DNA Sequence Classification Using CNN and Hybrid Models

Venkatakrishnan R
CH.EN.U4AIE20078
Computer Science and Engineering (AI)
Amrita Vishwa Vidyapeetham, Chennai

Siva Jyothi Nath Reddy B
CH.EN.U4AIE20063
Computer Science and Engineering (AI)
Amrita Vishwa Vidyapeetham, Chennai

Sarthak Yadav
CH.EN.U4AIE20058
Computer Science and Engineering (AI)
Amrita Vishwa Vidyapeetham, Chennai

Shaik Huziafa Fazil
CH.EN.U4AIE20060
Computer Science and Engineering (AI)
Amrita Vishwa Vidyapeetham, Chennai

Pravine Mukesh
CH.EN.U4AIE20050
Computer Science and Engineering (AI)
Amrita Vishwa Vidyapeetham, Chennai

Abstract

In recent years, Deep learning models have increased ability of extracting features of high-level abstraction from minimum preprocessing data has been widely used. In this work, we employed CNN, CNN-LSTM, and CNN-Bidirectional LSTM architectures for DNA sequence classification while considering these sequences as text data. We used one-hot vectors to represent sequences as input to the model; therefore, it conserves the essential position information of each nucleotide in sequences. Using DNA binding protein sequence dataset, models are evaluated on different classification metrics. From the experimental results, CNN-Bidirectional LSTM offers high accuracy with 99.5 %.

Keywords—DNA sequence classification; CNN ;CNN-LSTM; CNN-Bidirectional LSTM;

1.Introduction

DNA-binding proteins are proteins that have DNA-binding domains and thus have a specific or general affinity for single- or double-stranded DNA. A DNA-binding domain is an independently folded protein domain that contains at least one structural motif that recognizes double- or single-stranded DNA. This is a standard functional genomic question to detect of transcription-factor binding sites in DNA sequences. DNA-binding proteins include transcription factors which modulate the process of transcription, various polymerases, nucleases which cleave DNA molecules, and histones which are involved in chromosome packaging and transcription in the cell nucleus. Each transcription factor binds to one specific set of DNA sequences and activates or inhibits the transcription of genes that have these sequences near their promoters. The transcription factors do this in two ways. Firstly, they can bind the RNA polymerase responsible for transcription, either directly or through other mediator proteins; this locates the polymerase at the promoter and allows it to begin transcription.Alternatively, transcription factors can bind enzymes that modify the histones at the promoter. This alters the accessibility of the DNA template to the polymerase.In this project we will be classifying which DNA sequence is binding protein or not.

Suppose the DNA sequences increase exponentially, machine learning techniques are used for DNA sequence classification. Any living organism's blueprint is DNA (deoxyribonucleic acid). Adenine (A), cytosine (C), guanine (G), and thymine (T) are the four nucleotides that makeup DNA (T). These are called the building blocks of DNA. DNA appears as single-stranded or double-stranded (as shown in Figure 1). Each form of nucleotide binds to its complementary pair on the opposite strand in double-stranded DNA. Adenine and thymine form a pair, while cytosine and guanine form a pair. Ribonucleic acid (RNA) may be singlestranded or double-stranded. In RNA, uracil (U) replaces the thymine (T). Therefore, the genome is the sequence of nucleotides (A, C, G, T) for DNA virus and (A, C, U, G) for RNA sequence.

The raw DNA sequence cannot give as input to the CNN for feature extraction. It has to be converted into numerical representation before it is processed in the CNN. The encoding method also plays a vital role in classification accuracy. Encoding Method used in this work is one hot encoding , in which each nucleoid in a DNA sequence is identified by a unique index value, preserving the sequence's positional information. Feature engineering is fundamental for any model to produce good accuracy. In the machine learning models, feature extraction is done manually. But as the complexity of the data increases, the manual feature selection may lead to many problems like selecting features that do not lead to the best solution or missing out on essential features. Automatic feature selection can be used to overcome this issue. CNN is one of the best deep-learning techniques used to extract key features from the raw dataset.

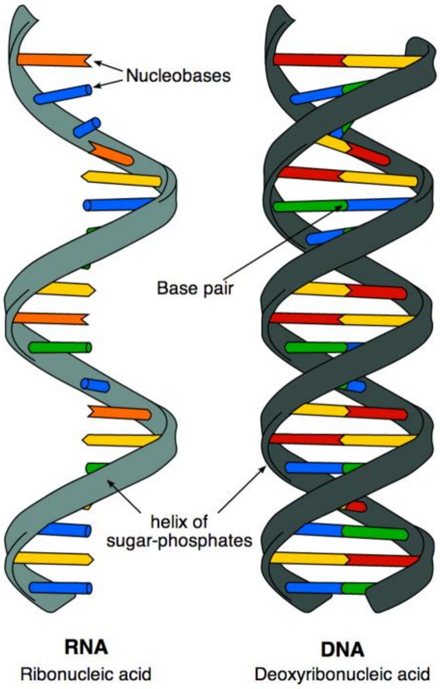


Figure 1: DNA and RNA structure

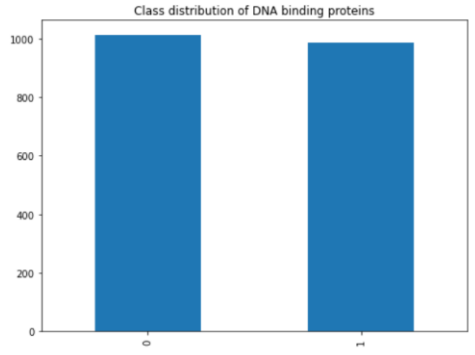
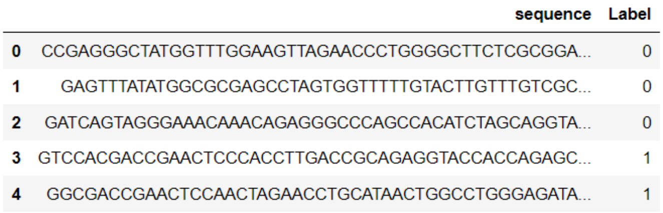


Figure 2: Distribution of each class and number of samples in a dataset.



	sequence	Label
0	CCGAGGGCTATGGTTTGGAAGTTAGAACCCTGGGGCTTCTCGCGGA...	0
1	GAGTTTATATGGCGCGAGCCTAGTGGTTTTTGTACTTGTTTGTCGC...	0
2	GATCAGTAGGGAAACAAACAGAGGGCCCAGCCACATCTAGCAGGTA...	0
3	GTCCACGACCGAACTCCCACCTTGACCGCAGAGGTACCACCAGAGC...	1
4	GGCGACCGAACTCCAACTAGAACCTGCATAACTGGCCTGGGAGATA...	1

Figure 3: Sample dataset with genomic sequences and their class labels.

2. Methodology

2.1. Data Collection: The complete dataset taken from kaggle. (https://www.kaggle.com/c/transcription-factors bindingprediction/overview).The class distribution of each class with the number of samples is shown in Figure 2. In Figure 2 it clearly shows both labels have completely balanced. The sample DNA sequences from the dataset with the complete genomic sequence of a virus, the length of the sequence, and the class labels are shown in Figure 3.

2.2 Data Pre-processing: Preprocessing data is the most critical step in most machine learning and deep learning algorithms that involve numerical rather than categorical data. The genomic sequence in the DNA dataset is categorical. There are many techniques available to convert the categorical data to numerical. The encoding technique is the process of converting the categorical data of nucleotide into numerical form. In this paper, one hot encoding are used to encode the DNA sequence.

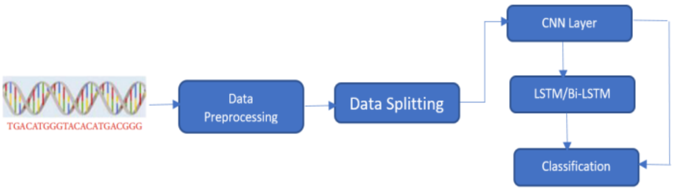


Figure 4: Workflow of the proposed model for the classification of DNA sequence.

2.3. Classification Models: In this work, three different classification models CNN, CNN-LSTM, and CNN-Bidirectional LSTM are used for DNA sequence classification. The one hot encoding technique is used to encrypt the DNA sequence, which preserves the position information of each nucleotide in the sequence.

The CNN layer is used as the feature extraction stage, and it is given as the input for LSTM and bidirectional LSTM for classification. The workflow for the proposed work is shown in Figure 4.

2.3.1. CNN. CNN is a common deep-learning technique that can yield cutting-edge results for most classification problems. CNN performs well not only on image classification, but it can also produce good accuracy on text data. Mainly, CNN is used to automatically extract the features from the input dataset, in contrast to machine learning models, where the user needs to select the features 2D CNN , and 3D CNN is used for image and video data, respectively, whereas 1D CNN is used for text classification. The DNA sequence is treated as a sequence of letters (nucleoids A, C, G, and T). Since CNN can work only with numerical data, the DNA sequence is converted into numerical values by applying one hot encoding. The CNN architecture uses a series of convolutional layers to extract features from the input dataset. Max pooling layer after each convolutional layer and the dimensions of extracted features are reduced. In the convolutional layer, the size of the kernel plays a significant role in function extraction. The model's hyper-parameters are the number of filters and kernel size.Table 1 shows the summary of the complete architecture of the proposed CNN model. Two convolutional layers are added to the model with filters of 128 and 64, along with the kernel of size (2×2) with ReLU as an activation function for feature extraction. The feature map dimensions are reduced by adding a max pooling layer of size (2×2). Finally, the feature maps are converted into single-column vector using the flatten layer. The output is passed to a dense layer with neurons 128 and 64, respectively. The Sigmoid function is used as the classification layer, which can perform well for the Binary classification problem.

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

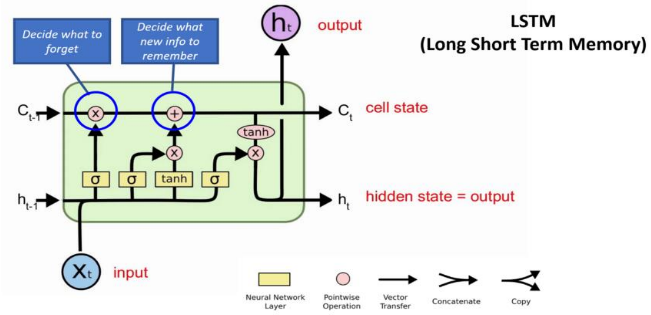


Figure 5: Architecture of the LSTM model.

2.3.2. CNN-LSTM: Long short-term memory (LSTM) is a recurrent neural network (RNN) that can learn long-term dependencies in a sequence and is used in sequence prediction and classification. It includes a series of memory blocks known as cells, each of which comprises three gates: input, output, and forget. The LSTM will selectively remember and forget things in this case. Figure 7 depicts the LSTM model's overall architecture. It is capable of learning and recognises the long sequence. The current state is calculated using Equation (2),

$$h_t = f(h_{t-1}, X_t) \qquad (2)$$

The first step is to decide what information we're going to throw away from the cell state. When detailed information becomes invalid for the sequence classification, the forget gate outputs a value of 0, indicating to remove the data from the memory cell.This gate takes two inputs, h_{t-1} (input from the previous state) and X_t (input from the current state). The input will be multiplied by a weight and added with bias. This decision is made by a sigmoid layer called the "Forget Gate layer."The input gate in the LSTM is responsible for adding all the relevant value to the cell state.The next step is to decide what *new* information we're going to store in the cell state. This has two parts. First, a sigmoid layer called the "Input Gate layer" decides which values we'll update. Next, a tanh layer creates a vector of new candidate values that could be added to the state with values ranging from -1 to 1.

The output gate in LSTM decides what value can be in the output by employing the sigmoid activation function and tanh activation function to the cell state. The LSTM layer with 100 memory units is added after the convolutional layers to predict the classification labels in our model. The features extracted by the convolutional layer is given as an input to the LSTM layer for classification. In many NLP tasks, CNN and LSTM are combined into a hybrid model to improve the accuracy of the classification. The LSTM model includes dropout layers and regularisation techniques to reduce the overfitting problem.

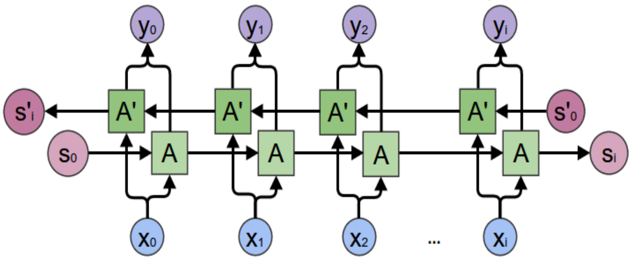
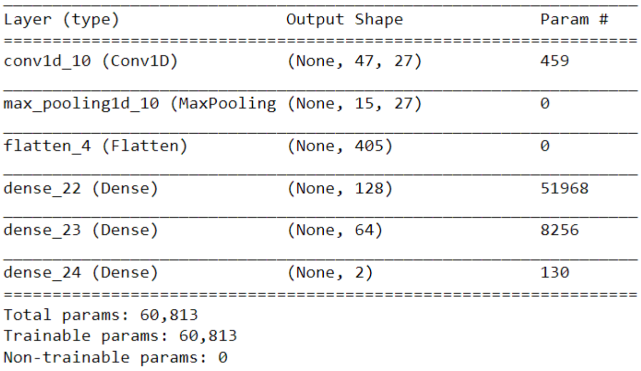


Figure 6: Architecture of the Bidirectional LSTM model.

2.3.3. CNN-Bidirectional LSTM: Bidirectional long-short term memory(bi-lstm) is the process of making any neural network o have the sequence information in both directions backwards (future to past) or forward(past to future).In bidirectional, our input flows in two directions, making a bi-lstm different from the regular LSTM. With the regular LSTM, we can make input flow in one direction, either backwards or forward. However, in bi-directional, we can make the input flow in both directions to preserve the future and the past information. The architecture of the bidirectional LSTM is given in Figure 6.

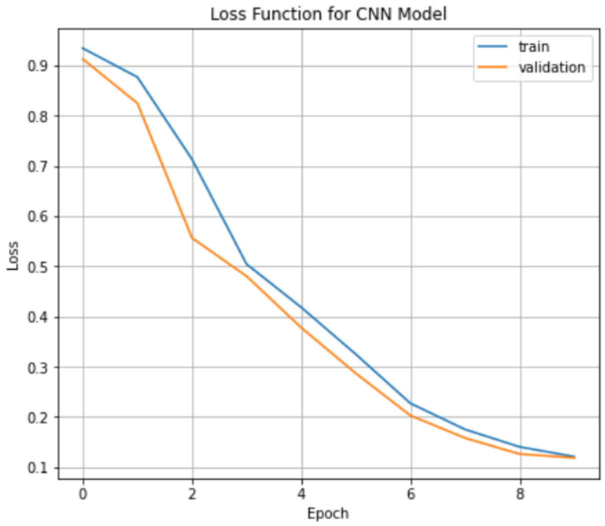


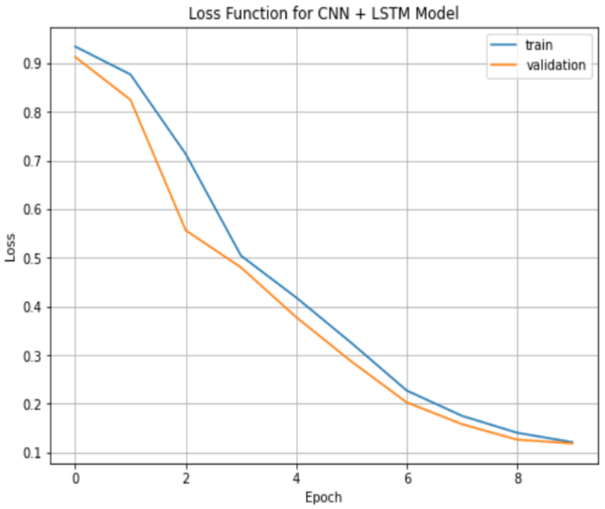
```
________________________________________________________________
Layer (type)                 Output Shape              Param #
================================================================
conv1d_10 (Conv1D)           (None, 47, 27)            459
________________________________________________________________
max_pooling1d_10 (MaxPooling (None, 15, 27)            0
________________________________________________________________
flatten_4 (Flatten)          (None, 405)               0
________________________________________________________________
dense_22 (Dense)             (None, 128)               51968
________________________________________________________________
dense_23 (Dense)             (None, 64)                8256
________________________________________________________________
dense_24 (Dense)             (None, 2)                 130
================================================================
Total params: 60,813
Trainable params: 60,813
Non-trainable params: 0
```

Table 1: Summary of CNN Architecture

3. Results and Discussion

The proposed models are experimented with using the NVDIA GeForce Gtx 1650i processor with a RAM size of 4 GB. The dataset consists of 2000 inputs divided into training, testing ratio of 80%, and 20%, respectively. In the training phase, the binary cross entropy function is used as the loss function. This loss function calculates the error between the actual output and the target label, on which the training and update of the weights are done. We tested the CNN, CNN-LSTM, and CNN bidirectional LSTM by varying the values of different hyper parameters like filters, filter size, the number of layers. Grid search cross validation is the most widely used parameter optimization method to select the best parameters for the model.





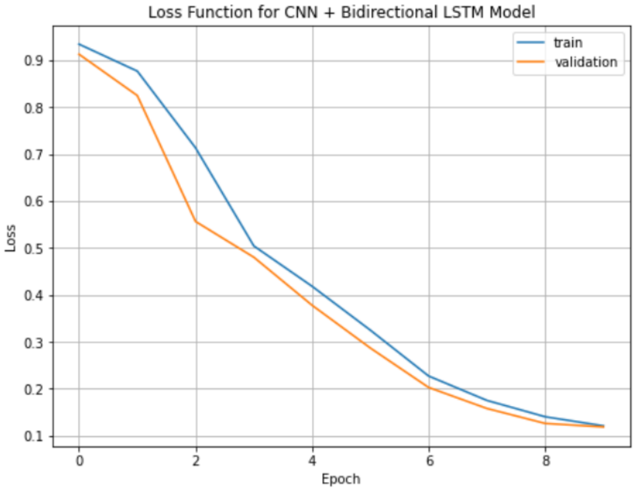


Figure 7 : Loss plots for Models

The best parameters of all three models are the numbers of filters 128, 64, and 32 in each layer. The size of the filter is 2×2, training batch size of 128, training epochs of 10. The classification models are evaluated using different classification metrics like accuracy, precision, recall, F1 score from confusion matrix.

Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision is the ratio of correctly predicted positive observations to the total predicted positive observations.

$$\text{precision} = \frac{TP}{TP + TN}$$

Recall or Sensitivity is the ratio of correctly predicted positive observations to the all observations in actual class - yes.

$$\text{Recall} = \frac{TP}{TP + FN}$$

F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. Intuitively it is not as easy to understand as accuracy, but F1 is usually more useful than accuracy.

$$\text{F1 score} = \frac{2 * (Recall\ *\ Precision)}{(Recall\ +\ Precision)}$$

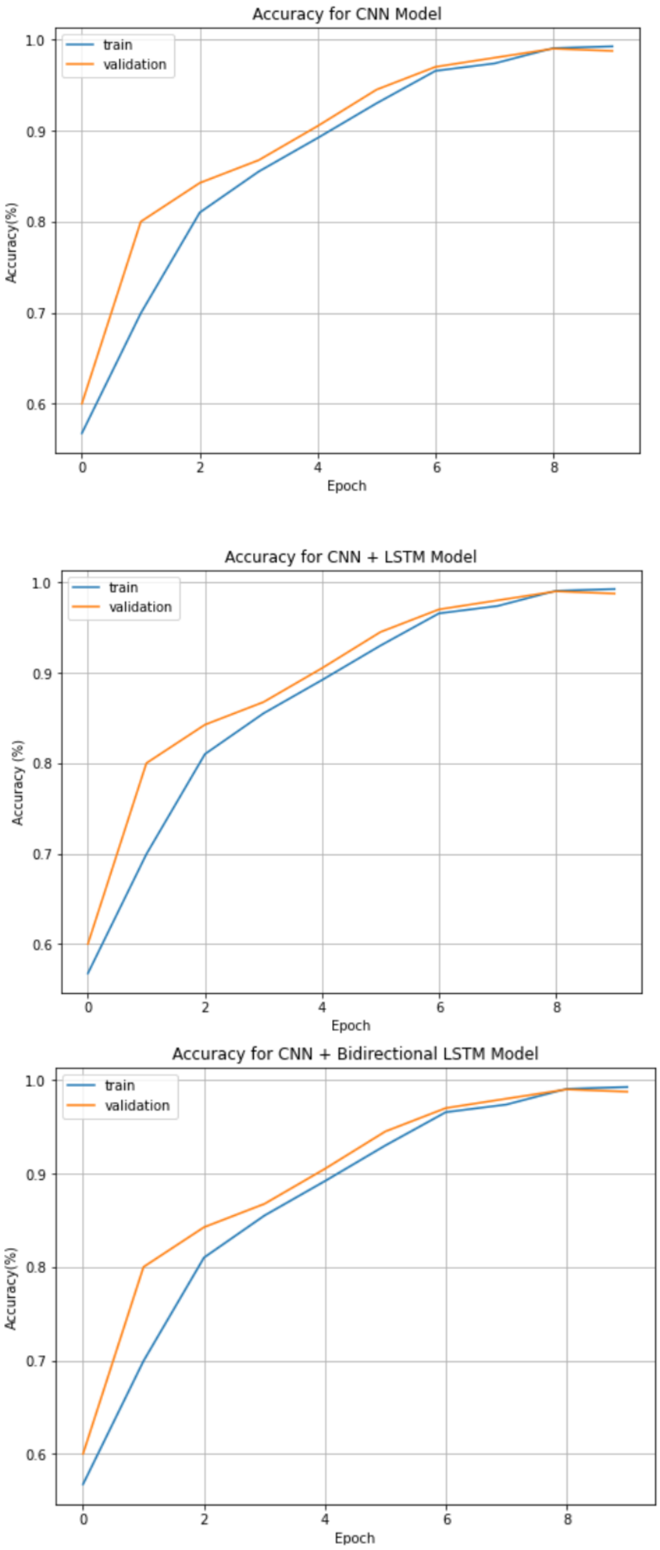


Figure 8 : Accuracy plots for All models

4.conclusion

This paper compared three deep-learning methods, namely, CNN, CNN-LSTM, and CNN-bidirectional LSTM, with one hot encoding. We found that CNN with Bidirectional LSTM outperforms the other models with 99%.we have evaluated classification on different metrics while precision of CNN with Bidirectional LSTM gave 97.5 %, Recall of CNN with Bidirectional LSTM gave 99.4 % and F1 score of CNN with Bidirectional LSTM gave 98.7%.

5.References

[1] M. F. Aslan, M. F. Unlersen, K. Sabanci, and A. Durdu, “CNN based transfer learning-BiLSTM network: a novel approach for COVID-19 infection detection,” Applied Soft Computing,vol. 98, article 106912, 2021.

[2] S. Shadab, M. T. Alam Khan, N. A. Neezi, S. Adilina, and S. Shatabda, “DeepDBP: deep neural networks for identification of DNA-binding proteins,” Informatics in Medicine Unlocked, vol. 19, article 100318, 2020.

[3] Hemalatha Gunasekaran, K. Ramalakshmi, A. Rex Macedo Arokiaraj, S. Deepa Kanmani, Chandran Venkatesan, C. Suresh Gnana Dhas, "Analysis of DNA Sequence Classification Using CNN and Hybrid Models", Computational and Mathematical Methods in Medicine, vol. 2021, Article ID 1835056, 12 pages, 2021

[4] Nguyen, N.G., Tran, V.A., Ngo, D.L., Phan, D., Lumbanraja, F.R., Faisal, M.R., Abapihi, B., Kubo, M. and Satou, K. (2016) DNA Sequence Classification by Convolutional Neural Network. J. Biomedical Science and Engineering, 9, 280-286