Assignment 1: Abstract and Introduction summary

➢ Abstract:

- We will talk about “analyzing the genome sequences of the human coronavirus: Predicting virulence and mutation”.
- this virus began to appear about two years ago, threatening the whole world.
- Covid-19 pandemic, caused by the SARS-CoV-2 genome sequence of coronavirus, has affected millions of people all over the world and taken millions of lives.
- It is of utmost importance that the character of this deadly virus be studied and its nature be analyzed.
- We use here an analysis pipeline comprising a classification exercise to identify the virulence of the genome sequences and extraction of important features from its genetic material that are used subsequently to predict mutation at those interesting sites using deep learning techniques.
- The accuracy of the SARS-CoV-2 genome sequencing and predicting mutations at some sites.
- We further use our mutation prediction pipeline to identify and analyze possible parents of a mutated genome sequence.
- We know that all human corona viruses have been traced to animal origins.

➢ Introduction:

Covid-19 was declared a global health pandemic on March 11, 2020. It is the biggest public health concern of this century.
It has already surpassed the previous two outbreaks due to the coronavirus, namely, Severe Acute Respiratory Syndrome Coronavirus (SARS Cov-2) and Middle East Respiratory Syndrome Coronavirus (MERS-Cov).
The virus acting behind this epidemic is known as Severe Acute Respiratory Syndrome Coronavirus 2 or SARS-CoV-2 virus, in short.
It is a single stranded RNA virus which is mainly 26,000 to 32,000 bases long on average.
The novel coronavirus is spherical in shape and has spike protein protruding from its surface.
These spikes assimilate into human cells, then undergo a structural change that allows the viral membrane to fuse with the cell membrane.
The host cell is then attacked by the viral gene through intrusion and it copies itself within the host cell, producing multiple new viruses.
The GISAID initiative database has been collecting high quality complete genome sequences of the SARS-CoV-2 virus from clinicians and researchers from around the world since the beginning of the COVID-19 outbreak.
To understand the virulence of the genome sequences and the nature of viral mutation, here, we present an analysis pipeline of the genome sequence leveraging the power of machine intelligence.

Overall, we present an analysis pipeline that can be further utilized as well as extended and revised to analyse its virulence, Example (with respect to the number of deaths its predecessors have caused in their respective countries) and to analyse the mutation at specific important sites of the viral genome.

Coronavirus infection spreads in clusters and some of the oldest and most effective containment measures such as social distancing, quarantine and isolation have been adopted to control the disease outbreak.

Early detection of new cases and the identification of factors associated with the spread of SARS-CoV-2 are important aspects to control the pandemic.

We aim to identify the representative viral genome sequences for the mutation prediction pipeline, to do that we have used an alignment-free genome sequence comparison method.

Training the models specifically for some South-Asian countries, India and Pakistan, We only used the best performing model for this analysis.