Global Journal of Computer and Engineering Technology (GJCET)

Vol. 2, No. 2, July, 2022 ISSN: 2767-1933

PREDICTING PROTEIN STRUCTURES USING DEEP LEARNING NEURAL NETWORK

Jeniffer S D, Manonmaniam Sundaranar University, India (jenifferdavid1996@gmail.com)
Senthamarai Kannan K, Manonmaniam Sundaranar University, India (senkannan2002@gmail.com)

ABSTRACT

This paper focuses mainly on the protein's secondary structure. Protein structure prediction is a critical step towards understanding disease processes and developing treatment options. Predicting the secondary structure of mutated AD12 protein sequences is the goal of this study. Because of recent successes with deep neural networks, a two-state convolutional neural network was used by the scientists to predict the secondary structure of proteins. Multiple aligned sequences were used to produce the Position Specific Score Matrix, which was then used to create the Deep Learning Network. From the enhanced feature map, the structure of the sequence has been retrieved. The performance of the trained model is assessed and checked against various Deep Learning models which are designed for biological sequence analysis.

Keywords: Agglomerative Clustering, Artificial Neural Network, Convolutional Neural Network, Multiple Sequence Alignment, PSSM, Protein Secondary Structure.

1. INTRODUCTION

Each cell contains thousands of proteins, each of which has a specific purpose. All proteins are made up of a sequence of amino acids, despite their vastly differing structures and functions.Protein has a vital part in each and every natural processes. Like cells, proteins also a base of life. They play a part in cell structural integrity, small patch transport and storehouse, catalysis, regulation, signaling, and the vulnerable system, to name a many. In nature, proteins are made up of twenty distinct amino acids. It is a major challenge in the study of protein structure and function to accurately predict the secondary structure of proteins. Despite the importance of this issue, protein structure can be difficult to discover experimentally since it has a considerable impact on its function. Significant progress has recently been made in the field of inheritable. In Alzheimer's disease, brain atrophy (shrinking) and loss of brain cells lead to cognitive decline. An Alzheimer's disease-related form of lunacy is marked by a decline in social interaction as well as a loss of cognitive and behavioral abilities. Around 5.8 million Americans 65 and older are afflicted by Alzheimer's disease. Eighty percent of them are beyond the age of 75. About 60 percent to 70 percent of the roughly 50 million people who suffer from mental illness are expected to be affected by Alzheimer's.

Early symptoms of Alzheimer's disease include an inability to remember recent events and conversations. Memory deteriorates as Alzheimer's disease progresses, and the ability to perform everyday tasks diminishes as a result.
In some cases, specifics might help alleviate or slow the progression of symptoms for a short time. People with Alzheimer's disease who take these medications may be able to maintain some level of independence while also improving their overall quality of life. People with Alzheimer's disease, as well as those who care for them, can benefit from a range of resources. In the case of Alzheimer's, there is no cure or treatment that modifies the brain's complaint process at all. Advanced stages of Alzheimer's might result in death due to complications such as dehumidification or malnutrition, as well as infection. However, the complaint's threat can be dropped or other options explored, If the complaint is read or the threat of attack is known before.

2. LITERATURE REVIEW

Secondary structure prediction for proteins is a far more difficult problem. Cheng et al.(2008) employs a variety of machine learning and statistical methodologies. Methods that extract statistical features from protein sequences often reach accuracy of no more than 65 percent until the study of Garnier et al.(1996) and Liu et al.(2017). Using PSIBLAST-based position specific scoring matrix (PSSM), Jones was able to better predict protein secondary

structure (1999).There have been several attempts to use Support Vector Machines (SVM), Neural networks (NN) and K-nearest neighbors, but the most significant advancement in protein secondary structure prediction has been made by Tan et al.(2015) and other machine learning algorithms employing PSSM profiles. This is a significant improvement in accuracy, with a prediction accuracy of 70-79 percent. To anticipate secondary structure, Zhou et al. (2014) suggest using deep hierarchical representations, a generative stochastic network (GSN).

The CB513 dataset yielded a Q8 accuracy of 66.4 percent. Yang et al. proposed DeepCNFs, a deep learning extension of conditional neural fields that may be used to predict secondary structure (CNF). Cross-validation investigations using the CASP test protein DeepCNF produced Q3 accuracy of 84 % and Q8 accuracy of 73.23%. "The theoretical limit of 88–90% accuracy is reached by employing structural templates, resulting in a three-state accuracy of 82–84 percent. As a result of more powerful training datasets, as well as more effective deep learning algorithms, these advancements have been made.Using CNNs with two convolutional layers and a new representation of the two-dimensional PSSM evolutionary matrix, features were retrieved in the work of Liu et al (2016 and 2017) The system can manage both long-range interactions of amino acid sequences and mutations in amino acid sequences, as shown by the restored CNN features. Yang et al. (2011) and Zhou et al. (2014) transformed the PSSM matrix into a one-dimensional vector even if the positional data for the amino acid sequence was not completely described.

This study predicts the three classes of protein secondary structure using a two-dimensional convolutional neural network with six convolutional layers and five max-pooling layers in order to handle large datasets. Using the input profile, a two-dimensional mixed PSSM matrix, the amino acid sequence is represented as a 20-bit binary vector.

3. RESEARCH METHODOLOGY

3.1 Preprocessing

Before building the Convolutional Neural Network, the data must be pre-processed. Pre-processing techniques include producing PSSM matrix, consensus sequence establishment, and hierarchical clustering using multiple sequence alignment.

Multiple Sequence Alignment

Determine the amount of relationship among the sequences, named as dissimilarity measures. Merge the sequences according to their similarity.

$$\delta(X,Y) = \max_{x_i \in X, x_j \in Y} d(x_i, x_j) \quad (1)$$

Where $\delta(X,Y)$ and $d(x_i, x_j)$ is the dissimilarity measures between the sequences X & Y and x_i & x_j respectively. From the seq – seq, seq – prof, prof – prof alignment, deduce consensus sequence.

PSSM matrix :

PSSM stands for position-specific score matrix. It is a representation of motifs in biological sequences. The numbers $\frac{log\, e_i\,(a)}{q_a}$ behave as if they were components of a score matrix s(a,b), where the second index is position i, rather than amino acid b. A position-specific scoring matrix approach is used because of this (PSSM). It is possible to utilize the PSSM to find an exact match in an indefinitely long sequence of length N, by assessing the score S_j at each beginning point of the series starting from 1 and going all the way to N-L +1, where L is the length of the sequence.

3.2 Convolutional Neural Network

Basic Diagram

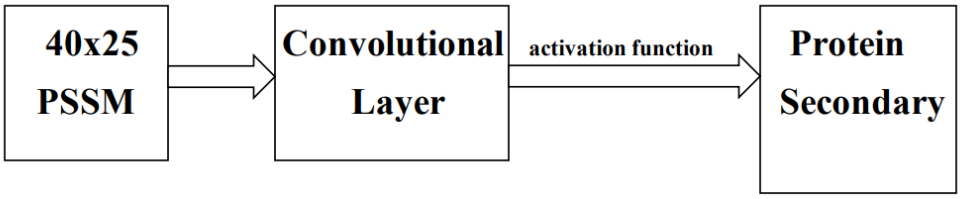

Figure 1: Basic diagram of CNN representing its workflow

Architecture of CNN :

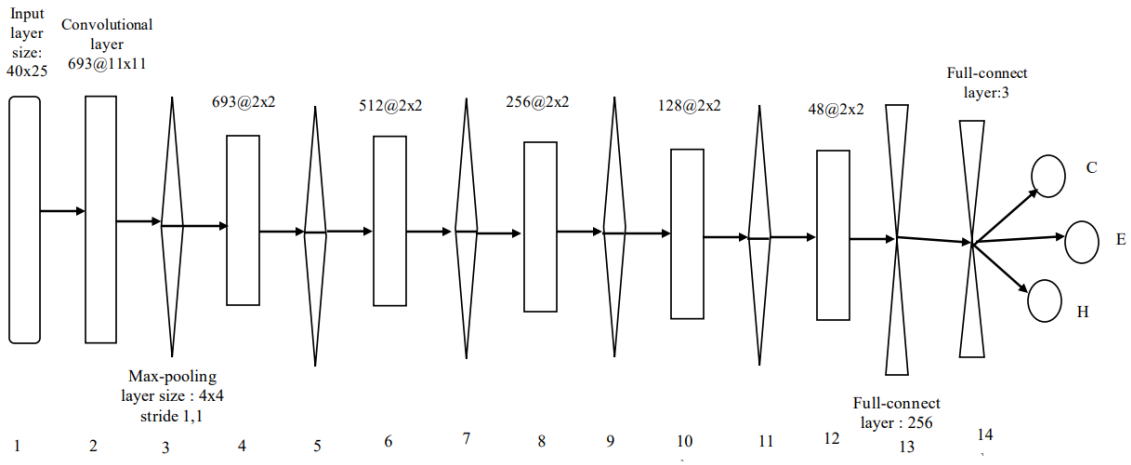

Figure 2: The convolutional neural network-based architecture of the prediction model. Convolutional layers are those with positions of 2, 4, 6, 8, 10, and 12. At locations 3,5,7,9, and 11, the maximum-pooling layers are located. Full-connected layers 13 and 14

Operator in CNN :

The convolution operation is a linear operation, and its symbol is an asterisk. It is used to combine two signals:

$$f[x,y]*g[x,y]=\sum_{n_1=-\infty}^{\infty}\sum_{n_2=-\infty}^{\infty}f[n_1,n_2].g[x-n_1,y-n_2] \quad (2)$$

Hyperparameters of CNN:

- Kernel size (K): How many pixels are in each sliding window. Smaller is preferable, and odd numbers like 1,3,5 or 7 are often used.
- Stride (S): Each convolution phase, the kernel window will move by a certain number of pixels in most cases, this is set to 1, which ensures that no areas in an image are missed, although it may be increased if the input size is reduced at the same time.

- Zero padding (pad): The number of zeros that should be added to the image's border. Padding enables the kernel to filter an input image down to its very edges.
- Number of filters (F): It defines how many filters will be used in the convolution layer. The amount of patterns or features that a convolution layer looks for may be adjusted using this parameter.

4. RESULTS

The taken protein sequences can be aligned as mentioned.

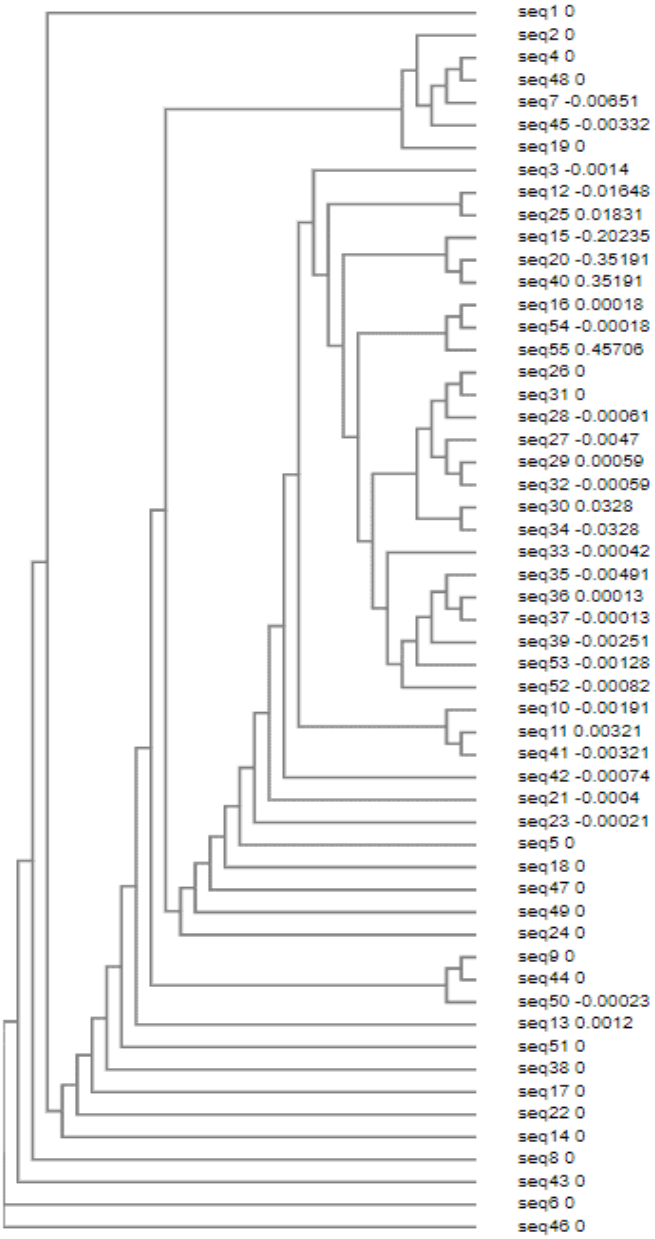

Figure 3: Phylogenetic tree for the alignment of protein sequences

Figure 4: Consensus Sequence Logo.

Construction of Input Matrix :

A convolutional neural network constructed in this study require two dimensional PSSM as input. It is possible to utilize PSSM as a feature vector for predicting secondary structure, and it may be generated using the PSI-BLAST. Input sequence characteristics into the BLOSUM62 matrix of evolution. This test uses a 0.001 cutoff value and three iterations to ensure correct. The PSSM matrix is 20x693 in this case; there are a total of 20 different types of amino acids. The consensus amino acid sequence, on the other hand, is 693 amino acids length. Consequently, the PSSM matrix contains twenty features for each residue in the protein sequence, each of which represents a probability that the residue will change to the corresponding amino acid. Multiple sequence alignment was used to build the mutation matrix. The amino acid sequences are represented in data using the quadrature encoding method. The amino acid type is identified by a 20-bit binary vector.

For each residue, there are 40 input characteristics available; 20 of them are provided by PSSM, while the other 20 are obtained from the different kinds of amino acids.
To predict the central residue secondary structure, amino acid consecutive sliding window may employ. The input matrix of the CNN model is as follows.

```
-1 -1 -2 -3 -1  0 -2 -3 -2  1  2 -1  6  0 -3 -2 -1 -1 -1  1
-2 -2 -4 -4 -1 -2 -3 -4 -3  2  4 -3  2  0 -3 -3 -1 -2 -1  1
-1 -2 -2 -2 -3 -1 -1 -2 -2 -3 -3 -1 -3 -4  8 -1 -1 -4 -3 -2
 0 -2  0 -1 -3 -2 -2  6 -2 -4 -4 -2 -3 -3 -2  0 -2 -3 -3 -3
-2 -2 -4 -4 -1 -2 -3 -4 -3  2  4 -3  2  0 -3 -3 -1 -2 -1  1
 4 -1 -2 -2  0 -1 -1  0 -2 -1 -2 -1 -1 -2 -1  1  0 -3 -2  0
-2 -2 -4 -4 -1 -2 -3 -4 -3  2  4 -3  2  0 -3 -3 -1 -2 -1  1
-2 -2 -4 -4 -1 -2 -3 -4 -3  2  4 -3  2  0 -3 -3 -1 -2 -1  1
-2 -2 -4 -4 -1 -2 -3 -4 -3  2  4 -3  2  0 -3 -3 -1 -2 -1  1
-2 -2 -4 -4 -1 -2 -3 -4 -3  2  4 -3  2  0 -3 -3 -1 -2 -1  1
 4 -1 -2 -2  0 -1 -1  0 -2 -1 -2 -1 -1 -2 -1  1  0 -3 -2  0
 4 -1 -2 -2  0 -1 -1  0 -2 -1 -2 -1 -1 -2 -1  1  0 -3 -2  0
-3 -3 -4 -4 -2 -2 -3 -3 -2 -3 -2 -3 -1  1 -4 -3 -3 11  2 -3
 0 -1  0 -1 -1 -1 -1 -2 -2 -1 -1 -1 -1 -2 -1  1  5 -3 -2  0
 4 -1 -2 -2  0 -1 -1  0 -2 -1 -2 -1 -1 -2 -1  1  0 -3 -2  0
-1  6  0 -2 -4  1  0 -2  0 -3 -2  2 -1 -3 -2 -1 -1 -3 -2 -3
 4 -1 -2 -2  0 -1 -1  0 -2 -1 -2 -1 -1 -2 -1  1  0 -3 -2  0
-2 -2 -4 -4 -1 -2 -3 -4 -3  2  4 -3  2  0 -3 -3 -1 -2 -1  1
-1  0  0  1 -2  1  4 -1  0 -2 -2  1 -1 -2 -1  0 -1 -2 -1 -2
 0 -1 -2 -2 -1 -1 -1 -2 -2  2  1 -1  0 -1 -1 -1  0 -2 -1  3
-1 -1 -1 -1 -2 -1 -1 -1 -1 -2 -2 -1 -2 -2  7  0 -1 -2 -2 -2
 0 -1  0 -1 -1  0  0 -1 -1 -1 -1  0 -1 -1  0  1  4 -1 -1  0
-1 -1  1  5 -2  0  1 -1 -1 -2 -2  0 -2 -2 -1  0 -1 -2 -2 -2
 0 -1  0 -1 -2 -1 -1  5 -1 -2 -2 -1 -2 -2 -1  0 -1 -1 -2 -2
-1  0  5  1 -2  0  0  0  0 -2 -2  0 -1 -2 -1  0  0 -2 -1 -2
 3 -1 -1 -1  0  0  0  0 -1 -1 -1  0 -1 -1  0  1  0 -2 -1  0
 0 -1  0 -1 -2 -1 -1  5 -1 -2 -2 -1 -2 -2 -1  0 -1 -1 -2 -2
-1 -1 -2 -2 -1 -1 -2 -2 -2  1  3 -1  2  0 -1 -1 -1 -1 -1  1
-1 -1 -2 -2 -1 -1 -2 -2 -2  1  3 -1  2  0 -1 -1 -1 -1 -1  1
 3 -1 -1 -1  0  0  0  0 -1 -1 -1  0 -1 -1  0  1  0 -2 -1  0
-1  0  0  1 -2  1  4 -1  0 -2 -2  1 -1 -2 -1  0 -1 -2 -1 -2
-1 -1 -1 -1 -2 -1 -1 -1 -1 -2 -2 -1 -2 -2  7  0 -1 -2 -2 -2
-1  1  0  0 -2  5  1 -1  0 -2 -1  1  0 -2 -1  0  0 -1 -1 -1
-1 -2 -2 -2 -1 -1 -2 -2 -2  3  1 -2  1  0 -1 -1  0 -2 -1  2
 3 -1 -1 -1  0  0  0  0 -1 -1 -1  0 -1 -1  0  1  0 -2 -1  0
-1 -1 -1 -2 -1  0 -1 -2 -1  1  2 -1  5  0 -1 -1  0 -1 -1  0
-1 -2 -2 -2 -2 -2 -2 -2 -1  0  0 -2  0  5 -2 -1 -1  1  2 -1
 0 -2 -2 -2  8 -2 -2 -1 -2 -1 -1 -2 -1 -2 -1 -1 -1 -1 -1 -1
 0 -1  0 -1 -2 -1 -1  5 -1 -2 -2 -1 -2 -2 -1  0 -1 -1 -2 -2
-1  5  0 -1 -2  1  0 -1  0 -2 -1  2 -1 -2 -1  0 -1 -2 -1 -2
-1 -1 -2 -2 -1 -1 -2 -2 -2  1  3 -1  2  0 -1 -1 -1 -1 -1  1
-1  0  5  1 -2  0  0  0  0 -2 -2  0 -1 -2 -1  0  0 -2 -1 -2
-1 -1 -1 -2 -1  0 -1 -2 -1  1  2 -1  5  0 -1 -1  0 -1 -1  0
-1  0  0 -1 -2  0  0 -1  7 -2 -2  0 -1 -1 -1 -1 -1 -1  1 -2
-1 -1 -1 -2 -1  0 -1 -2 -1  1  2 -1  5  0 -1 -1  0 -1 -1  0
```

Figure 5: Input matrix (40x25) for one slicing window of length 25.

Construction of CNN:

There are layers in a convolutional neural network called ‘‘convolutional layers’‘ that filter the inputs to those layers to locate valuable characteristics within those inputs. DSSP classifies proteins' secondary structures into eight categories, including α-helix, 3-helix, 5-helix, extended strand, isolated strand, turn, bend, and coil. The letters H, G, I, E, B, T, and S, as well as the sign ‘‘-,’‘ can all be used to denote these categories. As per the convenience, the eight categories can be reduced into 3 large categories. They are: H, E, C. The alphabet H includes the states H, G, and I; similarly, E contains the states E and B; then the remaining states can be lied into the category named as C.

The derived PSSM matrix of two dimension is act as the input to bring out the features of the proteins. For this, CNN is implemented. Along with the protein features, the information of amino acids also extracted. The convolutional neural network constructed in the study has several layers. Six layers are convolutional, five are max-pooling, two are fully linked, and the final one is a soft-max layer. Input layer: one (Figure 2). The layers mentioned above are essential for predicting the secondary structure of the protein and extracting its characteristics. To further reduce the feature's size, fully linked layers are used.

Soft-max divides protein secondary structure into three distinct states for final processing. In the first convolutional layer, a 1x1 stride filter is used, followed by a 4x4 max-pooling layer with a stride of 1x1. Convolutional layers with 512, 256, and 48 filters with 2x2 and 1x1 strides are discovered in positions 4 through 12. The 2x2 and 1x1 max-pooling activities were carried out by the max-pooling layers at locations 5, 7, 9, and 11. To further reduce the size, a second full-connected layer of 3 units is added on top of the first layer of 250 units. The soft-max classification model takes into account the characteristics of the final full-connected layer.
The extracted feature map of the convolutional neural network is as follows,

Figure 6: Extracted feature map of the Protein by two-layer convolutional neural network

Figure7: Fragment of the predicted individual secondary structures of a protein

In the above-mentioned figure, blue represents H, magenta represents S, and yellow represents the remaining secondary structure C. The secondary structure symbols also retrieved from the Figure 7, and listed a fragment of it as follows.

C,C,H,H,H,H,H,H,H,H,H,H,H,H,H,H,C,C,C,C,C,C,C,C,C,C,C,E,E,E,E,E,C,C,C,C,C,C,C,C,E,C,C,C,C,E,E,E,C,
C,C,C,C,C,C,C,C,C,C,C,H,H,H,H,H,H,H,H,H,H,H,C,C,C,C,C,E,E,E,E,E,E,C,C,C,E,E,E,C,C,C,C,C,C,C,C,C,C,C
,C,C,C,C,E,E,E,E,E,E,E,E,C,C,C,C,C,C,E,E,E,E,C,C,C,C,C,E,E,E,C,C,C,C,C,C,C,C,C,C,C,C,C,C,H,H,H,H,H,H,
C,C,C,C,C,C,C,E,E,E,C,C,C,C,C,C,C,C,C,C,C,E,E,E,E,C,C,C,C,C,C,C,C,C,C,C.

The test dataset can be used to verify the predicted secondary structures of the training model. Protein secondary structure predictions can be compared against real-world structures from the PDB. We determined the validation cumulus based on the results of Q3.The formula to determine Recall(Q_3) is,

$$Q_3 = \frac{TP}{TP+FN} \quad (3)$$

To determine H, E, and C, this criterion can be used individually. The Q3 criteria based on helix considers the number of genuinely identified helices in comparison to the total number of properly detected helices and incorrectly detected non-helices.

$$Q_3 - H = \frac{TP_H}{TP_H + FN_H} \quad (4)$$

Similarly,

$$Q_3 - E = \frac{TP_E}{TP_E + FN_E} \quad (5)$$

$$Q_3 - C = \frac{TP_C}{TP_C + FN_C} \quad (6)$$

The overall criteria such are computed by just averaging the above mentioned three criteria.

$$Q_3 - Overall = \frac{(Q_3 - H) + (Q_3 - E) + (Q_3 - C)}{3} \quad (7)$$

Some of the previous researchers, have considered Q_3 only for their analysis. In order to its inefficiency, the following measures also calculated. Using Q3 and precision and F-measure, researchers are able to perform a more fair and accurate evaluation. The F-measure is a harmonic average of Precision and Q3. An F-measure metric like this can be used to measure performance quality.

$$Precision = \frac{TP}{TP + TN} \quad (8)$$

$$F - measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (9)$$

The overall of performance of the two-layer Convolutional Neural Network developed in the study can be listed in the following table.

Table 1: Validation of the CNN

	Q3	Precision	F-measure
H	0.332	0.346	0.326
E	0.338	0.232	0.283
C	0.552	0.475	0.419
Overall	0.407	0.351	0.333

Using the F-measure, the table shows how each structure can be accurately measured. So, the ratio of H, E, and C is 32.6%, 25.4%, 41.9% respectively.

In overall, the performance of Convolutional Neural Network can be calculated as usual using Q_3 measure; and it can be found that 86.3%. It can be known as the model performed as good as the percentage.

5. DISCUSSION

A large number of perceptrons or neurons compose an Artificial Neural Network (ANN). ANNs are also referred to as ‘‘Feed-Forward Neural Networks’‘ since they solely process inputs in the forward direction. Models like convolutional neural networks (CNNs) have been popular in recent years (CNN). Convolutional layers can be joined or pooled together to form a multilayer perceptron variant in this neural network computational model. In nonlinear processing, the feature maps created by these convolutional layers are broken down into rectangles and used to capture more of the original image's features.

The performance of CNN may be upgraded in a number of ways, including with different extensions. One of these is the typically two neural networks. Network models that predict secondary structure of proteins may be compared using the suggested model's performance measure.

Initially, the Receiver Operating Characteristic curve has been constructed to assess the performance of the two-layer convolutional neural network proposed in this article.

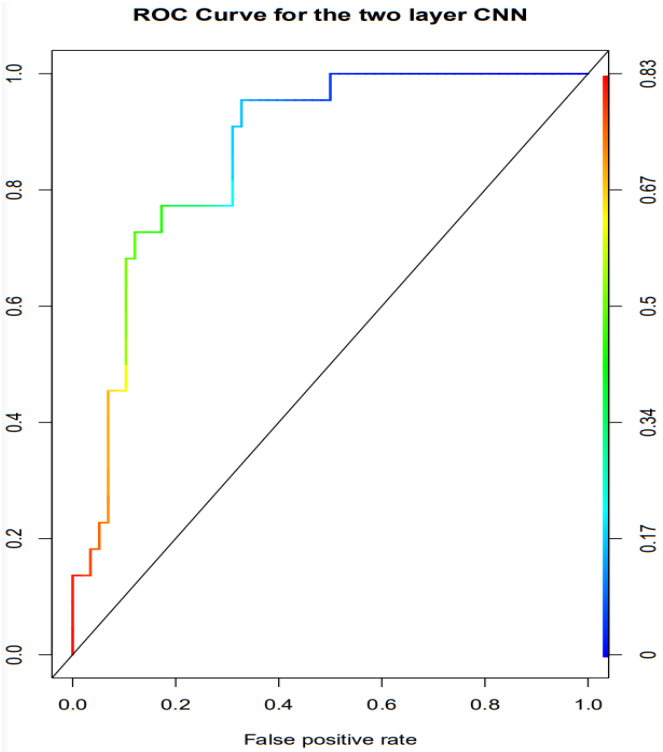

Figure 8: ROC curve for the two-layer CNN

Table 2: Comparison of various Techniques in predicting the protein secondary structure.					
	2-layer CNN	ANN	SVM	Dynamic CNN	Random Forest
AUC	0.863	0.72	0.81	0.83	0.76

From the above table, it is found to be that the two-layer convolutional neural network used in the study is the better performing one with 86.3% accuracy.

6. CONCLUSION

In response to targets, antibody development companies can synthesize hundreds of antibody sequences. They need to get rid of any antibodies that don't have the right qualities as quickly as possible. Predicting protein fold characteristics from amino acid sequence is one method for combating this quick-to-fail mentality. These methods of prediction will also be useful in the field of protein engineering. We may be able to predict a structure and use it to develop molecules with increased or decreased affinity to binding partners without the need for an experimental structure by using the increasing machine learning and deep learning techniques.

In this paper, 255 mutated gene sequences of AD12 protein, which causes the Alzheimer's disease have been taken. Aligned according to the dissimilarity measure among each pair of sequences. After the multiple sequence alignment, the PSSM matrix has been constructed from the deduced consensus sequence. Using the PSSM measures, the Convolutional Neural Network has been modelled and trained. The performance of the two-layer convolutional neural network model is assessed and compared; and found that the proposed model is the better performing one among the several machine learning and deep learning models.

REFERENCES

Bidargaddi, N. P., Chetty, M., & Kamruzzaman, J. (2009). Combining segmental semi-Markov models with neural networks for protein secondary structure prediction. Neurocomputing, 72(16-18), 3943-3950.

Chen, C., Tian, Y., Zou, X., Cai, P., & Mo, J. (2007). Prediction of protein secondary structure content using support vector machine. Talanta, 71(5), 2069-2073.

Cheng, J., Tegge, A. N., & Baldi, P. (2008). Machine learning methods for protein structure prediction. IEEE reviews in biomedical engineering, 1, 41-49.

Durbin, R., Eddy, S. R., Krogh, A., & Mitchison, G. (1998). Biological sequence analysis: probabilistic models of proteins and nucleic acids. Cambridge university press.

Garnier, J., Gibrat, J. F., & Robson, B. (1996). GOR method for predicting protein secondary structure from amino acid sequence. In Methods in enzymology, Vol. 266, pp. 540-553). Academic Press.

Islam, M. N., Iqbal, S., Katebi, A. R., & Hoque, M. T. (2016). A balanced secondary structure predictor. Journal of theoretical biology, 389, 60-71.

Jeniffer, S.D., & Senthamarai Kannan, K., (2021). Stochastic modelling for identifying malignant diseases, Advances and Applications of Mathematical Sciences, 20(9), 1923-1936.

Jones, D. T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. Journal of molecular biology, 292(2), 195-202.

Liu, Y., & Cheng, J. (2016, December). Protein secondary structure prediction based on wavelets and 2D convolutional neural network. In Proceedings of the 7th international conference on computational systems-biology and bioinformatics, pp. 53-57.

Liu, Y., Chen, Y., & Cheng, J. (2016, October). Feature extraction of protein secondary structure using 2D convolutional neural network. In 2016 9th International Congress on Image and Signal Processing, Biomedical Engineering and Informatics (CISP-BMEI), pp. 1771-1775).

Liu, Y., Cheng, J., Ma, Y., & Chen, Y. (2017, December). Protein secondary structure prediction based on two dimensional deep convolutional neural networks. In 2017 3rd IEEE International Conference on Computer and Communications (ICCC), pp. 1995-1999.

Malekpour, S. A., Naghizadeh, S., Pezeshk, H., Sadeghi, M., & Eslahchi, C. (2009). Protein secondary structure prediction using three neural networks and a segmental semi-Markov model. Mathematical Biosciences, 217(2), 145-150.

Petit, D., Montplaisir, J., & Boeve, B. F. (2010). Alzheimer's disease and other dementias. In Principles and Practice of Sleep Medicine: Fifth Edition , pp. 1038-1047.

Rastogi, S. C., Mendiratta, N., & Rastogi, P. (2008). Bioinformatics Methods and Applications: Genomics. Proteomics and Drug Discovery, PHI Learning Private Limited, New Delhi.

Reilly, C. (2009). Statistics in human genetics and molecular biology. CRC Press.

Roth, C. (2021). Statistical methods for biological sequence analysis for DNA binding motifs and protein contacts (Doctoral dissertation, Georg-August University).

Tan, Y. T., & Rosdi, B. A. (2015). Fpga-based hardware accelerator for the prediction of protein secondary class via fuzzy k-nearest neighbors with lempel–ziv complexity-based distance measure. Neurocomputing, 148, 409-419.

Yang, B., Wu, Q., Ying, Z., & Sui, H. (2011). Predicting protein secondary structure using a mixed-modal SVM method in a compound pyramid model. Knowledge-Based Systems, 24(2), 304-313.

Yang, Y., Gao, J., Wang, J., Heffernan, R., Hanson, J., Paliwal, K., & Zhou, Y. (2018). Sixty-five years of the long march in protein secondary structure prediction: the final stretch?. Briefings in bioinformatics, 19(3), 482-494.

Yao, X. Q., Zhu, H., & She, Z. S. (2008). A dynamic Bayesian network approach to protein secondary structure prediction. BMC bioinformatics, 9(1), 1-13.

Zafar, I., Tzanidou, G., Burton, R., Patel, N., & Araujo, L. (2018). Hands-on convolutional neural networks with TensorFlow: Solve computer vision problems with modeling in TensorFlow and Python. Packt Publishing Ltd.

Zhou, J., & Troyanskaya, O. (2014, January). Deep supervised and convolutional generative stochastic network for protein secondary structure prediction. In International conference on machine learning, pp. 745-753.