International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
***Article Received:** 05 July 2023 **Revised:** 25 August 2023 **Accepted:** 15 September 2023*

Proto-oncogene Protein Sequence Classification using RNN and CNN with Attention Mechanism

M.Vijayalakshmi[a]*, Dr.V.Vallinayagi[b]
[a]*Research Scholar, (Registration No.-20224542282029)
[b]Head and Associate Professor, [a*b]Department of Computer Science,
[a*b]Sri Sarada College for Woman(Autonomous), Tirunelveli-627 011
Affiliated to Manonmaniam Sundaranar University,
Abishekapatti, Tirunelveli-627 012, Tamil Nadu, India.
[a*]e-mail: vijimarthandam@gmail.com
[b]e-mail: vallinayagimahesh@gmail.com

Abstract— The fundamental tool for uncontrolled tumour progression is the lack of regulatory capacity of tumour suppression genes (TSG) and mutations in proto-oncogenes (OG). Even though tumour is a diverse complex of several diseases, discovering possibilities of genes connected to OG activity by computational research can aid in the development of medications that specifically target the disease. Attention mechanism in Deep learning has recently become an innovative approach for classifying protein sequences. The attention-based approach can offer a trustworthy and understandable method that aids in overcoming the existing difficulties in describing deep neural networks for classifying protein sequences. In this study, we classify proto-oncogenes (OG) with the help of CNN, Bi_LSTM and Bi_GRUwith attentionmechanisum. of all the three attention mechanisms, Bi_LSTM significantly performs far better than the other two approaches and achives F1-Score upto 97.3% and it is 3% more traditional ML Random Forest approach.

Keywords-Proto-Oncogenes; Deep learning; Attention Mechanism; Classification; Feature Extraction

I. INTRODUCTION

Our bodies are made up of billions of cells that must collaborate in order to keep us healthy. Our cells must be able to divide in order to produce new cells to aid in the growth of the body or to replace cells that have perished. Controlling cell development and division is necessary to prevent overcrowding of the surrounding cells due to excessive cell proliferation. Normal cell growth is regulated by certain genes that are present inside each cell. Every single gene is made up of a series of nucleotide bases that hold information for how cells develop and function. Every protein in the human body has a particular purpose. Proto-oncogenes are common genes that help in cell development and division to create new cells or maintain the viability of existing cells. Proteins produced by proto-oncogenes have a role in promoting cell proliferation, suppressing cell differentiation, and protecting against cell death. Proto-oncogenes become oncogenes as a consequence of mutations or gene amplification, which result in the changes that transform healthy cells into malignant ones [1]. Proto-oncogenes are frequently classed according to how closely their sequences resemble those of known proteins or according to how they typically behave inside cells [2].

Recent years have seen a lot of study on protein sequence analysis [3]. The main goal of protein sequence examination is to characterise protein sequences in silico and to predict the structures and functions of proteins. Protein sequence categorization is crucial to protein sequence analysis because it helps identify Members of the same protein superfamily's protein sequence that are linked to one another physically, functionally, and historically. Studies on protein functional detection often focus on all sorts of functions, whether or not they are connected to cancer. However, the category of cancer-related functional detection is very helpful for treating cancer. Several techniques, including prediction by evolutionary connections [4], sequence similarity [5, 6], gene-ontology hierarchy [7, 8, 9], genetic interactions [10], protein structures, and protein-protein interactions [11] are improving the prediction of the functional annotation of proteins.

In terms of extracting and classifying visual features, CNNs have demonstrated good results [12, 13]. This present work proposed a deep learning with attention based model for proto-oncogenes protein sequence classification.

II. PROTO ONCOGENES PREDICTION USING CNN MODEL WITH ATTENTION MODEL

A. Proto Oncogene prediction model using CNN and RNN with Attention

For image processing and recognition tasks, a Convolutional Neural Network (CNN) deep learning algorithm is the best choice. Of the various layers that collectively make up this framework are convolutional layers, pooling layers, and completely linked layers. They are widely employed in machine vision, image processing, and numerous other related

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *05 July 2023* ***Revised:*** *25 August 2023* ***Accepted:*** *15 September 2023*

fields. Today, attention is without a doubt one of the majorities of important concepts in the field of deep learning. Humans require the complicated cognitive function of attention, which is essential [14, 15].

Using limited processing resources, this enables people to quickly select the most essential data from an enormous quantity of information. The attention mechanism substantially improves the efficiency and accuracy of the interpretation of sensory data[16]. Deep learning's attention mechanisms [17,18] are mostly focused attention because they are created with specific objectives in mind.

Except for specific claims, the attention process described in this work typically refers to focused attention.

Bahdanau et al. [19], was designed to first apply the attention mechanism to the machine translation job. The RNN search consists of an encoder, a bidirectional recurrent neural network (BiRNN), and a decoder that mimics searching throughout a source sentence. Depends on the input sequence ($Y_1,\ldots,Y_T$), the encoder calculates annotations ($I_1,\ldots,I_T$). which constitute the hidden state of the BiRNN:

$$(I_1,\ldots,I_T) = BiRNN(Y_1,\ldots,Y_T) \qquad (1)$$

Regular RNNs have the drawback of just using past context. All input data that is available for the past and future of a given time frame can be employed to train the BiRNN. Particularly, as shown in Fig. 1, forward and backward RNNs are used to extract the hidden states $\{\tilde{I}_1,\ldots,\tilde{I}_T\}$ and $\{\tilde{I}_1,\ldots,\tilde{I}_T\}$, respectively. The forward hidden state $\tilde{I}$ and the backward hidden state $\tilde{I}$ are then concatenated by the encoder to provide an annotation for one word Y_i. An attention block and a recurrent neural network (RNN) make up the decoder.

Calculating the context vector *C*, which depicts the context link between each word in the entire input sequence and the current output symbol, is the task of the attention block. The context vector $CV_i$is calculated for each time step *ti* as a weighted sum of these annotations *Ij*.

$$CV_{ti} = \sum_{j=1}^{T} \alpha_{tij} I_j \qquad (2)$$

The attention weight α_{tj} of every annotation I_j is computed by e_{tj} = a(S_{t-1},I_j),and

$$\alpha_{tj} = \frac{exp\,(e_{tj})}{\sum_{K=1}^{T} exp(e_{tik})} \qquad (3)$$

whereα is a learnable function that, in accordance with the state st1, communicates the significance of the annotation $I_j$to the following concealed state S_{t-1}. The RNN then outputs the symbol at the current step that is most likely to be Z_t.

$$P(z_t|\, z_{1,\ldots\ldots}z_{t-1,}y) = RNN(CV_{ti}) \qquad (4)$$

Instead of encoding everything into a fixed-length vector, the encoder can do this and disperse the information from the source sentence over the whole sequence, allowing the decoder to get only the information it needs at each time step. With this formulation, the neural network can concentrate on important input components rather than unimportant ones.

B. CNN with Attention Model

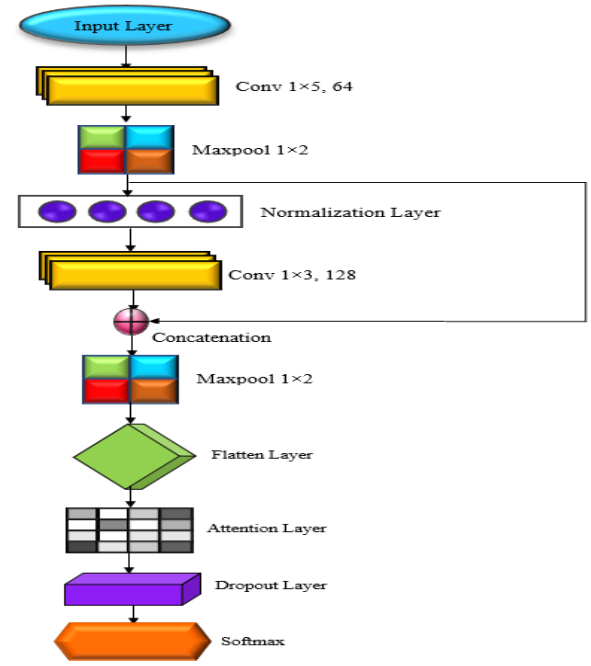

Figure 1. CNN Architecture

The above figure 1 shows the architecture of the CNN_ATT model for Proto Oncogene prediction. The convolution layer receives the input sequences Is and applies a 64-bit filter with a 1-bit kernel size. The output of the convolutional layers will be fed into size 2 and subsequently the MP1 maxpooling layer. The normalisation Layer will then be applied to the output.

The output will then be applied to a convolution layer with a 128-filter size and transmitted to a concatenation layer with a 1x3 kernel size. Concatenation layer output will be fed into MP2 of the maxpooling layer with size 2. The output will next be transferred to a flattening layer, followed by an attention layer. The dropout layer will be given the results of the attention layer. The SoftMax classification will be used in the dropout output.

III. BI-LSTM AND BI-GRU ATTENTION MODEL

The RNN's inadequacies led to the development of the LSTM. LSTM is used to resolve the neural network's vanishing gradient problem. Since the lost function roughly approaches

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *05 July 2023* ***Revised:*** *25 August 2023* ***Accepted:*** *15 September 2023*

zero, the disappearing gradient problem makes neural network training challenging [20,21,22]. The employment of memory cells and a gate mechanism to regulate them forms the basis of the LSTM structure, which uses it for keeping long-term historical data.

A single LSTM typically only processes one forward direction of information. Or, to put it another way, it can only access historical data. The BiLSTM design, on the other hand, places one layer of LSTM at the forward and the other at the retrogressive. The result in both hidden layers is then combined. The forward LSTM may access the data grouping's historical information data, while the opposite LSTM can access the information arrangement's prospective information data. Both of the forward and reverse hands are present in the secret state of the Bi-LSTM at the present time (t).

$$h_t = \vec{h}_t \oplus \overleftarrow{h}_t \qquad (5)$$

Where the forward and backward output components are combined using the component-based summation. Because it can use data from the past and present time, Bi-LSTM is more efficient than LSTM and RNN. There are two GRUs in the sequence processing paradigm referred to as a bidirectional GRU or BiGRU (Bi-directional Gate Recurrent Unit)[23]. The information is processed both forward and backward by one. With only input and forget gates, Bi-GRU can speed up and reduce the amount of computation required.

A structure called a bi-GRU combines two different directional GRUs. More precisely, one GRU travels from 1 to N and the other from N to 1 in a sequence of length N.

The same input was given to both directional GRUs at every time *ti*, and the result is decided jointly by both GRUs. At time *t*, the hidden states of the forward and backward GRUs are indicated by h_t and h_{t-1}, respectively. The two secret states are consolidated to work out the cell yield each time.

The architecture of Bi_LSTM and Attention model for proto oncogene prediction with Bi_LSTM model is depicted in figure 2. The input sequences I_s will be routed into the Bi_LSTM with size 128 filter. The outcome of Bi_LSTM is sent into max pooling layer MP1 with size 2. Then, the Relu Layer will be applied to the output.

The output of the Relu layer is then fed into the attention layer. The dropout layer will be given the results of the attention layer. The output of the BI_LSTM layers is fed into the max pooling layer MP2 with size2.

The outcome will then be applied to the Relu layer and then into the dense layer with size 128. The SoftMax classification will be used in the dense output.

Instead of BI_LSTM, the aforementioned approach is applied same for Bi_GRU.

A. Dataset Description

The well-known protein database Uniprot, whose data availability has been empirically validated, contains information on a variety of proteins. The Uniprot database's protein sequences with the word "proto-oncogene" in it were picked to gather information about. For our model evaluation, we used the pre-handled dataset that we obtained from [24] and the experimental section shows how the designed different deep model with attention mechanism boosts the performance better than the ProtoPred-RF approach [24]. Since Y+ and Y- are composed of 630 negative samples and 252 positive samples, respectively. we tested using independent testing and k-fold approaches.

The performance of three kinds of attention model is evaluated using Accuracy, Precision, Recall and F-measure. We used 30% testing and 70% training for the independent test. The results of the independent dataset training are displayed in table 1 below.

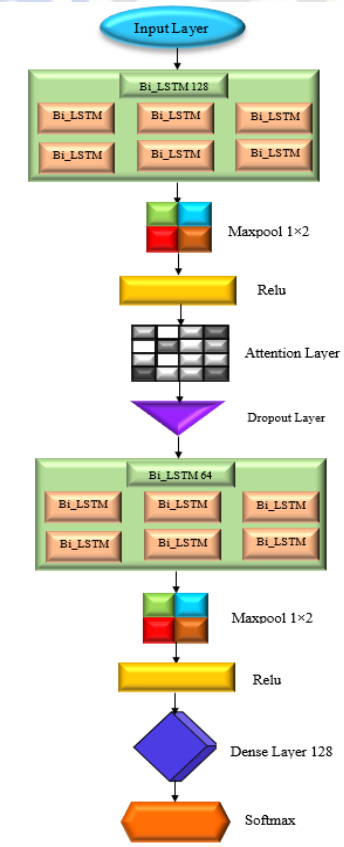

Figure 2. BILSTM and Attention model for Proto Oncogene Prediction

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *05 July 2023* ***Revised:*** *25 August 2023* ***Accepted:*** *15 September 2023*

IV. RESULTS AND DISCUSSION

The performance of this approach is analyzed with the help of classification metrics such as Accuracy, Precision Recall, F1-Score and Matthews correlation coefficient.

TABLE I. RESULTS OF THE INDEPENDENT DATASET TRAINING

Methods	Accuracy	F1-Score	Precision	Recall	Mcc
PSSM[25]	0.80754	0.77559	0.76624	0.79034	0.5561
PseAAC[26]	0.84529	0.81837	0.80844	0.83253	0.6405
ProtoPred_RF[24]	0.96969	0.95784	0.93939	0.98058	0.9191
CNN_ATT Model	0.966	0.9596	0.9491	0.9723	0.921
BiLSTM_ATT Model	0.9736	0.9686	0.9578	0.9815	0.939
BiGRU_ATT Model	0.9774	0.9727	0.966	0.9802	0.9461

From the Table I, it is found that, the BIGRU_ATT achieves +0.01 higher and BiLSTM_ATT Model achieves +0.01 higher accuracy than ProtoPred_RFmodel and CNN_ATT achieves same accuracy as ProtoPred_RF model.

The BIGRU_ATT and BiLSTM_ATTachieves same Recall as ProtoPred_RF model, and CNN_ATT receives -0.01 lesser precision than the ProtoPred_RF model. The BIGRU_ATT achieves +0.03 higher, BiLSTM_ATT achieves +0.02 higher, and CNN_ATT achieves +0.01 higher Mcc than ProtoPred_RF model.

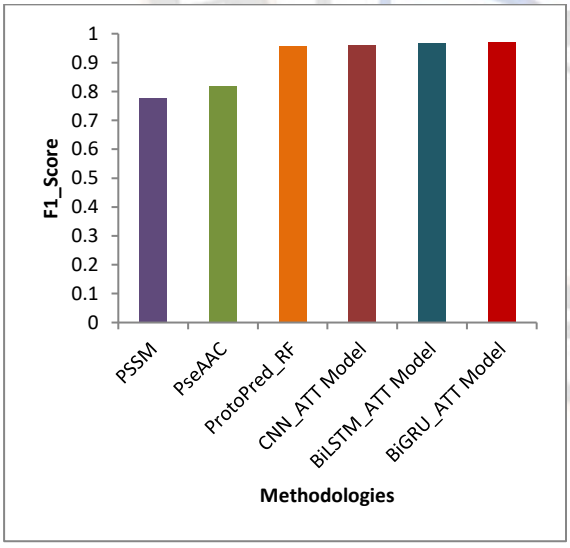

Figure 3. F1_Score comparison of ProtoPred_RF, CNN_ATT, BILSTM_ATT, and BIGRU_ATT

Figure 3 demonstrates that proposed methods CNN_ATT, BILSTM_ATT and BIGRU_ATT models performs better than ProtoPred_RF in terms of F1_Score. The BIGRU_ATT Model achieves +0.02 higher, BiLSTM_ATT achieves +0.01 higher F1_Score than ProtoPred_RF and CNN_ATT achieves same F1_Score as ProtoPred_RF model.

The benchmark data set is split into k(10) disjoint fold partitions for cross-validation. Table 2 and 3 displays the findings of the Kfold testing. Md1 stands for ProtoPred_RF, Md2 for CNN_ATT, Md3 for BILSTM_ATT, Md4 for BIGRU_ATT models that are represented in the Table II.

TABLE II. FINDINGS OF THE KFOLD TESTING FOR PRECISION AND RECALL

Fold #	Precision				Recall			
	Md1	*Md2*	*Md3*	*Md4*	*Md1*	*Md2*	*Md3*	*Md4*
1	96	93	96	100	90	89	90	100
2	98	97	98	98	96	94	96	96
3	96	96	96	93	90	90	90	89
4	98	98	96	98	96	96	95	96
5	95	96	96	95	88	90	90	88
6	99	99	97	97	98	98	97	97
7	96	97	97	97	90	92	92	92
8	97	97	98	98	94	94	96	96
9	99	99	99	97	98	98	98	97
10	100	100	100	100	100	100	100	100
Average	97	97	97	97	94	94	94	95

TABLE III. FINDINGS OF THE KFOLD TESTING FOR PRECISION AND RECALL

Fold #	Accuracy				F1_Score			
	Md1	*Md2*	*Md3*	*Md4*	*Md1*	*Md2*	*Md3*	*Md4*
1	94	93	94	1	92	91	92	1
2	97	96	97	97	97	95	97	97
3	94	94	94	93	92	92	92	91
4	97	97	96	97	97	97	95	97
5	93	94	94	93	91	92	92	91
6	98	98	97	97	98	98	97	97
7	94	95	95	95	92	94	94	94
8	96	96	97	97	95	95	97	97
9	98	98	98	97	98	98	98	97
10	100	100	100	100	100	100	100	100
Average	96	96	96	97	95	95	95	96

In our proposed model, Md4 obtains 2% higher F1_score than Md1, but only 3% more precision than Md1. Additionally, it outperforms Md1 in accuracy by 1% and

F1_score by 3%. The graphical comparison of the true positive rate and false positive rate for each of the approaches is shown in Figure 4.

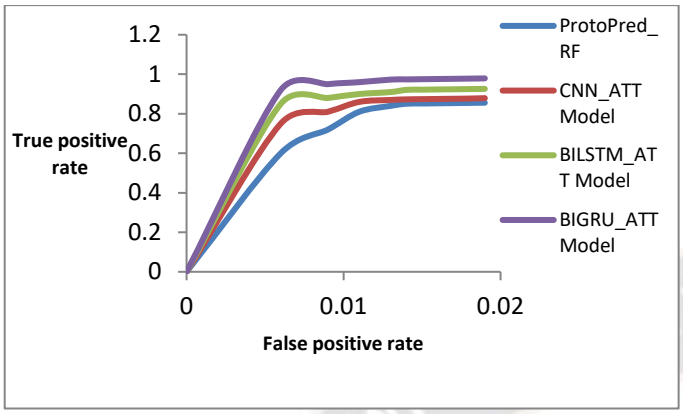

Figure 4. Comparative Analysis of proposed method with previous Models

V. CONCLUSION

Researchers are creating individualized and sophisticated ways for predicting the prognosis of cancer. This type of prediction includes the identification of proto-oncogene proteins. The proposed strategy includes each and every current idea in order to construct a computationally intelligent predictor. From the well-known Uniprot database, which has robust and non-homologous data that has been experimentally supported. We proposed three methods such as CNN_ATT, BILSTM_ATT, and BIGRU_ATT models which combines the strength of convolutional features and temporal features along with attention concept to find out the proto-oncogene. The results of the testing using the independent technique and kfold were obtained and show the proposed BIGRU_ATT model greatly outperforms the competition in terms of Accuracy, Precision, Recall, Mcc, and F1_score.

REFERENCES

[1] Williams, D.E., Eisenman, J., Baird, A., Rauch, C., Van Ness, K., March, C.J., Park, L.S., Martin, U., Mochizukl, D.Y., and Boswell, H.S.: ‘‘Identification of a ligand for the c-kit proto-oncogene’’, Cell,vol .63, no. 1, pp. 167-174,1990,doi:10.1016/0092-8674(90)90297-R.

[2] Cooper G M .‘‘Oncogenes”, II-nd edition. *Jones and Bartlett Publishers Inc*. Boston,pp.384,1995.

[3] H. Cong, M. Zhang, Q. Zhang et al., “Analysis of structures and epitopes of surface antigen glycoproteins expressed in bradyzoites of Toxoplasma gondii,” *BioMed Research International*, Article ID 165342, pp 9, 2013. doi: 10.1155/2013/16534.

[4] M. Kulmanov, M. A. Khan, and R. Hoehndorf, “Deepgo: Predicting protein functions from sequence and interactions using a deep ontology-aware classifier,” arXiv preprint arXiv:1705.05919, vol.34.pp.660-668.2017. doi:10.1093/bioinformatics/btx624.

[5] Y. Loewenstein, D. Raimondo, O. C. Redfern, J. Watson, D. Frishman, M. Linial, C. Orengo, J. Thornton, and A. Tramontano, “Protein function annotation by homology-based inference,” *Genome biology*, vol. 10, no. 2, pp. 207, 2009.doi:10.1186/gb-2009-10-2-207.

[6] P. Gaudet, M. S. Livstone, S. E. Lewis, and P. D. Thomas, “Phylogenetic-based propagation of functional annotations within the gene ontology consortium,” *Briefings in bioinformatics*, vol. 12, no. 5, pp. 449–462, 2011. doi: 10.1093/bib/bbr042.

[7] M. Costanzo, B. VanderSluis, E. N. Koch, A. Baryshnikova, C. Pons, G. Tan, W. Wang, M. Usaj, J. Hanchard, S. D. Lee et al., “A global genetic interaction network maps a wiring diagram of cellular function,” Science, vol. 353, no. 6306, pp.1420, 2016. doi: 10.1126/science.aaf1420.

[8] R. Sharan, I. Ulitsky, and R. Shamir, “Network-based prediction of protein function,” *Molecular systems biology*, vol. 3, no. 1, pp. 88, 2007. doi: 10.1038/msb4100129.

[9] J. Konc, M. Hodošcek, M. Ogrizek, J. T. Konc, and D. Janeži ˇ c, “Structure-based function ˇ prediction of uncharacterized protein using binding sites comparison,” *PLoS computational biology*, vol. 9, no. 11, p. e1003341, 2013.doi:10.1371/journal.pcbi.1003341.

[10] A. Sokolov and A. Ben-Hur, “Hierarchical classification of gene ontology terms using the GOstruct method,” *Journal of bioinformatics and computational biology*, vol. 8, no. 02, pp. 357–376, 2010. doi: 10.1142/s0219720010004744.

[11] A. Tavanaei, A. S. Maida, A. Kaniymattam, and R. Loganantharaj, “Towards recognition of protein function based on its structure using deep convolutional networks,” in Bioinformatics and Biomedicine (BIBM), IEEE International Conference on. IEEE, pp. 145–149,2016. doi: 10.1109/BIBM.2016.7822509.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” *in Advances in neural information processing systems*, vol.60,no.6.pp. 1097–1105,2012.doi:10.1145/3065386.

[13] J. Machado, A. C. Costa, and M. Quelhas, “Can power laws help us understand gene and proteome information?” *Advances in Mathematical Physics*, vol. 2013, Article ID 917153, pp.10 , 2013.doi:10.1155/2013/917153.

[14] R.A. Rensink, “The dynamic representation of scenes”, *Visual Cogn*. Vol.7. pp. 17–42,2000.doi:10.1080/135062800394667.

[15] M. Corbetta, G.L. Shulman, “Control of goal-directed and stimulus-driven attention in the brain”, Nat. Rev. Neurosci. Vol.3,no.3 ,pp. 201–215,2002.doi: 10.1038/nrn755.

[16] J.K. Tsotsos, S.M. Culhane, W.Y.K. Wai, Y. Lai, N. Davis, F. Nuflo, “Modeling visual attention via selective tuning”, *Artif. Intell.* Vol.78 ,pp.507–545,1995.doi:10.1016/0004-3702(95)00025-9.

[17] S.Hochreiter, J. Schmidhuber, “Long short-term memory Neural” Computer. vol.9, no.8.pp.1735–1780,1997.doi:10.1162/neco.1997.9.8.1735.

[18] K. Cho, B. van Merrienboer, Ç. Gülçehre, D. Bahdanau, F.

Bougares, H. Schwenk, Y. Bengio, "Learning phrase representations using RNN encoderdecoder",2014,doi :10.3115/v1/D14-1179.

[19] Bahdanau, Dzmitry&Cho,Kyunghyun&Bengio, Y." Neural Machine Translation by Jointly Learning to Align and Translate".https://doi.org/10.48550/arXiv.1409.0473.

[20] Hochreiter, S. The Vanishing Gradient Problem During Learning Recurrent Neural Nets and Problem Solutions. Int. J. Uncertain. Fuzziness Knowl.-Based Syst. 1998, Vol. 6, pp.107–116.

[21] Wang, H.; Chen, S.; Xu, F.; Jin, Y. Application of deep-learning algorithms to mstar data. In Proceedings of the 2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Milan, Italy, 26–31 July 2015; pp. 3743–3745.

[22] Lin, G.; Shen, W. Research on convolutional neural network based on improved Relu piecewise activation function. ProcediaComput. Sci. 2018, Vol. 131, 977–984.

[23] Meng, Saisi, et al. "Performance Evaluation of Channel Decoder based on Recurrent Neural Network." Journal of Physics: Conference Series. Vol. 1438. No. 1. IOP Publishing, 2020.

[24] S. J. Malebary, R. Khan and Y. D. Khan, "ProtoPred: Advancing Oncological Research through Identification of Proto-Oncogene Proteins," in IEEE Access, vol. 9, pp. 68788-68797, 2021, doi: 10.1109/ACCESS.2021.3076448

[25] Delorenzi, M., and Speed, T.: 'An HMM model for coiled-coil domains and a comparison with PSSM-based predictions', Bioinformatics, Vol. 18, no 4, pp. 617-625,2002.

[26] Jia, J., Liu, Z., Xiao, X., Liu, B., and Chou, K.-C.: 'iSuc-PseOpt: identifying lysine succinylation sites in proteins by incorporating sequence coupling effects into pseudo components and optimizing imbalanced training dataset', Analytical biochemistry, Vol. 497, pp. 48-56,2002.