IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Issue-2, Feb 2019 E-ISSN: 2347-2693

Classification of Breast Cancer Proteins Using DRNN Method

B Madhav Rao[1*], V Srinivasa Rao[2], K Srinivasa Rao[3]

[1]Rayalaseema University, Kurnool, India
[2,3]Department of CSE, V.R .Siddhartha Engineering College, Vijayawada 520007, India

[*]*Corresponding Author: madhavraob@gmail.com, Tel.: 9949582856*

DOI: https://doi.org/10.26438/ijcse/v7i2.106109 | Available online at: www.ijcseonline.org

Accepted: 13/Feb/2019, Published: 28/Feb/2019

Abstract- Classification of large amount data is one of the major difficult tasks in data science. This problem can be solved by using deep learning techniques like CNN and RNN. In computational bio informatics, protein sequence classification plays a crucial role to determine the accuracy. The proposed approach uses the RNN based architecture with GRU, LSTM, and basic LSTM and find the accuracy of training data and testing data by considering mean value of three methods. In this method the top fifteen proteins which are obtained by using preprocessing and sequence analyzer methods as one set of input and TCGA breast cancer dataset as second input to this proposed method. Every sequence in test dataset will compare with sequences in train dataset to get accurate classification results. Supervised learning requires complete labeled data where as unsupervised learning requires unlabelled data. In this approach semi supervised learning is used to get high throughput.

Keywords --- Deep Learning, Accuracy, Protein Sequence, Classification.

I. INTRODUCTION

Traditional machine learning algorithms succeeded in classifying of very small scale data, but failed to handle large amount of data. Hence, a new approach is introduced to handle large datasets. Various approaches have been developed to solve the protein classification problem. Most of the approaches make the model of protein families directly or indirectly. In this process, neural networks are the best option to classify the sequence datasets in large scale. CNN[9][2] (convolutional neural networks) is more suitable for classifying the images and objects where as RNN(recurrent neural networks) is used to handle large scale of data. In this approach, specifically DRNN (dynamic recurrent neural networks) method is proposed and it uses RNN and tensor flow methods to classify the protein sequences [5]. Eventually it produces accuracy.

In section I contains introduction to the protein classification, most used classification method in deep learning, section II contains related work of protein classification existing methods and their limitations, section III contains input datasets, section IV contains proposed work and step by step procedure to implement process of DRNN, section V contains Conclusion and future scope for this proposed work.

II. RELATED WORK

Most of the classification techniques use neural networks [11]. The features bio vector and protein vector are extracted from protein sequence and these factors are used for classifying different structures for which deep neural network [4][5] techniques can be used for finding the dimensional vectors. Some may use supervised learning and some may use unsupervised learning mechanism. In Supervised learning, complete labeled data set is needed to classify the sequences where as in unsupervised learning, unlabelled dataset is used to classify the sequences. Different types of features are extracted for performing the prediction by using machine learning methods such as KNN [11][1], Decision tree[6] and random forest[4].

The outputs of these methods are taken for protein classification. The sequence transformation of the various proteins are applied using set functions for analyzing the accuracy [10]. The classes are classified based on the advanced level of deep learning techniques by comparing with the old techniques using Back propagation neural networks.

II.I. MACHINE LEARNING

The data mining techniques consist of many classifiers such as Naive Bayes, SVM, random forest. They are easily scalable with linear parameters such as preprocessing and predictions which make us to learn about the problem. They are very efficient classifiers

with high level of accuracy. They take different attributes as parameters which continue and perform the maximum training by evaluating expression at a given liner time. These algorithms have capability of assign the label classes and identifying the problem instances called as unsupervised machine learning. The combination of supervised and unsupervised machine learning produces semi supervised learning. It results the high efficiency and accuracy. These are very advantageous to the users since they use little amount of training data.

II.II DEEP LEARNING TECHNIQUE

Deep learning is one of the most accurate model structure which can get feature representation from raw input. This DRNN architecture allows us to learn more complex structures of RNN models [15]. At the same time it can reduce error cost of the machine learning tools from a given nonlinear input. This uses sequence data in the internal memory to store the input information which generates the gradient error issues in long term. LSTM blocks are the main unit of building the layers of DRNN. This can handle the gradient error issues. There are different encoders and decoders which are responsible for memory to access the input and output gates since they take fixed weights.

II.III. LSTM (Long Short Term Memory)

A popular variant of RNN is LSTM[12]. It consists of memory to store the previously processed data along with repeated sequences to process the long sequences. The architecture of LSTM is shown in Figure 1.

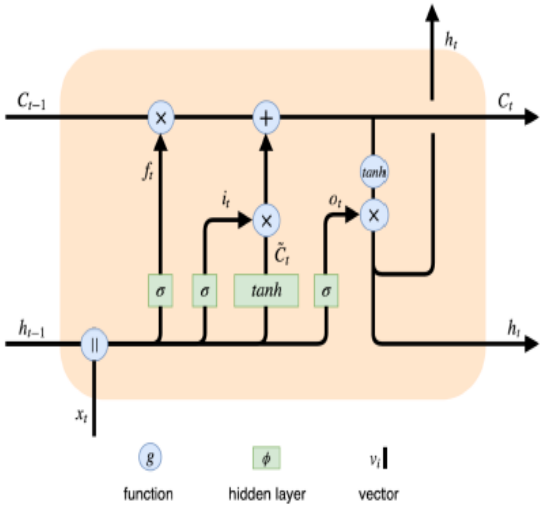

Figure 1. LSTM Architecture

In LSTM architecture, hidden layers give output as weight parameters like previous hidden state and cell state Ct . Input (it), Output (ot) and forget (ft) gates are responsible for changes in the cell state and hidden state(h_t).

$$f_t=\sigma(W_f . x_t+U_f.h_{t-1})$$

$$i_t= \sigma(W_i. x_t+U_f.h_{t-1})$$
$$\tilde{C}_t=\tan h(W_c.x_t+U_c.h_{t-1})$$
$$O_t= \sigma(W_o.x_t+U_o.h_{t-1})$$
$$C_t=f_t \text{ X } C_{t-1} +i_t \text{ X } \tilde{C}_t$$
$$h_t=o_t \text{ X } \tanh(C_t)$$

Dot(.) is the dot product, X is element wise product and $\sigma(x)=\frac{1}{1+exp(-x)}$ is the sigmoid function. Here U and W are different weight parameters for each hidden layer in the LSTM. There is no weight parameter consideration after the hidden layers.

II.IV. GRU

A gated recurrent unit (GRU)[14] was proposed in the year 2014 by Cho et al. Each recurrent unit adaptively captures the dependencies of different time scales. The graphical representation of GRU shown in Figure 2.

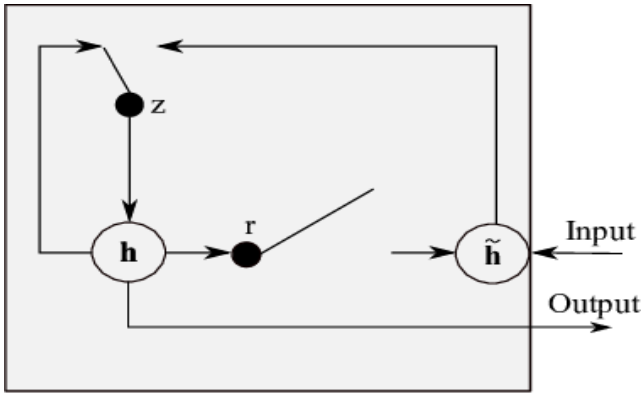

Figure 2. GRU architecture

The following are the GRU calculation functions.

$$r_t = \sigma(W_r x_t + U_r h_{t-1})$$
$$z_t = \sigma(W_z x_t + U_z h_{t-1})$$
$$\hat{h}_t = \tanh(W x_t + U(r_t \odot h_{t-1}))$$
$$h_t = (1 - z_t)h_{t-1} + z_t \hat{h}_t$$

Where $\sigma()$ is sigmoid function, (.) dot is element wise multiplication, z_t is update gate and r_t is reset gate and h denotes hidden layer.

III. DATA SET

785 TCGA breast cancer protein sequences and pruned protein sequences [3] are considered as an input datasets for the proposed DRNN .

Table 1. Top 15 proteins which are obtained by using sequence analyze

CANONICAL NAME	PROTEIN NAME	DISEASE SCORE	DIFFUSION OUTPUT HEAT	DIFFUSION OUTPUT RANK	DIFFUSION OUTPUT HEAT	DATABASE IDENTIFIER
P08183	ABCB1	5	0.943590371	13	0.943590371	9606.ENSP00000324856
Q99728	BARD1	5	0.980142536	8	0.980142536	9606.ENSP00000372088
P38398	BRCA1	5	0.981078564	6	0.981078564	9606.ENSP00000418960
P51587	BRCA2	5	0.98125451	3	0.98125451	9606.ENSP00000343392
Q9BX63	BRIP1	3.509493	0.98125451	5	0.98125451	9606.ENSP00000419471
P11802	CDK4	5	0.980142536	7	0.980142536	9606.ENSP00000336701
P07339	CTSD	2.590049	0.993062461	2	0.993062461	9606.ENSP00000265724
O15151	MDM4	2.757025	0.967160887	12	0.967160887	9606.ENSP00000236671
Q86YC2	PALB2	5	0.973790986	11	0.973790986	9606.ENSP00000369497
E9PI54	RAD51	5	1.0	1	1.0	9606.ENSP00000219746
Q86U92	RAD51B	3.574687	0.979811108	9	0.979811108	9606.ENSP00000356150
O43502	RAD51C	5	0.976763781	10	0.976763781	9606.ENSP00000259008
Q15831	STK11	5	0.98125451	4	0.98125451	9606.ENSP00000352271
O15405	TOX3	2.546027	0.942796205	14	0.942796205	9606.ENSP00000257904
O43542	XRCC3	5	0.936445107	15	0.936445107	9606.ENSP00000260947

Breast cancer dataset imported from TCGA live database are presented below.

```
Hybridization REF   TCGA-AO-A03P-01A-11R-A00Z-07
Composite Element REF   log2 lowess normalized (cy5/cy3) collapsed by protein symbol
MMP2    -0.148666666666667
C10orf90    -1.955
ZHX3    0.210833333333333
ERCC5   0.42975
GPR98   -0.970875
RXFP3   0.523
APBB2   1.458
PRO0478 -1.234
KLHL13  -3.43475
PRSSL1  -0.760333333333333
PDCL3   0.537333333333333
DECR1   -0.46275
SALL1   -3.7105
CADM4   -0.737166666666667
RPS18   0.603071428571429
HNRPD   -0.758833333333333
CFHR5   0.3075
SLC10A7 0.168625
OR2K2   1.047
LMAN1   0.0475
SUHW1   -2.38925
CHD8    0.8762
SUMO1   0.2691
GP1BA   -0.1218
DDB1    0.1055
```

Top pruned breast cancer causing proteins are shown in table 1. Figure 4 showes the TCGA protein dataset.

IV. METHODOLOGY

This proposed approach considers the parameter default layers count equal to the value 7. For calculating accuracy, this method uses tensor flow methods, GRU, and LSTM. This DRNN takes the input dataset which contains 785 protein sequences and it can be considered as a train dataset. This protein data set is compared with pruned dataset of top 15 protein sequences to get the accuracy. These top pruned protein sequences are test dataset.

IV.I ALGORITHMIC STEPS

1. Read the input breast cancer protein sequence dataset from TCGA, and protein Sequences which are obtained from sequence analyzer.
2. Initialize the basic parameters: number of layers=7, learning rate=0.001, protein sequence Length=40, Number of training epochs=20.
3. Consider TCGA dataset as a training dataset and 15 proteins as test dataset which are pruned using sequence analyzer.
4. Pass the input datasets to tensor flow methods.
5. Obtain the test accuracy and train accuracy.
6. Repeat step 4 and 5 with different approaches like GRU, LSTM, basicLSTM.
7. Calculate the mean value of 3 methods and display the accuracy.

V. RESULTS AND DISCUSSION

This method implements all the experiments using deep learning libraries such as tensor flow techniques, KNN models. A 10 cross fold validation has been implemented for getting the accuracy in DRNN model. The results are represented in the table2.

DRNN model implements 10 cross fold validation to analyze GRU and LSTM with range of 68 and 128 blocks with a learning rate 0.001. Our experimental results show RNN with LSTM which gives high accuracy . This proposed approach calculates the mean value of 3 implemented methods

$$\text{Accuracy} = \frac{\sum_{k=1}^{n} \mathrm{X(k)}}{n} \qquad \text{---------- (1)}$$

Where 'n' is total number of methods implemented. 'k' is constant value from 1… n and x(k) is accuracy of the current method. By using the formula (1) ,this DRNN approach produces the train accuracy 0.98 and test accuracy 0.95.

VI. CONCLUSION AND FUTURE SCOPE

In this proposed approach, the deep learning technique of RNN based architecture is used to classify the protein sequences and results with the accuracy of the given input dataset. For high throughput, this proposed DRNN method considers the mean of all methods implemented. This method produces the accuracy of 0.95. The study says that the pruned 15 protein sequences are the top most proteins which cause breast cancer. This research is more useful for inventing the drug and basis for future disease cause protein identification.

REFERENCES

[1] A. Bhola, S. K. Yadav, A. K. Tiwari, Machine Learning Based Approach For Protein Function Prediction Using Sequence Derived Properties, International Journal Of Computer Applications 105 (12).

[2] ABDALRAOUF HASSAN,AUSIF MAHMOOD," Convolutional Recurrent Deep Learning Model For Sentence Classification", VOLUME XX, 2017, 2169-3536,IEEE,2017

[3] B Madhav Rao, V Srinivasa Rao," Preprocessing Of Breast Cancer Protein Expressions Using Correlation Co-Efficient Factors", JASC: Journal Of Applied Science And Computations, Volume VI, Issue I, January/2019,Pp. 2198-2207,2019

[4] Chaitanya Gupte and Shruti Gadewar, "Diagnosis of Parkinson's Disease using Acoustic Analysis of Voice", International Journal of Scientific Research in Network Security and Communication, Vol 5, Issue 3, pp.14-18, June 2017

[5] Dr. S.Vijayarani And Ms. S.Deepa," PROTEIN SEQUENCE CLASSIFICATION IN DATA MINING– A STUDY", International Journal Of Information Technology, Modeling And Computing (IJITMC) Vol. 2, No.2, May 2014

[6] Deepika mallampati,"An efficient spam filtering using supervised learning techniques", International Journal of Scientific Research in Science and Engineering, vol.6 ,issue.2 , pp. 33-37, April 2018.

[7] H. Nielsen, S. Brunak, G. Von Heijne, Machine Learning Approaches For The Prediction Of Signal Peptides And Other Protein Sorting Signals, Protein Engineering 12 (1) Pp:3–9, 1999.

[8] I. Krasteva, N. Inglis, F. Sacchini, R. Nicholas, R. Ayling, C. Churchward, Et Al., Proteomic Characterisation Of Two Strains Of Mycoplasma Mycoides Subsp. Mycoides Of Differing Pathogenicity, J Proteomics Bioinform S 13, 2014 .

[9]Matthew D. Zeiler And Rob Fergus," Visualizing And Understanding Convolutional Networks", D. Fleet Et Al. (Eds.): ECCV 2014, Part I, LNCS 8689, Pp. 818–833, 2014. ,Springer International Publishing Switzerland 2014.

[10] P. Johansson, M. Ringner, Classification Of Genomic And Pro- ´ Teomic Data Using Support Vector Machines, In: Fundamentals Of Data Mining In Genomics And Proteomics, Springer, Pp. 187–202, 2007.

[11] S. Saha, R. Chaki, Application Of Data Mining In Protein Sequence Classification, Arxiv Preprint Arxiv:1211.4654.

[12] Sofia Edström Josefin Ondrus," Sequence Classification Applied To User Log Data",Pp.9-12,2016,SOFIA EDSTRÖM, JOSEFIN ONDRUS, June 2017.

[13] Timothy K. Lee, Tuan Nguyen," Protein Family Classification With Neural Networks", Pp: 1-9.

[14] Xingyou Wang , Weijie Jiang , Zhiyong Luo," Combination Of Convolutional And Recurrent Neural Network For Sentiment Analysis Of Short Texts", Proceedings Of COLING 2016, The 26th International Conference On Computational Linguistics: Technical Papers, Pages 2428–2437, Osaka, Japan, December 11-17 2016.

[15]Y. Lecun, Y. Bengio, G. Hinton, "Deep Learning", Nature International Journal of Science,521 (7553) (2015) pp. 436-444, 2015.

