Analyzing hCov genome sequences: Applying Machine Intelligence and beyond

Shashata Sawmya*, Arpita Saha*, Sadia Tasnim*, Naser Anjum, Md. Toufikuzzaman, Ali Haisam Muhammad Rafid, Mohammad Saifur Rahman, M. Sohel Rahman

Department of CSE, BUET, ECE Building, West palashi, Dhaka-1205, Bangladesh

*** Equal Contribution**

Abstract

Covid-19 pandemic, caused by the sars-cov-2 strain of coronavirus, has affected millions of people all over the world and taken thousands of lives. It is of utmost importance that the character of this deadly virus be studied and its nature be analysed. We present here an analysis pipeline comprising phylogenetic analysis on strains of this novel virus to track its evolutionary history among the countries uncovering several interesting relationships, followed by a classification exercise to identify the virulence of the strains and extraction of important features from its genetic material that are used subsequently to predict mutation at those interesting sites using deep learning techniques. In a nutshell, we have prepared an analysis pipeline for hCov genome sequences leveraging the power of machine intelligence and uncovered what remained apparently shrouded by raw data.

1. Introduction

Covid-19 was declared a global health pandemic on March 11, 2020 [1]. It is the biggest public health concern of this century [22]. It has already surpassed the previous two outbreaks due to the coronavirus, namely, Severe Acute Respiratory Syndrome Coronavirus (SARS-Cov) and Middle East Respiratory Syndrome Coronavirus (MERS-Cov). The virus acting behind this epidemic is known as Severe Acute Respiratory Syndrome Coronavirus 2 or in short sars-cov-2 virus. It is a single stranded RNA virus

which is mainly 26,000 to 32,000 bases long in average [2]. 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 [3].

As of mid-April, 2020, about 10,000 of high-quality complete genome sequences were present in the GISAID initiative database [23] collected from clinicians and researchers from around the world. To understand the viral evolution and its nature of spread among the different countries, we present an analysis pipeline of the genome sequence leveraging the power of machine intelligence.

This paper makes the following key contributions.

A. An alignment-free phylogenetic analysis is carried out with a goal to uncover the evolutionary history of sars-cov-2. The resulting phylogenetic tree is able to highlight evolutionary relationships that can be explained by facts and figures and has further identified some mysterious relationships.
B. Several Machine Learning and Deep learning models are used to identify the virulence of the strains (i.e., to classify a virus strain as either severe or mild). Additionally, from the classification pipeline, important features are identified as Sites of Interest (SoIs) in the virus strains for further analysis.
C. Several CNN-RNN based models are used to predict mutations at specific Sites of Interest (SoIs) of the sars-cov-2 genome sequence followed by further analyses of the same on several South-Asian countries.
D. Overall, we present an analysis pipeline that can be further utilized as well as extended and revised (a) to study where a newly discovered genome sequence lies in relation to its predecessors in different regions of the world; (b) to analyse its virulence with respect to the number of deaths its predecessors have caused in their respective countries and (c) to analyse the mutation at specific important sites of the viral genome.

2. Methods

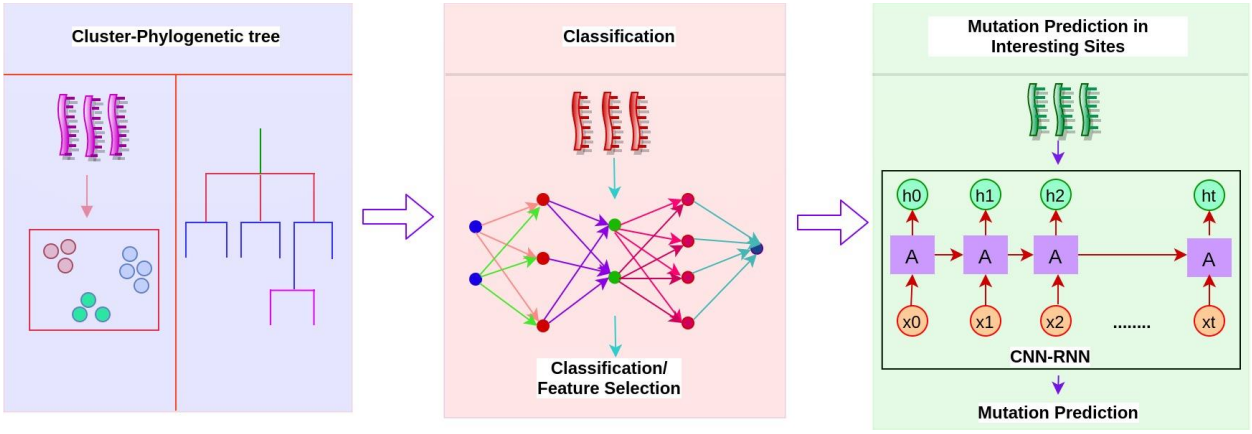


Figure 1: The whole analysis pipeline consisted of three phases. In the first phase, the genome sequences are divided into subsets based on country and a phylogenetic tree is constructed considering only the "representative" sequences of each such subset using an alignment-free sequence comparison approach. In the second phase, we employed state of the art classification algorithms, leveraging both traditional and deep learning pipelines to learn to discriminate the viral strains of many countries as either mild or severe. We also identify the features that contributed the most as the discriminant factor in the classification pipeline. Finally, we use the identified features from the previous stare to predict the mutation of the interesting sites in the viral strain using a deep learning model.

Figure 1 presents our overall analysis pipeline. Below we present the details of the pipeline.

2.1 Data Collection and Preprocessing

We have collected 10179 hCov genome sequences upto the date 24 April, 2020 (cut-off date) from the GISAID initiative dataset [23]. These are high quality complete viral genome sequences submitted by the scientists and scientific institutes of individual countries.

We also have collected country wise death statistics (upto cut-off date) from the official site of WHO [6]. The label was assigned based on a threshold of deaths which is the estimated median of the number of deaths in the data points. Any genome sequence of a country having deaths below (above) the threshold were considered a mild (severe) strain, i.e., assigned a label 0 (1). A sample labelling is shown in the supplementary

Table 1. Informatively, we have also considered some other metrics for labeling purposes albeit with unsatisfactory output (please see supplementary file for details). We divided the whole dataset into training and testing subset in 80/20 ratio with a balanced number of data points per class for traditional machine learning pipeline and for deep learning classification routine, we created the subsets training/validation/testing in 68/12/20 ratio.

2.2 Identifying Representative Viral Strains

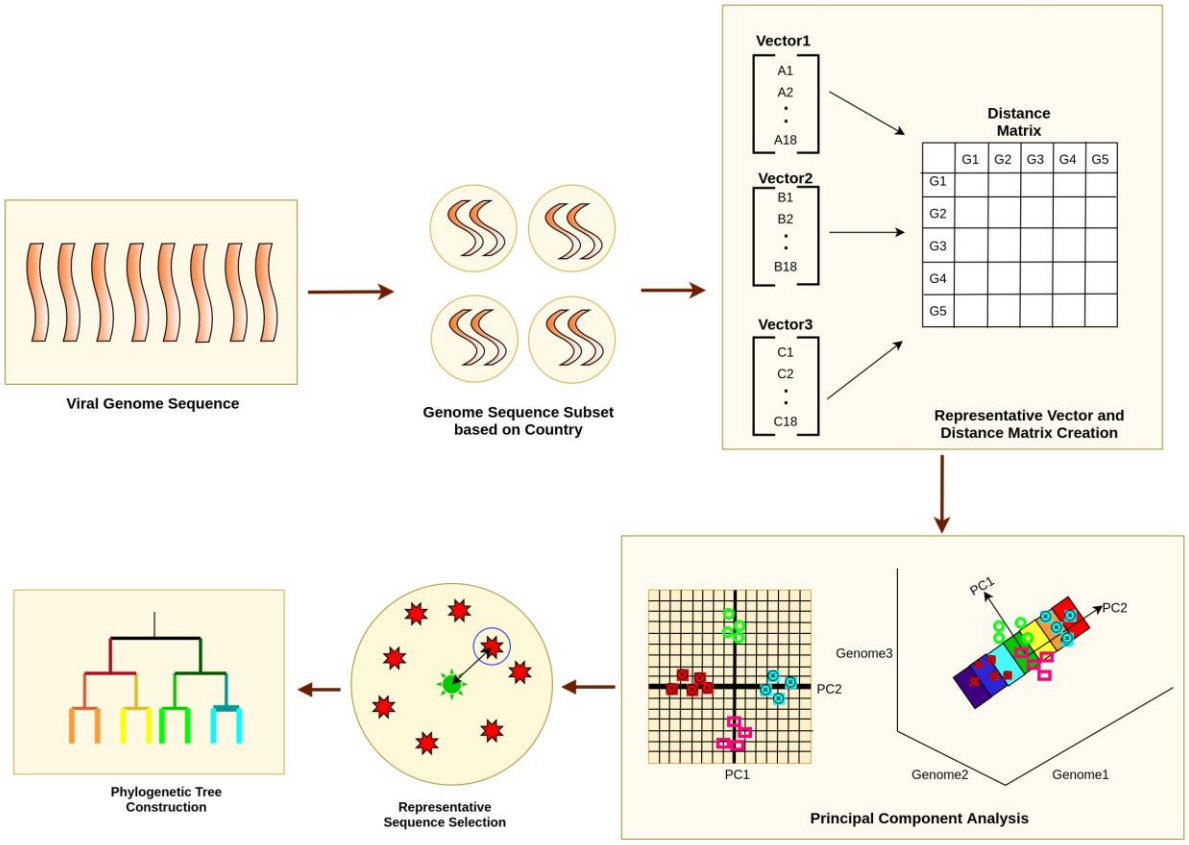


Figure 2: The Viral Genome Sequences were divided into subsets of sequences based on country. For each subset, each Viral Genome Sequence is converted into a vector representation and pairwise euclidean distance was calculated among the vectors to create the distance matrix. As the matrix is very high-dimensional, we used principal component analysis to find the principal component matrix from the distance matrix. Representative sequences were identified through K-means clustering on the PCA Matrix, and a phylogenetic tree was constructed from the representative sequence of each country.

We aim to identify and interpret the evolutionary relationships among the hCov Genome sequences uploaded at GISAID from different regions around the globe (Figure 2). To

do that we have used an alignment-free genome sequence comparison method as proposed in [5] as briefly described below. Notably, we do not consider any alignment-based method since it is not computationally feasible for us to align thousands of viral sequences for analysis and clustering purposes [4].

At first the sequence set is divided into subsets of sequences based on the location. All sequences are converted into representative $\mathbb{R}^{18}$vector. Pairwise distance among vectors derived from the fast vector method [5] are computed using Euclidean distance. Due to the high dimensionality of the resulting distance matrix, we resort to Principal Component Analysis (PCA) technique [9] to reduce the dimension of the matrix. Subsequently, we use K-means clustering [43] to identify the corresponding cluster centers. For the K-means clustering algorithm, we have used the implementation of [38] and used the default parameters except for the number of clusters which were set to 1 for determining the cluster center for each of the subsets. For each location-based cluster, the representative sequence (i.e., the “centroid” of the cluster) is then identified and used in the subsequent step of the pipeline.

2.3 Phylogeny Estimation

The evolutionary relationship among the representative sequences of different clusters (from Section 2.2) has been estimated by constructing a phylogenetic tree. We have used the Neighbor Joining algorithm [37] for phylogenetic tree construction since it is more reliable [25]. We have used Euclidean distance among the vectors, as described in the Section 2.2, to prepare the distance matrix.
While we predominantly have used the alignment-free method of [5], in this stage, we have only 67 representative sequences and hence we have also attempted a few other alignment-free and alignment-based methods to estimate the phylogenetic tree; however, these didn’t produce satisfactory results (more details are in supplementary file).

2.4 Classification Models

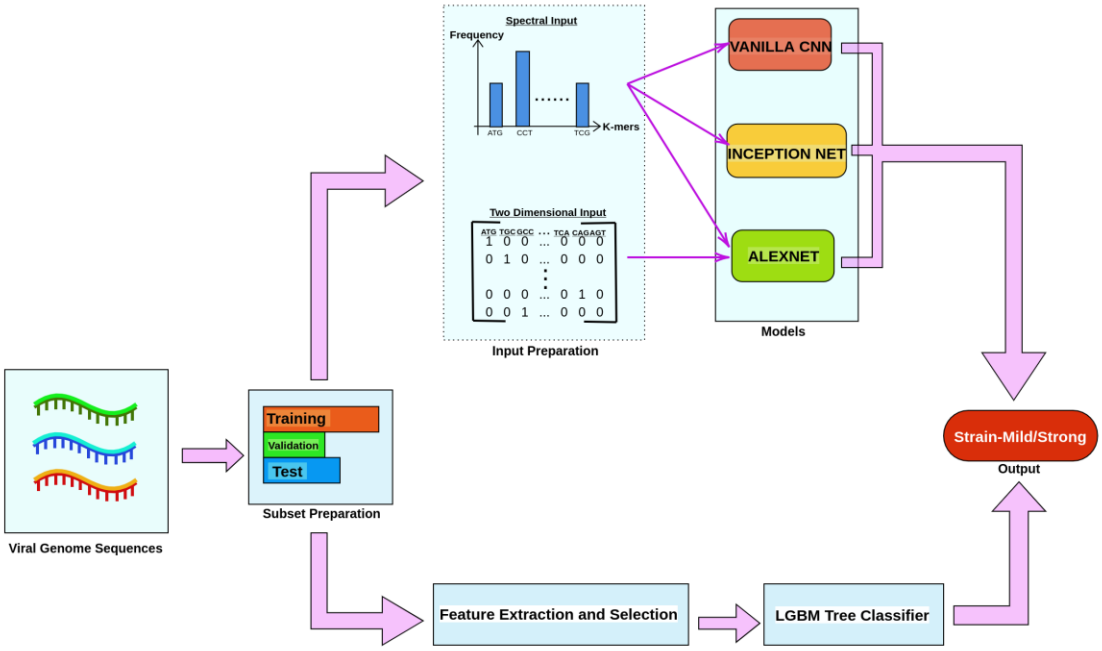


Figure 3: This figure illustrates our classification pipelines. We have leveraged the potential of both traditional and deep learning pipelines and in addition to doing an interesting classification task (mild vs. severe) we aim to identify, more importantly, the (sequence based) features and hence the corresponding sites of the the viral genome that are performing as the most discriminant feature with respect to classification.

Figure 3 illustrates our overall classification pipeline. We describe each stage in detail in the following subsections.

Traditional Machine Learning Pipeline:

For traditional machine learning, we use a pipeline similar to [12] (See Figure 3 in Supplementary file). We extracted three types of features from the genomic sequence of novel sars-cov-2. Inspired by the recent works [12][14][64][65] that focus only on sequences, we also extract only sequence-based features. These features are: position independent features, n-gapped dinucleotides and position specific features (see details in Section 3 of supplementary file). We use the gini value of the Extremely Randomized Tree (Extra Tree) classifier [13] to rank the features. Subsequently, only the features with gini value greater than the mean of the gini values are selected for training a LightGBM classifier model [15] (with default parameters) and performed 10-fold cross

validation. LightGBM is a highly efficient and fast gradient boosting framework which uses tree-based algorithms.

Important feature identification:

We use SHAP values and Univariate feature selection to compare the importance of the features. SHAP (SHapley Additive exPlanations) is a game theoretic approach which is used to explain the output of a model [44]. Univariate feature selection works by selecting the best features based on univariate statistical tests [50]. We use SelectKBest univariate feature selection to get the top K highest scoring features according to ANOVA f_classif feature scoring [56] function.

Deep Learning Models:

We leverage the power of 3 different deep learning (DL) classification models, namely, vanilla CNN [7], AlexNet [40] and InceptionNet [41]. We transform the raw viral genome sequences into two different representations, namely, K-mers spectral representation [7] and one hot vectorization [8] to feed those into the DL networks in a seamless manner. Details of these representations are given in Section 5.2 of the Supplementary File. For K-mers spectral representation we experimented with different values of K (K = 3,5,7 for Vanila CNN and K = 3 & 5 only for the rest due to resource limitation). For one hot vectorization, we have trained InceptionNet for 150 epochs for both 3- and 5-mers and trained AlexNet for 135, 100 and 100 epochs for 3-,4- and 5-mers respectively.

2.5 Mutation Prediction

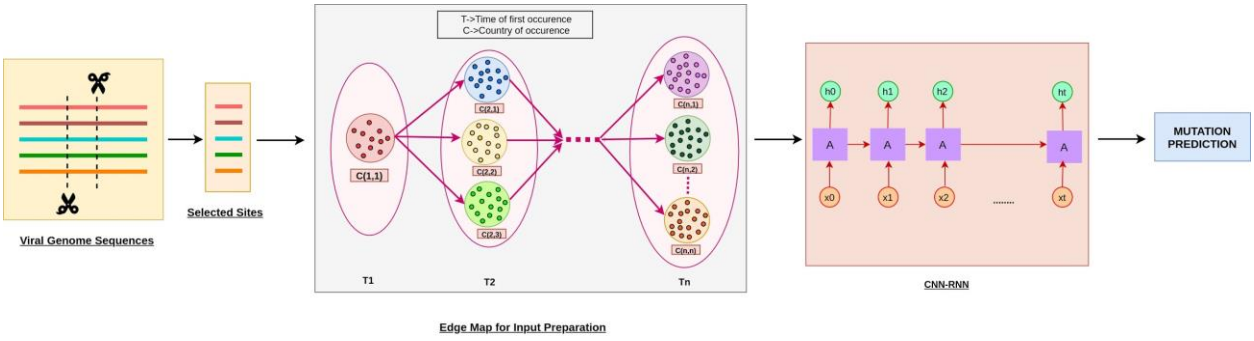


Figure 4: The interesting sites were selected from the viral genome sequences using the feature selection routine described in the classification methods. The sites were then divided by geographical locations and clustered by the time of the first occurrence at that respective region. Time-series sequences were created by

concatenating random strains from the closest sub-clusters and trained in an CNN-RNN network for predicting mutation in the sites of the final time-step.

We design a pipeline to predict mutation on specific sites (chosen in an earlier stage of the pipeline) in the sars-cov-2 genome (Figure 4). We follow a similar protocol followed by [10] and adopt it to fit our setting as follows. We divide all the available countries and the states of the USA into different time-steps by the date of the first reported incidence of sars-cov-2 infected patients of that location. Thus, every resulting time-step represents a date (T_k for Cluster k) and contains the clusters of genome sequences of the countries/states. Then the time series samples are generated by concatenating sites from different time-step one-by-one that represent the evolutionary path of the sars-cov-2 viral strain. For example, T_1 is the very first date when the virus is discovered in China. So, the time-step 1 contains only one country, China. Likewise, time-step T_2 contains clusters for those countries where the virus is discovered on date T_2 and so on. (Check Table 3 in supplementary file for more details).
We generate 300000 time series sequences by concatenating genome sites from $T_1,T_2,....,T_n$ (in our case, n = 40) and then fed the samples to the model which consists of a convolutional one dimensional layer and a recurrent neural network layer [34]. We experiment with both pure LSTM and bidirectional LSTM as our RNN layer (see section 4.3 of supplementary file). The model has a dense layer of 4 neurons in the end which predicts the probability of the next base pair of the next time-step. So, in a nut-shell the model takes concatenated genome sequences from $T_1,T_2,....,T_{n-1}$ as input and predicts the mutation for time T_n.

We further use our mutation prediction pipeline to identify and analyze possible parents of a mutated strain. For this particular analysis, we trained the models specifically for some South-Asian countries, namely, Bangladesh, India and Pakistan. We only used the best performing model for this analysis and generated five time series samples. At the time of generating these samples, the country/location having the minimal euclidean distance was taken for each time-step.

2.6 Coding and Experimental Environment

We have implemented our experiments mostly in python. We have used scikit-learn library [38] for clustering and plotting the graphs. For deep learning models, scikit-learn, tensorflow and keras neural network libraries are used and for LightGBM classifier, python LightGBM framework has been used. The phylogenetic trees are constructed using the Dendropy library of python [57] keeping default parameters. We use the tree visualizer tools Dendroscope [11] and Evolview [24] for tree visualization and annotation. The experiments have been conducted in the following machines:

a) Clustering and phylogenetic analyses have been carried out in a machine with Intel(R) Core (TM) i7-6500U CPU @ 2.50GHz, Ubuntu 19.04 OS and 8 GB RAM.
b) Experiments involving the deep learning pipelines (i.e., both classification and mutation prediction) have been conducted in the work-stations of Galileo Cloud Computing Platform [35] and the default GPU provided by the Google Colaboratory Cloud Computing Platform [36].
c) The LightGBM classifier model was trained in a machine with Intel Core i5-4010U CPU @ 1.70GHz x 4, Windows 10 OS and 16 GB RAM.

All the codes and data (except for the Genome Sequences) of our pipeline can be found at the following link: https://github.com/pythonLoader/Analyzing-hCov-Genome-Sequence.
The Genome Sequence data have been extracted from and are publicly available at GISAID [23].

3. Results:

Evolutionary Relationship of the virus strains:

We identify the representative sequence of each of the 67 countries as present in the GISAID dataset (upto cut-off date). The estimated phylogenetic tree constructed from the representative sequences is shown in Figure 5. In what follows, we will be referring to this tree as the SC2 (sars-cov-2) Tree.

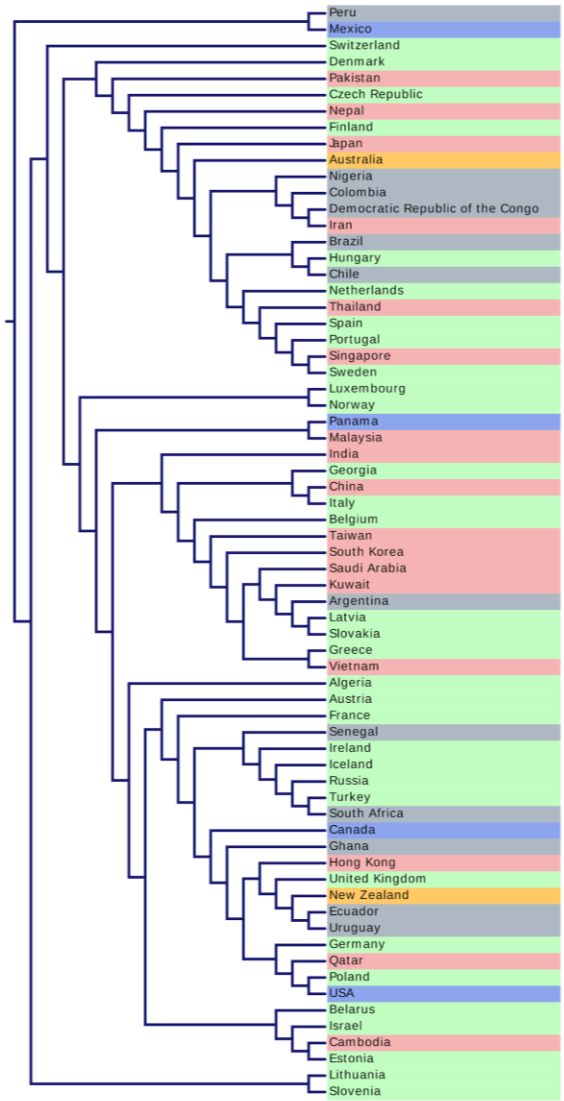


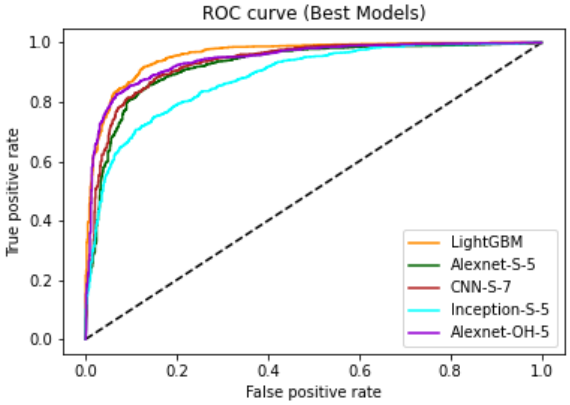
Figure 5: The estimated phylogenetic tree of all the representative sequences (i.e., SC2 Tree). The representative sequence identified in the previous steps for each country were used in phylogeny estimation. In total 67 representative sequences from 67 countries were taken.

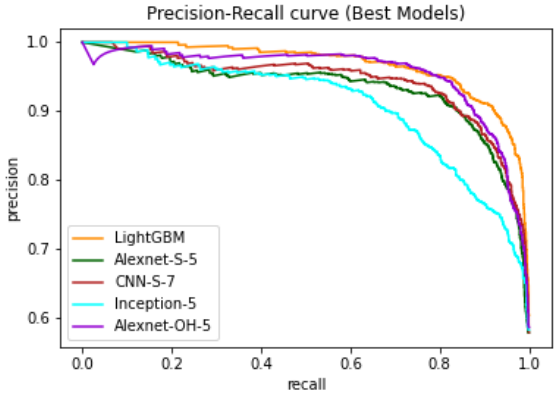
The phylogenetic tree generated is expected to reveal the evolutionary relationship of the viral strains. However, with careful scrutiny we have some apparently unusual but interesting observations. For example, it is generally expected that the countries sharing (open) borders (e.g., countries in Europe) should be either neighbours or at least in the same clade in the tree. However, surprisingly from the tree, we do not notice geographically adjacent countries in Europe as neighbors; rather we see for example that China and Italy are immediate neighbors. It is to be noted that these two countries

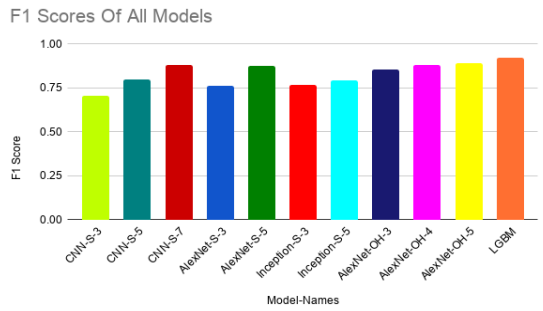
are also the first countries to get hit by the first pandemic wave. In addition to that, although the USA and Canada share the longest un-militarized international border in the world, representative strains do not appear to be sister branches as they should have been. Also, we notice that the USA, UK, Canada, Turkey and Russia are in the same clade which have a higher number of deaths than most of the other countries.

Mild/Severe Strain Classification:

All our classifiers are trained to learn whether a given strain is mild or severe. The classification accuracy of the LightGBM classifier (~91%) is superior to that of the deep learning classifiers (~84-89%), which, while is somewhat surprising, is in line with the recent findings of [12]. It should be noted that LightGBM had produced better results in significantly less time than deep learning models for this dataset. The results of the classifier models are shown in Figure 6.







Abbreviation	Model Name Description
CNN-S-3	CNN Spectral (Input: 3-mer)
CNN-S-5	CNN Spectral (Input: 5-mer)
CNN-S-7	CNN Spectral (Input: 7-mer)
AlexNet-S-3	AlexNet Spectral (Input: 3-mer)
AlexNet-S-5	AlexNet Spectral (Input: 5-mer)
Inception-S-3	InceptionNet Spectral (Input: 3-mer)
Inception-S-5	InceptionNet Spectral (Input: 5-mer)
AlexNet-OH-3	AlexNet One-Hot (Input: 3-mer)
AlexNet-OH-4	AlexNet One-Hot (Input: 4-mer)
AlexNet-OH-5	AlexNet One-Hot (Input: 5-mer)
LGBM	LightGBM Classifier

Figure 6: The Receiver Operating Characteristic Curves (ROC) show the diagnostic capability of the ML and the DL classifiers used in our experiments. In the first figure (top-left), we have shown the ROC curves of the best performing models used to classify the genome sequences, in the second figure we have shown the Precision-Recall curve of the same. The third figure shows the F1-scores. Please refer to Table 5 and Figure 4 of the supplementary file for detailed results.

Quantitative results aside, we also have applied our classifiers on the sequences that have been deposited at GISAID after the cut-off date (i.e. April 18, 2020). Since the cut-off date, the country wise death statistics [6] has certainly changed significantly and this has pushed a few countries, particularly from Asian regions and several states of the United States of America transition from mild to severe state (based on our predefined threshold). Interesting, our classifiers have been able to predict the severity of the new strains submitted from these countries/states correctly. Table 6 in the supplementary file shows a snapshot of a few such countries/states with the relevant information.

Sites of Interest (SoI):

We preliminarily identify the top 10 features of SHAP and SelectKBest feature selection (with K=10). From these features, as SoIs, we have selected the features that are also biologically significant, i.e., cover different significant gene expression regions (Figure 7). In particular, we have selected the position specific features pos_8445_8449, pos_19610_19614, pos_24065_24069 and pos_23825_23829 as the SoIs for the mutation prediction analyses down the pipeline. Here, pos_X_Y indicates the site from Positions X to Y of the virus strains. The reason for selecting these features as SoIs are outlined below.

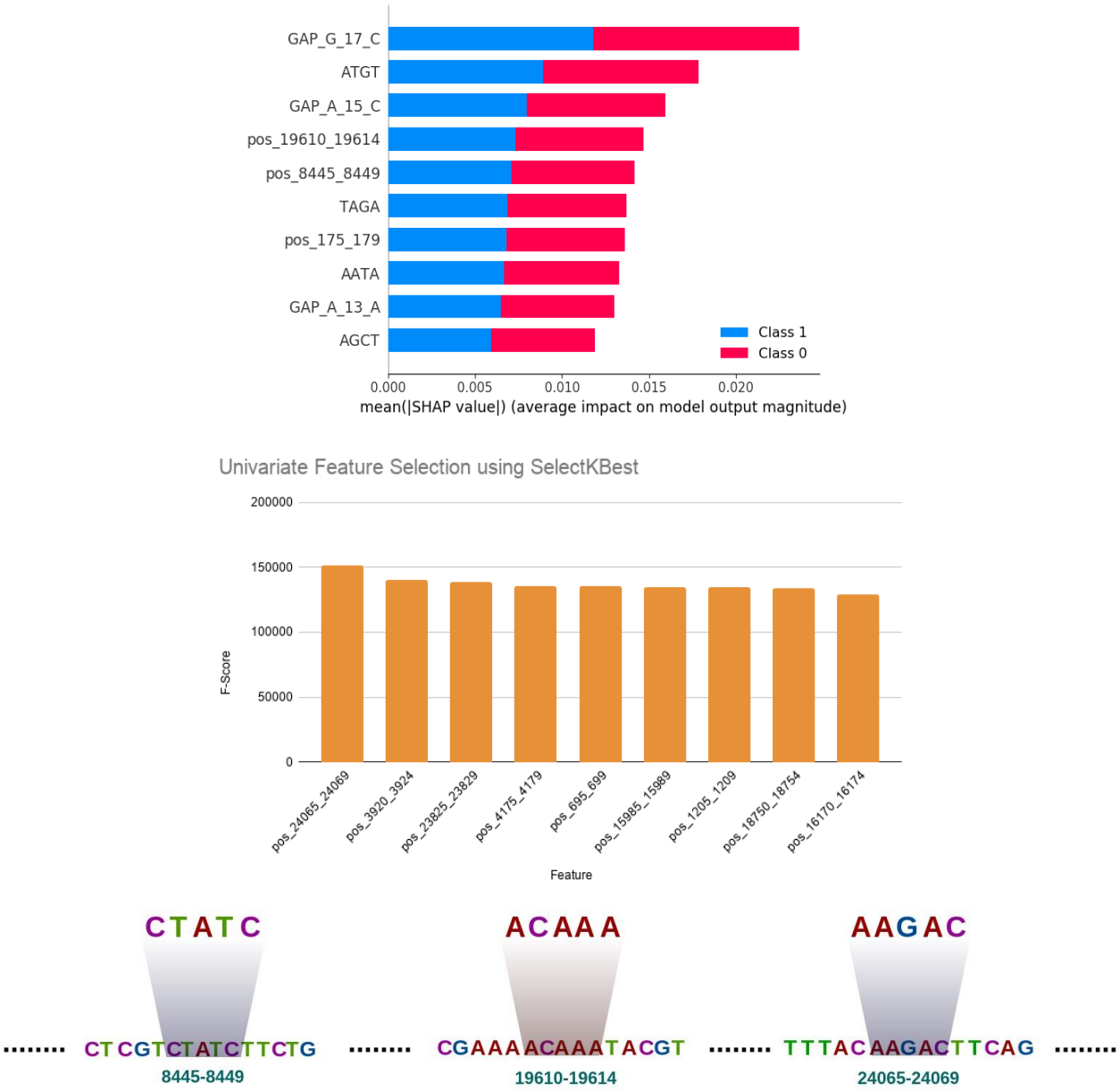


Figure 7: Top 10 features based on SHAP values and SelectKBest and identifying the position specific feature(s) from the genome sequence as site(s) of interest. These SOIs will be used for mutation prediction down the pipeline.

According to gene expression studies [62][63], our SoIs, namely, pos_8445_8449 and pos_19610_19614 encode to two Non-structural Proteins, Nsp3 and Nsp11, respectively. And, our other two SoIs, namely, pos_24065_24069 and pos_23825_23829 correspond to the Spike Protein of sars-cov-2. Nsp3 binds to viral RNA, nucleocapsid protein, as well as other viral proteins, and participates in polyprotein processing. It is an essential component of the replication/transcription

complex [51]. So, the mutation in this protein is expected to affect the replication process of the sars-cov-2 in host bodies. On the other hand, the spike protein sticks out from the envelope of the virion and plays a pivotal role in the receptor host selectivity and cellular attachment. According to Wan et al. there exists strong scientific evidence that SARS and sars-cov-2 spike proteins interact with angiotensin-converting enzyme 2 (ACE2) [52]. The mutation on this protein is expected to have a significant impact on the human to human transmission [53]. Therefore, it is certainly interesting and useful to predict the mutation of such SoIs.

Mutation Prediction Results:

CNN-LSTM and CNN-bidirectional LSTM performed in a similar manner for different SoIs of the genome registering 94.98% and 95% accuracy, respectively, considering all SoIs together. For detailed results please check Table 7 and Table 8 of the supplementary material.

Analyzing Parent Strains:

For the model involving only Bangladesh, we applied the CNN-bidirectional LSTM model (as this is the best performer among the two) and achieved almost 100% accuracy. Then we analyzed the ancestors in the time series test samples and noticed that some of the states of the USA are present in these samples. These states are California, Massachusetts, Texas, New Jersey and Maryland.
For India and Pakistan, we got similar results for some sites but for other sites, accuracy was not as high as Bangladesh (Check Table 9 of the supplementary file for details).

4. Discussion

Evolutionary Relationships among the Virus Strains:

Our analyses reveal a very close (evolutionary) relationship between the genome sequences of China and Italy. Also, similarity was found among the virus strains of the USA, Germany, Qatar and Poland. These countries have similar numbers of deaths and although not geographically directly adjacent (except for Germany and Poland) they have strong air connectivity among them. In fact, a number of interesting relationships can be inferred from the estimated phylogenetic tree as follows.

1. The Italian strain of the virus is believed to be transmitted from China via two Chinese tourists [26]. This relationship is clearly portrayed in the SC2 tree where the two strains appear to be immediate siblings.
2. Poland's strain is in the same clade as that of Germany, which can be explained by the fact that its strain (through Poland's Patient Zero) came from Germany [27].
3. Taiwan is geographically very close to China. The virus was confirmed to have spread to Taiwan on January 21, 2020, through a 55-year-old woman who had been teaching in Wuhan, China [28]. The virus strains from these regions are also close together as can be seen from the SC2 tree, about 6 branches apart. Similar relationship can also be inferred from the tree between China and South Korea: the strain of the virus in South Korea is believed to be transmitted from China firstly through a 35-year old Chinese woman and secondly by a 55-year old South Korean national [29]. Interestingly, from the SC2 tree it can also be deduced that the South Korean strain is very close to that of Taiwan and also near to the strain from China. The incident of a Taiwanese woman being deported from South Korea after refusing to stay at a quarantine facility can be a probable explanation as to how the South Korean strain might have found its path to Taiwan [46].
4. On March 2, 2020, the virus was confirmed to have reached Portugal, when it was reported that a Portuguese 33 year-old man working in Spain was tested positive for COVID-19 after returning home [49]. Subsequently, within a span of 9 days, 5 more cases were reported all originating from Spain [49][61]. The fact that the first cases of COVID-19 in Portugal originated from Spain is clearly captured in our SC2 tree.

5. The SC2 tree suggests that India's strain is closely related to that from China and also Italy (around 4 branches) and that it is also connected to that from Saudi Arabia. These relationships can be explained as follows.

a. On January 30, India reported its first case of COVID-19 in Kerala, which rose to three cases by February 3; all were students who had returned from Wuhan, China [17][18].
b. On March 12, 2020, a 76-year-old man returning from Saudi Arabia became the first death case of the virus in the country [19]. India didn't impose a travel ban on Saudi Arabia at that point also.
c. A Sikh preacher that returned from travel to Italy and Germany, carrying the virus, turned into a "super spreader" by attending a Sikh festival in Anandpur Sahib during 10–12 March [20][21].

6. Strains from Austria and Greece are quite close to each other and near to that of Italy in the SC2 tree as they are believed to be transmitted from Italy. The two people diagnosed with coronavirus in the Tyrol region of Austria were both Italian citizens [30] and a Greek woman who recently returned home from northern Italy became Greece's first coronavirus case [31].

7. Turkey's first identified case was a man who was travelling Europe [33]. Turkey also announced a huge number of cases and subsequent deaths, which were originating from Europe [47]. In our inferred relationship, we can see that the Turkish representative strain is quite close to several Central and Western European countries like Russia, Iceland and Ireland which can be backed up by the two facts stated above.

8. It is visible from the SC2 tree that the strain of Germany is very close to the strains of both Poland and the USA. It might be the case that the community transmission occurred concurrently in both USA and Poland from Germany which hit the peak of pandemic before both USA and Poland [42].

9. Qatar has the second highest number of Covid-19 patients in the Middle-East [48]. The first case of Qatar was reported on February 27,2020 to be a man working in Iran [55]. Qatar introduced a travel ban to and from Germany and the USA as precautionary measures in Mid-March, quite a while later following the first occurrence. Qatar has 5 air-routes with Germany and USA, with more than 10 airlines operating in that route [59][60]. Though the first case has originated from Iran, it might be the case that subsequent patients were found to be travelling from the aforementioned countries as a result of which the travel ban was introduced. Our estimated SC2 tree places Qatar very close to both the USA and Germany.

10. While we can certainly explain many of the relationships identified by the estimated SC2 tree a above, there are some relationships which are not that apparent. One such example is the direct relationship between Vietnam and Greece. While apparently, there exists no direct relationship, when investigated further, we identified something interesting. Patient Zero of Greece is believed to have been contaminated during her trip to the Milan Fashion week which took place during February 18-24, 2020 [45]. Interestingly, the first COVID-19 patient in Hanoi [16] left Hanoi on February 15 to visit family members living in London, England and three days later, she traveled from London to Milan City. Could she be in contact with Patient Zero of Greece or any other who had been contaminated by the latter, before returning to London on February 20? We can't be certain, but our inferred relationship between Vietnam and Greece certainly put a lot of legitimacy to that question.

11. Finally, we are unable to find any apparent explanation analyzing the reported news sources for a few other strong relationships inferred by the tree (e.g., Congo-Iran, Panama-Malaysia, Sweden-Singapore, Japan-Australia, etc). This could be because of the inherent inaccuracies of the distance matrices as well as the limitations of the tree estimation algorithms: none of these algorithms are 100% accurate. From another angle, perhaps, the tree did identify these relationships correctly; but the relevant incidences were not accurately identified or not documented.

Severity of some recent virus strains:

In recent times, the number of deaths is increasing rapidly in India. We have been closely following the change in the virus strains of India before and after the cut-off date. A genome sequence (EPI_ISL_435050) was collected on April 13, 2020 (before our cut-off date) from a patient in Ahmedabad, Gujrat, India. It was predicted to be a severe strain (with low confidence) even though at that time we trained the classifier to consider the Indian sequences as mild. According to our evolutionary relationship, India is very close to both Italy and China. So, we calculated the distance between the representative sequence of both Italy and China with this strain. We considered another strain (EPI_ISL_437447) which was collected from another patient from the same place in India on April 26, 2020 (after our cut-off date) and predicted the severity thereof. The classifiers declared this isolate to be severe with very high confidence (about 98%). We did the distance calculation like before. Interestingly, it was identified that this isolate is closer to both Italy and China's representative sequence than the previous less severe one. This strongly suggests that there were some mutations that turned the Indian sequences from mild or less severe to severe or highly severe, respectively.

Also, the sequences from the US states of Pennsylvania, Maryland, Indiana, Illinois and Florida that were collected on May 25, 2020 (about one month after our cut-off date) were analyzed and our classifiers could correctly capture the severity of the genome sequences (see Table 4 in the supplementary file).

Possible parents for some South Asian strains:

We conduct an analysis to predict possible parents of the (mutated) virus strains of the South Asian Region (Bangladesh, India and Pakistan). Our mutation prediction pipeline suggests that the strains of some states of the USA, namely, California, Massachusetts, Texas, New Jersey and Maryland could be the parents/ancestors of these South Asian strains. Now, the total deaths in these states up to June 1, 2020 are 4240, 6846, 1686, 11711 and 2532 respectively [58] and the strains thereof are also classified to be severe

by our classification pipeline. it thus seems quite likely that the sars-cov-2 situation in these South-Asian countries will worsen in near future.

Bangladesh, India and Pakistan are ranked 88th, 112th and 122nd in global health performance compared to the United States of America which is at the 37th position [54]. In the majority of lower middle-income countries such as Bangladesh, India and Pakistan, available hospital beds are < 1 bed per 1000 population and ICU beds are < 1 bed per 100,000 population [39]. Additionally, an uncontrolled epidemic is predicted to have 6,000,220 deaths having a duration of nearly 200 days in the majority of these countries [39]. These predictions coupled with our findings call for stern actions (i.e., interventions) on part of these countries.

Bibliography**:**

[1] Coronavirus disease (COVID-19) outbreak situation, https://www.who.int/emergencies/diseases/novel-coronavirus-2019

[2] Lu, R., Zhao, X., Li, J., Niu, P., Yang, B., Wu, H. et. al (2020). Genomic characterisation and epidemiology of 2019 novel coronavirus: Implications for virus origins and receptor binding. The Lancet, 395(10224), 565-574. doi:10.1016/s0140-6736(20)30251-8

[3] Wrapp D, Wang N, Corbett KS, et al. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science*. 2020;367(6483):1260-1263. doi:10.1126/science.abb2507

[4] Zielezinski A, Vinga S, Almeida J, Karlowski WM. Alignment-free sequence comparison: benefits, applications, and tools. *Genome Biol*. 2017;18(1):186. Published 2017 Oct 3. doi:10.1186/s13059-017-1319-7

[5] Li, Y., He, L., Lucy He, R. *et al.* A novel fast vector method for genetic sequence comparison. *Sci Rep* **7,** 12226 (2017). https://doi.org/10.1038/s41598-017-12493-2

[6] WHO coronavirus disease (COVID-19) dashboard. https://covid19.who.int/

[7] Rizzo R., Fiannaca A., La Rosa M., Urso A. (2016) A Deep Learning Approach to DNA Sequence Classification. In: Angelini C., Rancoita P., Rovetta S. (eds)

Computational Intelligence Methods for Bioinformatics and Biostatistics. CIBB 2015. Lecture Notes in Computer Science, vol 9874. Springer, Cham. https://doi.org/10.1007/978-3-319-44332-4_10

[8] Nguyen, N. , Tran, V. , Ngo, D. , Phan, D. , Lumbanraja, F. , Faisal, M. , Abapihi, B. , Kubo, M. and Satou, K. (2016) DNA Sequence Classification by Convolutional Neural Network. *Journal of Biomedical Science and Engineering*, **9**, 280-286. doi: 10.4236/jbise.2016.95021.

[9] Principal Component Analysis and Factor Analysis. (n.d.). *Principal Component Analysis Springer Series in Statistics,* 150-166. doi:10.1007/0-387-22440-8_7

[10] Yin R, Luusua E, Dabrowski J, Zhang Y, Kwoh CK. Tempel: time-series mutation prediction of influenza A viruses via attention-based recurrent neural networks. *Bioinformatics*. 2020;36(9):2697-2704. doi:10.1093/bioinformatics/btaa050

[11] Daniel H. Huson and Celine Scornavacca. Dendroscope 3: An interactive tool for rooted phy-logenetic trees and networks, Systematic Biology (2012), http://sysbio.oxfordjournals.org/cgi/content/abstract/sys062?ijkey=ZCxPRbYt74aQJhR&keytype=ref, software freely available from www.dendroscope.org

[12] Ali Haisam Muhammad Rafid; Md. Toufikuzzaman; Mohammad Saifur Rahman; M. Sohel Rahman. CRISPRpred(SEQ): A Sequence-Based Method for sgRNA On Target Activity Prediction Using Traditional Machine Learning. BMC Bioinformatics, To Appear.

[13] Maier, Oskar, et al. "Extra tree forests for sub-acute ischemic stroke lesion segmentation in MR sequences." *Journal of neuroscience methods* 240 (2015): 89-100.

[14] Rahman, M. Saifur, et al. "isGPT: An optimized model to identify sub-Golgi protein types using SVM and Random Forest based feature selection." Artificial intelligence in medicine 84 (2018): 90-100.

[15] Ke, Guolin, et al. "Lightgbm: A highly efficient gradient boosting decision tree." Advances in neural information processing systems. 2017

[16] Person. “Vietnam Confirms 17th Covid-19 Patient - VnExpress International.” *VnExpress International – Latest News, Business, Travel and Analysis from Vietnam*, VnExpress.net, 10 Mar. 2020, a.vnexpress.net/news/news/vietnam-confirms-17th-covid-19-patient-4065517.html

[17] davyreid73. “India Confirms Its First Coronavirus Case.” *CNBC*, CNBC, 30 Jan. 2020, www.cnbc.com/2020/01/30/india-confirms-first-case-of-the-coronavirus.html

[18] TWC India Edit Team. “Kerala Defeats Coronavirus; India's Three COVID-19 Patients Successfully Recover.” *The Weather Channel*, The Weather Channel, 15 Feb.

2020, weather.com/en-IN/india/news/news/2020-02-14-kerala-defeats-coronavirus-indias-three-covid-19-patients-successfully

[19] "India's First Coronavirus Death Is Confirmed in Karnataka." *Hindustan Times*, 12 Mar. 2020, www.hindustantimes.com/india-news/india-s-first-coronavirus-death-in-karnataka-confirmed/story-2ZJ6IuxJ38EiGndBq5pfHO.html

[20] "Coronavirus: India 'Super Spreader' Quarantines 40,000 People." BBC News, BBC, 27 Mar. 2020, www.bbc.com/news/world-asia-india-52061915

[21] Wallen, Joe. "40,000 Indians Quarantined after 'Super Spreader' Ignores Government Advice ." The Telegraph, Telegraph Media Group, 28 Mar. 2020, www.telegraph.co.uk/news/2020/03/28/40000-quarantined-punjab-super-spreader-ignores-government-advice/

[22] Gates, B. (2020). Responding to Covid-19 — A Once-in-a-Century Pandemic? *New England Journal of Medicine, 382*(18), 1677-1679. doi:10.1056/nejmp2003762

[23] Elbe, S., and Buckland-Merrett, G. (2017) Data, disease and diplomacy: GISAID's innovative contribution to global health. *Global Challenges,* 1:33-46. doi:10.1002/gch2.1018 PMCID: 31565258

[24] Zhang H, Gao S, Lercher MJ, Hu S, Chen WH (July 2012). "EvolView, an online tool for visualizing, annotating and managing phylogenetic trees". Nucleic Acids Research. 40(Web Server issue): W569–72

[25] Mihaescu R, Levy D, Pachter L (2009). "Why neighbor-joining works". Algorithmica. 54 (1): 1–24. arXiv:cs/0602041

[26] Severgnini, Chiara. "Coronavirus, Primi Due Casi in Italia: Sono Due Turisti Cinesi." *Corriere Della Sera*, Corriere Della Sera, 31 Jan. 2020, www.corriere.it/cronache/20_gennaio_30/coronavirus-italia-corona-9d6dc436-4343-11ea-bdc8-faf1f56f19b7.shtml?refresh_ce-cp.

[27] Maciejewicz, Andrzej, and Wojciech M. "Koronawirus w Lubuskiem. 44 Godziny, Dwa Razy Za Wolno. Daleko Do Laboratorium." *Oko.press*, 6 Mar. 2020, oko.press/koronawirus-wojewodztwo-lubuskie/.

[28] Cna. "Taiwan Confirms 1st Wuhan Coronavirus Case (Update)." *Focus Taiwan*, Focus Taiwan - CNA English News, 23 Jan. 2020, focustaiwan.tw/society/202001210019.

[29] https://www.mk.co.kr/news/society/view/2020/01/80017/

[30] Renton, Adam, and Mike Hayes. “Austria's 2 Coronavirus Cases Are Italian Citizens.” *CNN*, Cable News Network, 26 Feb. 2020, edition.cnn.com/asia/live-news/coronavirus-outbreak-02-25-20-hnk-intl/h_66850dfa7d9602fb4b0d38efdd19d939.

[31] “Greece Confirms First Coronavirus Case, a Woman Back from Milan.” *Reuters*, Thomson Reuters, 26 Feb. 2020, www.reuters.com/article/us-china-health-greece-idUSKCN20K1IA

[32] Kambas, Michele. “As Coronavirus Takes Hold, Greece Worries about Migrant Camps.” *Reuters*, Thomson Reuters, 27 Feb. 2020, www.reuters.com/article/us-china-health-greece-idUSKCN20L1F3.

[33] Daily Sabah. “Turkey Remains Firm, Calm as First Coronavirus Case Confirmed.” *Daily Sabah*, Daily Sabah, 11 Mar. 2020, www.dailysabah.com/turkey/turkey-remains-firm-calm-as-first-coronavirus-case-confirmed/news

[34] Wang, R., Zang, T. & Wang, Y. Human mitochondrial genome compression using machine learning techniques. *Hum Genomics* 13, 49 (2019). https://doi.org/10.1186/s40246-019-0225-3

[35] Galileo. (n.d.). Retrieved from http://galileo.io/

[36] Bisong E. (2019) Google Colaboratory. In: Building Machine Learning and Deep Learning Models on Google Cloud Platform. Apress, Berkeley, CA

[37] Saitou, N.; Nei, M. (1 July 1987). "The neighbor-joining method: a new method for reconstructing phylogenetic trees". Molecular Biology and Evolution. 4 (4): 406–425

[38] “Sklearn.cluster.KMeans¶.” *Scikit*, scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html.

[39] Chowdhury, R., Heng, K., Shawon, M.S.R. *et al.* Dynamic interventions to control COVID-19 pandemic: a multivariate prediction modelling study comparing 16 worldwide countries. *Eur J Epidemiol* (2020). https://doi.org/10.1007/s10654-020-00649-w

[40] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM, 60*(6), 84-90. doi:10.1145/3065386

[41] C. Szegedy et al., "Going deeper with convolutions," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, 2015, pp. 1-9, doi: 10.1109/CVPR.2015.7298594.

[42] Rourke, Alison. “Europe's Coronavirus Numbers Offer Hope as US Enters 'Peak of Terrible Pandemic'.” *The Guardian*, Guardian News and Media, 6 Apr. 2020,

www.theguardian.com/world/2020/apr/06/europes-coronavirus-numbers-offer-hope-as-us-enters-peak-of-terrible-pandemic.

[43] Hartigan, J. A., Wong, M.A. Algorithm AS 136: A K-Means Clustering Algorithm. Journal of the Royal Statistical Society. Series C (Applied Statistics), Vol. 28, No. 1 (1979), pp. 100-108

[44] Lundberg, Scott M., Gabriel G. Erion, and Su-In Lee. "Consistent individualized feature attribution for tree ensembles." arXiv preprint arXiv:1802.03888 (2018).

[45] Natalie, et al. “Greece's 'Patient Zero' Shares Coronavirus Experience.” *Greek City Times*, 1 May 2020, greekcitytimes.com/2020/05/01/greeces-patient-zero-shares-coronavirus-experience/.

[46] 이민지 . “(LEAD) Taiwanese Woman Deported for Refusing to Stay at Quarantine Facility.” *Yonhap News Agency*, 이민지, 6 Apr. 2020, en.yna.co.kr/view/AEN20200406004451315?fbclid=IwAR0oghjTEZrak6g803tAjGSlrE_qZGJu_FqqWoiW2biIiLDXQOwbPyweyi8.

[47] “Sağlık Bakanı Fahrettin Koca: Pozitif Çıkan Yeni Vakalarımız Var.” *Sağlık Bakanı Fahrettin Koca: Pozitif Çıkan Yeni Vakalarımız Var - Türkiye Haberleri*, www.posta.com.tr/amp/saglik-bakani-fahrettin-koca-pozitif-cikan-yeni-vakalarimiz-var-2244301.

[48] “Qatar to United States Flights.” *Flights from Qatar*, www.qatar.to/United-States/Qatar-to-United-States.php.

[49] Expresso. “Ministra Confirma Primeiro Caso Positivo De Coronavírus Em Portugal.” *Jornal Expresso*, Expresso, 2 Mar. 2020, expresso.pt/sociedade/2020-03-02-Ministra-confirma-primeiro-caso-positivo-de-coronavirus-em-Portugal.

[50] “1.13. Feature Selection¶.” *Scikit*, scikit-learn.org/stable/modules/feature_selection.html#univariate-feature-selection.

[51] Lei, Jian, et al. “Nsp3 Of Coronaviruses: Structures and Functions of a Large Multi-Domain Protein.” *Antiviral Research*, vol. 149, 2018, pp. 58–74., doi:10.1016/j.antiviral.2017.11.001.

[52] Wan, Yushun, et al. “Receptor Recognition by the Novel Coronavirus from Wuhan: an Analysis Based on Decade-Long Structural Studies of SARS Coronavirus.” *Journal of Virology*, vol. 94, no. 7, 2020, doi:10.1128/jvi.00127-20.

[53] Ortega, J. T., Serrano, M. L., Pujol, F. H., & Rangel, H. R. (2020). Role of changes in SARS-CoV-2 spike protein in the interaction with the human ACE2 receptor: An *in silico* analysis. *EXCLI journal*, *19*, 410–417. https://doi.org/10.17179/excli2020-1167

[54] Tandon, Ajay & Murray, Christopher & Lauer, Jeremy & Evans, David. (2000). Measuring Overall Health System Performance for 191 Countries. Global Programme on Evidence forHealth Policy Discussion Paper No. 30.

[55] "Qatar Reports First Case of Coronavirus." *The Peninsula Qatar*, www.thepeninsulaqatar.com/article/29/02/2020/Qatar-reports-first-case-of-coronavirus.

[56] "Sklearn.feature_selection.f_classif¶." *Scikit*, scikit-learn.org/stable/modules/generated/sklearn.feature_selection.f_classif.html.

[57] Sukumaran, J. and Mark T. Holder. 2010. DendroPy: A Python library for phylogenetic computing. *Bioinformatics* 26: 1569-1571

[58] https://www.worldometers.info/coronavirus/country/us/

[59] "Qatar to Germany Flights." *Flights from Qatar*, www.qatar.to/Germany/Qatar-to-Germany.php.

[60] "Qatar to United States Flights." *Flights from Qatar*, www.qatar.to/United-States/Qatar-to-United-States.php.

[61] "RELATÓRIO DE SITUAÇÃO". *Direção-Geral da Saúde* (in Portuguese). 2020-03-12.

[62] Cascella, Marco. "[Figure, Single-Stranded RNA Genome of SARS-CoV2...] - StatPearls - NCBI Bookshelf." *StatPearls [Internet].*, U.S. National Library of Medicine, 6 Apr. 2020, www.ncbi.nlm.nih.gov/books/NBK554776/figure/article-52171.image.f5/.

[63] "SARS-CoV-2 (Severe Acute Respiratory Syndrome Coronavirus 2) Sequences." *National Center for Biotechnology Information*, U.S. National Library of Medicine, www.ncbi.nlm.nih.gov/genbank/sars-cov-2-seqs/.

[64] M. Saifur Rahman, Md. Khaledur Rahman, Sanjay Saha, M. Kaykobad and M. Sohel Rahman: Antigenic: An improved prediction model of protective antigens. Artificial Intelligence in Medicine 94 (2019), 28-41.

[65] M. Saifur Rahman, Swakkhar Shatabda, Sanjay Saha, M. Kaykobad and M. Sohel Rahman:DPP-PseAAC: A DNA-binding protein prediction model using Chou's general PseAAC. Journal of Theoretical Biology 452: 22-34 (2018).