ISSN:0975-1459
Journal of Pharmaceutical Sciences and Research
www.jpsr.pharmainfo.in

ADAMMLP: An Adaptive Moment Based Hybrid Multi-Layer Perceptron for Protein Secondary Structure Prediction

Bishnupriya Panda [a*], Babita Majhi [b*]

[a]*Department of Computer Science and Engineering, Institute of Technical Education and Research, Siksha 'O' Anusandhan University, Bhubaneswar, Orissa, India*

[b]*Department of Computer Science and Information Technology, Guru Ghashidas Vishwavidyalaya (A Central University), Bilaspur, Chhattishgarh, India*

Abstract

Protein is the crucial bio- molecule in living cells, involved in all the life processes and, controlling most of the biochemical reactions, signal transduction and transmission or the correct expression of genetic information. In the current study, we have developed a hybrid multi-layer perceptron network (ADAMMLP) using an adaptive moment based stochastic gradient descent optimization technique, which makes the model noise resistant, takes very less training time and suitable for both small and large sized datasets. We have evaluated the model using 4 benchmark datasets,204,277,498 and PDB25 with sequence similarity as low as 25%. The model shows high class wise prediction accuracies for all the datasets. This model shows highest prediction accuracy of 90.23%, 91.64%, 73.68% and 79.8% for 204,277,498 and PDB25 dataset respectively in presence of 20% noise. The maximum training times in presence of 20% noise are 3.05sec, 2.64sec, 3.54 sec and 10.47sec for 204,277,498 and PDB25 datasets respectively.

Keyword: Protein secondary structure prediction, Computational biology, Adaptive moment, optimization, Bioinformatics

1. Introduction

Protein secondary structure prediction (PSSP) is a critical task in protein science and computational biology, and can be applied to understand protein 3-dimensional (3-D) and quaternary structure further, to predict their biological functions. With exponential growth of protein sequences in Protein Data Bank (PDB), out of which only 0.2% with known structure and much lesser are with known functions. Existing experimental methods are extremely time consuming and hence efficient computational methods for protein structure prediction are needed to mitigate these problems. The pioneering work of (Levitt M *et al.*, 1976) distinguishes four structural classes of globular proteins: (1) all-α class that encompasses proteins with strands only, (2) all-β class that incorporates proteins with only small amount of helices, (3) α/β class with proteins that include both helices and mostly parallel strands and (4) $\alpha + \beta$ class, which includes proteins with both helices and antiparallel strands. Application of machine learning techniques in PSSP has been the primary focus in the last three decades. The principle of computational methods for PSSP is that the methods can learn some rules based on the analysis of known protein sequences and their secondary structures to predict the secondary structures of unknown protein sequences.

Many methods have been proposed for predicting the secondary structure of protein, such as methods by exploiting the physical and chemical properties of amino acids, methods based on sequence homology and statistical analysis, etc. (Chen *et al.*,2007; Ding S *et al.*, 2012). The problem of protein secondary structural class prediction from the primary sequence is mainly three fold: (i) to construct a reliable benchmark dataset for training and testing the predictor (ii) to design a hybrid feature extraction method to achieve structural information from amino acid sequences (iii) to lay out a amalgamated classification algorithm which could be trained to predict the secondary structures with boosted accuracy. Here an adaptive moment based feed forward neural network is proposed, which has fewer training time and works efficiently even with less training data.

The rest of the paper has been organized as follows: section 2 describes dataset used and feature extraction methods, section 3 provides description of a feed forward network based on first-order gradient-based optimization, based on adaptive estimates of lower-order moments. Section 4 presents results and discussion, and section 5 provides conclusion and limitations.

2. Data Set and Feature Extraction

2.1 Data Set

We have used 4 publicly available dataset. (1) First data set constructed by (Chou KC *et al.*, 1999) has 204 residues (2) Two other data set constructed by (Zhou GP *et al.*, 1998) have 277 and 498 residues respectively. The average sequence similarity scores in these protein classes are 21% for all α, 30% for all β, 15% for α/β and 14% for α+β class.(3) Dataset PDB25 (Kurgan LA *et al.*, 2006) has 25% sequence similarity.

2.2 Feature Extraction

2.2.1Preserving the basic amino acid composition information

The primary amino acid sequence is the skeleton of the protein molecule, which subsequently undergoes many post translational changes to stabilize. Amino acid composition (AAC) as described in (Sahu *et al.*, 2010) is the normalized count of amino acids in a protein chain. AAC is useful in the sense AAC has these benefits: (a) AAC is a simple, yet powerful feature which performs surprisingly well given its simplicity, (b) AAC is independent of the classification framework, and, (c) it is computationally tractable.

2.2.2 Preserving the sequence order information

It is indispensable to perpetuate the sequence order information as protein molecules having equivalent AAC may have different functionality. In this context Pseudo Amino Acid (*PseAA*) composition has been used widely in literature in varied bioinformatics application. Sequence order using *PseAA* along a protein chain can be formulated as in Eq.(1) (Zhou *et al*.,2007)

$$\theta_m = \frac{1}{L-m}\sum_{i-1}^{L-m}\theta(R_i, R_{i+m}), \quad m = 1,2,\ldots,\lambda \; and \; \lambda \prec L \qquad (1)$$

Where $\theta_m = m^{th}$ order correlation coefficient of amino acid sequence R of length L.

Correlation coefficient θ_m can be computed in different ways by taking different operators as computed in Eqn. (2).

$$\theta(R_i, R_j) = H(R_i) * H(R_j) \qquad (2)$$

Where $H(R_i)$ and $H(R_j)$ are either hydrophobicity (Tanford C *et al*, 1962) or hydrophilicity (Hopp TP *et al.*, 1981) indices of different amino acids. The hydrophobicity and hydrophilicity values were subjected to standard normalization procedure as given in Eq. (3) (Zhou *et al.*, 2007).

$$H(R_i) = \frac{H(R_i) - \sum_{k=1}^{20} H_0(R_k)/20}{\sqrt{\sum_{u=1}^{20}[H_0(R_u) - \sum_{k=1}^{20}(R_k)/20]^2 /20}} \qquad (3)$$

R_i : i^{th} Amino acid residue

H_0 : Original values of hydrophobicity or hydrophilicity.

2.2.3 Spectrotemporal analysis of protein

Amino acids in protein chain are not rigidly fixed rather in continuous internal motions. These internal movement of amino acids play vital role in formation of both α helix and β sheets. Low frequency components have higher biological significance (Chou KC *et al.*, 1999). Raw amino acid sequence can be thought of as a discrete signal. The significant low frequency components can be extracted by transforming the discrete amino acid sequence to frequency domain using discrete Fourier transform (DFT) (Liu *et al.*, 2005) using Eqn.(4).

$$X(k) = \sum_{n=1}^{L} H(P_n) e^{\left(\frac{-j2\pi nk}{L}\right)}, \; k=1,2\ldots\ldots L \qquad (4)$$

P_n :n^{th} Protein sequence.

(Sahu S *et al.*, 2010) suggested discrete cosine transformation (DCT) (Ahmed *et al.*, 1974) to extract spectral information, which overcomes the limitations of DFT. The advantages of DCT over DFT can be summarized as (1) it doesn't add any discontinuity and artifact to the original signal. (2) Computational complexity is lower. The discrete cosine transformation of the protein sequence P_n can be derived as in Eq. (5)

$$G(k) = a(k)\sum_{n=0}^{L-1} H(R_n)\, cos\left[\frac{(2n+1)k\pi}{2L}\right], k = 0,1,2,\ldots\ldots, L-1 \qquad (5)$$

$$a(k) = \begin{cases} \sqrt{\frac{2}{L}} & k \neq 0 \\ \sqrt{\frac{1}{L}} & k = 0 \end{cases}$$

L : Length of amino acid sequence

G_0 : DC component of the signal

Both DCT and DFT are incapable of dealing with noise, which creeps into biomedical signals and both the approach deal with only the frequency domain of the signal and neglecting time domain. Stockwell transform is a mechanism for analyzing signals in a small time window (Stockwell RG *et al.*, 1996). In this study, protein molecules are represented by replacing each amino acid by their corresponding electron ion interaction potential (EIIP) values as described previously (Veljkovic V *et. al.*, 1985). Stockwell-transform is applied to each EIIP encoded protein sequence, to compute four different features as discussed in (Mishra S *et al.*, 2008). The sequence representation can be summarized as a hybrid of amino acid composition, amphiphillic correlation factors, and Stockwell transform based features extracted from equivalent EIIP representation of raw amino acid sequence.

P=$[P_1 \ldots\ldots P_{20} P_{21} \ldots\ldots P_{30} P_{31} \ldots\ldots P_{40} F_1 F_2 F_3 F_4]$

Where,

$P_1 \ldots\ldots P_{20}$: 20 Features -amino acid composition.

$P_{21} \ldots\ldots P_{30}$: 10-Correlation coefficients based on hydrophobicity.

$P_{31} \ldots\ldots P_{40}$: 10-Correlation coefficients based on hydrophilicity.

$[F_1 F_2 F_3 F_4]$: Four features were extracted using S-transform on equivalent EIIP coded sequence.

3. Design of feed forward network based on first order gradient-based stochastic optimization

3.1. Feed-forward neural networks (Jain *et al.*, 1996) can be applied to a wide range of classification problems with a high degree of accuracy. Feed-forward neural networks are nonlinear, self- adaptive methods and make very few a priori assumptions about the models for problems under study with high generalization ability. One distinguished advantage of the feed-forward neural networks over other classes of nonlinear models is that they are universal approximators, which is due to their ability to process data in parallel and the network model is built upon the characteristic from the data. A network model is characterized by 3 layers, one input layer, and one output layer and can be many hidden layers. Inter layer nodes have acyclic network links but intra-layer nodes are not connected. The relationship between the output y_t and the inputs $(y_{t-1} \ldots\ldots y_{t-p})$ can be represented as in Eqn. (7).

$$y_t = w_0 + \sum_{j=1}^{Q} w_j sig\left(w_0 + \sum_{i=1}^{P} w_{i,j} y_{t-i}\right) + e_t \qquad (7)$$

Where w_0 is the initial weight vector, w_j are the network connecting weights, P is the no. of input nodes, Q is the no. of output nodes, sig is the sigmoidal or logistic transfer function which is often used in literature as activation function or transfer function as there is no established linear mapping between input and output signal.

$$sig(x) = \frac{1}{1+e^{-x}}$$

e_t : Error estimated in time step t for input signal x.

A defined feed forward network is used for training which is a process of parameter estimation. Different optimization procedures can be used for network parameters, which aims at overall minimization of the cost function of the network. Generally mean squared error is choosen as the cost function. The cost function E can be given as in the Eq. (9).

$$E = \frac{1}{N}\left(\sum_{i=1}^{N}(e_i)^2\right) = \frac{1}{N}\sum_{i=1}^{N}\left(y_t - \left(w_0 + \sum_{j=1}^{Q} w_j sig\left(w_{0j} + \sum_{i=1}^{P} w_{i,j} y_{t-i}\right)\right)\right)^2 \qquad (9)$$

Where N is the no. of epoches. The network free parameters are the connecting weights, which are minimized by some efficient non-linear optimization algorithm like back propagation, which is based on gradient-descent algorithm as in Eq. (10).

$$\Delta w_{i,j} = -\eta \frac{\partial E}{\partial w_{i,j}}$$

Where η is the learning rate parameter and $\frac{\partial E}{\partial w_{i,j}}$ is the direction of optimal change in the weight vector.The problem with the gradient descent approach is that, it can fall in local optima. Therefore another optimization method based on adaptive estimates of lower order moments was suggested by (Kingma D (8) *et al.*,2014).

3.2. Adaptive moment based stochastic optimization

Biomedical signals have inherent source of noise, which may be due to experimental error or due to human error. Gradient-descent optimization is a simple and efficient optimization approach,that works well if the objective function is differentiable, data sample is non-noisy. If the objective function is stochastic, data is high dimensional and noisy an alternative optimaztion approach is desirable,as reported by (Kingma D *et al.*,2014).This is based on first-order gradients. This can be used in feed forward neural network for weight vector optimization.

Modified Algorithm for feed forward network:

```
Function adapative-moment-based-BackPropagation(input data)
  Input: Data set with input vecor X and output vector Y
         Intial weight vector W and tanh activation function
        repeat
          for each p in dataset do
             for each node j in the input layer do $a_j \leftarrow x_j[p]$
             for $l = 2$ to $M$ do
               $in_i = \sum_j W_{j,i} a_j$
                $a_i = tanh(in_i)$
            for each node i in the output layer do
               $\Delta_i = tanh'(in_i) * (y_i[p] - a_i)$
              for $l = M - 1$ to 1 do
                for each node j in layer l do
                   $\Delta_j = tanh'(in_j) \sum_i W_{j,i} \Delta_i$
                     For each node i in layer $l + 1$ do
                         Update weight W := Adam $(\Delta_i)$
        Until stopping criteria is satisfied
        end
  Function Adam $(\Delta_i)$:
  Input: Requires $\beta_1$, $\beta_2$, step size $\alpha, \epsilon$ and current gradient value $\Delta_i$, $W_t$ is the weight at epoch t, Initially $\beta_1 = 0.9, \beta_2 = 0.999$ and $\epsilon = 10^{-8}$
    Compute:
        $m_t = \beta_1.m_{t-1} + (1-\beta_1).\Delta_i$
        $v_t = \beta_2.v_{t-1} + (1-\beta_2).(\Delta_i)^2$
        $\widehat{m}_t = m_t/(1-{\beta_1}^t)$
        $\hat{v}_t = v_t/(1-{\beta_2}^t)$
        $W_t = W_{t-1} - \alpha.\widehat{m}_t/(\sqrt{\hat{v}_t} + \epsilon)$
        return $W_t$
end
```

3.3. Network structure and performance measurement

We decided to choose a constant number (40) hidden nodes, learning rate (0.001) and tangent hyperbolic $(tanh)\ activationfunction$ for ADAMMLP for all data sets after several run of the algorithm.β_1 and β_2 were taken as 0.9 and 0.999 respectively and initial weight vector $W_o$was chosen randomly.

Prediction accuracy of the ADAMMLP was measured using F_1score, *Recall* and *Precision* score.

4. Result and Discussion

In Figure 1(A) for 204 dataset, the error rate per epoch falls consistently and converges after 500 epochs even with addition of noise and the ADAMMLP exhibits a stable behaviour. In 1(B) for 277 dataset, the model converges smoothly upto 15% noise addition but after 20% noise, though the network is not unstable, the fall in error rate is rather sharp. However in 1(C) and 1(D) the error rate per epoch donot fall smoothly.The model predicts nicely for α and β classes but its performance falls for classes α+β and α/β and the model requires 1000 epoches to converge.With higher epoches, the performance donot increase further. In 204 and 277 dataset the prediction accuracy of the classifier is unaffected as the relative amount of α+β class proteins in these datasets are significantly lower as compared to large dataset such as 498.For PDB25 dataset, even the error rate donot fall smoothly,the class wise prediction accuracy is significantly higher. In all the cases the prediction accuracy of α+β class has remained lower. The model requires less training time for all datasets as low as in the order of seconds. In Table [1], the class wise prediction results of the ADAMMLP has been listed for degree of noise addition.

In Figure 2(B-C), the prediction results for α and β classes for all data sets are significantly better but for α/β and α+β classes, only 204 and 277 data sets have better accuracy. Addition of noise have not affected the prediction accuracy of the model for small data sets like 277 and 204, despite only 500 epochs have been used for training of the network. For 498 dataset, the network performs well after 1000 epochs for α and β despite having constant learning rate and network structure. For PDB25 dataset, prediction accuracy is satisfactory for all the classes.

In fig 3(A), the overall prediction accuracy for all datasets for different noise level indicates that the noise addition doesnot affect significantly in the performance of the network except for the 498 dataset. In fig 3(B), for 204 and 277 data set the F1-score remain consistently above 0.75 and for PDB25, it remains above 0.55 except for 10% noise addition. Recall score remain above 0.8 for 204 and 277 dataset and above 0.6 for PDB25 data set.For 498 data set,it is around 0.5 but falls after 10% noise addition as shown in fig 3(C), while specificity score is consistently high above 0.8 for all tha data set as shown in fig3(D).

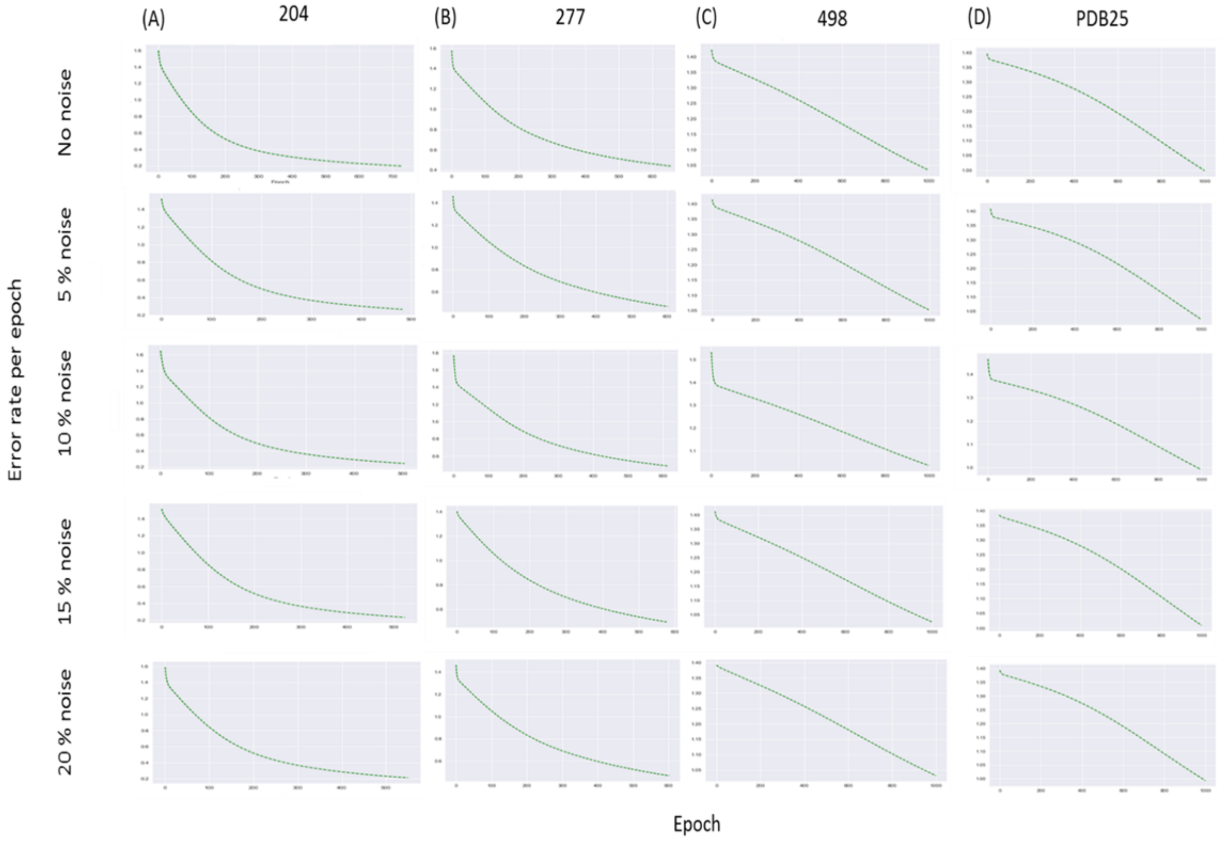

Figure 1. Impact of noise 5%,10%,15% and 20% on the prediction accuracy of the ADAMMLP model on datasets 207,277,498 and, PDB25.

Table 1: Prediction of ADAMMLP for differenent classes of all considered datasets

Data Set		**Class wise Prediction Accuracy-Without Noise**		**Time for Training in seconds**	
Class	**α**	**β**	**α+β**	**α/β**	
204	95.12	87.8	78.04	95.12	3.45
277	96.29	88.9	88.9	92.59	3.03
498	86.86	87.8	78.7	**55.6**	3.53
PDB25	83.28	78.80	79.40	86.8	12.96
Data Set		**Class wise Prediction Accuracy-5% Noise**			
Class	**α**	**β**	**α+β**	**α/β**	
204	95.12	92.68	82.92	95.12	3.07
277	94.5	88.8	88.89	94.4	3.35
498	86.8	86.8	80.8	**58.6**	3.22
PDB25	87.4	81.7	80	88.03	10.75
Data Set		**Class wise Prediction Accuracy-10% Noise**			
Class	**α**	**β**	**α+β**	**α/β**	
204	92.68	87.80	78.04	92.68	3.41
277	96.2	88.9	85.1	88.9	3.32
498	86.8	86.8	77.8	**53.5**	3.81
PDB25	79.1	71.34	76.11	86.26	13.96
Data Set		**Class wise Prediction Accuracy-15% Noise**			
Class	**α**	**β**	**α+β**	**α/β**	
204	97.56	90.24	82.92	95.12	3.32
277	96.2	90.7	87.03	88.9	2.62
498	86.8	86.8	75.7	**55.6**	3.82
PDB25	83.28	73.8	78.5	87.4	10.65
Data Set		**Class wise Prediction Accuracy20 % Noise**			
Class	**α**	**β**	**α+β**	**α/β**	
204	95.12	90.24	82.9	92.68	3.05
277	96.29	88.9	88.9	92.5	2.64
498	84.8	87.8	73.73	**48.4**	3.54
PDB25	80.8	73.7	77.6	87.1	10.47

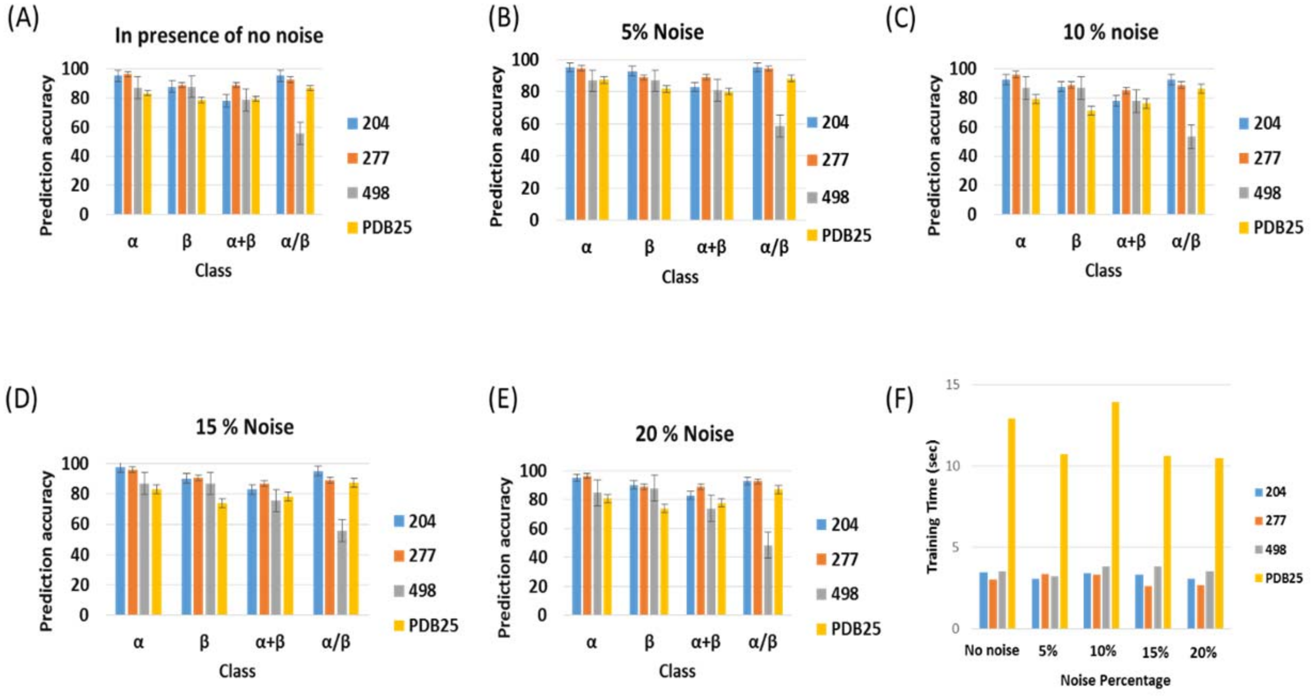

Figure 2. Class wise prediction accuracy of the ADAMMLP for datasets 204,277,498 and PDB25 in presence of 5%,10%,15% and 20% Gaussian noise and effect on training time.

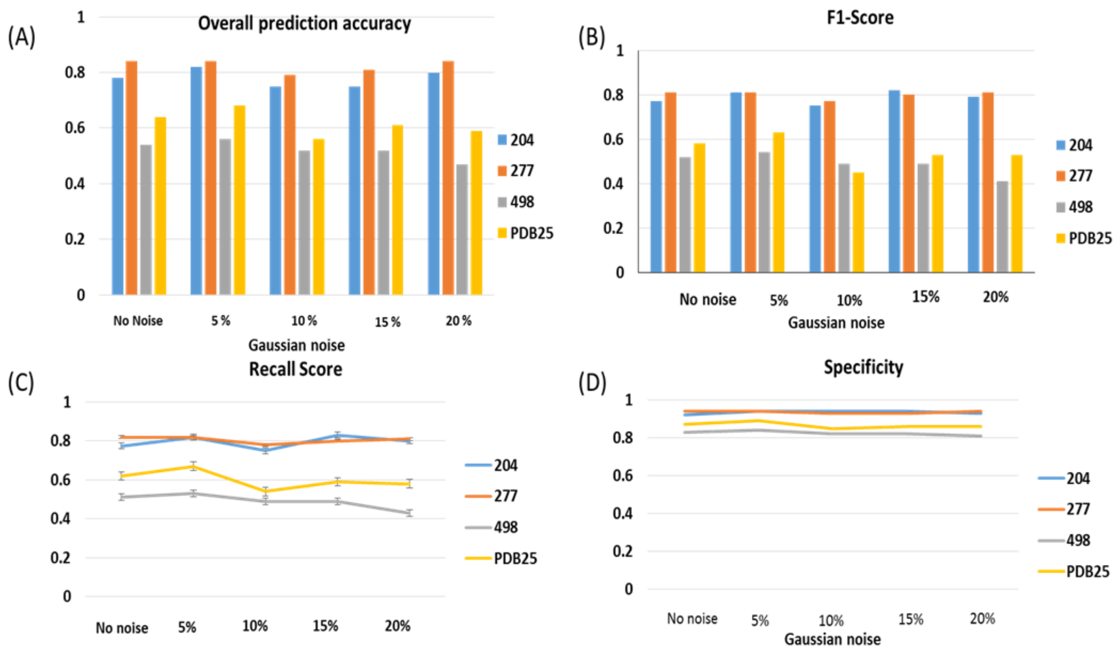

Figure 3. (A) shows the overall prediction accuracy of ADAMMLP, F1-score, Recall and precision are shown in (B),(C) and (D) in presence of different degree of noise in 204,277,498, and PDB25 datasets.

5. Conclusion

In this work, a hybrid adapative moment stochastic gradient descent based muliti layer perceptron network-ADAMMLP is suggested and verified for stable prediction accuracy in presence of Gaussian noise, which takes significantly less training time but with high prediction accuracy both for small and large sized datasets.The robustness of the model is evident from its ability to predict as high as 91.64% and minimum 73.64% in presence of 20% Gaussian noise. This model can be employed in other classification problems involving huge amount of data and noise.

Acknowledgements:

This project received no funding from any external organization. The authors are thankful to the “Department of Computer Science and Engineering, Institute of Technical Education and Research, Siksha 'O' Anusandhan University, Bhubaneswar, India” for providing high performance computational facility for accomplishing the project.

Conflict of Interests

We declare that we have no competing interests as well as conflict of interests.

References:

1. Ahmed N, Natarajan T, Rao KR. Discrete cosine transform. IEEE transactions on Computers. 1974 Jan;100(1):90-3.
2. Bhende CN, Mishra S, Panigrahi BK. Detection and classification of power quality disturbances using S-transform and modular neural network. Electric Power Systems Research. 2008 Jan 31;78(1):122-8.
3. Chen J, Chaudhari NS. Bidirectional segmented-memory recurrent neural network for protein secondary structure prediction. Soft Computing-A Fusion of Foundations, Methodologies and Applications. 2006 Feb 1;10(4):315-24.
4. Chou KC. A key driving force in determination of protein structural classes. Biochemical and biophysical research communications. 1999 Oct 14;264(1):216-24.
5. Ding S, Zhang S, Li Y, Wang T. A novel protein structural classes prediction method based on predicted secondary structure. Biochimie. 2012 May 31;94(5):1166-71.
6. Hopp TP, Woods KR. Prediction of protein antigenic determinants from amino acid sequences. Proceedings of the National Academy of Sciences. 1981 Jun 1;78(6):3824-8.
7. Jain AK, Mao J, Mohiuddin KM. Artificial neural networks: A tutorial. Computer. 1996 Mar;29(3):31-44.
8. Kingma D, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980. 2014 Dec 22.
9. Kurgan LA, Homaeian L. Prediction of structural classes for protein sequences and domains—impact of prediction algorithms, sequence representation and homology, and test procedures on accuracy. Pattern Recognition. 2006 Dec 31;39(12):2323-43.
10. Levitt M, Chothia C. Structural patterns in globular proteins. Nature. 1976 Jun 17;261(5561):552-8.
11. Liu H, Wang M, Chou KC. Low-frequency Fourier spectrum for predicting membrane protein types. Biochemical and biophysical research communications. 2005 Oct 28;336(3):737-9.
12. Sahu SS, Panda G. A novel feature representation method based on Chou's pseudo amino acid composition for protein structural class prediction. Computational biology and chemistry. 2010 Dec 31;34(5):320-7.
13. Stockwell RG, Mansinha L, Lowe RP. Localization of the complex spectrum: the S transform. IEEE transactions on signal processing. 1996 Apr;44(4):998-1001.
14. Tanford C. Contribution of hydrophobic interactions to the stability of the globular conformation of proteins. Journal of the American Chemical Society. 1962 Nov;84(22):4240-7.
15. Veljkovic, V., Cosic, I., Lalovic, D. (1985). Is it possible to analyze DNA and protein sequences by the methods of digital signal processing? IEEE Transactions on Biomedical Engineering. 5, 337-341.
16. Zhou GP. An intriguing controversy over protein structural class prediction. Journal of protein chemistry. 1998 Nov 1;17(8):729-38.
17. Zhou XB, Chen C, Li ZC, Zou XY. Using Chou's amphiphilic pseudo-amino acid composition and support vector machine for prediction of enzyme subfamily classes. Journal of theoretical biology. 2007 Oct 7;248(3):546-51.