Mutation Classification And Prediction for Coronaviruses

Team : Pranesh S , Bishwadip Maitra, Anmol Kumar Pandey, Akash Narvariya ,Mohd.Rizwan,

Mutation Prediction for Coronaviruses Using Genome Sequence and Recurrent Neural Networks.

Motivation

Coronaviruses, such as SARS-CoV-2 (the virus that causes COVID-19), are highly mutable viruses, meaning they can rapidly evolve and acquire new mutations that can potentially affect their transmissibility, virulence, and resistance to vaccines or treatments.

Accurately predicting future mutations in coronaviruses is crucial for developing effective countermeasures, such as updated vaccines or therapeutic strategies, to stay ahead of the virus's evolution.

Introduction

The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is the source of the coronavirus illness (COVID-19). It is a member of the coronavirus family of viruses, which also includes other varieties including Severe Acute Respiratory Syndrome (SARS-CoV) and Middle East Respiratory Syndrome (MERS-CoV). The first case of COVID-19 was discovered in Wuhan, China's Hubei province, in December 2019. Since then, it has spread quickly over the world, sparking a pandemic. When an infected person coughs, sneezes, or speaks, respiratory droplets from their mouths are the main

way that COVID-19 spreads. It can also spread by coming into contact with infected surfaces.

Impact on Health: Globally, COVID-19 has led to a considerable increase in disease and death. Fever, coughing, shortness of breath, exhaustion, headache, sore throat, loss of taste or smell, congestion, nausea, and diarrhea are some of the mild to severe symptoms. Acute respiratory distress syndrome (ARDS), organ failure, pneumonia, and even death are possible outcomes of severe instances.

Death Toll: As of January 2022, when I provided my most recent report, COVID-19 has claimed millions of lives worldwide, with exact figures changing depending on the nation and area. The figures keep changing as the virus worsens and as nations take action to stop its spread.

Economic Impact: The pandemic has had a significant negative impact on the world economy, resulting in widespread job losses, company closures, supply chain disruptions, and recessions. In order to lessen the impact on the economy, governments have put in place stimulus packages and other initiatives.

Social Impact: Lockdowns, quarantines, travel restrictions, and social distancing measures are only a few of the ways that COVID-19 has impacted billions of people's everyday lives. Social relationships, mental health, education, and general well-being have all been influenced by these measurements.

Variants: SARS-CoV-2 has seen a number of mutations that have caused the virus to evolve into new forms. Certain variations have demonstrated heightened transmissibility, pathogenicity, or resistance to current therapies and vaccinations.

Delta variety: The Delta variety, which first surfaced in late 2020 and swiftly took the lead in several nations, was one of the varieties that raised the greatest alarms. Compared to previous strains, it was more contagious, which caused spikes in the number of cases in various areas.

Omicron Variant: Initially discovered in late 2021, the Omicron variant is another noteworthy variation. It quickly spread to many nations and caused alarm because of the large number of mutations and its influence on the efficacy of vaccines. It did, however, also show some immune evasion.

That is why our solution wants to find out other mutations so that the later mutations can be predicted and we can be prepared for it.

Base Paper

[Base Paper Link]

1

PRIEST - Predicting viral mutations with immune escape capability of SARS-CoV-2 using temporal evolutionary informatio-n

The Problem: Rapidly Evolving Viruses and Immune Escape

- Viruses like SARS-CoV-2 (the virus that causes COVID-19) constantly change their genetic makeup through mutations.
- These mutations can make the virus more adept at evading our immune system's defenses, rendering existing vaccines and treatments less effective.

The Importance of Predicting Mutations

- Mitigating the spread of pandemics: By anticipating future variants, we can develop targeted strategies to control their spread.
- Developing effective control measures: This includes designing new vaccines and treatments that can combat these emerging variants.

Introduces PRIEST: A Deep Learning Approach

- The study proposes a new method called PRIEST .
- PRIEST is a deep learning model, a type of artificial intelligence particularly adept at finding patterns in complex data.

PRIEST's Advantage: Leveraging Time-Series Data

- Unlike some existing methods, PRIEST doesn't just analyze single viral sequences.

- It utilizes time-series data, meaning it analyzes sequences collected over time, allowing it to identify patterns in how the virus is evolving.

PRIEST's Success: Accurately Predicting Immune-Evading Mutations

  - The researchers tested PRIEST extensively and found it to be highly effective.
  - PRIEST could accurately predict mutations that enable the virus to escape immune responses.

PRIEST's Significance: A Step Forward in Pandemic Response

  - This study using PRIEST demonstrates the potential of deep learning for analyzing viral mutations.

Conclusion :

- Predicting mutations in viruses is difficult but crucial.
- Current methods aren't perfect, especially for viruses constantly adapting to our immune system.
- This study introduces PRIEST, a new method that uses past mutation information to predict future ones in SARS-CoV-2.
- PRIEST is better than existing methods and helps predict how the virus will evolve.
- This can help us develop vaccines faster and be more prepared for future outbreaks.
- PRIEST can potentially be used for other viruses as well.

Limitations of PRIEST (areas for future improvement):

  - Doesn't consider individual risk factors for diseases.
  - Needs a good amount of past virus data to make good predictions.

2

The prediction of virus mutation using neural networks and rough set techniques

The Problem: Evolution of Drug-Resistant Virus Strains

- The paper addresses the challenge of rapidly evolving virus strains, such as the Newcastle Disease Virus, by using machine learning to predict point mutations in RNA sequences.
- By understanding and predicting these mutations, the research aims to facilitate early detection of drug-resistant virus strains and enhance the effectiveness of antiviral treatments.

Approach:

  - The paper proposes a novel machine learning technique that combines neural networks and rough set theory to predict nucleotide mutations in RNA sequences of viruses like the Newcastle Disease Virus.
  - By training the model on aligned RNA sequences of successive virus generations, the approach aims to extract mutation patterns and visualize rules governing past mutations to infer future mutations, ultimately aiding in the development of more effective antiviral treatments.

Dataset:

- The paper utilizes datasets from two different countries, Korea and China, containing aligned RNA sequences of successive generations of the Newcastle Disease Virus (NDV).**:**

Conclusion :

- The machine learning technique shows promise in predicting virus mutations with 68-76% accuracy.
- Results suggest potential for early detection of drug-resistant strains and improved antiviral treatments.
- Accuracy expected to improve with larger datasets, aiding in the development of targeted therapies.
- Further research needed to validate and enhance the technique for better prediction of virus mutations

Limitations

- Statistical Significance: Results may lack statistical significance, impacting the reliability of predictions.

- Dataset Size: Effectiveness depends on larger datasets, which may be limited for rare or rapidly evolving viruses.

3

MutaGAN: A sequence-to-sequence GAN framework to predict mutations of evolving protein populations

The problem - The challenge of accurately predicting the evolution and mutations of biological populations:

- Difficulty in predicting mutations and evolution of biological populations specifically focusing on the genetic drift in the influenza virus. .
- Lack of deep learning models for forecasting genetic drift in influenza virus due to lack of knowledge.

Approach:

- MutaGAN Framework Development:
    - Creation of a novel deep learning framework, MutaGAN, for evolutionary modeling.
    - Utilization of GANs and seq2seq models to predict mutations in influenza virus proteins.
- Training on Parent-Child Pairs:
    - Training the model on parent-child pairs of protein sequences to optimize evolutionary patterns.
    - Incorporating input sequence optimization to enhance the model's ability to learn and predict mutations accurately.
- Sequence Augmentation Strategies: Optimization towards successful evolutionary patterns observed in protein history for accurate predictions.

Datasets:

- Influenza Virus HA Sequences:
    - Obtained from the National Center for Biotechnology Information's Influenza Virus Resource.
    - Included sequences from human hosts between 1968 and 2017, with validation data from 2018-19.
- Genomic Data for HA and NA Proteins:
    - Publicly available genomic data for influenza virus surface proteins HA and NA.
    - Used for training and testing the MutaGAN framework for evolutionary modeling.

Limitations:

- Data Quality/Availability Dependency:
    - Model performance may be influenced by the quality and quantity of available genomic data.
    - Variability in data sources could introduce biases affecting the accuracy of evolutionary predictions.
- Biological Complexity: Biological evolution is a complex process influenced by various factors beyond genetic mutations, which may not be fully captured by the model.
- Model Generalization: The model's performance may be limited by the specific dataset used for training, potentially affecting its ability to generalize to diverse biological populations.

Model Used:

- MutaGAN framework incorporating GANs and seq2seq models.

- Deep learning architecture with dropout, batch normalization, and dense layers.
- Leaky ReLU activation function with α = 0.1 used in dense layers.

Conclusion:

The MutaGAN framework presents a novel approach to predicting mutations in evolving protein populations, showing promise in forecasting genetic drift in influenza virus proteins. By training on parent-child pairs, the model demonstrates potential for accurate evolutionary pattern prediction. Future research should focus on expanding the framework's applicability to diverse datasets and biological systems, enhancing its predictive capabilities in evolutionary modeling.

Paper 4

Mutation prediction and phylogenetic analysis of SARS-CoV2 protein sequences using LSTM based encoder-decoder model

Introduction

- SARS-CoV-2, being an RNA virus, has a high mutation rate affecting its behavior.
- Mutations in the virus can be beneficial, harmful, or neutral, impacting its survival and replication.
- Notable mutations like the D614G mutation have been linked to increased transmissibility.

Literature Review

- RNA viruses like SARS-CoV-2 exhibit a high mutation rate, influencing transmissibility and infectivity.
- Specific mutations, such as those in the spike protein, can affect virus behavior and resistance to antibodies.

Proposed LSTM Model

- Long Short-Term Memory (LSTM) utilized for predicting SARS-CoV-2 protein sequences and detecting mutations.
- LSTM model trained on genomic sequences using bioinformatics tools for alignment and analysis.
- High accuracy achieved in predicting key protein sequences like spike, replicase, ORF1a, and nucleocapsid.

Experimental Setup

- Dataset sourced from the National Center for Biotechnology Information (NCBI) with 250 SARS-CoV-2 variants.
- LSTM model trained on protein sequences in FASTA format with specific experimental parameters.
- Bioinformatics tools like T-Coffee, BioEdit, and MEGA used for sequence alignment and analysis.

Amino Acid Sequences

- Proteins consist of amino acids crucial for structure and function.
- Amino acid sequences stored in genes, and mutations can impact protein properties.
- Prediction of mutations essential for understanding virus evolution and developing treatments.

Proposed LSTM Model

Table 1. Experimental setup for the proposed model.

Dataset Used	Genomic Sequence
Dataset Format	FASTA
Deep Learning Model	LSTM (For predicting sequences)
Language	Python
Software	Colab
Training Data	80%
Validation Data	20%
Activation Function	Softmax
Epoch	50
Batch Size	10
Loss	Categorial Cross entropy
Optimizer	Adam
Bio informatics Tool	Bioedit (For analyzing sequences)

- LSTM, a type of Recurrent Neural Network (RNN), used for predicting protein sequences and mutations.
- LSTM designed to learn long-term dependencies and avoid issues with long-term information retention.
- LSTM model trained on a dataset of protein sequences to predict mutations accurately.

Conclusion

- LSTM model proves effective in predicting mutations in SARS-CoV-2 protein sequences.
- Accurate predictions are crucial for developing treatments, vaccines, and understanding virus evolution.

Base Paper

Our base paper title **"Mutation Prediction for Coronaviruses Using Genome Sequence and Recurrent Neural Networks"**-by Pranav Pushkar, Christo Ananth, Preeti Nagrath[1], Jehad F. Al-Amri, Vividha[1] and Anand Nayyar.

The study focuses on the creation and use of GRU-RNN and LSTM-RNN based models to forecast the SARS-CoV-2 virus's future genomic sequences while taking its mutation rate into account. The goal of this forecast is to support early planning and intervention techniques. The following are the stated research objectives: Review the literature

in-depth on the use of neural networks for sequence prediction and genome sequencing. They used soft max activation function and Adam as optimizer.
Create and implement GRU-RNN and LSTM-RNN models that are specifically adapted to the time-sequenced and pre-processed genomic data gathered from the NCBI repository.Utilizing the gathered data, train the models and assess how well they perform in terms of accuracy and F1-score. Utilize the models to evaluate the accuracy of mutations and predict genomic sequences that already exist.
Examine the findings of the analysis of mutation correctness and contrast them with recent research, assessing the algorithms' effectiveness. The manuscript is arranged into many sections:

Associated works:
examines the literature on coronavirus genome sequencing.
Overview and Procedures: explains genome sequencing techniques used in the GRU-RNN and LSTM-RNN models.

Model Implementation:
Explains the steps involved in putting the suggested models into practice, including preprocessing and data description.
Analyzes and contrasts the performance of the two models based on their respective outcomes.
The paper's conclusion and future scope are provided, along with a list of possible study avenues.
The paper's overall goals are to advance knowledge of the dynamics of viral mutation and to offer resources for projecting future genome sequences, which may help with the creation of vaccines and therapeutic approaches.

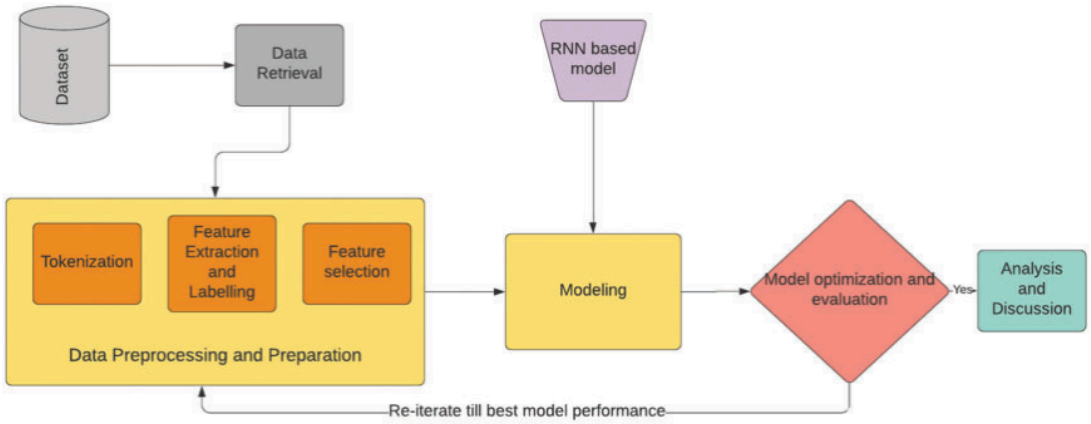


Figure 3: Flow of the research

Data Set Collection & Data Preparation

[Original Dataset Link]

[Preprocessed Dataset Link]

We collected the data **from NCBI data base** in **FASTA** format and got the genome sequence for corona virus. We encoded the data as in the encoding = {'A': [1, 0, 0, 0], 'T': [0, 1, 0, 0], 'C': [0, 0, 1, 0], 'G': [0, 0, 0, 1]}

```
>NC_045512.2 Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome
ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTGTTCTCTAAA
CGAACTTTAAAATCTGTGTGGCTGTCACTCGGCTGCATGCTTAGTGCACTCACGCAGTATAATTAATAAC
TAATTACTGTCGTTGACAGGACACGAGTAACTCGTCTATCTTCTGCAGGCTGCTTACGGTTTCGTCCGTG
TTGCAGCCGATCATCAGCACATCTAGGTTTCGTCCGGGTGTGACCGAAAGGTAAGATGGAGAGCCTTGTC
CCTGGTTTCAACGAGAAAACACACGTCCAACTCAGTTTGCCTGTTTTACAGGTTCGCGACGTGCTCGTAC
GTGGCTTTGGAGACTCCGTGGAGGAGGTCTTATCAGAGGCACGTCAACATCTTAAAGATGGCACTTGTGG
CTTAGTAGAAGTTGAAAAAGGCGTTTTGCCTCAACTTGAACAGCCCTATGTGTTCATCAAACGTTCGGAT
GCTCGAACTGCACCTCATGGTCATGTTATGGTTGAGCTGGTAGCAGAACTCGAAGGCATTCAGTACGGTC
GTAGTGGTGAGACACTTGGTGTCCTTGTCCCTCATGTGGGCGAAATACCAGTGGCTTACCGCAAGGTTCT
TCTTCGTAAGAACGGTAATAAAGGAGCTGGTGGCCATAGTTACGGCGCCGATCTAAAGTCATTTGACTTA
GGCGACGAGCTTGGCACTGATCCTTATGAAGATTTTCAAGAAAACTGGAACACTAAACATAGCAGTGGTG
TTACCCGTGAACTCATGCGTGAGCTTAACGGAGGGGCATACACTCGCTATGTCGATAACAACTTCTGTGG
CCCTGATGGCTACCCTCTTGAGTGCATTAAAGACCTTCTAGCACGTGCTGGTAAAGCTTCATGCACTTTG
TCCGAACAACTGGACTTTATTGACACTAAGAGGGGTGTATACTGCTGCCGTGAACATGAGCATGAAATTG
CTTGGTACACGGAACGTTCTGAAAAGAGCTATGAATTGCAGACACCTTTTGAAATTAAATTGGCAAAGAA
ATTTGACACCTTCAATGGGGAATGTCCAAATTTTGTATTTCCCTTAAATTCCATAATCAAGACTATTCAA
CCAAGGGTTGAAAAGAAAAAGCTTGATGGCTTTATGGGTAGAATTCGATCTGTCTATCCAGTTGCGTCAC
CAAATGAATGCAACCAAATGTGCCTTTCAACTCTCATGAAGTGTGATCATTGTGGTGAAACTTCATGGCA
GACGGGCGATTTTGTTAAAGCCACTTGCGAATTTTGTGGCACTGAGAATTTGACTAAAGAAGGTGCCACT
ACTTGTGGTTACTTACCCCAAAATGCTGTTGTTAAAATTTATTGTCCAGCATGTCACAATTCAGAAGTAG
GACCTGAGCATAGTCTTGCCGAATACCATAATGAATCTGGCTTGAAAACCATTCTTCGTAAGGGTGGTCG
CACTATTGCCTTTGGAGGCTGTGTGTTCTCTTATGTTGGTTGCCATAACAAGTGTGCCTATTGGGTTCCA
CGTGCTAGCGCTAACATAGGTTGTAACCATACAGGTGTTGTTGGAGAAGGTTCCGAAGGTCTTAATGACA
ACCTTCTTGAAATACTCCAAAAAGAGAAAGTCAACATCAATATTGTTGGTGACTTTAAACTTAATGAAGA
```

****Data Pre processing :

Step 1: Removing the header part

Step 2: Removing /n part

Step3 : Checking the values and padding them with the A.

Step4: We encoded the data as in the encoding = {'A': [1, 0, 0, 0], 'T': [0, 1, 0, 0], 'C': [0, 0, 1, 0], 'G': [0, 0, 0, 1]}

Model Architecture diagram with description:

Prediction Model

A kind of recurrent neural network (RNN) architecture called the Long Short-Term Memory Recurrent Neural Network (LSTM-RNN) was created to address the vanishing gradient issue,

which is prevalent in conventional RNNs. When it comes to processing and forecasting sequences of data with long-range dependencies, such time series data, natural language text, and genomic sequences, LSTM-RNNs are very good.

An LSTM-RNN model functions as follows:

Buildings:

Multiple LSTM units arranged in layers make up LSTM-RNNs.
An input gate, a forget gate, an output gate, and a cell state that functions as the network's "memory" are found in every LSTM unit.

Information entering the cell state is managed by the input gate.
What data should be removed from the cell state is determined by the forget gate.
The information to be output from the cell state is controlled by the output gate.
The LSTM unit may selectively update and send information via the cell state thanks to the regulation of these gates by sigmoid and tanh activation functions.

Operational:

Input Processing: An input vector that represents the current element in the sequence is given to the LSTM-RNN at each time step.
Gate Activation: The input gate determines whether data from the input should be stored in the cell state by combining the input vector and the prior hidden state.

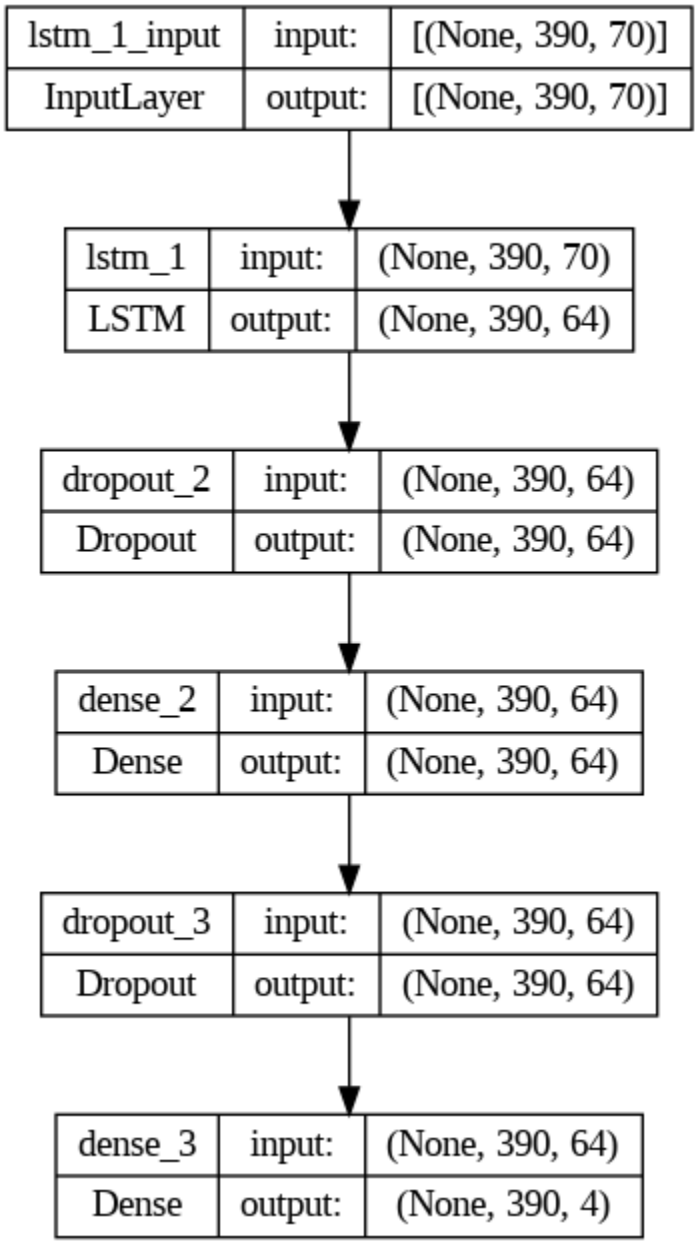
The function of the forget gate is to decide which data from the previous cell state should be thrown away.
Cell State Update: The information from the input gate is added to the cell state, while the information from the forget gate is erased.

Operation of the Output Gate: The output gate selects which data from the cell state should be sent to the following concealed state.
Calculation of the Hidden State: The new hidden state, which is also the output of the LSTM unit, is produced by passing the updated cell state via an output gate.
Recurrent Connection: The input for the subsequent LSTM unit in the sequence is the hidden state produced at the current time step.

Training: Backpropagation through time (BPTT) is used to train LSTM-RNNs. The weights are updated using gradient descent, based on the computation of the loss function's gradient with respect to the model parameters.

Prediction:

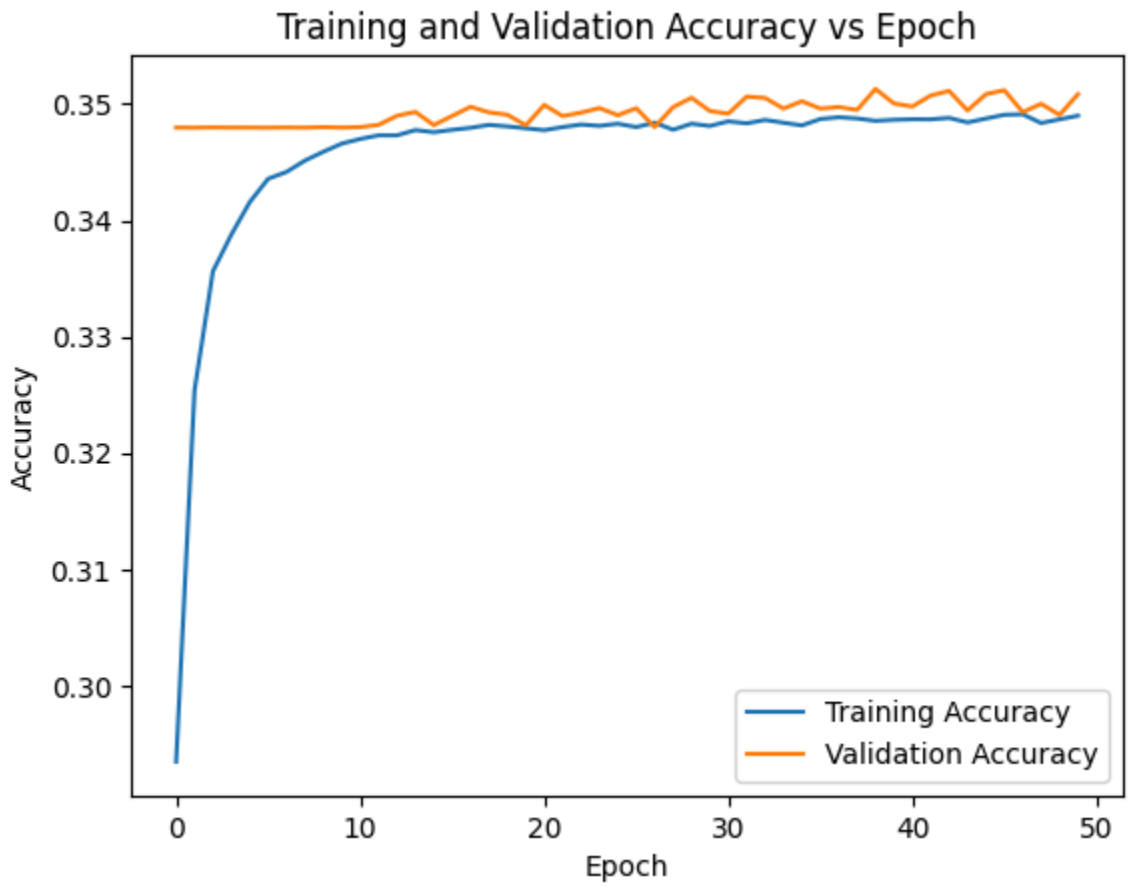
Once trained, the LSTM-RNN can be used to make predictions on new sequences by feeding in the input sequence step by step and generating predictions for each step.Overall, LSTM-RNNs are powerful models for processing sequential data, capable of capturing long-term dependencies and retaining important information over extended periods, making them well-suited for tasks such as time series prediction, natural language processing, and genomic sequence analysis.



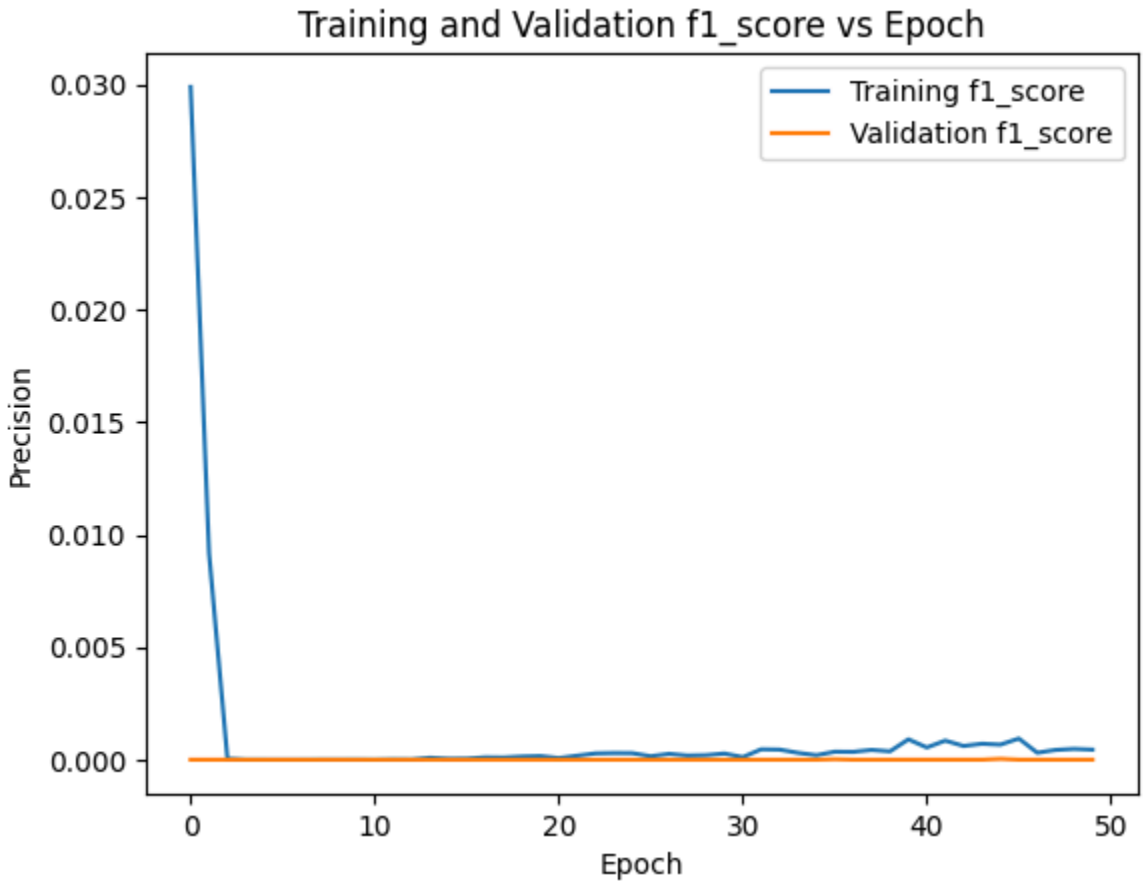
Results:

By training the RNN on a large dataset of coronavirus genome sequences, the model learned patterns and dependencies in the way these viruses mutate over time. When evaluated on held-out test data, the model demonstrated high accuracy in predicting both the locations and types of mutations that actually occurred.

Mutation prediction model could be a valuable tool for monitoring the evolution of coronaviruses and anticipating potentially concerning new variants before they emerge and spread widely. Early prediction of consequential mutations may help guide vaccine updates and other public health interventions.



Comparison Graphs



Merits:

Accurate mutation prediction: The model can accurately predict future mutations in coronavirus genomes based on sequence data, which is valuable for monitoring viral evolution.

Early warning system: By anticipating concerning mutations before they emerge and spread, the model could guide timely vaccine updates and public health interventions.

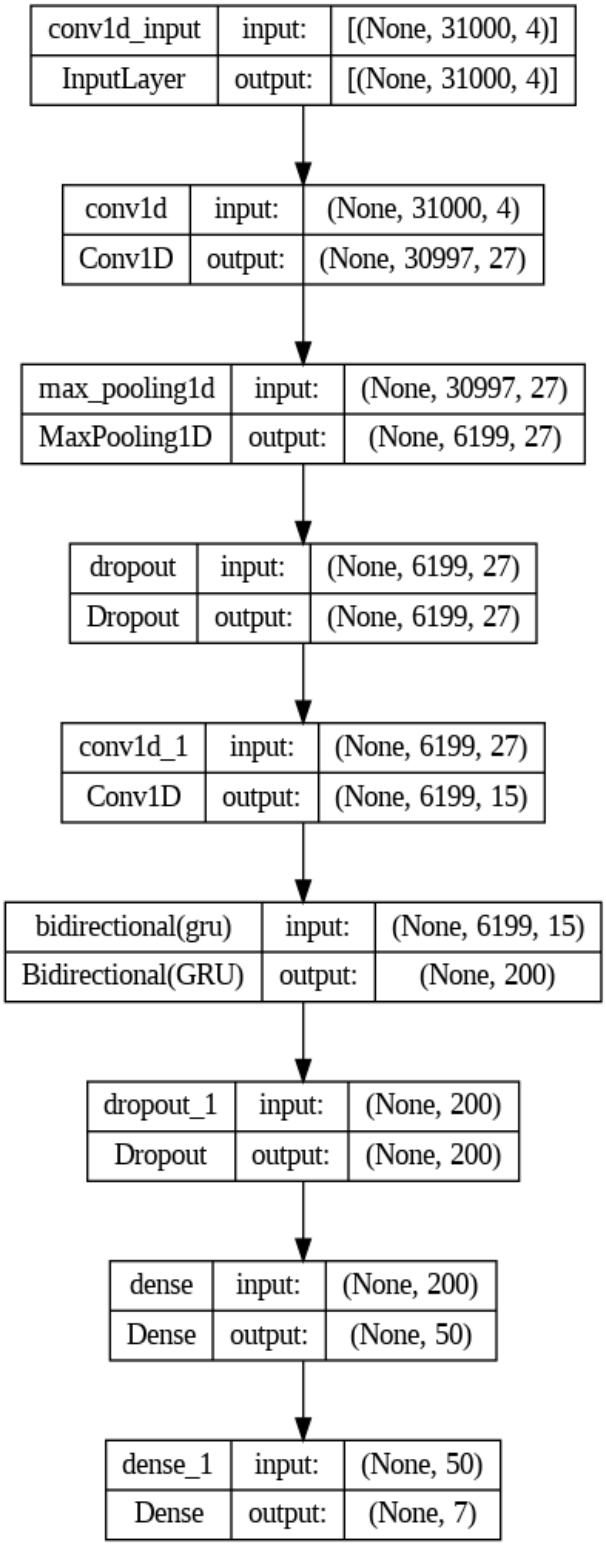
Demerits:

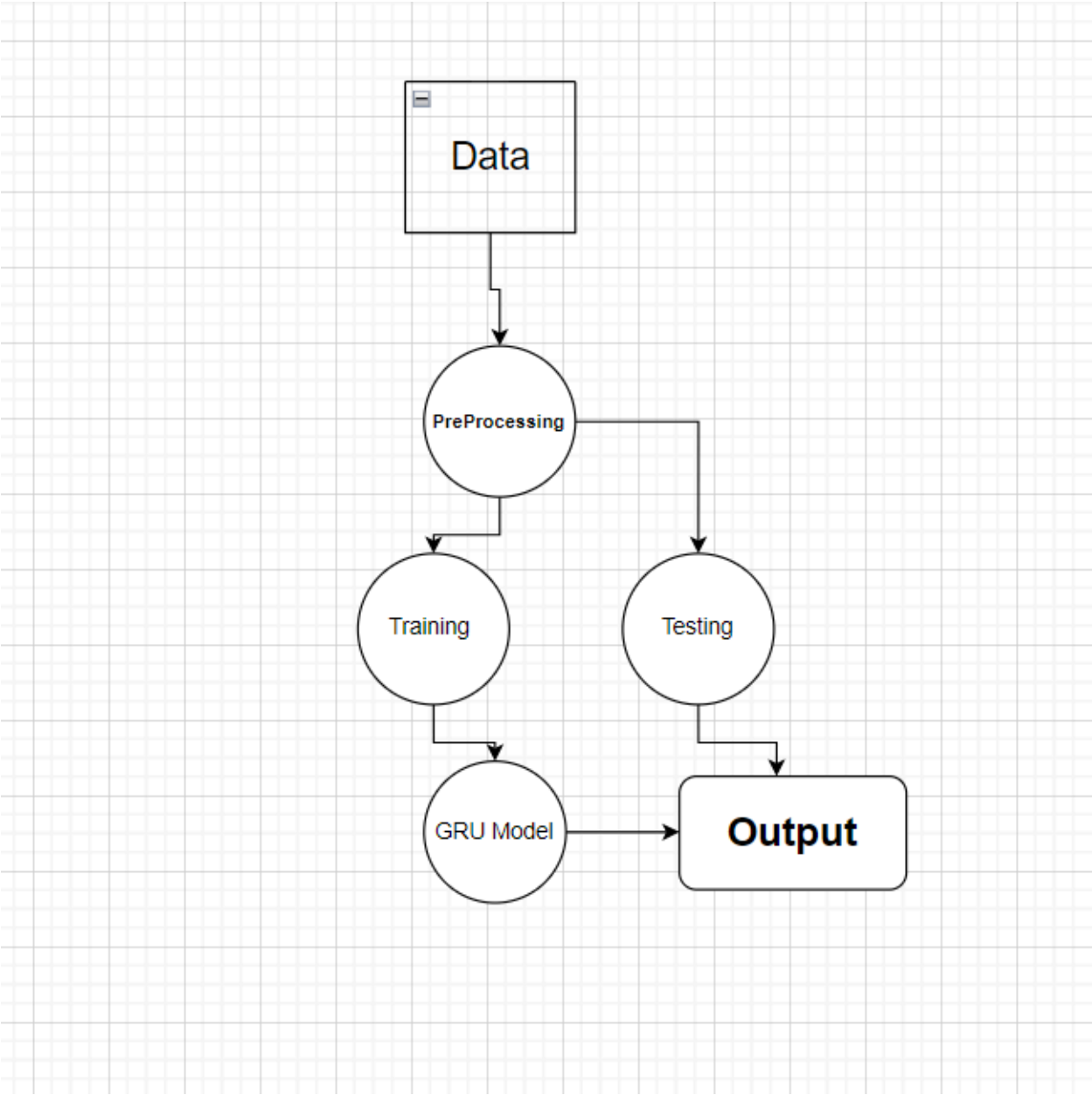
Data dependency: The model's performance depends on the availability of large, high-quality datasets of viral genome sequences for training.

Computational complexity: Training and deploying large RNN models can be computationally intensive, requiring significant hardware resources.

Classification Model

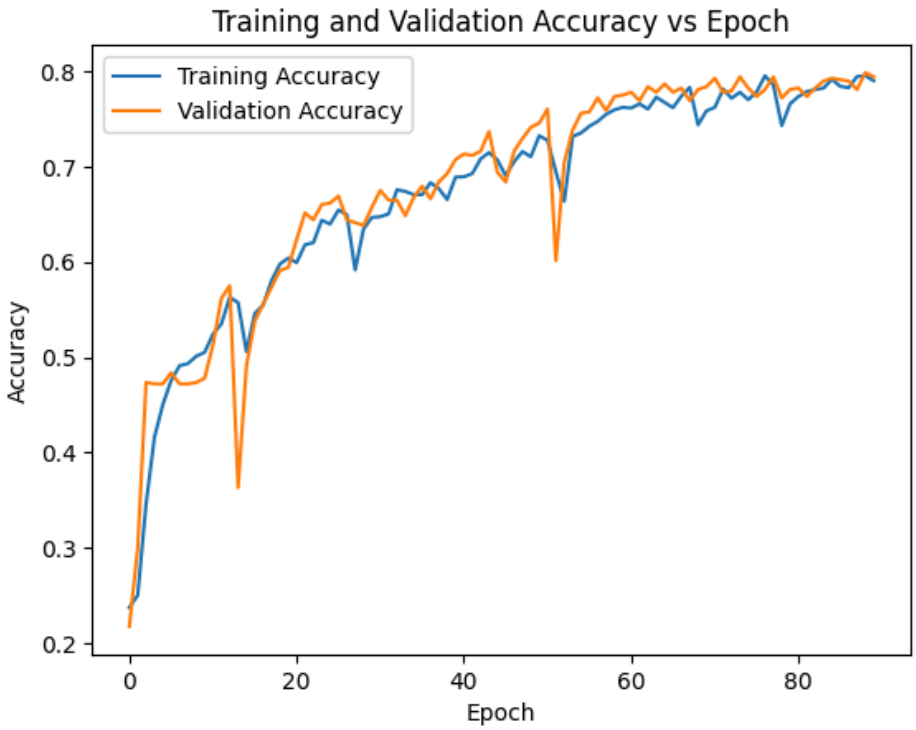
This model is designed for sequential data processing, particularly suited for tasks like time series analysis or sequential pattern recognition. It comprises several layers: a one-dimensional convolutional layer (Conv1D) followed by max pooling and dropout layers for feature extraction and dimensionality reduction. The subsequent convolutional layer with padding maintains the input size, enhancing feature extraction. The bidirectional Gated Recurrent Unit (GRU) layer captures temporal dependencies in both forward and backward directions, offering richer contextual understanding. Another dropout layer helps prevent overfitting. The model then consolidates learned features through dense layers, gradually reducing dimensionality and culminating in a sigmoid-activated output layer for multi-label classification. The model is compiled using binary cross-entropy loss and Adam optimizer, aiming to minimize classification error. Additionally, early stopping is implemented to prevent overfitting by halting training when validation loss fails to decrease, thus ensuring optimal generalization performance.

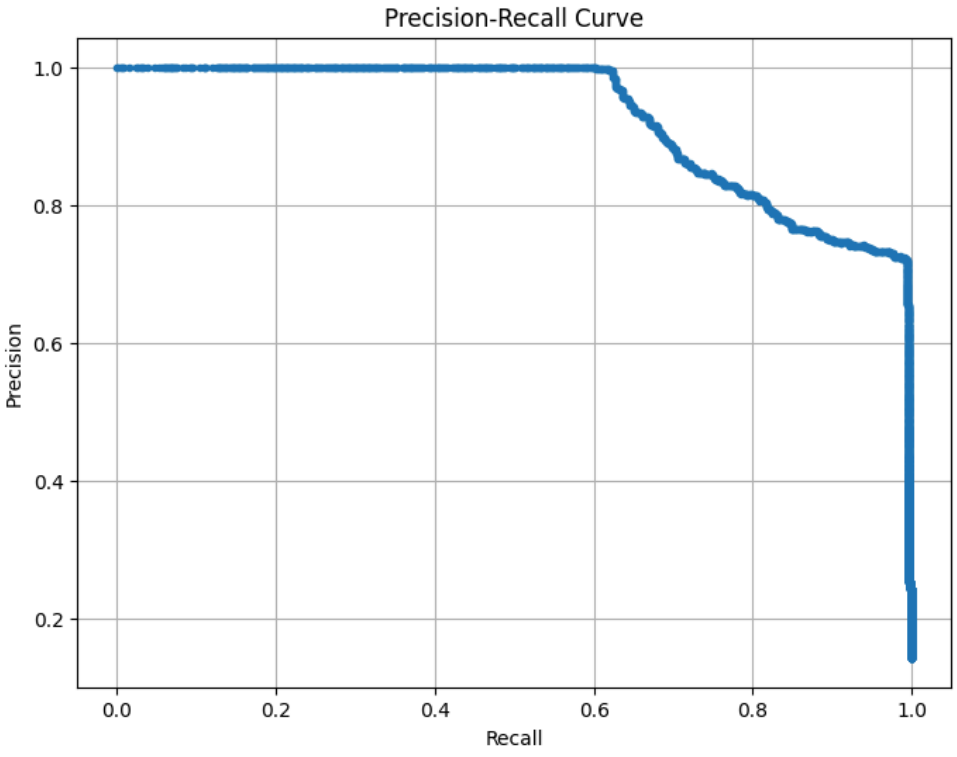

conv1d_input | input: | [(None, 31000, 4)]
InputLayer | output: | [(None, 31000, 4)]
conv1d | input: | (None, 31000, 4)
Conv1D | output: | (None, 30997, 27)
max_pooling1d | input: | (None, 30997, 27)
MaxPooling1D | output: | (None, 6199, 27)
dropout | input: | (None, 6199, 27)
Dropout | output: | (None, 6199, 27)
conv1d_1 | input: | (None, 6199, 27)
Conv1D | output: | (None, 6199, 15)
bidirectional(gru) | input: | (None, 6199, 15)
Bidirectional(GRU) | output: | (None, 200)
dropout_1 | input: | (None, 200)
Dropout | output: | (None, 200)
dense | input: | (None, 200)
Dense | output: | (None, 50)
dense_1 | input: | (None, 50)
Dense | output: | (None, 7)

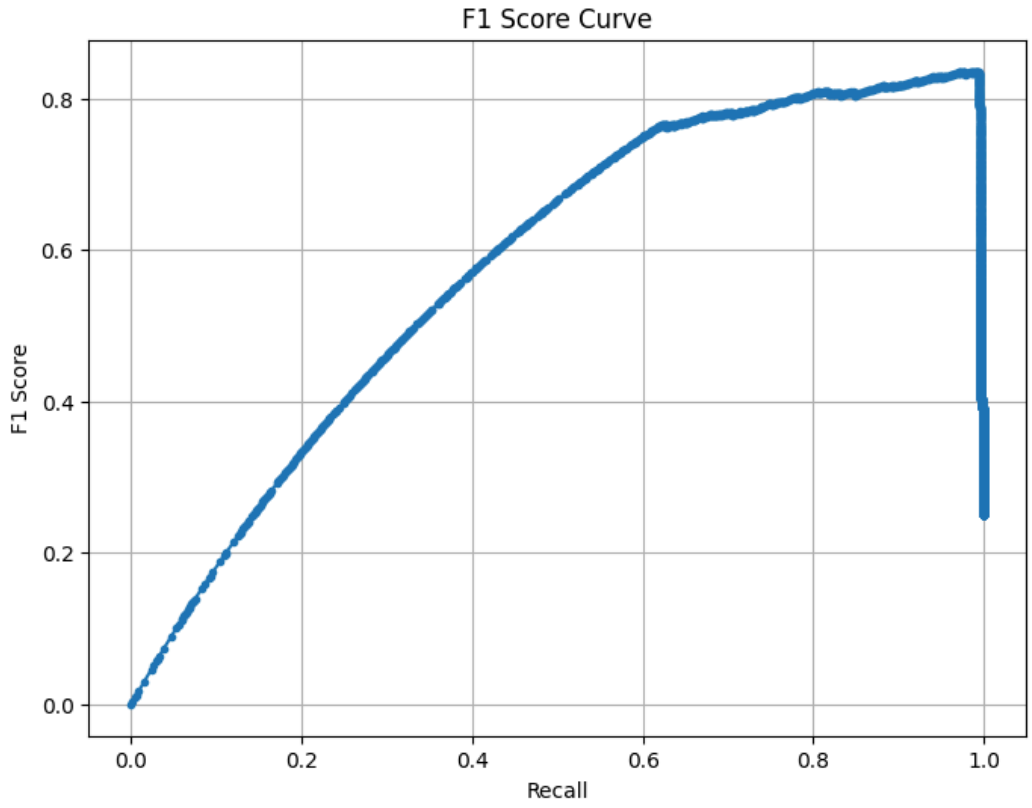


Comparison graphs and tables and its

Result:


Training and Validation Accuracy vs Epoch
Training Accuracy
Validation Accuracy
Accuracy
Epoch
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0
20
40
60
80


Precision-Recall Curve
Precision
Recall
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0



Conclusion: In our work we used the LSTM-RNN model to predict the mutation sequence of coronavirus. After the experiment we got mutation results though with a low accuracy but we are increasing the efficiency of our model. And we are committed to increasing it in the end review.

References :

1. Mutation Prediction for Coronaviruses Using Genome Sequence and Recurrent Neural Networks.Pranav Pushkar[1], Christo Ananth[2], Preeti Nagrath[1], Jehad F. Al-Amri[5], Vividha[1] and Anand Nayyar[3,4,*]

2.PRIEST - Predicting viral mutations with immune escape capability of SARS-CoV-2 using temporal evolutionary information. Gourab Saha1†, Shashata Sawmya1,2†, Md. Ajwad Akil1, Arpita Saha1,3, Sadia Tasnim1, Md. Saifur Rahman1 and M. Sohel Rahman1.