AMRITA VISHWA VIDYAPEETHAM

PREDICTION OF RNA PROTEIN INTERACTION

22BIO201-INTELLIGENCE OF BIOLOGICAL SYSTEMS

Group 11
Syed Anees Ashraf [23068]
C V Shreyaas Aditya [23018]
B Sai Saran Goud [23017]
S G Sathwik [23064]

INTRODUCTION

- The project aims to predict RNA-protein interactions, which are essential in biological processes like gene regulation, RNA splicing, and translation.
- Understanding these interactions can provide insights into disease mechanisms, aid in drug discovery, and advance synthetic biology.
- Experimental methods to study RNA-protein interactions are costly and time-consuming, making computational predictions an efficient alternative.
- The project uses deep learning models—specifically hybrid CNN-LSTM with Attention and a BERT-based sequence classifier —to predict whether an RNA sequence binds to a protein.
- These models leverage advanced neural network architectures to capture complex sequence patterns, offering accurate and interpretable predictions for RNA-protein interactions.

PROBLEM STATEMENT

- RNA-protein interactions are vital but difficult to study using traditional experimental methods.
- Techniques like EMSA and CLIP are time-consuming, expensive, and labor-intensive, restricting the pace of research.
- The slow identification of RNA-protein interactions hampers advancements in understanding diseases and developing targeted treatments.
- There is a growing demand for accurate, efficient computational models to predict RNA-protein interactions.
- This project focuses on developing deep learning models to predict RNA-protein binding, providing scalable tools for accelerating biomedical research and therapeutic discoveries.

DATASET :

Benchmark datasets for RBP-binding linear RNAs

We downloaded the benchmark dataset for RBP binding linear RNAs of RBPSuite. This benchmark dataset consists of 353 RBPs and their binding sites (RNAs) are derived from POSTAR3 database.

Methodology:

Data Preparation and Preprocessing:

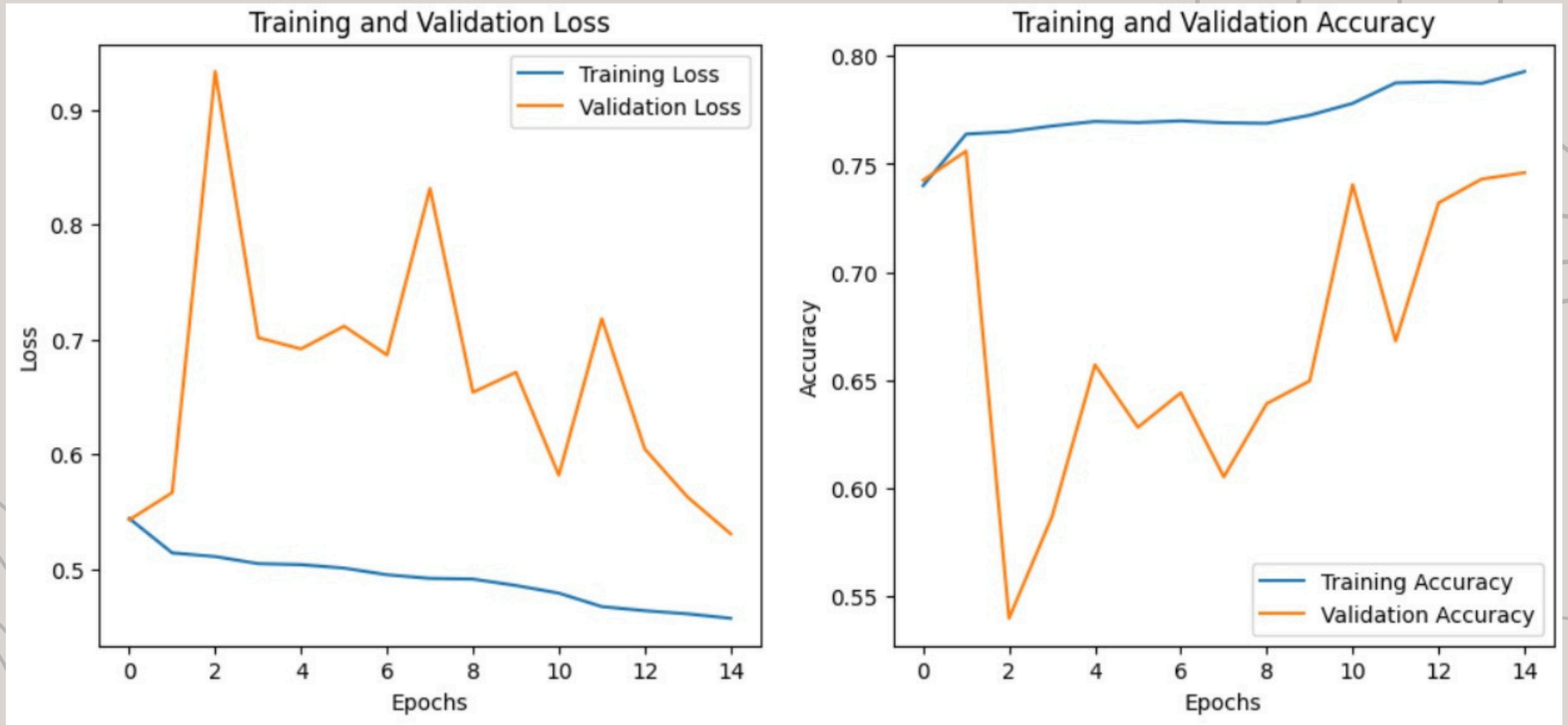
- The input data consists of RNA sequences in FASTA format, with positive sequences indicating binding interactions with a specific protein and negative sequences indicating non-binding interactions.
- one-hot format using the characters 'A', 'C', 'G', and 'U' (the four nucleotides in RNA). It initializes a one-hot encoded matrix of shape (sequence_length, 4). For each character in the sequence, it assigns a 1 at the corresponding nucleotide index (A = 0, C = 1, G = 2, U = 3).
- Positive and negative RNA sequences are loaded separately using the prepare_dataset function. The sequences are padded to a fixed length to ensure consistent input size for the model. The labels are assigned as 1 for positive interactions and 0 for negative interactions.

Methodology:

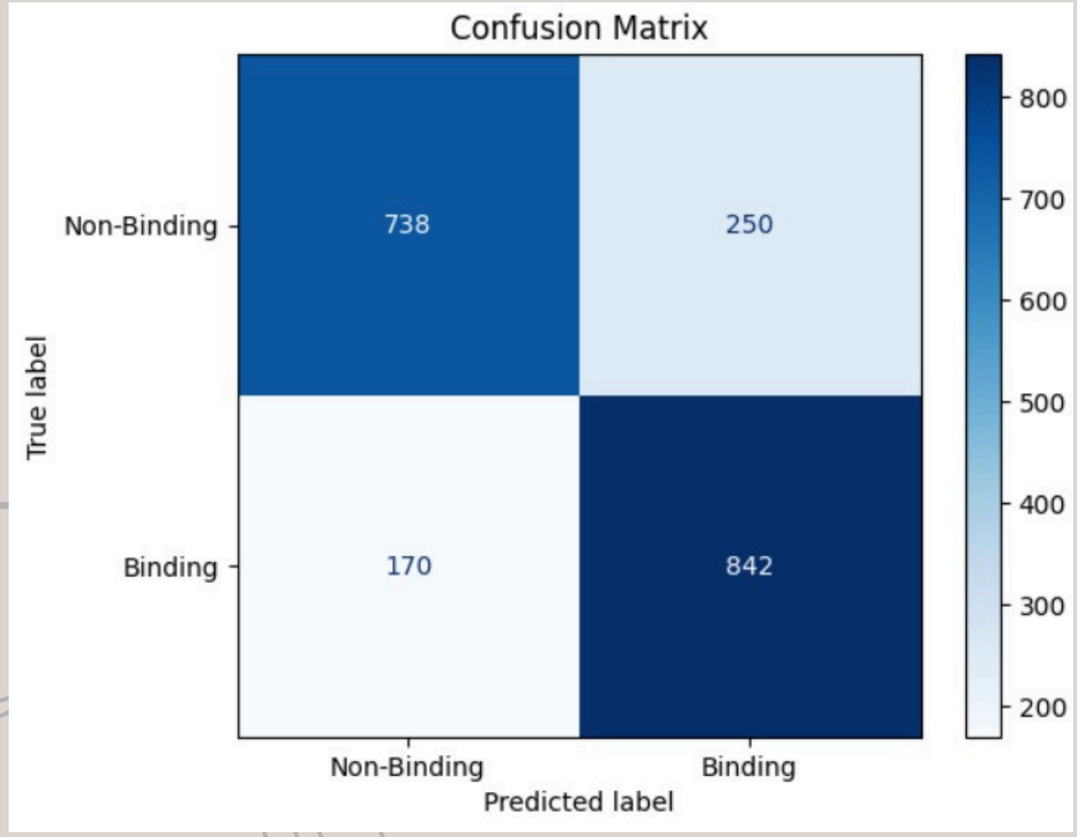
Model:

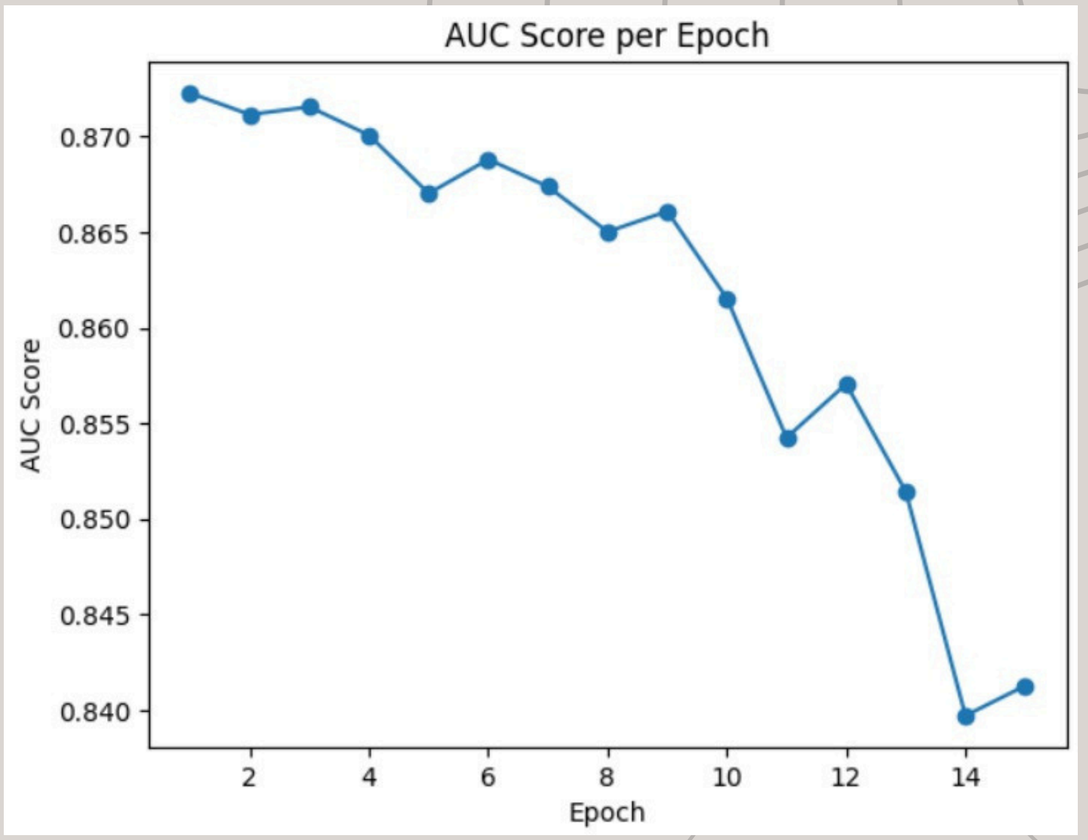
- The model takes sequences of one-hot encoded RNA sequences as input with the shape (sequence_length, 4). Two Conv1D layers are applied sequentially to extract features from the RNA sequence. The first convolutional layer uses 64 filters with a kernel size of 3, followed by a ReLU activation function and padding to maintain the sequence length.
- A Bidirectional LSTM layer with 64 units is used to capture long-term dependencies in the RNA sequence in both forward and backward directions. This layer outputs a sequence, which is then fed into the Attention layer.
- The Attention layer allows the model to focus on different parts of the sequence when making predictions. The attention mechanism helps the model weigh the importance of various sequence positions and capture complex relationships between nucleotides.
- The final layer is a dense layer with a single unit and a sigmoid activation function. This outputs a value between 0 and 1, representing the probability of the RNA sequence binding to the protein.

Results :

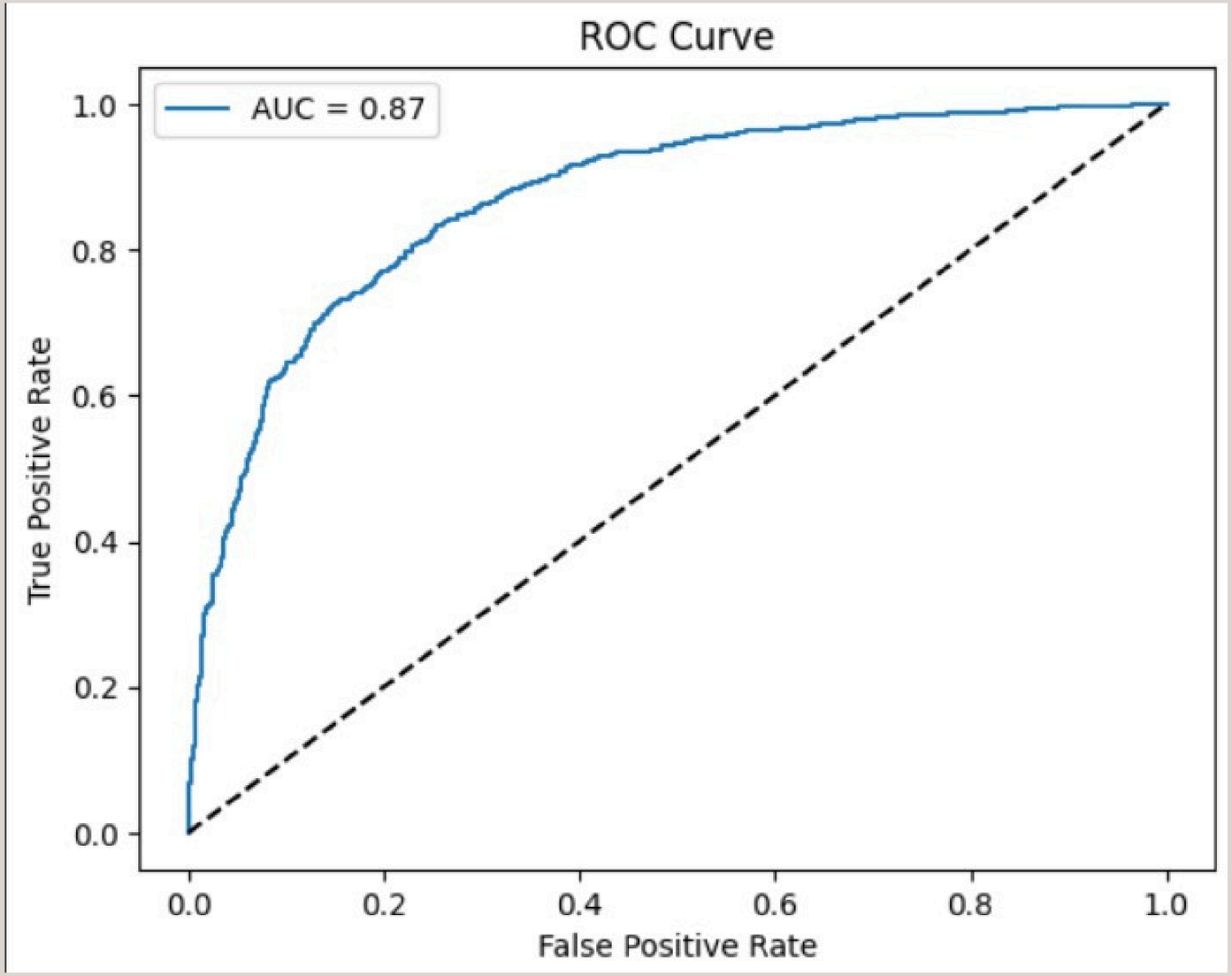

Training and Validation Loss
Training Loss
Validation Loss
Loss
Epochs
0.9
0.8
0.7
0.6
0.5
0
2
4
6
8
10
12
14
Training and Validation Accuracy
Training Accuracy
Validation Accuracy
Accuracy
Epochs
0.80
0.75
0.70
0.65
0.60
0.55

Results :


Confusion Matrix
Non-Binding
Binding
738
250
170
842
True label
Non-Binding
Binding
Predicted label
800
700
600
500
400
300
200


AUC Score per Epoch
AUC Score
0.870
0.865
0.860
0.855
0.850
0.845
0.840
2
4
6
8
10
12
14
Epoch

Results:


ROC Curve
AUC = 0.87
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0

FUTURE SCOPE :

1. **Binding Site Prediction:**Extend the model to predict specific binding sites on RNA and proteins, providing detailed insights into which regions are responsible for the interaction. This could aid in targeted drug development and RNA engineering.
2. **Interaction Strength Estimation:**Develop models capable of predicting the strength of RNA-protein interactions, offering quantitative insights into binding affinities and helping prioritize biologically significant interactions.
3. **Multi-Protein Sequence Analysis:**Enhance the model to handle multiple protein sequences simultaneously, enabling the prediction of RNA interactions with protein complexes or families, which are common in many biological processes.

Reference Papers :

[1] Prediction of RNA–protein interactions using a nucleotide language model (Keisuke Yamada, Michiaki Hamada).
Available online: https://academic.oup.com/bioinformaticsadvances/article/2/1/vbac023/6564689

[2] Predictions of protein–RNA interactions (Davide Cirillo, Federico Agostini, Gian Gaetano Tartaglia).
Available online: https://wires.onlinelibrary.wiley.com/doi/full/10.1002/wcms.1119

[3] Protein–RNA interaction prediction with deep learning: structure matters (Junkang Wei, Siyuan Chen, Licheng Zong, Xin Gao, Yu Li).
Available online: https://academic.oup.com/bib/article/23/1/bbab540/6470965

[4] Predicting RNA–protein binding sites and motifs through combining local and global deep convolutional neural networks (Xiaoyong Pan, Hong-Bin Shen).
Available online: https://academic.oup.com/bib/article/23/1/bbab540/6470965

[5] Predicting DNA- and RNA-binding proteins from sequences with kernel methods
Xiaojian Shao a, Yingjie Tian b, Lingyun Wu c, Yong Wang c, Ling Jing a, Naiyang Deng
Available online: https://www.sciencedirect.com/science/article/abs/pii/S002251930900289

Thank You

Presented by Group 11