International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

Wrapper Fuzzy Approach with 3d Fast Convolution Neural Network (FCNN) Based Feature Selection in Protein Sequence Classification

***T. Sudha Rani[1], Dr. A. Yesu Babu[2], Dr. D. Haritha[3]**

Submitted: 17/08/2022 **Accepted:** 22/11/2022

***Abstract*:** In research area, an emerging field is Bioinformatics in the past decades. Biological data storage and management was the definite motivation of bioinformatics and the tools for computation are developed and analyzed for enhancing their understanding. The data size is gathered under different project sequence is exponentially increased, that provides the problems for the methods of experiment. Newly sequenced protein and known functions proteins have gap and this gap is reduced by several techniques of computation incorporating classification and algorithms of clustering were presented in the past. The sequences of protein are classified into superfamilies exists in literature is useful for the prediction of structure and function of huge proteins that are discovered newly. The existing classification's results are unacceptable because of larger feature size acquired by several approaches of feature encoding. This paper proposes noise removal technique depending on selection of feature for protein sequence classification. Here we use wrapper fuzzy model with fast convolution neural network (FCNN) for feature selection and remove the noise. This research involved in removal of noisy or unwanted data related to protein composite. To improve classification accuracy, wrapper fuzzy is utilized for selection of features. Wrapper algorithm involved in selection of protein features for accurate identification of protein composites. For classification we use 3D FCNN which can improve the accuracy of classification. The classification of protein proposed in this method proves momentous enhancement with respect to measuring the metrics of performance: accuracy, sensitivity, specificity, recall, F-measure, and etc.

Keywords: *Bioinformatics, protein sequences, classification, feature selection, noise removal, wrapper fuzzy, Classification using 3D-Fast Convolution neural network (3D- FCNN)*

1. Introduction

Protein 3D- structure was predicted by the sequence of amino acid was the major aim of bioinformatics for the past decades [1] and there was no definite solution. Highly consistent method recently incorporates modelling of homology, and permits for assigning the structure of protein which is defined already to given unknown protein which are similar among two detectable sequence. De novo approaches are required for viable homology modelling depending on either physical-based potential [2] or knowledge-based potentials. The function of energy is utilized for the approximation of free energy amount in the provided protein conformation present also with function of search that efforts various conformation of structure for minimising some function of energy [3]. Inappropriately, huge molecule structures are defined as protein with numerous conformations although for comparatively small protein making it unaffordable for folding them even on modified computer hardware.

[1]*Assoc. Professor and Research Scholar, Department of CSE, Aditya Engineering College and JNTUK-Kakinada, sudha.mahi84@gmail.com*
[2]*Professor, Department of CSE, Sir C R Reddy College of Engineering, Eluru*
[3]*Professor, Department of CSE, University of College of Engineering, JNTUK-Kakinada*
** Corresponding Author Email: sudha.mahi84@gmail.com*

Binded amino acid sequence is referred as protein having peptide bond playing important character in life maintenance [4].

Organ functioning and human tissues were improved by protein structures. Primary, secondary and tertiary are the three basic protein structure. Protein structure assists in the determination of protein's functional behaviour and predicting functions. Similarity of sequence are found for clustering the protein's similar kinds for finding protein interaction among protein and several other literatures. Prediction of protein function are the basis for biological research and this prediction are performed by sequence or similarity structure. Also this prediction type picks huge resources and time of computation. For improving the accuracy of computation with resource reduction and time of computation and using the approach of classifying machine learning [5].

In bioinformatics, biological sequences are the classification of important task. Family identification are uninterruptedly interested by biologists. Protein evolution is possibly studied and its biological functions are discovered. In general, for classification of new biological sequences into classes or families that are known already through the similarity search and sequences homologies, some alignments were used by the biologists. Though, this method is frequently inefficient e.g: Metagenomics, which is a main issue in run into the application of these method which is in between 25% and 65% of sequences without homologous in databases, and this is a

reason for unusable sequences [6] and these problems are rectified only by the Machine learning approaches. Characterisation of these sequences format are encoded by its own, and it is not probable for using familiar algorithms of classification which are effective in actual task of data mining having determined data in the format of relational data. Therefore, it is essential to introduce a pre-processing step for parsing the biological data into to new suitable format for various tools of data mining. Phase in pre-processing is performed by motif extraction which indicates the solution that is use widely for sequences of protein and every residue of amino acid present in the protein structure is indicated by a 20 alphabet size character. Motifs discovery from the sequences are gentle task motivated towards substring findings or words known as descriptors. Feature space yields descriptors allowing the transformation of sequences into value vectors, which facilitates the data processing through tools of data mining. This phase of pre-processing is the main step for consistent knowledge discovery process as it is affected directly by result quality obtained [7].

Organization of the paper is given below: Related works are presented in section 2. Proposed work is defined in section 3. Results of experiments and discussions are described in section 4. Conclusion and future scope are presented in section 5.

2. Related Works

Some of classification approaches for the protein classification are as follows:

Feature selection algorithm stability are stated in [8] according to data with random perturbation. The subsets of selected feature's probability distributions generated properties are used for algorithms of feature selection stability. Algorithms for selecting features generate probability distribution are away from the uniform and nearer to peak value. Instance sub-sampling are performed in [9] for the simulating the perturbation of data. Each n sub-samples perform feature selection. Feature selection output f are put into every sub-sample is contrasted with other subsample's outputs through the coefficient of Pearson correlation, coefficient of Spearman rank correlation similarity by Jaccard index are measured for noting S. Measure of higher stability are mostly similar to each and every outputs are presented in [10]. Various sub-samples (or training sets) are produced by the similar generating distribution. Various training sets are quantified by stability that disturbs the output of feature selection. Three representation types for feature subsets are accounted by the authors. Each feature is assigned by score or weight in first type which is indicated by its significance. Features are assigned with ranks by the representation in second type. Feature set without weight or rank are taken in third type. Stability measurement needs measurement of similarity for representing features. This clearly based on the representation utilized through a provided algorithm of feature selection for describing its feature subset. Three measures of similarity are utilized by authors: they are correlation coefficient of Pearson, correlation coefficient for spearman and distance of Tanimoto. In many applications like biology, approaches of machine learning are used and agonizes from the dimensionality curse for huge feature space which are data available become sparse and resulted in degradation of performance. Thus, information wealth is required to be filtered out to acquire feature's final set that are appropriate for the issue called as reduction in dimensionality. Feature's original set have subsets are present in selection of features is a special case for reducing dimensionality. Simply, every possible feature subset is tested for subset selection in which error are minimized as per ground truth. Brute-force technique are feasible computation for small feature sets. Feature Selection Algorithms (FSAs) was presented in [11] that are involved in search problem characterization in the space of hypothesis (ie, candidate feature subset's space) based on three conditions: strategy of searching, used for exploring the hypothesis space; successor candidate generation and measures evaluation, used for the evaluating the function through successor candidates and comparison of various hypotheses are compared for guiding the process of searching. Bioinformatics with feature selection having particular applications of these techniques in the analysis of sequence, microarray and mass spectra are reviewed as well as the algorithms of feature selection are categorized for big data bioinformatics into exhaustive, heuristic and hybrid search methods [12].

All the features available are fitted initially in this predictive model, and the removal of the weakest feature happens till the reaching the minimum number which is predetermined. [13,14] includes RFE as an example of this work. At the same time, every feature of individuals is evaluated from that one selected from the beginning of forward feature selection resulted in the model with best performance. All selected feature combination are probable and evaluation of the subsequent features are performed with respect to the second feature selection and this process is repeated iteratively with maximum number. In [15] selection of forward features are used for performing feature ranking in WEKA tool88 utilizing SVM as an evaluator. Subsets of candidate feature are evaluated by the methods of filtering through the proxy measure rather than feature selection algorithms which obtains error rate. Due to this method cost effectiveness, the measures are selected and feature set usefulness are selected.

3. System Model

The 1672 proteins of 25% identified sequence are comprised in 25PDB; 640 proteins of about 25% identified sequence are comprised in 640 PDB; 513 proteins of below 25% identified sequence are comprised in CB513, these are the datasets utilized for this paper. PSSM encodes all the datasets of protein stated above through PSI-BLAST. For the model learning prediction, protein information evolution is introduced by PSSM. The homologous secondary structure of protein having the known structure of protein was predicted on the theoretical basis [13]. NCBI's NR database (ftp://ftp.ncbi.nih.gov /blast/db/nr) obtains homogenous sequences that are aligned and searched by using every protein sequences through iterative databank searching tool BLAST (PSI-BLAST) (http://blast.ncbi.nlm.. nih.gov/Blast.cgi) having 3 iterations and Evalue cut off is 0.001. BLOSUM62 Substitution Matrix isused in this experiment which is a score matrix for measuring the reflection similarity in the sequence of amino acids. At last, PSSM profile obtained for the sequence of protein is in the form of L×20 matrix, where protein instance length is represented by L. The problem of predicting the secondary structure is formulated usually in the sliding window concept. The sequence of amino acids within the window is used for the prediction of center position's secondary structure. The sequence of protein will have head and tail part, which are inferred by the reflection in the edge approaches and because of this location of amino acid at head and tail, protein sequence is

forecasted by same size of window. The L amino acid having entire protein sequence is divided into L L nonoverlapping intervals of n base pairs. 5,13 and 35 are size of window for the prediction of secondary structure. Dictionary of Secondary Structure of Proteins (DSSP) is a tertiary structure which assigns the secondary structure of protein [14]. The secondary structures contents are defined by utilizing protein's DSSP file with eight classes: H (α-helix), G (310-helix), I (л-helix), E (β-strand), B (β-bridge), T (turn), S (bend) and C (rest random coil). The α-helix and β-strand denotes H and E and C is denoted by all the other elements are incorporated in coil. The proposed model Architecture is represented below in figure 1:

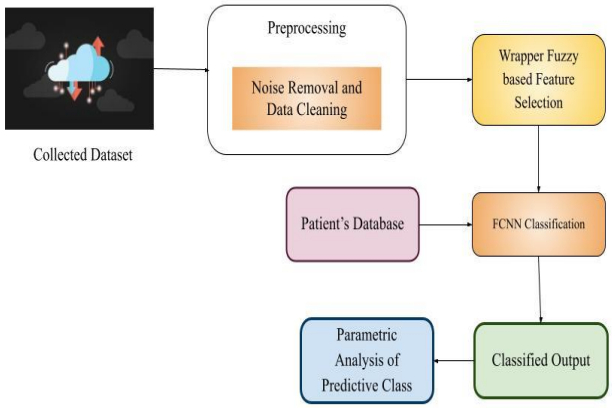

Figure-1 Proposed Architecture

Next to the technique of pre-processing, vector of 8420 features are used to represent every sequence of protein. Several features for vectors of features contains empty or zero values. Many irrelevant or features of redundancy are present in this vector without protein sequence information is provided [15]. The performance and sequence classification algorithm's running time are affected greatly by the redundant features. The proposed approach for the selection of feature subset, every feature's statistical significance of superfamily from each and every other super families are calculated. The sequence representation is not contributed by features and are eliminated from the unique feature space which is used for the lowering the dimension of feature vectors [16]. The features with various subsets are extracted from original feature space by feature selection approach proposed and subset of best feature was selected and this will produce the results of maximum accuracy [17]. The best subset and relevant features are utilized for the discrimination among various classes of protein or super families [18]. The feature's original representation is not altered by this technique and only the best feature subset was selected from it. The technique of feature selection in the classification process surely obtains 5% to 10% of accuracy increment and lower complexity of computation. The selection of feature subset is used successfully by supervised and unsupervised learning algorithms. The system training time is highly reduced by this technique and the over fitting chances are also lowered [19]. Feature selection follows the steps described below and these feature selection process has the tendency to separate various super families. Mathematical steps for this technique are expressed as:

The *i*th sequence's superfamily is denoted by X^i. *k*th superfamily's sequence X_k is represented by X_k^i. k = 1, 2, . . . , Ni, and the sequence number is represented by Ni in the superfamily i. The sequence K's feature vector are represented by $X_k^i(j), j$ = 1, 2, . . . , 8420 for superfamily i. Mean vector computed for every superfamily is expressed below eq. (1):

$$\bar{X}^i(j) = \frac{\sum_{k=1}^{N_t} x_k^i(j)}{N_i}, j = 1,2,\dots,8420 \quad (1)$$

The every superfamily's variance is computed below in eq (2):

$$S_i^2(j) = \frac{\sum_{k=1}^{N_i} \left(\vec{X}^i(j) - x_k^i(j)\right)^2}{N_i - 1} \quad (2)$$

For super families having every pair (say p and q) a distance vector is computed by utilizing the definition of metrics stated as follows:

$$vd_{p,q}(j) = \frac{\left|\bar{X}^p(j) - \bar{X}^q(j)\right|}{\sqrt{\left(S_p^2(j)/N_{\text{Total}}\right) + \left(S_q^2(j)/N_{\text{Total}}\right)}} \quad (3)$$

For every column, choosing least of the 3 distances as the final metric, as in (7). As it is motivated towards searching the capable best features capable which is discerning the various super families, the final metric's highest values represented in columns are chosen. Let,

$$vd(j) = \text{Min}_{p \neq q}\{vd_{p,q}(j)\} \quad (4)$$

3.1 Wrapper-fuzzy model (WrFuz):

The bases of the fuzzy rules with WM approach are generated sequentially with the combination of dataset features by using wrapper is the main idea in this method. Complexity of exponent is avoided by the best-first heuristic method. The following steps are used for the demonstration of Wrapper-fuzzy method: 1) All FRBs are generated by WM approach for a domain having m features and each and every probable (m-1) features are combined. 2) FRB features are removed by the best i.e, from the dataset, low rate of error are removed and each and every are generated by WM approach and each and every probable (m-2) features are combined. 3) The above step is repeated till only one remaining feature 4) FRB's rank are generated by their rate of errors 5) FRB features are selected with (lowest) best error rate. The fuzzy rules are generated with maximum number in the 1st step of WM methods is restricted to the presented number of examples (n). Therefore, this method's complexity which is described by the feature m is demonstrated as O(n × m2). All-important fuzzy logic characteristics are considered in the base WM method for wrapper approach used in the FRB's generation rather than using general parameters or characteristics

The WrFuzfeature selection proposed are proceeded along with steps followed:

3.2 Initialization:

1.1 Provided the feature set $A = \{z_1, \dots, z_j, \dots \dots, z_n\}$, the feature FPVs is computed $G(z_j), j = 1, \dots \dots, n$

1.2 RS(0)=A, FS(0)= $\emptyset$ and CS=(0)= $\emptyset$ are Set.

1.3 The feature z_{l_1} is found, so $z_{l_1} = arg \max_{l=1,\dots n} \{|G(z_l)|\}$ and the first feature is selected

1.4 Set CS(1)=G(z_{l_1}), FS(1)=FS(0)+{ z_{l_1}} and RS(1)=RS(0)-{ z_{l_1}}

2. Sequential Forward Selection loop For iterations $\rho = 2,3, \dots$ perform the following:

2.1 FuzCoC feature Selection

2.2 In a filter manner, remaining features are evaluated, i.e feature's additional contribution are computed in RS(p-1) (loop for feature evaluation)

2.3 $AC(p,j) = G(z_j)| - |CS(p-1)j = 1, \dots n, z_j \in RS(p-1)$

2.4 Find $z_{l_p} \epsilon RS(p-1)$ so $l_{p=arg \max_{z_l \epsilon RS(p-1)} |AC(p,j)|} = \max_{z_l \epsilon RS(p-1)} |CS(p-1)U|G(z_j)$

2.5 The z_{l_p}'s percentage improvement is computed as per CS(p-1):

$$h_{l_p} = \frac{|AC(p,l_p)|}{CS(p-1)} \times 100\%$$

2.6 If $h_{l_p} > e_z$ THEN

4.3.1 Set FS(p)=FS(p-1)+ { z_{l_p}}, RS(p)=RS(p-1)- { z_{l_p}}

4.3.2 Compute the FPV G(FS(p))by applying FO-SVM on the selected feature space FS(p)

4.3.3 Set CS(p)=G(FS(p))

4.3.4 Increment $p \leftarrow p+1$ and go to step 4.1

ELSE Terminate FuzCoC procedure

Output: The set FS(m)= { z_{l_1},...., z_{l_m}}of m finally selected features.

4. Classification using 3D- Fast Convolution Neural Network (3D- FCNN)

Linear classifier is 3D FCNN. For label, weighted softmax cross entropy loss is used and (label weight is equal to the label frequency inverse in the training set) and adam optimizer and activation of ReLU are used by CNNs and every convolutional layer is dropout in the process of training. The network is contracted for the prevention parameter overload for small comparative Kaggle dataset. The 3D protein stack evolution is provided in figure 2 represented below.

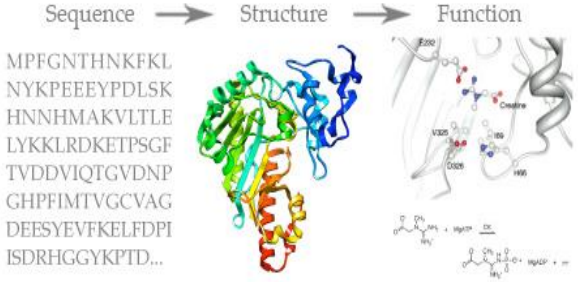

Figure-2 3D protein stack evolution

Frame size 60 * 40 centered on the present frame which are seven in number are considered as an input to 3D CNN method is used in this architecture. Hardwired kernels set for the generation of multiple information channels obtained from the input frame are applied first. This 33 feature map result in the 2nd layer in 5 various channels are given as gray, gradient-x, gradient-y, optflow-x, and optflow-y. The value of gray pixel of seven input frames are present in the gray channel. CNN comprises few number of convolutional layers, that follows one or more fully connected layers and at last an output layer. Protein sequence classification with 3D architecture is provided in figure-3.

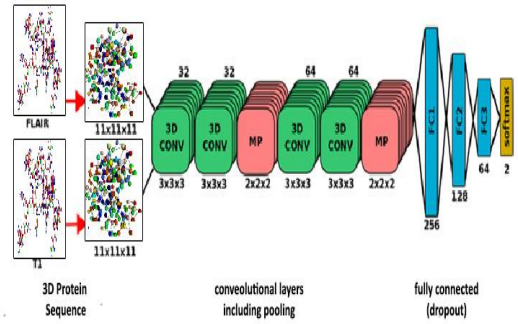

Figure-3 3D architecture for protein sequence

Network's layer m of input is denoted by $I^{(m)}$ formerly. The conventional layer with 3D input of NN is $n_1^{(m-1)} \times n_2^{(m-1)} \times n_3^{(m-1)}$ three dimensional object having $n_c^{(m-1)}$ and also $I^{(m-1)} \epsilon \mathbb{R}^{n1(m-1) \times n_2^{(m-1) \times n_3^{(m-1)}}}$ and the elements ate given by I (m,`). The 3D volume index are i, j and k index are used for channel selection. The convolutional layer m's output is represented by its dimension $n_1^{(m)} \times n_2^{(m)} \times n_3^{(m)}$ and also the filter numbers or produced channel is $n_c^{(m)}$. The layer m's output is convoluted input having filter and it is calculated as $I_{i,j,k}^{(m,l)} = f_{tanh}\left(b^{(m,l)}\right) + \sum_{i,j,k,l} I_{i,j,k}^{(m-1,l)} W_{i-i,j-j,k-k,l}^{(m-l)}$ in this equation, $b^{(m,l)}$ and $W^{(m,l)}$ are the defined parameters are defined by mth layer of lth filter. The location for the evaluation of filters (i.e., I,j, and k values for $I_{i,j,k}^{(m,l)}$ are calculated) and the filter size ($W^{(m,l)}$) non-zero values are the architectural parameters of network. At last, function of hyperbolic tangent activation having $f_{\tanh(a)} = \tanh(a)$. Input's spatial structure are preserved by the convolutional layers and by utilizing several layers several complex input representations are constructed. The convolutional layer's output is utilized as an input for the layer of fully connected network. For this, channel and spatial structure is neglected and the convolutional layer output is considered as only one vector. The fully connected output is single dimension vector $I^{(m)}$ and dimension is the network architecture parameter. The ith neuron output in mth layer is proceeded by

$$I_i^{(m)} = f_{ReLu}\left(b^{(m,i)} + \sum_j I_j^{(m-1)} W_j^{(m,i)}\right) \quad (5)$$

In the above equation, mth layer of ith neuron parameters are $b^{(m,i)}$ and $W^{(m,i)}$ and the jth sum over is the sum over all input dimension. The $f_{ReLu}(.)$ is the activation function is selected here as Rectified Linear Unit (ReLU) having $f_{ReLu}(a) = \max(0,a)$. Function of activation was used widely in domain number and is supposed to be specifically useful in task of classification because it induces sparsity in the result which is used for creating the gap between classes in the process of learning. Fully connected layer in the last is utilized as an input to the layer of output. The structure and the layer output form is based on specific task. Two various kinds of function of output are considered. K classes are present in the problem of classification, function of softmax is general function of output:

$$f_i = \frac{\exp(I_i^{(o)})}{\sum_j \exp(I_j^{(o)})} \quad (6)$$

$$I_i^{(o)} = b^{(o,i)} + \sum_{k=1}^{k} W_k^{(o,i)} I_k^{(N)} \quad (7)$$

In the above equation, last fully connected layer index is given by N, Output unit of i has the parameters $b^{(o,i)}$ and $W^{(o,i)}$ and ith class output is $f_i \epsilon [0,1]$ this is taken as the probability of that class with input provided. Logistic output function with variation are considered as:

$$f = a + (b-a)(1 + \exp(b^{(o)} + \sum_j W_j^{(o)} I_j^{(N)})^{-1} \quad (8)$$

From the above equation, f is continuous output limited lies in the range (a, b) having $b^{(o)}$ and $W^{(o)}$ parameters. The scaled logistic output function is used in this function. The problem of multi-class classification with ranking type such as malignancy level prediction is considered in this output function that is probable for better performance.

Table-1 3D FCNN Architecture (Dropout with 0.2, Adam Optimizer, Learning Rate = 0.0001)

Layer	Params	Activation	Output
Input			28 × 28 × 28
Conv1	5 × 5× 5	ReLu	28 × 28 × 28× 7
Max Pool	1× 1× 1, stride 2× 2× 4		14 × 14 × 7 × 7
Conv2	5 × 5× 3	ReLu	14 × 14 × 7 × 17
Max Pool	2× 2 × 2, stride 1× 1 × 0		6 × 6 × 3 × 17
Dense			256
Dense			2

4.1 Training:

Provided with data collection and an architecture of network, the major motivation is fitting the network parameters to the data. For this function of objective was defined and gradient based optimization is used for searching the parameters of network, and are used for the minimization of objective function. Consider $D = n_i, y_i \; i = 1$ is the D set (augmented potentially) are examples for training, in this equation input is presented by n and output is given by y. All weight's W collection is denoted by Θ and b as biases for each and every network layer. The objective function is represented as

$$E(\theta) = \sum_{i=1}^{D} L\big(y_i, f(n_i, \Theta)\big) + \lambda E_{prior}(\Theta) \quad (9)$$

In the above equation, network output is f(ni , Θ)) is the estimated on n input having parameters Θ, loss function is L(yi , f(ni , Θ)) in which differences among the network's desired output y and the network prediction yˆ are penalized. The weight decay prior is $E_{prior}(\Theta) = ||W||^2$ function used in the prevention of over-fitting by penalizing the weights norm and prior strength is controlled by λ.

Two various objective functions are considered in this paper based on the output function's choice. For the function of softmax, loss function for standard cross-entropy $L(y_i, \hat{y}) = -\sum_{k=1}^{k} y_k \log(\hat{y}k)$ is used in this equation, binary indicator vector is assumed by y and probabilities vector for every K classes is given by yˆ. The cross entropy loss limitation is that each and every errors of class are equally considered, so malignancy level 1 is mislabelled as level 2 and are taken the worst mislabel of 5. For this problem, in the scaled logistic function, loss function of squared error is used for capturing this. $L(y_i, \hat{y}) = (y_i, \hat{y})^2$ is used formally and real values of y and yˆ are assumed. Particularly in the tth iteration, updated parameters are given as $\Theta_{t+1} = \Theta_t + \Delta\Theta_{t+1}$, $\Delta\Theta_{t+1} = \rho\Delta\Theta_t - \varepsilon\nabla E_t(\Theta_t)$

From the above equation, Momentum parameter is given by ρ = 0.9, vector of momentum is denoted by $\Delta\Theta_{t+1}$, the rate of learning is t and objective function's gradient is given by $\nabla E_t(\Theta_t)$ which is estimated by utilizing only the examples of training chosen at tth iteration. In the 0th iteration, 0 is the value set for all biases and the filters values and weights are reset by homogeneously sampling from the interval $\left(-\sqrt[n]{\frac{6}{f_{an_{in}} + f_{an_{out}}}}, \sqrt[n]{\frac{6}{f_{an_{in}} + f_{an_{out}}}}\right)$ where fan_in denote the number of nodes in the previous hidden layer and the number of nodes in the current layer is denoted by fan_out. Providing the setting and initialization values ϵ_t = 0.01 for 2000 epochs, in the process of 10% decrement in ϵ_t for each 25 epochs for the convergence ensurance.

4.2 Performance Analysis:

MATLAB 2014a processing on Intel® Core™ i7-4790 3.60 GHz CPU with 32 GB RAM is used for the implementation of this paper. CB513, 640 and 25PDB are matrix generated inside the scope of search with datasets nr are taken as actual input data protein sequence and also secondary structure of protein with H [1, 0, 0], E [0, 1, 0] and C [0, 0, 1] as 3 states. Sliding window concept is utilized to pill up the samples of these protein sequence and size of window are 5, 13 and 35 utilized in this experiment. Predicting the secondary structure of protein

is evaluated usually by Q3, in which residue in percentage are measured through this secondary structure in third state is predicted correctly. The softmax classifier at the top layer is evaluated by using Q3. Fivefold cross-validation method is used in simulation in order to consider the consumption of time and convenience, Performance evaluation of softmax classifier is evaluated by this fivefold cross-validation method. From each results of evaluation, Q3 mean value is above five times applied in running below fivefold partition, four out of five proteins are selected randomly as data for training and left over proteins are taken as samples for testing.

5. Feature Selection based on Convolution and Pooling

CB513 and 25PDB produces 35-polypeptides using the approach of sliding window and are utilized for the model of prediction. In LSFE, 35-polypeptide's representation is constructed depending on filters of self-taught feature which are unsupervised and 3D FCNN. The size of pool for 3D FCNN is six. Information about position are reflected by new representation for two local structure types. Evolutionary information are not only reflected by this new representation and also it reflects residue's interaction of sequence. The dimension for new representation are high, because of the operation of pooling and sparse auto encoder's constraints sparsity. Very sparse new representation neglects the problem of over fitting prediction. So, LSFE becomes influential feature extraction technique for prediction of secondary structure. Accuracy, precision, recall and F1 score are the parameters use in the performance analysis and are defined below.

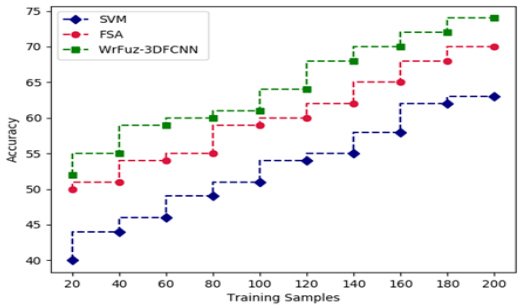

Figure- 4: Accuracy Comparison

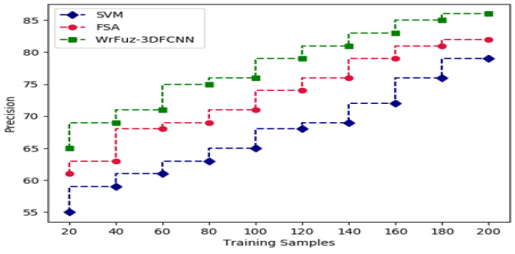

Figure- 5: Precision Comparison

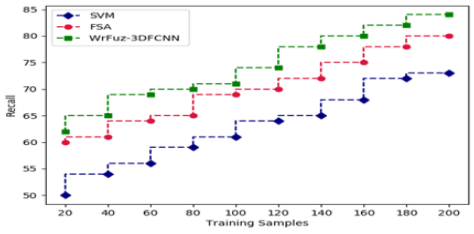

Figure-6: Recall Comparison

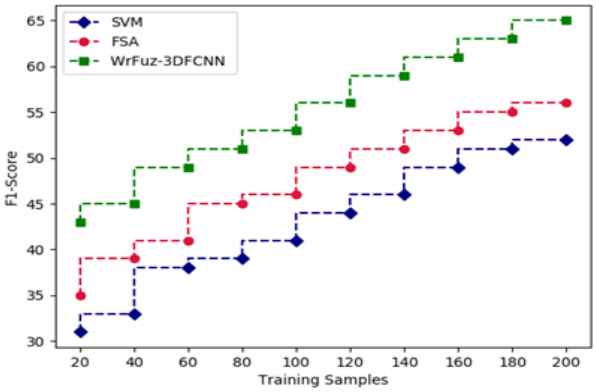

Figure-7: F1- score Comparison

The figure 4,5,6, and 7 presented above represents the accuracy, precision, recall, f1- score comparison between existing and proposed technique in classifying the protein sequence.

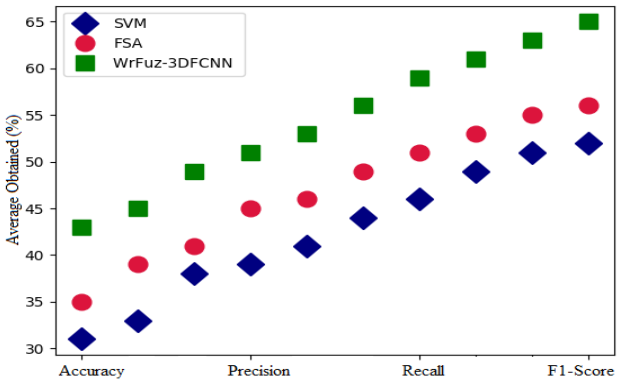

Figure-8 Overall Comparison for protein sequence feature selection and classification

The above figure-8 shows the overall parametric comparison of proposed technique in comparison with existing technique for protein sequence feature selection and classification

6. Conclusion

Protein is considered as essential factor which describes the status of bones and health related concern. In existing, several techniques are designed for protein detection and classification based on machine learning. However, those techniques subjected certain limitation due to noisy data. This research aimed to develop appropriate machine learning scheme for improving classification and prediction rate of protein composite in human. For achieving higher accuracy optimization and regression model is designed for classification and prediction of protein. The proposed approach for the selection of feature subset utilizes a threshold for the selection of the extremely informative and significant features. This technique results were authenticated by the classification and learning algorithms which are well-recognized. The sequence of protein with three various datasets are classified efficiently into applicable super families with considerably increased accuracy of classification. The method of classification introduced is free in alignment, easy, quick and consistent. Bioinformatics and machine learning uses this feature selection technique for the data dimension reduction in the process of structure prediction for sequences of protein which are unknown. This proposed work is extended in future in the pattern recognition areas such as various types of proteomic

classification and genetic diseases.

Reference

[1] Xing, Zhengzheng, Jian Pei, and Eamonn Keogh. "A brief survey on sequence classification." ACM Sigkdd Explorations Newsletter 12.1 (2010): 40-48.

[2] Cai, Jie, et al. "Feature selection in machine learning: A new perspective." Neurocomputing 300 (2018): 70-79.

[3] Wang, Lipo, Yaoli Wang, and Qing Chang. "Feature selection methods for big data bioinformatics: A survey from the search perspective." Methods 111 (2016): 21-31.

[4] Yang, Wuritu, et al. "A brief survey of machine learning methods in protein sub-Golgi localization." Current Bioinformatics 14.3 (2019): 234-240.

[5] Saeys, Yvan, InakiInza, and Pedro Larranaga. "A review of feature selection techniques in bioinformatics." bioinformatics 23.19 (2007): 2507-2517.

[6] Nemati, Shahla, et al. "A novel ACO–GA hybrid algorithm for feature selection in protein function prediction." Expert systems with applications 36.10 (2009): 12086-12094.

[7] Ma, Shuangge, and Jian Huang. "Penalized feature selection and classification in bioinformatics." Briefings in bioinformatics 9.5 (2008): 392-403.

[8] Qin, Xinyi, et al. "Structural protein fold recognition based on secondary structure and evolutionary information using machine learning algorithms." Computational Biology and Chemistry 91 (2021): 107456.

[9] Zhao, Xudong, et al. "ECFS-DEA: an ensemble classifier-based feature selection for differential expression analysis on expression profiles." BMC bioinformatics 21.1 (2020): 43.

[10] Guannoni, Naoual, FaouziMhamdi, and MouradElloumi. "Improved Feature Selection Algorithm for Biological Sequences Classification." International Conference on Knowledge Science, Engineering and Management. Springer, Cham, 2019.

[11] Sang, Xiuzhi, et al. "HMMPred: accurate Prediction of DNA-binding proteins based on HMM Profiles and XGBoost feature selection." Computational and mathematical methods in medicine 2020 (2020).

[12] Sarkar, Jnanendra Prasad, et al. "Machine learning integrated ensemble of feature selection methods followed by survival analysis for predicting breast cancer subtype specific miRNA biomarkers." Computers in Biology and Medicine 131 (2021): 104244.

[13] Cinelli, Mattia, et al. "Feature selection using a one dimensional naïve Bayes' classifier increases the accuracy of support vector machine classification of CDR3 repertoires." Bioinformatics 33.7 (2017): 951-955.

[14] Blanco, Jose Liñares, et al. "Prediction of high anti-angiogenic activity peptides in silico using a generalized linear model and feature selection." Scientific reports 8.1 (2018): 1-11.

[15] Rangasamy, Ranjani Rani, and RamyachitraDuraisamy. "Ensemble of Artificial Bee Colony Optimization and Random Forest Technique for Feature Selection and Classification of Protein Function Family Prediction." Soft Computing in Data Analytics. Springer, Singapore, 2019. 165-173.

[16] Sequeira, Ana Marta, Diana Lousa, and Miguel Rocha. "ProPythia: A Python Automated Platform for the Classification of Proteins Using Machine Learning." International Conference on Practical Applications of Computational Biology & Bioinformatics. Springer, Cham, 2020.

[17] Saikumar, K., Rajesh, V., Babu, B.S. (2022). Heart disease detection based on feature fusion technique with augmented classification using deep learning technology. Traitement du Signal, Vol. 39, No. 1, pp. 31-42. https://doi.org/10.18280/ts.390104.

[18] Kailasam, S., Achanta, S.D.M., Rama Koteswara Rao, P., Vatambeti, R., Kayam, S. (2022). An IoT-based agriculture maintenance using pervasive computing with machine learning technique. International Journal of Intelligent Computing and Cybernetics, 15(2), pp. 184–197.

[19] Rao, K. S., Reddy, B. V., Sarada, K., & Saikumar, K. (2021). A Sequential Data Mining Technique for Identification of Fault Zone Using FACTS-Based Transmission. In Handbook of Research on Innovations and Applications of AI, IoT, and Cognitive Technologies (pp. 408-419). IGI Global.

Authors Profile

Mrs. T. Sudha Rani, working as Assoc. Professor, Department of Computer Science and Engineering, Aditya Engineering College. Currently Pursuing Ph. D at JNTUK, Kakinada. Her Area of Research are Bioinformatics and Data Mining.
E-mail: sudha.mahi84@gmail.com

Dr. A. Yesu Babu, Currently working as a Professor, Department of Computer Science and Engineering, Sir C R Reddy College of Engineering, Eluru. He is having 31 years of Academic, Research & Academic Administration experience. Published 43 Research Papers in International journals and 6 chapters. Reviewer of Research publications for premier publishing groups like Springer, Elsevier, Inderscience and a number of SCOPUS and SCI indexed journals.

Dr. D. Haritha, She is working as Associate Professor in Computer science and Engineering Department at Jawaharlal Nehru Technological University Kakinada. She has 17+ years of experience. She guided 50 M.Tech students and 15 MCA students for their project. Her research interest is on Image Processing, Data Structures, Software Engineering and Networking. She published 12 research papers in international journals. She published 11 research papers in international conferences