Viral Disease prediction And Mutation Pattern Extraction

Saiteja, Rupesh, Varun,Rahul Promod,Kusuma
Group No-3

Problem Statement

Viral disease prediction and mutation pattern extraction using deep learning

The study aims to develop a deep learning model that can accurately predict the viral disease by DNA sequences and also extracting the mutation pattern.

Motivation

Early and accurate diagnosis of viral diseases is crucial for patient treatment and controlling outbreaks.Understanding how mutations affect the virus can be instrumental in designing drugs and vaccines.Deep learning models excel at finding complex patterns in large datasets, making them well-suited for analyzing DNA sequences

Keywords: LSTM, CNN, one-hot encoding.

I. INTRODUCTION

In an era where viral diseases continue to pose significant threats to global public health, the ability to predict outbreaks and understand mutation patterns becomes paramount. Our project, "Viral Disease Prediction and Mutation Pattern Extraction," delves into the realm of bioinformatics to address this pressing need.

With the emergence of diseases such as SARS-CoV-1, MERS-CoV, SARS-CoV-2, Ebola, Dengue, and Influenza, there is a critical demand for sophisticated tools that can not only forecast the spread of these diseases but also unravel the intricate mutation patterns that shape their evolution. Leveraging cutting-edge technologies including Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) models, our project endeavors to provide accurate predictions of disease outbreaks, aiding in proactive public health measures and resource allocation.
Moreover, our focus extends beyond prediction to mutation pattern extraction. By dissecting the genetic sequences of these viral pathogens, we aim to uncover the underlying patterns of mutation that drive their evolution. Understanding these mutations is crucial for developing targeted therapeutics, vaccines, and diagnostic tools that can effectively combat these diseases.

This documentation serves as a comprehensive guide to our project, detailing our methodologies, findings, and implications for the field of bioinformatics and public health. Through our collective efforts, we strive to contribute to the ongoing battle against viral diseases, ultimately safeguarding the health and well-being of populations worldwide.

II. Base Paper Link AND Related papers

Our base paper is COVID-DeepPredictor: Recurrent Neural Network to Predict SARS-CoV-2 and Other Pathogenic Viruses.

- Frontiers | COVID-DeepPredictor: Recurrent Neural Network to Predict SARS-CoV-2 and Other Pathogenic Viruses (frontiersin.org)

Below mentioned are the other related papers.

- A Deep Learning Approach for Viral DNA Sequence Classification using Genetic Algorithm (thesai.org)
- https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0222271
- https://www.hindawi.com/journals/cmmm/2021/1835056/

III. Preprocessed Dataset source link

There are a whole of 6 datasets.out of that one is used for training purpose and other 5 are used for testing purpose.

TABLE I
DATASET SIZE AND LINKS

s.no	I. Appearance (in Time New Roman or Times)		
	Data set Name	**size**	**Link**
1	Training dataset	1500	Link-1
2	Testdata-1	3143	Link-2
3	Testdata-2	1090	Link-3
4	Testdata-3	4000	Link-4
5	Testdata-4	3500	Link-5
6	Testdata-5	5010	Link-6

IV. Features extracted from dataset

We found out features of datasets like ,size of datasets distribution of the datasets etc..

A. Training dataset

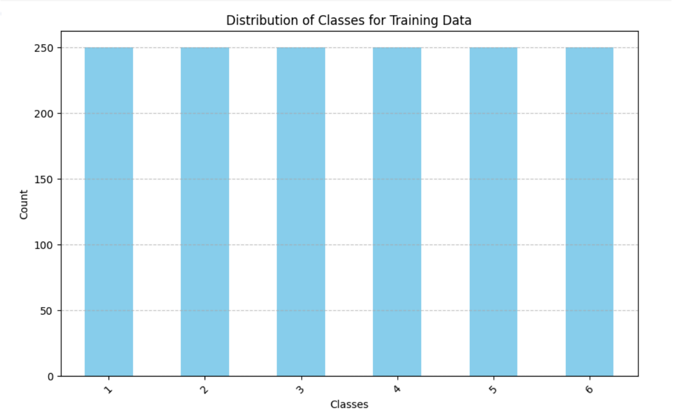


Figure 1: Distribution of classes in Training data set

B. Testdata-1

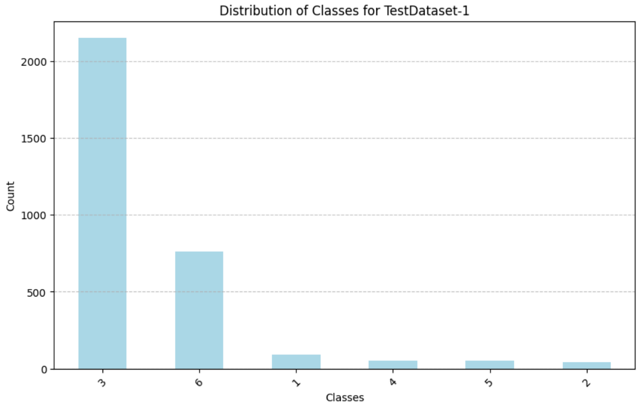


Figure 2: Distribution of classes in Test data set-1

C. Testdata-2

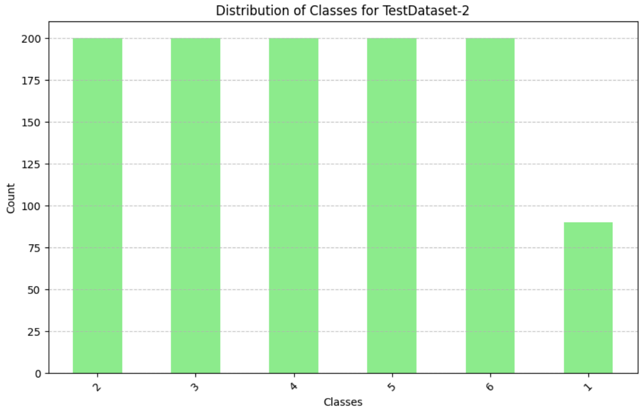


Figure 3: Distribution of classes in Test data set-2

D. Testdata-3

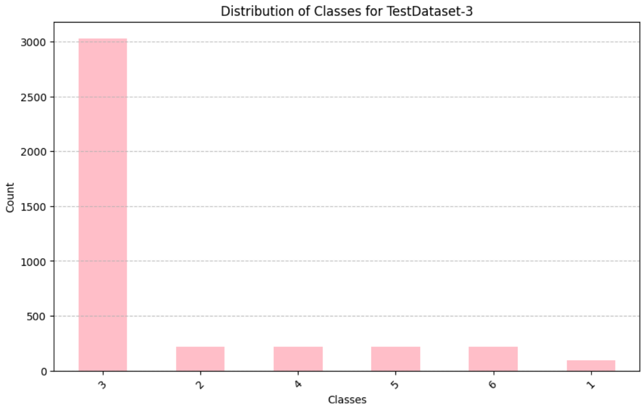


Figure 4: Distribution of classes in Test data set-3

E. Testdata-4

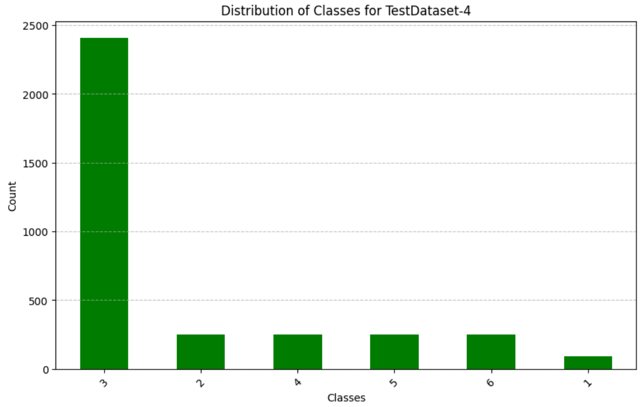


Figure 5: Distribution of classes in Test data set-4

F. Testdata-5

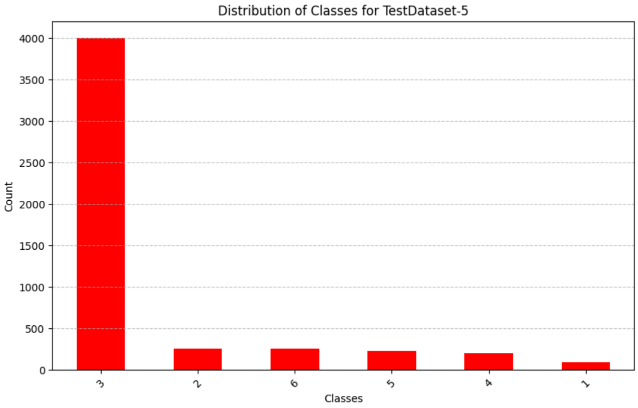
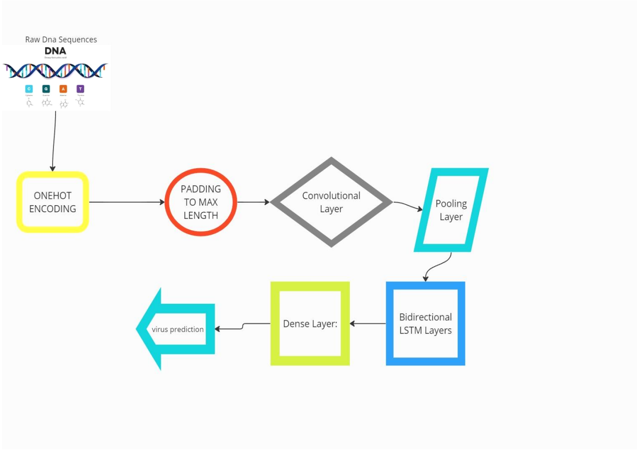


Figure 6: Distribution of classes in Test data set-5

V. Model Architecture



In our approach, we employed a combination of Convolutional Neural Network (CNN) and Bidirectional Long Short-Term Memory (Bi-LSTM) models for training. Initially, the raw DNA sequences underwent conversion into numerical format using a one-hot encoder. Each sequence was then padded to match the length of the longest DNA sequence in the dataset.

These preprocessed sequences were subsequently passed through a convolutional layer. The output from this convolutional layer was then fed into a pooling layer for feature extraction. Following this, the features were passed onto a Bi-LSTM network, which is capable of capturing sequential patterns in both forward and backward directions. Finally, a dense layer was added to the network for classification, enabling the prediction of the correct class label.

VI. Results & Comparison graphs

We conducted evaluations on five distinct test datasets and obtained varying results. The highest accuracy was achieved on test data 5, reaching 95%. Conversely, the lowest accuracy was recorded on test data 2, with a performance of 80%.

S.No	Data Set Name	Accuracy
1	Testdata-1	93.0957

2	Testdata-2	80.0000
3	Testdata-3	90.4999
4	Testdata-4	91.4285
5	Testdata-5	95.3692

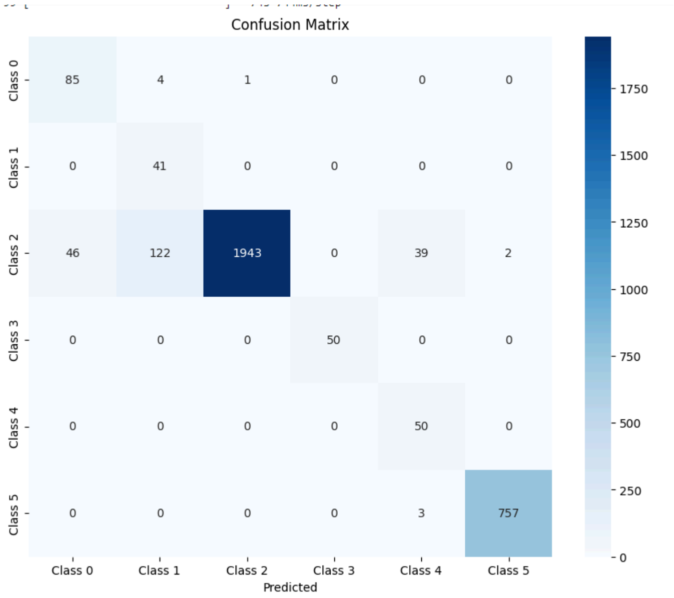


Figure 7: Confusion matrix of Test data 1

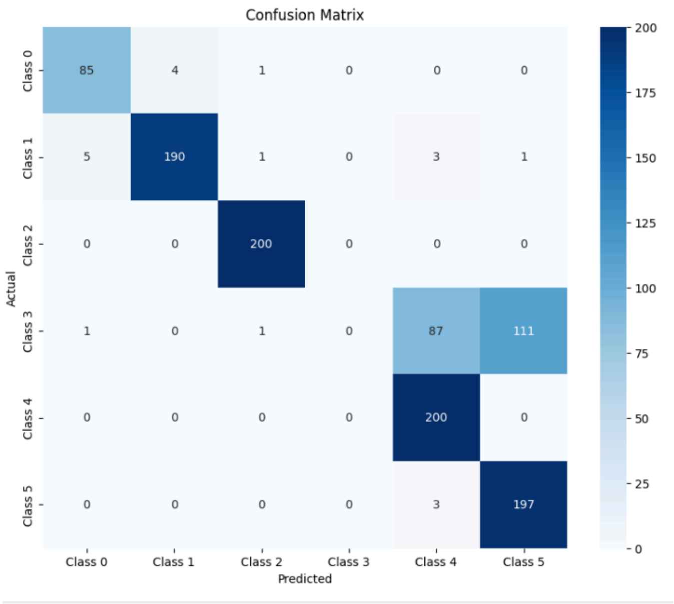


Figure 8: Confusion matrix of Test data 2

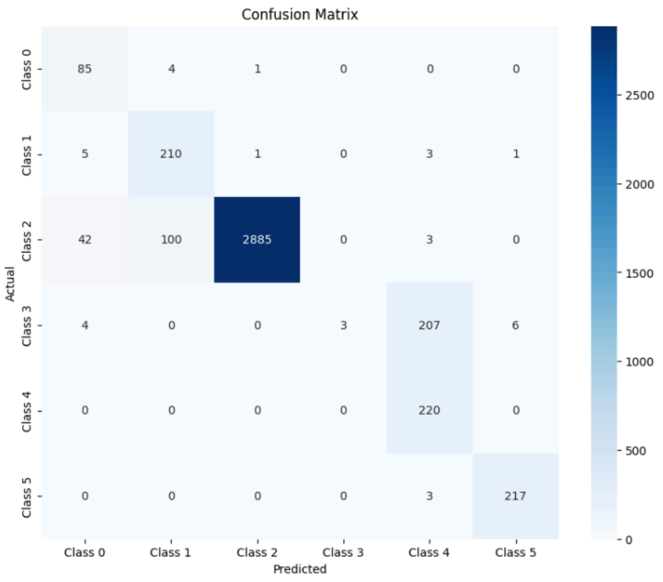


Figure 9: Confusion matrix of Test data 3

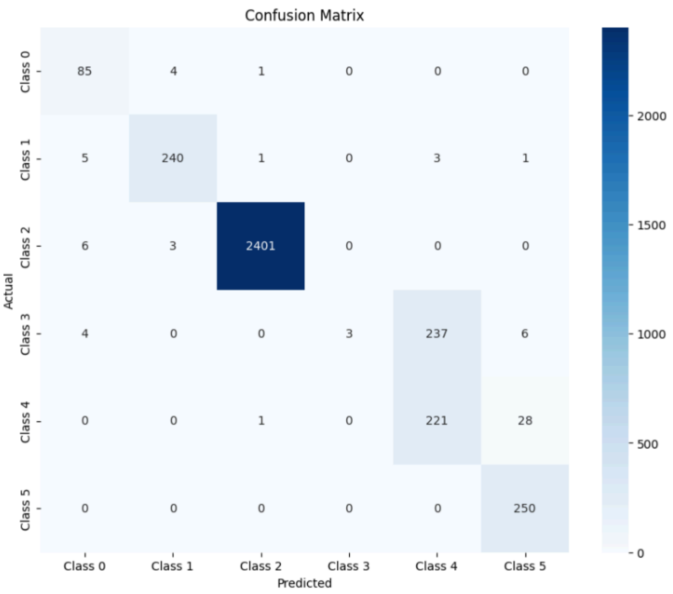


Figure 10: Confusion matrix of Test data 3

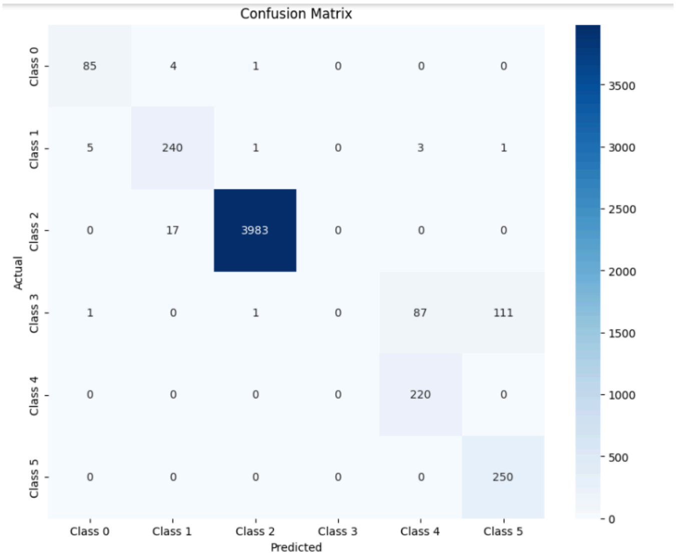


Figure 10: Confusion matrix of Test data 3

VII. Merits and Demerits

We achieved favorable results in accurately predicting virus classes. The padding technique employed ensured that all features of the sequences were retained, unlike other sequence truncation methods where valuable information might be lost.

However, despite our overall success, there were notable drawbacks. Particularly, the prediction performance for the Ebola virus was notably poor. This deficiency can be attributed to factors such as the length of sequences and the limited number of samples available. Consequently, these factors constitute a significant limitation to our approach.

VIII. Full Code and Execution Procedure

Performs one-hot encoding of DNA sequences and encodes labels into categorical form.
Constructs a neural network architecture comprising Conv1D, MaxPooling1D, Dropout, Bidirectional LSTM, and Dense layers.
Compiles the model with categorical cross-entropy loss and the Adam optimizer.
Trains the model on the training data for a specified number of epochs

.

Loads test data from another CSV file.

Preprocesses the test data similarly to the training data. and test it on the model and
generate a confusion matrix to evaluate the model's performance on the test data

.

IX. Mutation Analysis

In this mutation analysis, DNA sequences from a reference file and a query file are compared base by base. For each position, if the nucleotide in the query sequence differs from the corresponding nucleotide in the reference sequence, a mutation is recorded. These mutations are tallied, and the frequency of each mutation type is calculated. The top 10 mutation types are visualized using a horizontal bar plot, showing the frequency of occurrence for each mutation type.

Mutation Analysis has been done for 4 viruses namely **MERS-Cov-2, SARS-Cov-2, Influenza, Dengue.** Workflow of mutation analysis, code snippet and some visualizations are presented below.

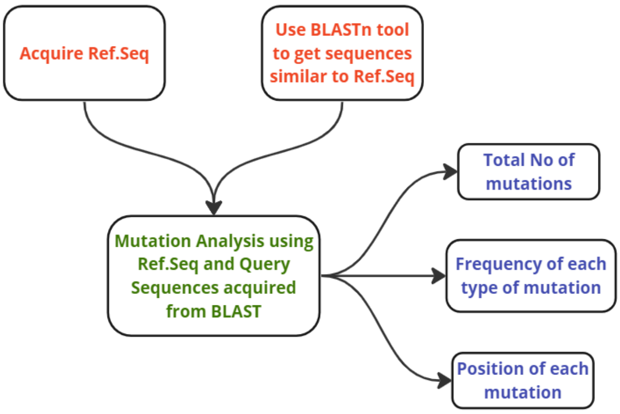
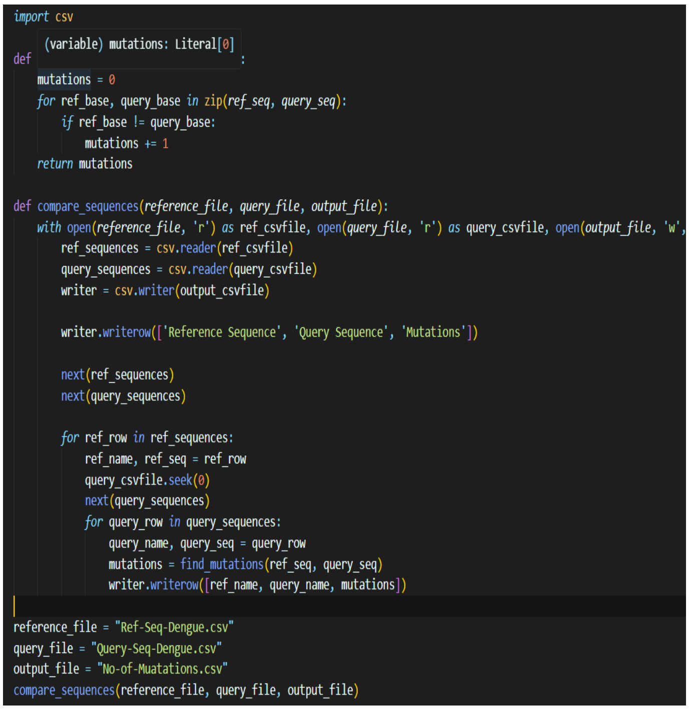

Acquire Ref.Seq
Use BLASTn tool to get sequences similar to Ref.Seq
Mutation Analysis using Ref.Seq and Query Sequences acquired from BLAST
Total No of mutations
Frequency of each type of mutation
Position of each mutation

Figure 11: Workflow of mutation extraction



```
import csv
    (variable) mutations: Literal[0]
def                                       :
    mutations = 0
    for ref_base, query_base in zip(ref_seq, query_seq):
        if ref_base != query_base:
            mutations += 1
    return mutations

def compare_sequences(reference_file, query_file, output_file):
    with open(reference_file, 'r') as ref_csvfile, open(query_file, 'r') as query_csvfile, open(output_file, 'w',
        ref_sequences = csv.reader(ref_csvfile)
        query_sequences = csv.reader(query_csvfile)
        writer = csv.writer(output_csvfile)

        writer.writerow(['Reference Sequence', 'Query Sequence', 'Mutations'])

        next(ref_sequences)
        next(query_sequences)

        for ref_row in ref_sequences:
            ref_name, ref_seq = ref_row
            query_csvfile.seek(0)
            next(query_sequences)
            for query_row in query_sequences:
                query_name, query_seq = query_row
                mutations = find_mutations(ref_seq, query_seq)
                writer.writerow([ref_name, query_name, mutations])

reference_file = "Ref-Seq-Dengue.csv"
query_file = "Query-Seq-Dengue.csv"
output_file = "No-of-Muatations.csv"
compare_sequences(reference_file, query_file, output_file)
```

Figure 12: Code snippet of mutation Analysis

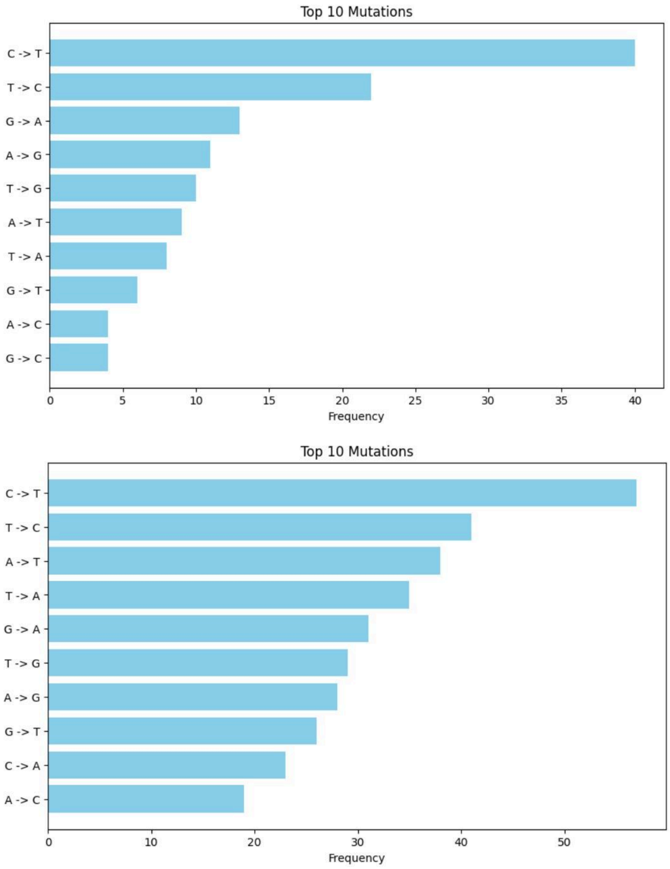


Figure 13, 14: Top 10 mutations of a random query sequence for MERS-CoV

X. CONCLUSION

In conclusion, our bioinformatics project on Viral Disease Prediction and Mutation Pattern Extraction represents a significant advancement in the field. By employing CNN and LSTM models, coupled with one-hot encoding technique, we have successfully developed robust predictive algorithms for six major diseases including SARS-CoV-1, MERS-CoV, SARS-CoV-2, Ebola, Dengue, and Influenza. Furthermore, our comprehensive analysis of mutation patterns across these diseases sheds light on their evolutionary dynamics, providing valuable insights for the development of targeted interventions and therapies. Through our meticulous approach and innovative methodologies, we have laid the foundation for proactive strategies in combating viral outbreaks and advancing public health efforts on a global scale.

XI. REFERENCES

[1] Saha I, Ghosh N, Maity D, Seal A and Plewczynski D (2021) COVID-DeepPredictor: Recurrent Neural Network to Predict SARS-CoV-2 and Other Pathogenic Viruses. Front. Genet. 12:569120. doi: 10.3389/fgene.2021.569120

[2] Ahmed El-Tohamy, Huda Amin Maghwary and Nagwa Badr, "A Deep Learning Approach for Viral DNA Sequence Classification using Genetic Algorithm" International Journal of Advanced Computer Science and Applications(ijacsa), 13(8), 2022. http://dx.doi.org/10.14569/IJACSA.2022.0130861

[3] Tampuu, Ardi, et al. "ViraMiner: Deep learning on raw DNA sequences for identifying viral genomes in human samples." PloS one 14.9 (2019): e0222271.

[4] Gunasekaran, H., Ramalakshmi, K., Arokiaraj, A. R., Kanmani, S. D., Venkatesan, C., & Dhas, C.S. G. (2021). Analysis of DNA sequence classification using CNN and hybrid models. Computational and Mathematical Methods in Medicine, 2021, Article ID 1835056