Deep Learning-Based Prediction of RNA Secondary Structures

This project focuses on developing an advanced computational model to predict RNA secondary structures from raw RNA sequences. RNA folding determines how RNA molecules perform their biological functions, making accurate prediction of these structures essential in molecular biology and medical research. Unlike traditional models, this work emphasizes recognizing complex RNA formations, such as pseudoknots, which are critical for understanding RNA behavior but challenging to model.

The model integrates modern deep learning techniques to analyze sequences and predict base pairing probabilities. By combining convolutional neural networks with recurrent neural networks, it captures both local sequence motifs and long-range interactions. The outcome is a detailed prediction of how the RNA folds, represented in standard notations useful for researchers studying RNA function and designing RNA-targeted therapeutics.

Aim

The aim of this project is to build a deep learning-based predictor that accurately forecasts the secondary structure of RNA sequences, including pseudoknots. The goal is to convert raw RNA sequences into interpretable structural formats that reveal base pairing patterns, aiding biological research and applications such as drug development and viral studies.

Methodology

Dataset Preparation

The dataset used in this project is a curated subset of the bpRNA-1m(90) dataset, comprising 28,174 RNA sequences, each paired with its experimentally known secondary structure. This dataset provides a reliable foundation for training deep learning models to understand the structural behavior of RNA molecules.

To prepare the data for training:

- Each RNA sequence is converted into a one-hot encoded format, representing the four nucleotides:
  - A = [1, 0, 0, 0]
  - U = [0, 1, 0, 0]
  - G = [0, 0, 1, 0]
  - C = [0, 0, 0, 1]

- Dot-bracket notation is used to label the known secondary structures (e.g., stems, loops, and pseudoknots).

- To ensure uniform tensor dimensions for model processing, all sequences are truncated or zero-padded to a fixed length of 2000 nucleotides.

- The dataset is split and loaded using a custom PyTorch Dataset class, allowing efficient batching and sampling.

This structured and preprocessed dataset enables the model to learn meaningful base-pairing patterns and generalize effectively across diverse RNA types.

Training Procedure

- The model is trained using supervised learning with known RNA secondary structures as labels.

- Loss functions compare predicted base-pair probabilities with true pairings, guiding model optimization.

- Adam optimizer is used for parameter updates.

- Dropout and layer normalization help prevent overfitting.

- Training runs on GPU when available to speed up computation.

- Early stopping is applied to halt training when validation performance stops improving.

Evaluation

- Performance is measured by precision, recall, and F1-score on predicted base pairs against ground truth.

- Special attention is paid to accurately predicting complex structures like pseudoknots.

- Visualization tools such as dot-bracket notation and Forna URLs help interpret the folding patterns predicted by the model.

- Training curves (loss and accuracy) are monitored for model convergence and stability**.**

RESULT :

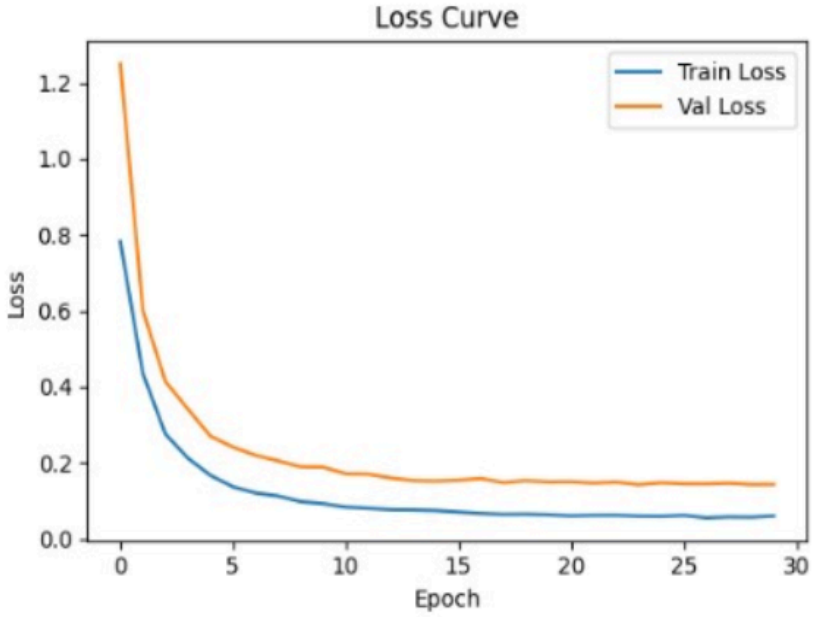
Over the course of 30 epochs, the model demonstrated consistent improvements in both training and validation performance.

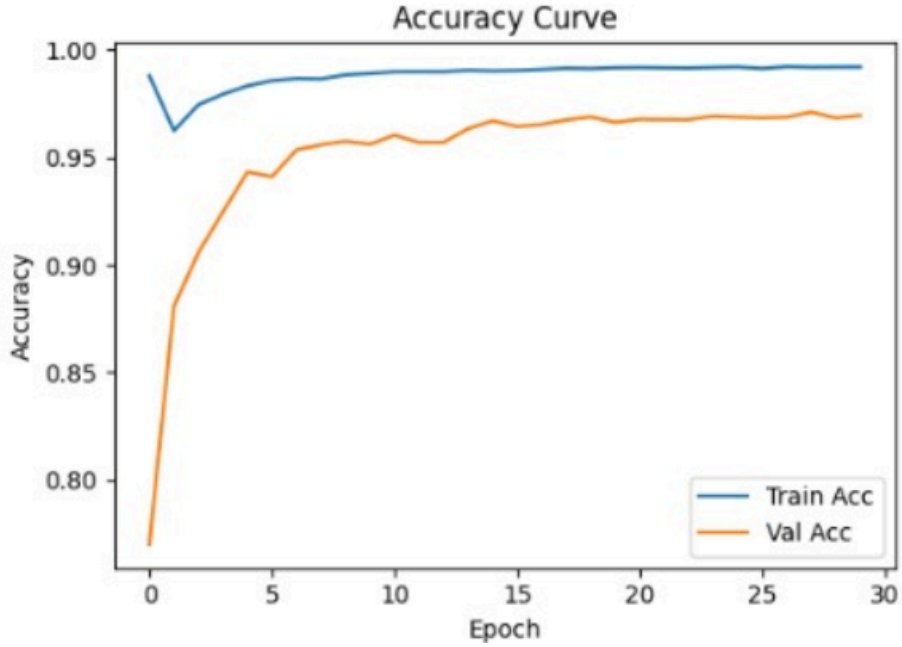
- Training Accuracy started at 98.79% and gradually increased, reaching 99.21% by the final epoch. The training loss reduced from 0.7820 in the first epoch to 0.0600 in the last, indicating effective learning and convergence.

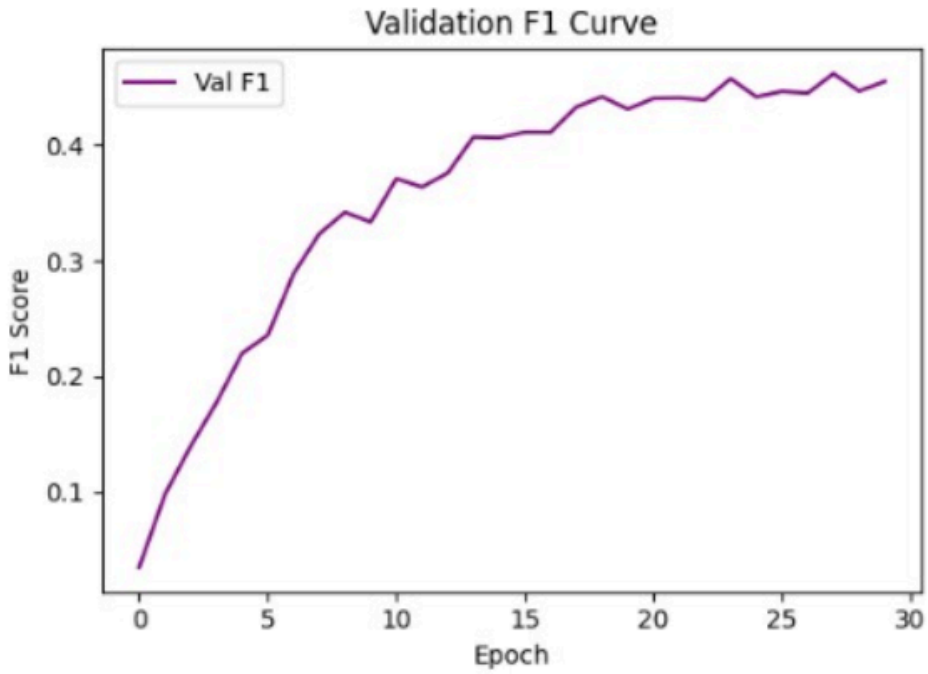
- On the validation set, the model's performance showed significant improvement:

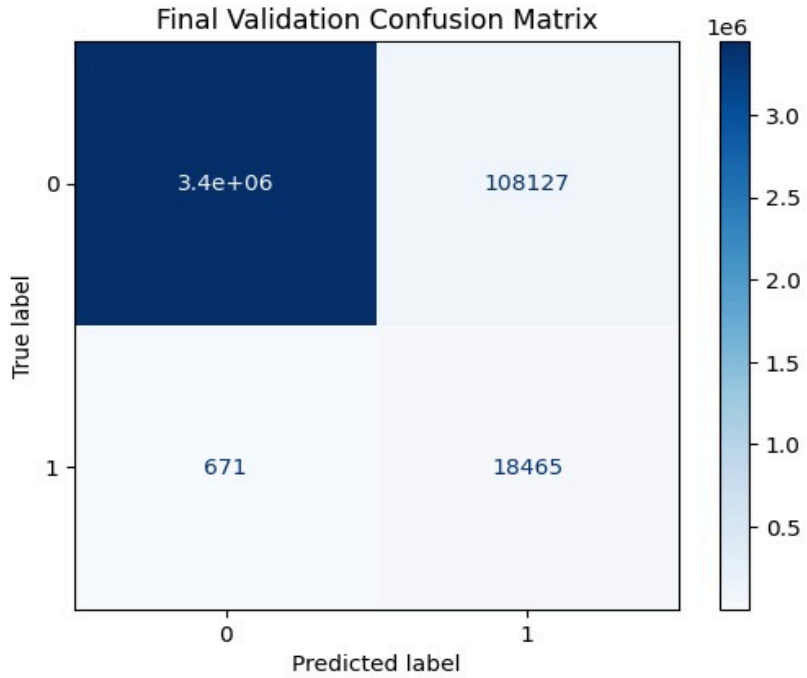
    - Precision rose from a low of 1.79% in the first epoch to 31.63% by epoch 28.
    - Recall maintained a consistently high value throughout training, ranging from 69.10% to 94.32%, showing the model's ability to identify relevant samples well.
    - The F1 score, a balance between precision and recall, started at 0.0348 and improved steadily to reach a peak of 0.4623, indicating better harmonic balance between precision and recall.
    - Validation Accuracy increased from 71.93% to 98.06%, and the raw validation accuracy also followed this upward trend, reaching 97.11%.

The best F1 score achieved was 0.4623, at which point the model was saved. This suggests that the model not only learned to generalize better over time but also maintained high recall while steadily improving precision.


Loss Curve
Train Loss
Val Loss
Loss
Epoch
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0
5
10
15
20
25
30


Accuracy Curve
Train Acc
Val Acc
Accuracy
Epoch
0.80
0.85
0.90
0.95
1.00
0
5
10
15
20
25
30


Validation F1 Curve
Val F1
F1 Score
Epoch
0.1
0.2
0.3
0.4
0
5
10
15
20
25
30


Final Validation Confusion Matrix
3.4e+06
108127
671
18465
True label
Predicted label
0
1
0
1
1e6
3.0
2.5
2.0
1.5
1.0
0.5

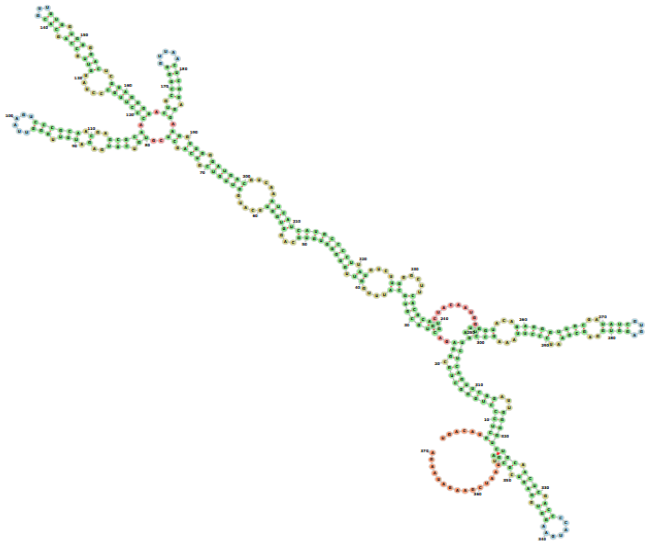


Image generated by forna by the sequence from provided by the model

LINK TO THE github CODE :

github

https://github.com/Aswinpt2004/RNA_PREDICTION