International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10s
DOI: https://doi.org/10.17762/ijritcc.v11i10s.7693
***Article Received:** 08 June 2023 **Revised:** 27 July 2023 **Accepted:** 17 August 2023*

Colorectal Cancer Classification from Protein Sequences Using Several RNN Pre-Trained Models

Madhav Rao B[1], Kunjam Nageswara Rao[2]
[1]Research Scholar, Department of CS & SE, A.U College of Engineering(A),
Andhra University, Vizag,India
e-mail: madhavraob@gmail.com
[2]Professor, Department of CS & SE, A.U.College of Engineering(A),
Andhra University, Vizag, India
e-mail:kunjamnag@gmail.com

Abstract— Bioinformatics is one field that can integrate health information with data mining applications in order to predict or analyze the patient's information. One among the several diseases is colorectal cancer,which is horrible and rated the third leading disease in the area of cancer which leads to death in both men and women.In general, there are several intelligent methods to identify several kinds of problems in gene selection for predicting cancer,but there is no method to give a solution for patients who are diagnosed in the advanced stage.This motivated me to design theproposed approach in which we try to identify all the homo protein sequences and then train these sequences corresponding to one which causes colorectal cancer.InBioinformatics,a protein acts as one of the main agent or source to perform a biological function by interacting with molecules like Deoxyribonucleic acid (DNA), Ribonucleic acid (RNA). The function of a protein determines the healthy or diseased states of an organism. Protein interaction with other proteins can be visualized through a network called Protein-Protein Interaction Network (PPIN).In general classification of protein sequences is a very complex task. Deep learning techniques like CNN and RNN can be used to solve the problem.In computational bioinformatics,the classification of protein sequence plays an important role in determining accuracy.To improve the accuracy of our current model, the suggested method incorporates GRU, LSTM, RNN, and Customized LSTM into an RNN based architecture by optimizing the parameters in a two-way direction.Here we try to test all the models on sample protein sequences that are collected from TCGA and then determine the correctness of testing data and training data.

Keywords-Deep Learning, Deoxyribo Nucleic Acid, Ribonucleic Acid, Bioinformatics, Protein-Protein Interaction Network, Sequence Classification.

I. INTRODUCTION

The capacity to extract relevant information from a vast number of data sets is known as data mining. This will help in transforming the huge amount of data into different rules and patterns,which are further used for several fields like pattern recognition, optimization of statistical information, and Bioinformatics. It is also called Knowledge Discovery in Databases (KDD)[8]. This is mainly derived as the way to find out the best associations and patterns from the large data. In general, the mining of data is not exact to any business industry. It involves smart innovations and the motivation to investigate the chance of hidden facts that exist in the information. The knowledge extracted from data mining helps the user in the procedure of decision making. Nowadays there is a lot of scope for medical data mining for learning hidden patterns from the input dataset and trying to explore all the hidden truths for medical field improvement.

A. Bioinformatics

Bioinformatics is the art of integrating the Information Technology domain into biomedical data to extract knowledge in the area of computer science and biology. Bioinformatics and biomedical informatics are two main domains that are becoming popular in several fields like applied mathematics, biology and health care technology to grab more improvements in the medical domain. In general, the biological dataset is having a huge volume of data, there is a basic necessityto construct,develop,design and test some novel methods to obtain the desired results.There are several pre-trained models in the IT department to work on the huge volume of data, but most of them have overfitting problems when applied toolarge dataset.Hence the main fundamental principle in medical bioinformatics is to create a unique model by optimizing the parameters and enhancing the model's accuracy over a set of pre-trained models.

Since there was a huge amount of biological data available on the various websites,most of the data is associated with noisy and irrelevant sample information.If the information is not relevant or properly pre-processed, the model will not generate accurate results. The ability to choose the best model plays a vital role in dealing with the unknown

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10s
DOI: https://doi.org/10.17762/ijritcc.v11i10s.7693
***Article Received:** 08 June 2023 **Revised:** 27 July 2023 **Accepted:** 17 August 2023*

factors of the neural networks, for predicting the diseases[2].In our proposed work we try to gather raw protein sequences from the TCGA database as one input to compare 17 protein sequences which comes from diffusion as second input[11][7].Once the data pre-processing is applied all the missing or irrelevant sequences are removed and only those which have correct information are collected and assigned with some family id.

B. Data mining in bioinformatics:

a) The capacity to extract relevant information from a vast number of data sets is known as data mining. This is converted into patterns and rules by extracting the most useful information from that large data set.In bioinformatics, there is a huge data that requires many mining applications for the analysis of patient's diseases. The data mining domain is recently launched in the medical domain for diagnosis of cancer based on protein sequence and to warn the patient by explaining the risk factors. Figure-1 shows different Data Mining Techniques used in the Field of bioinformatics.

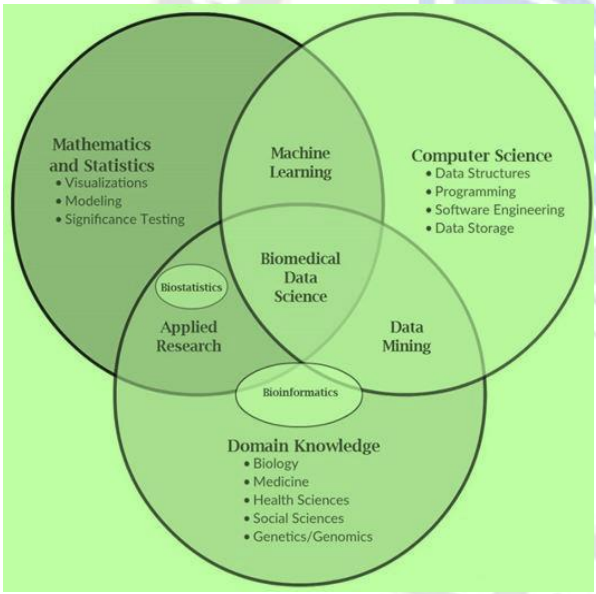

Figure 1. Represents the Data Mining Techniques in the Field of bioinformatics

II. MATERIALS AND METHODS

This section focuses on the background research that has been done in order to construct the proposed method for protein accuracy classification.

A. Classification:

Classification is defined as one of the main task in the area of data mining,which is used for discovering new classes from ambiguous data. This acts as an interface that can assemble a large volume of data based on features.A group of objects is identified as one group with distinct attributes and functionality is present.Classification technique is mainly used to form one function to enhance the classes.In order to train a system,one should identify all the important fields present in the dataset and they need to classify them into separate clusters and train the model on several parameters.

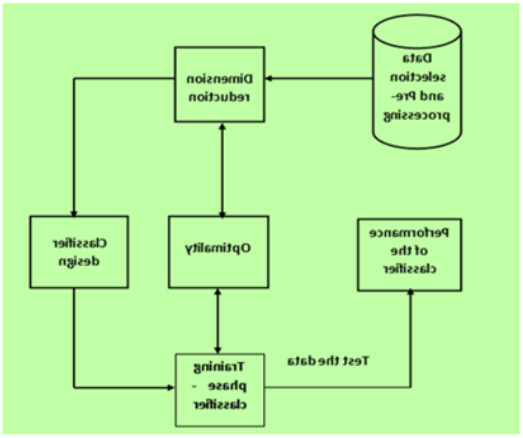

Figure 2 Stages in data classification.

From the above figure 2,we can identify several stages which are present in data classification[2].Initially, we try to select input data and apply the data pre-processing technique on that input data.Now the data is pre-processed and dimensions are reduced in order to optimize the training classifier to increase the performance of proposed models.Ingeneral,there are several methods available for classifying the data such as Bayesian Classifier,Rule Based Classifier,Support Vector Machine and a lot more.In order to apply any mining algorithm,we need to apply pre-processed data by removing all irrelevant or redundant information.In order to remove such irrelevant information, we need to apply feature selection methods to increase the accuracy of the classifier.In general, the classification technique can be applied to hundreds of applications ranging from small scale to real world applications such as medical data,textclassification,data classification and a lot more.In our current application, we try to classify the homo protein sequences and then train these sequences corresponding to one which causes colorectal cancer.

B. Applications of disease classification:

In this section, we try to discuss several applications which come under disease classification especially colorectal cancer identification.Normally the cancer disease is classified based on two techniques like:

(1) Imaging Technique.
(2) Classification of Genetic Data[10].

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10s
DOI: https://doi.org/10.17762/ijritcc.v11i10s.7693
***Article Received:** 08 June 2023 **Revised:** 27 July 2023 **Accepted:** 17 August 2023*

Compared with the above two methods,classification of genetic data is mostly used for many years and this can able to predict the patient based on his genetic data.Normally this process is applied based on filter based and wrapper based techniques.First, if we look across filter basedtechniques,this technique will gather all the gene data from the database and extract top ranked genetic data, and then test the corresponding patient data with that top ranked data in order to classify the record.Next, if we look at wrapper based technique in this we can use several pre-defined classifiers and then check the type of appropriate gene.In general, using filter based classification technique gives accurate results in very less time and with more accuracy[1].The wrapper based techniques havesome limitations compared to filter based techniques in terms of dimensionality problems (i.e If the dimensions are more or less compared with predefined classification technique,this may not generate the result accurately and hence we need to work more to process the dimensions into normal manner.

In this current work, we try to apply RNN models by optimizing the dimensions in order to reduce the computation cost for protein classification [3][4][9].Here we try to collect the most suitable samples for training the system and we try to eliminate the unnecessary data which are present in our dataset, we only use relevant information to increase the performance of our model.

In this section, we try to discuss several well-known RNN models and then find the accuracy of every individual model for protein accuracy classification [IV].Thewell known protein classification models are as follows:

(1) GRU Model
(2) RNN Model
(3) Basic LSTM Model
(4) Optimized Bi-LSTM Model

(1) Gated Recurrent Unit(GRU) Model

A gated process is used to manage and control the information flow between cells in neural networks through a gated recurrent unit, which is a variation on the RNN design[7]. GRU, which was developed in 2014 by Cho et al., makes it easier to capture dependencies from large sequential data sets without omitting information from the previous segment. This is accomplished through its gated units, which address the standard RNN's exploding/vanishing gradient concerns. These gates regulate the information that must be kept or destroyed at each step. GRU employs two gates: an update gate and a reset gate.The graphical representation of GRU is shown in figure 3.

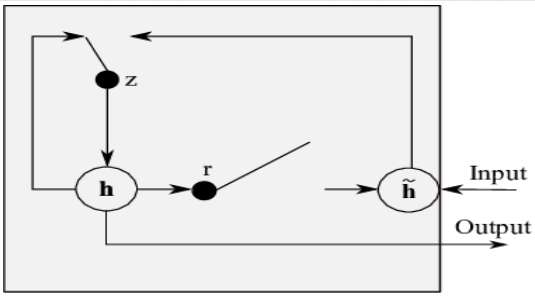

Figure 3.Represent the GRU architecture

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

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

(2) Recurrent Neural Networks (RNN) Model

Recurrent Neural Networks is the most important technique in machine learning applications Feedforward neural networks considers only fixed length data as an input, whereas RNN can handle variable length sequences as input by recurrent hidden state. Each state in RNN depends on the previous state output. The RNN architecture is shown in figure 4. For better understanding consider the example, the input contains four word sentence, then the network forms a 4-layer artificial neural network. For every word, it constructs a layer.

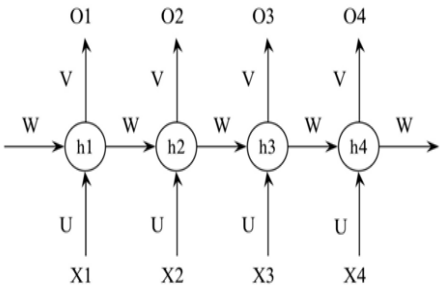

Figure 4. The flow of RNN Model

(3) Basic Long Short Term Memory (LSTM) Model

Long Short Term Memoryisapopular variant of RNN[6]. 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-5.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10s
DOI: https://doi.org/10.17762/ijritcc.v11i10s.7693
***Article Received:** 08 June 2023 **Revised:** 27 July 2023 **Accepted:** 17 August 2023*

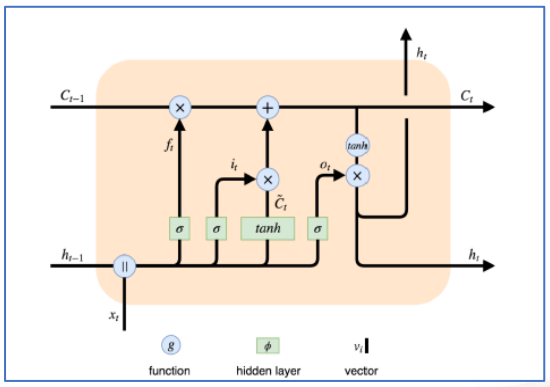

Figure 5. LSTM Architecture

In LSTM architecture, hidden layers give output as weight parameters like previous hidden state and cell state C_t . Forget gate (f_t), Input gate (i_t) andOutput gate (o_t) are responsible for changes in the cell state and hidden state(h_t).

$$f_t = \sigma(W_f \cdot x_t + U_f \cdot h_{t-1})$$
$$i_t = \sigma(W_i \cdot x_t + U_i \cdot h_{t-1})$$
$$\tilde{C}_t = tanh(W_C \cdot x_t + U_C \cdot h_{t-1})$$
$$o_t = \sigma(W_o \cdot x_t + U_o \cdot h_{t-1})$$
$$C_t = f_t \times C_{t-1} + i_t \times \tilde{C}_t$$
$$h_t = o_t \times tanh(C_t).$$

III. PROPOSED METHODOLOGY

The suggested method incorporates GRU, LSTM, RNN, and Customized LSTM into an RNN based architecture by optimizing the parameters in a two-way direction to increase the accuracy of our current model.Here we try to test all the models on sample protein sequences that are collected from TCGA as one input and data obtained from the sequence analyzer as a second input and then find the accuracy of testing and trainingdata[11]. Initially, we collected a raw dataset from the TCGA database and this data has not contained any information related to family id and family types.Hence in order to process that dataset into meaningful data,we try to load the 32469protein sequences which are classified with family sequence id.In general, proteins are made from twenty different kinds of amino acids and every protein differs in structure and function.It is a big challenge for classifying every protein function corresponding to an amino acid sequence in molecular biology.Hence we try to gather the processed protein sequence data from the KAGGLE Pfamdatabase.

On the input dataset, data cleaning techniques were used to extract all of the unique characteristics included in the entire text dataset and attempt to tokenize each with separate tokens, as well as count the number of attributes present in the input dataset. Following feature extraction, the entire dataset is partitioned into two parts: 20% and 80%. We allocate 20% of the data present from the input dataset for testing reasons, while the remaining 80% of data is used for training purposes.

Bidirectional LSTMs are a sort of LSTMs that can be employed in sequence classification issues to improve model performance [5]. Bidirectional LSTMs train two LSTMs on the input sequence instead of on LSTM in instances where all time steps of the input sequence are known. The suggested model, known as BiLSTM is organized in a modular fashion to handle multi-dimensionality concerns that develop as the number of layers or parameters increases. The optimized Bi-LSTM is introduced in addition to the fundamental LSTM model to address the problem with the intrinsic structure, characterizing the operation of multi-state parameters. This model isusedtoselect the appropriate configuration for the proposed regression model, resulting in improved performance. In order to improve protein classification accuracy and efficiency,we try to reduce or optimize some parameters so that the Overfitting problem is removed for the current dataset and we can achieve a high level of accuracy compared with other pre-trained models. Figure-6 shows the architecture of the Bi-LSTM.

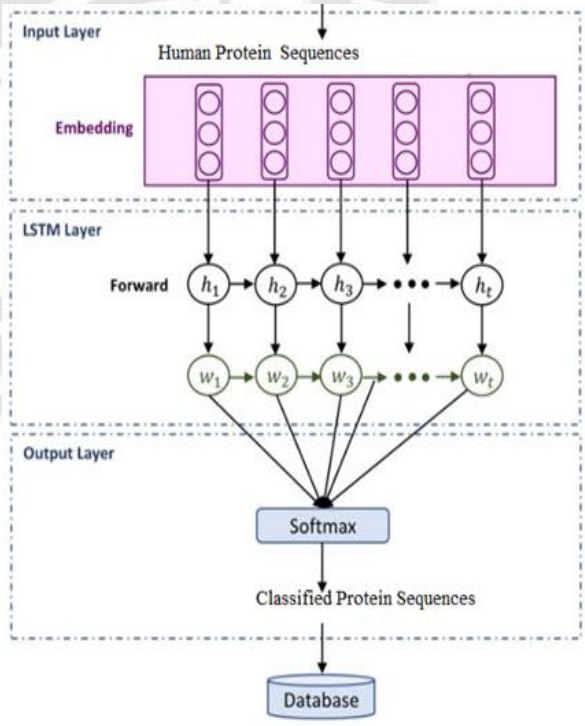

Figure 6. Represents the Optimized Bi-LSTM Architecture

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10s
DOI: https://doi.org/10.17762/ijritcc.v11i10s.7693
***Article Received:** 08 June 2023 **Revised:** 27 July 2023 **Accepted:** 17 August 2023*

The input dataset is loaded first, which is obtained from the STRING database. Nearly 2000 unique live protein sequences gathered from humans make up this input collection. The data was collected in the form of a.csv document. The input sequence.txt file is first converted to a.csv file, which comprises almost 2000 sequences that are used for testing. The dataset consists of a series of numerous columns, each of which has its own set of values, while the main column contains 49 different family ids.

Figure-7 shows the dataset contains multiple attributes like a label with distinct label ids and where every label contains a unique family id.

```
import pandas as pd
import os
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
seq=pd.read_csv('Sequence_data_with_label_v3.o.csv')
seq.head()
```

	Unnamed: 0	Unnamed: 0.1		Sequence	Label
0	0	0	0	MVTEFIFLGLSDSQELQTFLFMLFFVFYGGIVFGNLLIVTVVSDS...	PF13692.6
1	1	1	1	MDGENHSVVSEFLFLGLTHSWEIQLLLLVFSSVLYVASITGNILIV...	PF01255.19
2	2	2	2	MDGENHSVVSEFLFLGLTHSWEIQLLLLVFSSVLYVASITGNILIV...	PF01255.19
3	3	3	3	MWLVFHRPHPRPSWPLRAALGFGRRQSSLRCFPVLPSARPYVSANP...	PF01368.20
4	4	4	4	MSKGILQVHPPICDCPGCRISSPVNRGRLADKRTVALPAARNLKKE...	PF01687.17

Figure 7. Input Dataset

A. Algorithm Foraccuracy Classification

Here we try to find out the step-by-step procedure for protein accuracy classification by comparing it with several RNN models.

1. Read the colorectal cancer protein sequence dataset from TCGA, as well as the proteinSequenceswereobtainedbythe sequence analyzer along with family id's asan input.
2. Set the fundamental parameters, such as attributes=4, protein sequenceMax_Length, Max Number oftraining epochs=22 andPadding Value =100.
3. Consider the TCGA dataset to be the training dataset and the test dataset as 17 proteinsprunedwithsequenceanalyzer.
4. Tensor flow methods are used to process the input datasets.
5. Apply multiple well-known RNN models to obtain test and training accuracy.
6. Steps 4 and 5 should be repeated using alternative approaches such as GRU, RNN,BasicLSTM, andOptimized Bi-LSTM.

IV. RESULTS

In this proposed approach the performance of all four models: GRU, Optimized Bi-Directional LSTM,RNN and Basic LSTM Models, as well as sample human protein sequences collected from the STRING database are evaluated. The findings clearly show that the Optimized Bi-LSTM model has more accurate and efficient than other algorithms in detecting protein classification accuracy.The model is tested using the trained data set to determine its accuracy during training, whereas the model is tested using a new or different data set to determine its accuracy during validation.

A. GRU Model Accuracy

From figure-8, we can calculate the model accuracy as 24 Percent,which is tested on some embedding parameters (100,128).

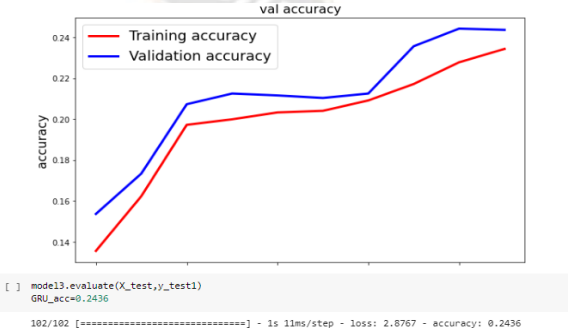

Figure 8. Output graph for GRU Model

B. Basic RNN Model Accuracy:

From figure-9, we can calculate the model accuracy as23 Percent,which is tested on some embedding parameters (100,128).

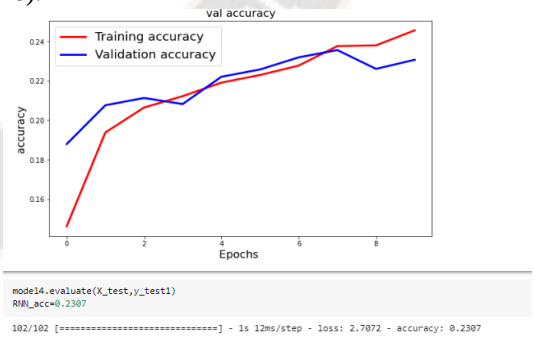

Figure 9. Output graph for RNN Model

C. Basic LSTM Model Model Accuracy:

From the figure-10, we can calculate the model accuracy as 20 Percent,which is tested on some embedding parameters (100,128).

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10s
DOI: https://doi.org/10.17762/ijritcc.v11i10s.7693
***Article Received:** 08 June 2023 **Revised:** 27 July 2023 **Accepted:** 17 August 2023*

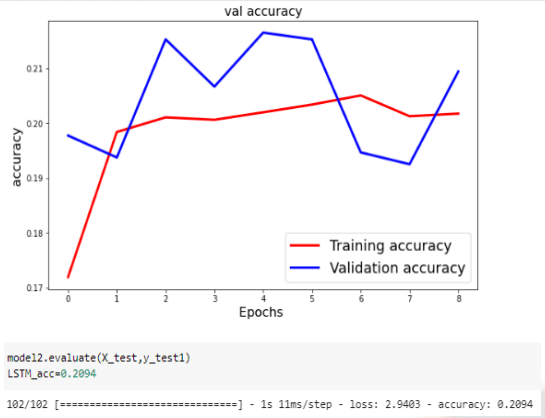

Figure 10. Output graph for BASIC LSTM Model

D. Bi-LSTM Model Accuracy:

From the figure-11, we can calculate the model accuracy as 80 Percent, which is tested on some embedding parameters (100,128) and drop outvalues (None,256) and tested on 49 Family Sequences.

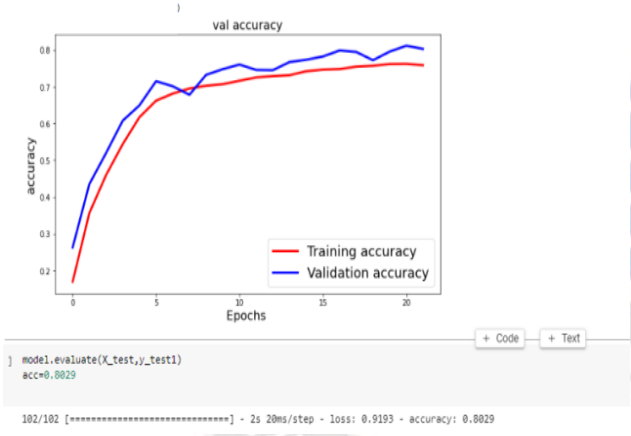

Figure 11.Output graph of Optimized Bi-LSTM Model

Figure-12 shows the validation accuracies of different RNN models.

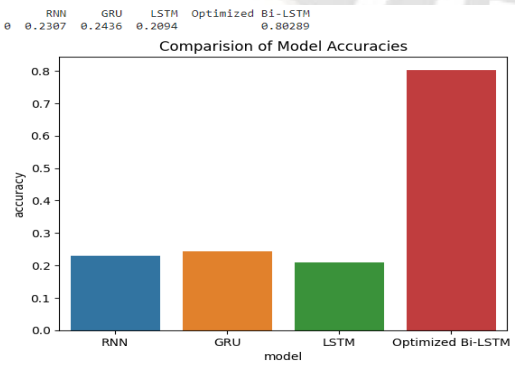

Figure 12.Accuracies of different RNN models.

The following table-1 represents the four model's accuracies and from the Table 1, we can conclude Optimized Bi-LSTM model is very efficient in the effective detection of colorectal cancer protein from given sequences.

TABLE 1. MODELS AND THEIR ACCURACIES.

Model	Training Accuracy	Validation Accuracy
RNN	0.24	0.23
GRU	0.23	0.24
LSTM	0.19	0.20
Optimized Bi-LSTM	0.75	0.80

V. CONCLUSION AND FUTURE ENHANCEMENT

We conducted a comparative examination of multiple pre-trained models necessary for successful protein classification in our proposed approach. In general, a lot of studiesare being done on the subject of Artificial Intelligence, which mostly focuses on recognizingthe state rather than delving into the underlying reasons behind such sequences. There were several flaws in the early approaches for accurate protein classification like overfittingproblems and so on. In this suggested 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 method produces the training accuracy of 0.75 and validation accuracy 0.80 the study says that the pruned 17 protein sequences are the top most proteins which cause colorectal cancer. We aim to mine several types of protein sequences and assign family ids for the corresponding live sequences by using some pre-processed data set which is collected from KAGGLE PfamDatabase.Here we try to apply some well known models such as RNN,GRU,LSTM and Optimized Bi-LSTM. Thesuggested method incorporates GRU,LSTM,RNN, and Customized LSTM into an RNN based architecture by optimizing the parameters in a two-way direction to increase the accuracy of our current model.Here we test all the models on sample protein sequences that are collected from TCGA and then determine the correctness of testing data and training data.we can conclude Optimized Bi-LSTM model is very efficient in the effective detection of colorectal cancer protein from given sequences. To improve the accuracy rate for this Optimized Bi-LSTM model, we intend to incorporate a few more parameters in the future.

REFERENCES

[1] Hatim Z Almarzouki, "Deep-Learning-Based Cancer Profiles Classification Using GeneExpressionData Profile", Journal of Healthcare Engineering, vol. 2022, Article ID 4715998, 2022. https://doi.org/10.1155/2022/4715998.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10s
DOI: https://doi.org/10.17762/ijritcc.v11i10s.7693
***Article Received:** 08 June 2023 **Revised:** 27 July 2023 **Accepted:** 17 August 2023*

[2] Sam Gelman, Sarah A. Fahlberg, Pete Heinzelman, Philip A. Romero, Anthony Gitter,"Neuralnetworks to learn protein sequence–function relationships from deep mutationalscanning data",Proceedings of the National Academy of Sciences Nov 2021,118 (48)e2104878118;DOI: 10.1073/pnas.2104878118.

[3] Tran KA, Kondrashova O, Bradley A, Williams ED, Pearson JV, Waddell N. Deep learningin cancerdiagnosis, prognosis and treatment selection. Genome Med. 2021;13(1):152.Published2021 Sep 27.doi:10.1186/s13073-021-00968-x.

[4] Alzubaidi, L., Zhang, J., Humaidi, A.J. et al. Review of deep learning: concepts, CNNarchitectures,challenges,applications, future directions. J Big Data 8, 53 (2021).https://doi.org/10.1186/s40537-021-00444-8.

[5] Unnam, A. K. ., & Rao, B. S. . (2023). An Extended Clusters Assessment Method with the Multi-Viewpoints for Effective Visualization of Data Partitions. International Journal of Intelligent Systems and Applications in Engineering, 11(1s), 51–56. Retrieved from https://ijisae.org/index.php/IJISAE/article/view/2476

[6] Guo, L., Wang, S., Li, M. et al. Accurate classification of membrane protein types based onsequenceand evolutionary information using deep learning. BMC Bioinformatics 20, 700(2019).https://doi.org/10.1186/s12859-019-3275-6.

[7] Prof. Sharayu Waghmare. (2012). Vedic Multiplier Implementation for High Speed Factorial Computation. International Journal of New Practices in Management and Engineering, 1(04), 01 - 06. Retrieved from http://ijnpme.org/index.php/IJNPME/article/view/8

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

[9] Anna, G., Hernandez, M., García, M., Fernández, M., & González, M. Optimizing Course Recommendations for Engineering Students Using Machine Learning. Kuwait Journal of Machine Learning, 1(1). Retrieved from http://kuwaitjournals.com/index.php/kjml/article/view/104

[10] XingyouWang ,Weijie Jiang , Zhiyong Luo," Combination Of Convolutional And RecurrentNeuralNetwork For Sentiment Analysis Of Short Texts", Proceedings Of COLING 2016,The 26thInternationalConference On Computational Linguistics: Technical Papers, Osaka, Japan,December 11-17 2016, Pages 2428–2437.

[11] Susan P. Imberman," Effective Use of The Kdd Process and Data Mining for ComputerPerformanceProfessionals":https://www.researchgate.net/ publication/ 221445402, NOV-2014.

[12] Sofia Martinez, Machine Learning-based Fraud Detection in Financial Transactions , Machine Learning Applications Conference Proceedings, Vol 1 2021.

[13] Muhammad Javed Iqbal, Ibrahima Faye, BrahimBelhaouari Samir, Abas MdSaid, "Efficient FeatureSelection and Classification of Protein Sequence Data inBioinformatics", The Scientific WorldJournal, vol. 2014, Article ID 173869, 12 pages,2014.https://doi.org/10.1155/2014/ 173869.

[14] Khatri, K. ., & Sharma, D. A. . (2020). ECG Signal Analysis for Heart Disease Detection Based on Sensor Data Analysis with Signal Processing by Deep Learning Architectures. Research Journal of Computer Systems and Engineering, 1(1), 06–10. Retrieved from https://technicaljournals.org/RJCSE/index.php/journal/article/view/11

[15] Hussein Hijazi[1] and Christina Chan,A Classification Framework Applied to Cancer GeneExpressionProfiles,PMID: 23778014,doi: 10.1260/2040-2295.4.2.255,J Healthc Eng. 2013; 4(2): 10.1260/2040-2295.4.2.255.

[16] John N. Weinstein, Eric A. Collisson, Gordon B. Mills, Kenna M. Shaw, Brad A.Ozenberger, KyleEllrott, Ilya Shmulevich, Chris Sander, Joshua M. Stuart,The CancerGenome Atlas Pan-CancerAnalysisProject,PMID: 24071849,Nat Genet. 2013Oct;45(10): pp1113–1120.doi: 10.1038/ng.2764