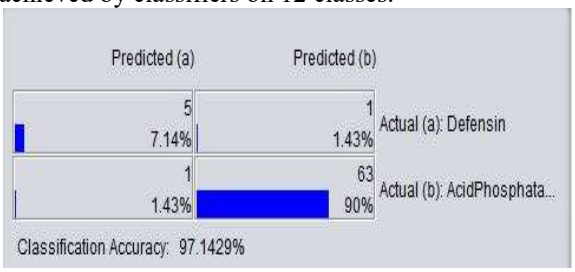


Figure 4: *Protein classification accuracy for 'Defensin' class*

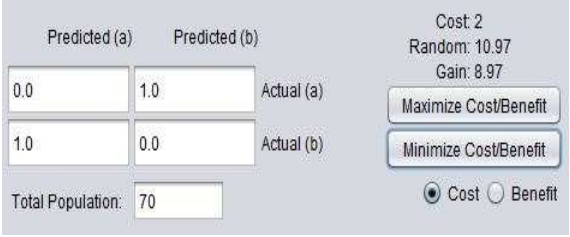


Figure 5: *Cost/benefit Analysis for the 'Defensin' class*

Table 1: *Accuracy comparison of ML approaches with 20, 24, 34 feature and 12 classes variations*

Classifiers	20 Feature 12 classes 70 sequences	24 Feature 12 classes 70 sequences	34 Feature 12 classes 70 sequences
Random Forest	57.14	57.14	64.29
J48	42.85	45.71	50
PART	40	48.57	52.86
BayesNet	40	44.29	40
Logistic	32.85	42.86	42.86
IBK	54.28	55.71	54.29
Bagging	45.71	47.14	50

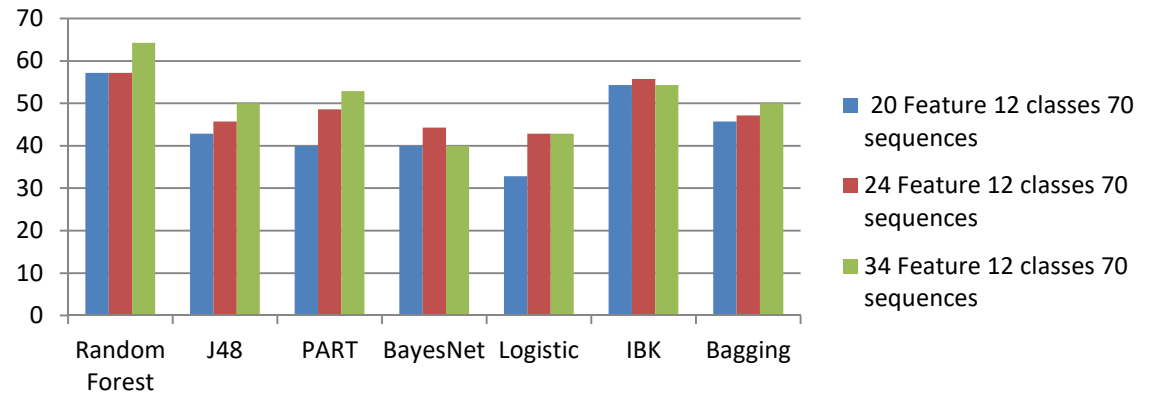


Figure 6: *Accuracy comparisons of ML approaches with 20, 24, 34 feature and 12 classes variations*

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 2 shows the accuracies achieved by different classifiers with 20, 24, 34 features with 10 classes and 58 sequences. It also depicts how accuracies are increasing in the case of random forest algorithm and J48.

The Figure 7 shows its graphical representation of the accuracies achieved by classifiers on 10 classes.

The vibrant observation of this analysis can also be depicted from the configuration changes confusion matrix. It shows how the change in configurations increases the accuracies as well as decreases the particular class classification sometimes.

It also depicts that for the resultant class prediction, the particular feature series can be elected, which gives true positive results for particular class prediction.
The following Table 3 shows confusion matrix for classes with an increase in the features, which is generated by SDFES while processing on random forest method. The Table 3 depicts that to predict or to classify the classes defensin, decarboxylase and waterchannel the configuration of 34 features, 12 classes and 70 sequence should be used because using this setting the classification accuracy is increased but for BCellAntigenReceptor prediction this configuration should not be used otherwise the result will be biased.

Table 2: *Accuracy Comparison of ML approaches with 20, 24, 34 feature and 10 classes variations*

Classifiers	20 Feature 10 classes 58 sequences	24 Feature 10 classes 58 sequences	34 Feature 10 classes 58 sequences
Random Forest	68.96	65.52	70.69
J48	50	51.72	53.44
PART	43.1	46.55	44.83
BayesNet	55.17	58.62	50
Logistic	41.38	46.55	44.83
IBK	63.79	65.55	63.79
Bagging	53.45	55.17	55.17

This work explore ML approaches and find out pitfalls in ML approaches that increase in overall accuracy give better prediction results. Novelty of the approach is that it can be used to improve results of prediction just by trying various settings in the existing system only so it highlighted that ML has a flexibility which can be harnessed for many research areas. Impact of this research is very far reaching as this is not only improving results of this specific domain of protein class prediction but it is equally applicable on other versatile research areas.

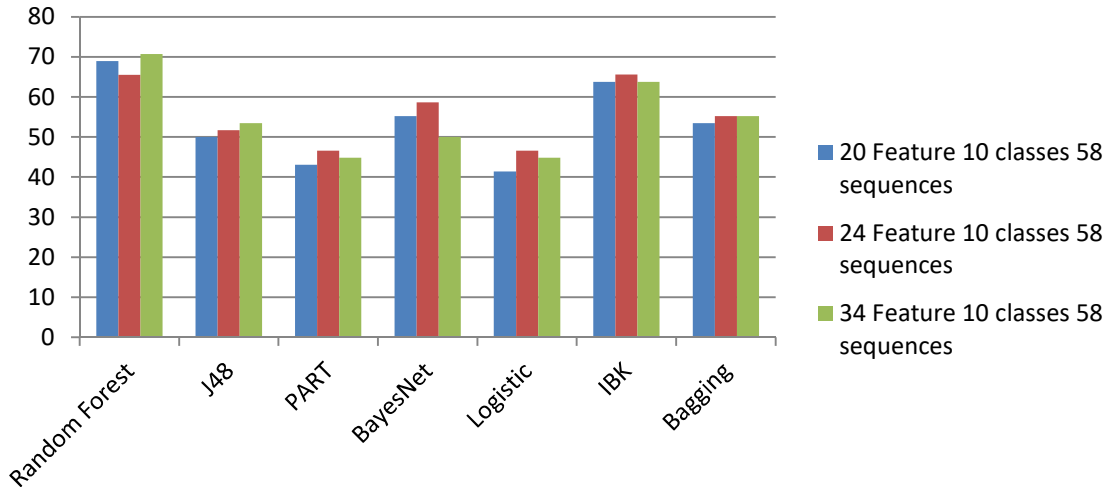


Figure 7: *Accuracy Comparison of ML approaches with 20, 24, 34 feature and 10 classes variations*

Table 3: *Confusion Matrix*

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

CLASSES	24 features, 12 classes and 70 sequences	24 features 10 classes and 58 sequences	34 features, 12 classes and 70 sequences	34 features 10 classes and 58 sequences
	a b c d e f g h i j k l	a b c d e f g h i j k l	a b c d e f g h i j k l	a b c d e f g h i j k l
Defensin (a)	5 x 1 x x x x x x x x x | a	5 x 1 x x x x x x x x x | a	6 x x x x x x x x x x x | a	5 x x x x x 1 x x x x x | a
AcidPhosphatase (b)	x 3 x x x x x 1 2 x x x | b	x 3 x x x x x 1 2 x x x | b	x 4 x x 1 x x x 1 x x x | b	x 4 x x x x x 1 1 x x x | b
VoltageGatedChannel (c)	x x 3 x x x 2 x 1 x x x | c	x x 3 x x x 2 x 1 x x x | c	x x 3 x x x 2 x 1 x x x | c	x x 3 x x x 2 x 1 x x x | c
DNARepairProtein (d)	x x x 1 1 1 x 1 x 2 x x | d	x x x x x x x x x x x x | d	x x x 1 1 2 x 1 x 1 x x | d	x x x x x x x x x x x x | d
Decarboxylase (e)	x 1 x x 1 1 1 2 x x x x | e	x x x x x x x x x x x x | e	x x x x 6 x x x x x x x | e	x x x x x x x x x x x x | e
HeatShockProtein (f)	x x x x x 4 x 1 x x x x | f	x x x x x 4 x 1 x x x x | f	x x x x x 4 x 1 x x x x | f	x x x x x 4 x 1 x x x x | f
Aminopeptidase (g)	x x 1 x x x 4 x x x x 1 | g	x x 1 x x x 4 x x x x 1 | g	x x 1 x x x 5 x x x x x | g	x x 1 x x x 5 x x x x x | g
G-Protein (h)	x x x x 2 x x 3 x 1 x x | h	x x x x x x x 4 x 2 x x | h	x x x x 1 1 x 3 x 1 x x | h	x x x x x x x 4 x 2 x x | h
WaterChannel (i)	x 1 x x x x x x 5 x x x | i	x 1 x x x x x x 5 x x x | i	x x x x x x x x 6 x x x | i	x 1 x x x x x x 5 x x x | i
NucleotidylTransferase (j)	x 1 x 1 1 x x 1 x 2 x x | j	x x 1 x x x x 2 x 3 x x | j	x 1 x 1 1 x x 1 x 2 x x | j	x 1 x x x x x 1 x 4 x x | j
BCellAntigenReceptor (k)	x x x x x x x x x x 5 x | k	x x x x x x x x x x 5 x | k	x x x x x x x x x x 4 1 | k	x x x x x x x x x x 5 x | k
CellSurfaceReceptor (l)	1 x x x x x x x x 1 x 4 | l	1 x x x x x x 1 x 1 1 2 | l	1 x x 1 x x x x x 1 1 2 | l	1 x x x x x x 1 x 1 1 2 | l

5. CONCLUSION

The implementation effort of developing a SDFES proved successful and worthwhile as the overall time and effort to derive sequence-based features from human protein sequence was considerably reduced hence it will facilitate the overall prediction process in fields like disease detection and drug discovery. The developed Extraction server was tested and validated for accuracy by comparing its results with online tools and was found accurate.

Second goal of finding efficient ML algorithms revealed some limitation of ML application and various ML configurations were tried to overcome this problem. The results witnessed that critical analysis through the random forest ML approach with individual protein-class classification provides an accuracy of 97% shown in figure-4 and 70.69% combined classification accuracy as depicted in table-2. Research results clearly show that for selective prediction interest as in this case was finding a specific protein-class; different ML configurations can be used rather than a single optimal setting. The weaknesses governed by confusion matrix were suggested for particular classes of prediction. The different configuration settings of the features as well as sequences should be used for the prediction of the particular class accurately, while its overall accuracy may be less. These configurations should be used at the early stages of ML training for the resultant implementation else the results will be biased and could feed error for future investigations in domains like disease detection, drug discovery, crop hybridization etc. The proposed methodology is fast, reliable, efficient and is applicable to other research domains with similar requirements.

Future work can incorporate inclusion of more features and automation of configuration selection process best suited for a specific scenario. Use of this approach can be incorporated in discovery of new frameworks for disease detection as well as for personalized medication.

REFERENCES

[1] A. Shehu, D. Barbara and K. Molloy, "A Survey of Computational Methods for Protein Function Prediction," *Springer,* no. Big Data Analytics in Genomics, october 2016, pp. 225-298.

[2] W.-F. HUO, N. Gao, Y. Yan, J.-Y. LI, J.-H. YU and R.-R. XU, "Decision Trees Com-bined with Feature Selection for the Rational Synthesis of Aluminophos-phate AlPO4-5," *National Natural Science Foundation of China,* vol. 7, no. 9, September 2011, pp. 2111-2117.

[3] I. o. See5/C5.0, "Information on See5/C5.0," [Online]. Available: http://rulequest.com/see5-info.html. [Accessed 05 January 2016].

[4] 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, october 2005, pp. 387-393.

[5] B. Bergeron, Bioinformatics computing, Delhi, Upper Saddle River: Prentice Hall PTR, 2002.

[6] TMHMM-Server, "TMHMM Server, v. 2.0," [Online]. Available: http://www.cbs.dtu.dk/services/TMHMM/. [Access 05 December 2017].

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[7] SignalP-4.1-Server, "SignalP-4.1-Server," [Online]. Available: http://www.cbs.dtu.dk/services/SignalP/. [Access 05 June 2016].

[8] NetNGlyc-Server, "NetNGlyc-1.0-Server," [Online]. Available: http://www.cbs.dtu.dk/services/NetNGlyc/. [Access 12 September 2017].

[9] ExPASy, "ExPASy:SIB Bioinformatics Resource Portal - Home," [Online]. Available: http://expasy.org/. [Access 15 November 2018].

[10] PSORT-WWW-Server, "PSORT WWW Server," [Online]. Available: http://psort.hgc.jp/. [Access 10 November 2017].

[11] PROFEAT-Server, "PROFEAT 2015 HOME," [Online]. Available: http://bidd2.nus.edu.sg/cgi-bin/prof2015/prof_home.cgi. [Accessed 16 December 2017].

[12] MATLAB-Mathworks, "MATLAB," [Online]. Available: https://www.mathworks.com/products/matlab.html. [Accessed 02 January 2016].

[13] S. Sharma, A. Singh and R. Singh, "Enhancing Usability of See5 (Incorporating C5 Algorithm) for Prediction of HPF from SDF," *International Journal of Computing and Technology,* vol. 3, no. 4, 2016.

[14] M. Kanakubo and M. Hagiwara, "Speed up technique for Associative Rule Mining based on an Artificial Algorithm," *GRC book on granular computing,* vol. 38, no. 12, pp. 319-322, 2007.

[15] H. P. R. Database-HPRD, "Human Protein Reference Database," [Online]. Available: http://www.hprd.org/. [Accessed 20 January 2014].

[16] S. C.Gill and P. H. Hippel, "Calculation of protein extinction coefficients from amino acid sequence data," *Analytical Biochemistry,* vol. 182, no. 2:283, 1 November 1989, pp. 319-326.

[17] C. N. Pace, F. Vajdos, L. Fee, G. Grimsley and T. Gray, "How to measure and predict the molar absorption coefficient of a protein.," *Protein Sci.,* vol. 4, no. 11, November 1995, pp. 2411-2423.

[18] A. IKAI, "Thermostability and aliphatic index of globular proteins," *The Journal of Biochemistry,* vol. 88, no. 6, 1 October 1980, pp. 1895-1898.

[19] J. Kyte and R. F. Doolittle, "A simple method for displaying the hydropathic character of a protein," *J. Mol. Bio,* vol. 157, no. 1, 5 May 1982, pp. 105-132,.

[20] K. Guruprasad, B. Reddy and M. W.Pandit, "Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence," *Protein Engineering.,* vol. 4, December 1990, pp. 155-161.

[21] Y. Harpaz, M. Gerstein and C. Chothia, "Volume changes on protein folding," *Structure,* vol. 2, no. 7, July 1994, pp. 641-649.

[22] J. L., "Prediction of Protein Function from Sequence Derived Protein Features," L.Jensen, Denmark, 2002.

[23] C. CZ, L. Han, Z.L.Ji, X.Chen and Y. Chen, "SVM-Prot: Web-based support vector machine software for functional classification of a protein from its primary sequence," *Nucleic Acids Research,* vol. 31, no. 13, July 2003.

[24] I. Friedberg, "Automated protein function prediction--the genomic challenge," *Briefings in Bioinformatics,* vol. 7, no. 3, 1, September 2006, pp. 225-242,.

[25] A. Al-Shahib, R. Breitling and D. R. Gilbert, "Predicting protein function by machine learning on amino acid sequences-A critical evaluation," *BMC Genomics,* vol. 8, no. 78, March 2007, pp. 1-10.

[26] A. Lobley, M. B. Swindells, C. A. Orengo and D. T. Jones, "Inferring function using patterns of native disorder in proteins," *PLoS Computational Biology,* vol. 3, no. 8, 24, Auggust 2007, pp. 030162.

[27] ". A. D. M. Nofal and S. Bani-Ahmad, "Classification based on association rule mining techniques: A general survey and empirical comparative evaluation," *Ubiquitous Computing and Communication Journal,* vol. 5, no. 3, 2010.

[28] A. Jaiswal, A. Chhabra, U. Malhotra, S. Kohli and V. Rani, "Comparative analysis of human matrix metalloproteinases: Emerging therapeutic targets in diseases," *Bioinformation,* vol. 6, no. 1, 2, March 2011, pp. 23-30.

[29] S. Saha and P. Chatterjee, "Protein function prediction from protein interaction network using physico-chemical properties of amino acids," *International Journal of pharmacy and Biological Sciences India,* vol. 4, no. 2, June 2014, pp. 55-65.

[30] A. Bhola, S. K. Yadav and A. k. Tiwari, "Machine Learning based Approach for Protein function prediction using sequence derived properties," *A. Bhola, S.K. Yadav, A.K.Tiwari 2014. Machine LearnInternational journal of computer applications,* vol. 105, no. 12, November 2014, pp. 17-21.

[31] D. Ofer and M. Linial, "ProFET: Feature engineering captures high-level protein functions," *Bioinformatics,* vol. 31, no. 21, 1, November 2015, pp. 3429-3436.

[32] R. Singh, R. Singh and D. P. Kaur, "Improved protein function classification using support vector machine," *International journal of computer science and information technologies,* vol. 6, no. 2, 2015, pp. 964-968.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[33] SayoniDas and C. A.Orengo, "Protein function annotation using protein domain family resources," *Methods,* vol. 93, January 2016, pp. 24-34, 15.

[34] WEKA-Machine, "Weka Machine Learning," [Online]. Available: https://en.wikipedia.org/wiki/Weka_machine_learning. [Accessed 05 January 2016].

[35] M. Singh and G. Singh, "Development of Predictor for Sequence Derived Features from Amino Acid Sequence using Associative Rule Mining," *International journal of computer science and security,* vol. 5, no. 1, 2011.

[36] M. Singh and G. Singh, "Cluster Analysis Technique based on Bipartite Graph for Human Protein Class Prediction," *International Journal of Computer Applications,* vol. 20, no. 3, April, 2011, pp. 22-27.

[37] W. M. Learning, "Weka Machine Learning," [Online]. Available: https://en.wikipedia.org/wiki/Weka_machine_learning. [Accessed 05 January 2016].