Multimedia Systems

Applying Deep Learning for Genome Detection of Coronavirus

--Manuscript Draft--

Manuscript Number:	
Full Title:	Applying Deep Learning for Genome Detection of Coronavirus
Article Type:	S.I. : Deep Learning for Multimedia Healthcare
Keywords:	- COVID-19; Deep Learning; CNN; Drug; Genome Matching; SARS-CoV-2
Corresponding Author:	Vijaypal Singh Dhaka Manipal University - Jaipur Campus INDIA
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	Manipal University - Jaipur Campus
Corresponding Author's Secondary Institution:	
First Author:	Geeta Rani
First Author Secondary Information:	
Order of Authors:	Geeta Rani
	Meet Ganpatlal Oza
	Vijaypal Singh Dhaka
	Nitesh Pradhan
	Sahil Verma
	Joel J. P. C. Rodrigues
Order of Authors Secondary Information:	
Funding Information:	FCT/MCTES through national funds (309335/2017-5.) Dr Joel J. P. C. Rodrigues
Abstract:	Amidst the global pandemic and catastrophe created by 'COVID-19', every research institution and scientists are doing their best efforts to invent or find the vaccine or medicine for the disease. The objective of this research is to design and develop a deep learning model for finding the degree of similarity of the genome of the Severe Acute Respiratory Syndrome-Coronavirus 2 ('SARS-CoV-2') with a given genome. This research also aims at detecting the genome of 'SARS-CoV-2' in the host human beings. The experimental results on the dataset publicly available at National Centre for Biotechnology Information, show that the model is effective in predicting the similarity score of the genomic sequence of 'SARS-CoV-2' and other prevalent viruses such as Severe Acute Respiratory Syndrome-Coronavirus, Middle East Respiratory Syndrome Coronavirus, Human Immunodeficiency Virus, and Human T- cell Leukaemia Virus. This is successful in detecting the genome of 'SARS-CoV-2' in the host genome with an accuracy of 99.27%. It may prove a useful tool for doctors to quickly classify the infected and non-infected genomes. It can also be useful in finding the most effective drug from the available drugs for the treatment of 'COVID-19'.

Applying Deep Learning for Genome Detection of Coronavirus

Geeta Rani[1], Meet Ganpatlal Oza[1], Vijaypal Singh Dhaka[1], Nitesh Pradhan[2], Sahil Verma[3],
Joel J. P. C. Rodrigues[4,5]

[1] Department of Computer and Communication Engineering,

Manipal University Jaipur, Rajasthan, India

Email:Geeta.rani@jaipur.manipal.edu, meetoza08@gmail.com,
vijaypalsingh.dhaka@jaipur.manipal.edu

[2] Department of Computer Science and Engineering,

Manipal University Jaipur, Rajasthan, India

nitsh.pradhan@jaipur.manipal.edu

[3]School of Computer Science and Engineering, Lovely Professional University, Phagwara, India,
144411

sahilverma@ieee.org

[4]Federal University of Piauí (UFPI) Teresina - PI, Brazil

[5]Instituto de Telecomunicações, Portugal
joeljr@ieee.org

Abstract – Amidst the global pandemic and catastrophe created by 'COVID-19', every research institution and scientists are doing their best efforts to invent or find the vaccine or medicine for the disease. The objective of this research is to design and develop a deep learning model for finding the degree of similarity of the genome of the Severe Acute Respiratory Syndrome-Coronavirus 2 ('SARS-CoV-2') with a given genome. This research also aims at detecting the genome of 'SARS-CoV-2' in the host human beings. The experimental results on the dataset publicly available at National Centre for Biotechnology Information, show that the model is effective in predicting the similarity score of the genomic sequence of 'SARS-CoV-2' and other prevalent viruses such as Severe Acute Respiratory Syndrome-Coronavirus, Middle East Respiratory Syndrome Coronavirus, Human Immunodeficiency Virus, and Human T- cell Leukaemia Virus. This is successful in detecting the genome of 'SARS-CoV-2' in the host genome with an accuracy of 99.27%. It may prove a useful tool for doctors to quickly classify the infected and non-infected genomes. It can also be useful in finding the most effective drug from the available drugs for the treatment of 'COVID-19'.

Keywords - COVID-19, Deep Learning, CNN, Drug, Genome Matching, SARS-CoV-2

1. Introduction

The world is facing serious health pandemic for the last seven months due to the newly identified Coronavirus. The Coronaviruses are responsible for the common cold, mild respiratory problems, gastrointestinal problems, and infections in the throat [1]. The newly identified virus is a type of human Coronaviruses [1] named as ‘SARS-CoV-2’. This is the causing agent for the disease ‘COVID-19’.

The first instance of ‘COVID-19’ was reported in Wuhan city of China in January 2020. The number of cases is increasing rapidly among people of different age groups and genders. As per the data available at the web portal of the World Health Organization (WHO) ‘SARS-CoV-2’ has infected 19,811,134 people and caused 729,653 deaths till 9 August 2020, across 215 countries [2]. The majority of patients have shown the symptoms of varying degrees of severe pneumonia. The study reported in [3] states that the human to human transmission is possible even though the infected person is not showing any symptoms of respiratory problems. The authors in [3] discussed that the number of people affected by ‘COVID-19’ exceeds the epidemics caused by Severe Acute Respiratory Syndrome (SARS) in 2002-2003 and Median East Respiratory Syndrome (MERS) in 2012. Therefore, the WHO declared ‘COVID-19’ as a Global-Pandemic. This pandemic has increased the burden on the health services of the world.

Expedite testing of patients is necessary for controlling the outbreak caused by ‘COVID-19’. WHO validated the Nucleic Acid Amplification Tests (NAAT) for the diagnosis of this disease. The health experts use the sample of fluid from the nose, a swab from the throat, a sample of mucus from the lungs (sputum) or blood sample for detection of the presence of Ribo Nucleic Acid (RNA) of ‘SARS-CoV-2’ in the human genome. The real-time Reverse Transcription Polymerase Chain Reaction (RT-PCR) is performed for the diagnosis of ‘COVID-19’. [4]. The average accuracy of this test is reported as 60-70% [5]. Low accuracy of available testing kits [5] and the limited testing capacity of labs are challenges in screening the huge populace. Moreover, doctors manually read the reports of the tests which is a time-consuming task. Thus, it becomes a need of the hour to find a computer-based solution for providing swift assistance to clinicians to deal with the outbreak of ‘COVID-19’.

The ability of deep learning neural networks in pattern recognition, learning, and matching can be used for the detection of mutation caused by ‘SARS-CoV-2’ in the human genome [6,30]. This motivated the authors to provide a deep learning based solution for quick and mass screening of ‘COVID-19’ using the genomic sequences.

The main objective of this research is to develop a quick and reliable tool for mass screening of patients infected with Coronavirus ‘SARS-CoV-2’. It also aims at predicting the similarity score of the genome of ‘SARS-CoV-2’ with other viruses namely SARS-CoV, MERS-CoV, Human Immunodeficiency Virus (HIV), and Human T- cell Leukaemia Virus (HTLV). This similarity score will pave the way for biotechnology experts and other researchers to contribute in dealing with the pandemic caused by ‘COVID-19’ across the globe. This research focuses on utilizing the strengths of Convolutional Neural Networks (CNN) and Long Short Trem Memory (LSTM) for improving the accuracy of classification and similarity score prediction. Thus, the authors propose a deep learning model ‘GenomeSimilarityPredictor’ for detecting the mutation caused by ‘SARS-CoV-2’ in the genome of human beings and for detecting the similarity score of ‘SARS-CoV-2’ with the other viruses namely ‘SARS’, ‘MERS’, ‘HIV’ and ‘HTLV’. Its illustration of the workflow is shown in Fig 1. The main contributions of this research are.

- Precisely detecting the presence of ‘SARS-CoV-2’ in the human genome.
- Quickly classifying the host genomes into infected and non-infected categories.
- Finding the similarity of ‘SARS-CoV-2’ with other viruses.
- Providing insights for finding the effective drug/vaccine for the treatment of ‘COVID-19’ from the available drugs or vaccines.
- Conducting the extensive analysis of the existing models for finding the genomic similarity.
- Finding the advantages of developing the problem specific CNN and LSTM model.
- Analysing the impact of optimizing the hyperparameters on the performance of the model.
- The use of K-Cross validation to validate the performance of the ‘GenomeSimilarityPredictor’. The experiments conducted and analysis done prove that the model outperforms the existing models available in the state of the art.

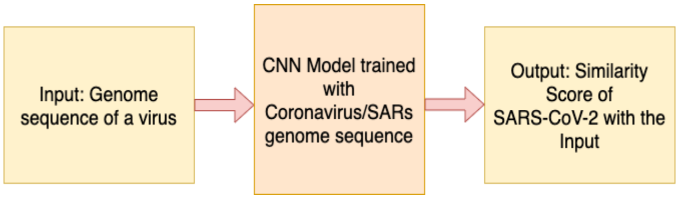
Input: Genome sequence of a virus
CNN Model trained with Coronavirus/SARs genome sequence
Output: Similarity Score of SARS-CoV-2 with the Input

Fig1: Illustration of the Model flow.

The remaining paper is organized as follows. Section 2 presents the related work and a brief description of the background of the proposed work is given in Section 3. Section 4 provides a detailed discussion of the methodology used to design and develop the model. Section 5 demonstrates the experiments. In section 6, the authors give a discussion on the contributions of this research and the state of the art. They conclude the work and give directions for future work in Section 7.

2. Related Work

The study of literature gave us information about the biological, biotechnological, statistical, and technical aspects related to 'COVID-19' and its causing agent 'SARS-CoV-2'. The research community is contributing to find the solutions for early diagnosis and treatment of 'COVID-19'. The researchers in [7] focus on analyzing the data available at web-based platforms to demonstrate the trends about the effect of 'SARS-CoV-2' across the globe. The authors in [8] reviewed the findings of the recent literature published on 'COVID-19'. They highlighted the challenges in its diagnosis and treatment. They presented the role of Artificial Intelligence (AI) to resolve the identified challenges. The authors have shown the application of Generative Adversial Networks (GANs) in detecting the genome of 'SARS-CoV-2' in the human genome. They also focused on the LSTM used for identifying the precise treatment for patients of 'COVID-19'. The researchers in [9] discussed different strategies for the diagnosis of 'COVID-19'. They claimed that genomic sequence detection is a fast and reliable technique for diagnosis of the disease.

The researchers in [10-21] proposed the computer-based solutions for the detection of the viral genome or predicting the genomic similarity among viruses. The work proposed in [10] employs the hidden Markov's model and identifies the viral genome from the host cell with an accuracy of 87%. The authors claimed that the model overcomes the challenges of Polymerase Chain Reaction (PCR) and achieved higher accuracy. Their model is effective in identifying even the elusive genomic sequences. The authors in [11] developed a web-based software named as 'Coronavirus Typing Tool' for rapid detection of the genome of Coronavirus 'SARS-CoV-2' from the given genomic sequences. The tool identifies the phylogenetic clusters as well as the genotype of the virus. It adds the identified phylogenetic and genomic information in the database. Using this information, the tool recognizes the genomic sequence of the virus with an accuracy of 87.5%. They improved the accuracy reported in [12] by 0.5%. This tool is freely available online. But, its low accuracy is a point of concern for its use in the screening of patients in real-time.

The authors in [12] applied natural language processing to detect the genome of a viral sequence. They considered the genome as a string and detected the sequence of base pairs as a sub-string. They also compared their model with the existing models and claimed that the machine learning and deep learning models outperform the Term Frequency (TF) and Inverse Document Frequency methods in the detection of a viral genome. They reported the highest Area Under Curve (AUC) score of 85%. The researchers in [13] applied the K-tuple word frequencies for the detection of a viral genome from the metagenomics sequence. Their model outperformed the model proposed in [14] and achieved the AUC score of 91.4%.

The researchers in [14-21] took advantage of object detection, classification and pattern matching ability of the deep learning models. They proposed the deep learning models for the detection of the genome of the virus from the given genomic sequence. The researchers in [14] predicted the tendency of Coronavirus to infect the human host cell. They applied the deep learning model on the genomic sequences collected and claimed that the Coronavirus 'SARS-CoV-2' has a high tendency to combine with the human genome and cause infection. The researchers in [15] applied the random forest and Artificial Neural Network (ANN) based model for the classification of viral genomes present in the metagenomics sequences. They achieved the AUC score of 79%. The authors in [16] developed the deep learning model 'Basset' for the multi-class classification of viral genomes. They improved the accuracy reported in [15] by 10% and achieved the highest accuracy of 89%. The authors in [17] developed the CNN based model 'ViraMiner' for the detection of the viral genome in the human genome. They reported the AUC score of 92%. The researchers in [18] modified their model and applied CNN based k-mer classification. Their model worked on the top k tuples for classification. They reported an accuracy of 93%.

The authors in [19] applied UNet for detecting the 2019 novel Coronavirus. They have done a significant improvement of 2.24% over the work proposed in [18] and achieved the highest accuracy of 95.24%. But, they predicted the region of infection under the confidence cut off value of 0.50. The authors did not justify how did they select the value of confidence cut. The researchers in [20] applied the Bi-path CNN and improved the accuracy of detection of the novel Coronavirus 'SARS-CoV-2'. They applied their model on the genomic sequences of infected people in Wuhan, China. They claimed that their model is effective in identifying the genomes of the 2019-nCoV or 'SARS-CoV-2', and 'SARS-CoV'. They predicted that 2019-nCoV is present in most of the samples. The authors reported an accuracy of 97.05%. But, the high computation complexity of Bi-path CNN is a limitation in its use. The authors in [21] proposed the deep learning model that distinguishes the 'SARS-CoV-2' from other Coronaviruses with an accuracy of 98.17%. They reported the improvement of 1.12% over the work proposed in [20]

The analysis of related research works shows that the limited techniques are available for the screening of 'COVID-19' using the genomic similarity. The computer based techniques proposed to detect the genome of the virus or classification of viruses based on their genomic sequences face the challenge of low accuracy in multi-class classification and high computation complexity. Thus, there is a scope to improve the accuracy and reduce the time complexity of the existing models. Also, there is a lack of research works in finding the genomic similarity score of the 'SARS-CoV-2' with other viruses. The genomic similarity of Coronavirus with other viruses may give insights to find the drug effective for the treatment of 'COVID-19' from the pre-discovered drugs for the treatment of the infections caused by viruses similar to 'SARS-CoV-2'.

3. Background

This section gives a brief description of the techniques involved in designing and developing the model for the prediction of genomic similarity.

Convolutional Neural Network: A CNN works well for identifying the simple patterns present in the dataset. The identified patterns are used to form the complex patterns in higher layers of the network. The 1-Dimensional (1-D), CNN is very effective in deriving the interesting features from small and fixed-length segments of the data set in which the location of the feature within the segment has minimum impact on the performance of the model [22].

Long Short-Term Memory: It works on the current instance as well as the previously passed input instances. LSTM uses its hidden state and preserves the selective information from the previously received inputs [23]. LSTM can be categorized as unidirectional and bidirectional. Unidirectional LSTM performs the processing based on the information preserved only from the past. Bidirectional LSTM is useful when all time steps of the input are recorded. In this LSTM, the model is trained on the forward as well as a reverse copy of the input sequence. It is useful in teaching the additional context to the network. Moreover, it takes a short time duration to give the results.

Optimizer: An optimizer is a connecting link between the loss function and parameters of the model. It updates the parameters in response to the values of the loss function. Therefore, it works for enhancing the accuracy of the model by continuously updating the weights of parameters. The optimizers such as Stochastic Gradient Descent (SGD), Adaptive Gradient (AdaGrad), Root Mean Square Propagation (RMSProp), and Adaptive Moment Estimation (Adam) are popularly used for training the neural networks. The optimizer is selected based on the values of gradients and types of parameters. The RMSProp [22] works well when there is a large variation in the values of gradients. The AdaGrad [24] is applied for the sparse dataset to improve the learning rate of its parameters. The SGD trains the model only on a randomly selected sample rather than a full dataset. Therefore, it minimizes the time required for training the model. Adam optimizer is an adaptive learning rate optimization algorithm. It includes the advantages of AdaGrad and RMSProp optimizers [24]. It computes the individual weights for each parameter. This optimizer determines the individual learning rate for each gradient and also effective in dealing with sparse parameters.

4. Methodology

In this section, the authors present the architecture, training parameters, and working of the CNN and LSTM based model 'GenomeSimilarityPredictor' applied for genome classification and prediction of the genomic similarity score of 'SARS-CoV-2'and other viruses.

4.1 The Architecture of the Model

The model consists of three branches of 1-D convolutional layers viz. C_1, $C_{2,}$ and C_3. Each convolutional layer includes 200 filters. These layers differ in their kernel size. The kernel size of C_1 is 2, C_2 is 3 and C_3 is 4. The convolution layers employ the Rectified Linear Unit (ReLU) activation function. Each convolution layer is further connected to a bidirectional LSTM layer individually. Each LSTM layer is now connected to a Global Max Pooling (GMP) layer individually. The outputs of all the GMP layers are concatenated and passed to the dropout layer. The dropout layer is connected to the dense layer which employs the sigmoid activation function. The complete architecture of the CNN model 'GenomeSimilarityPredictor' is shown in Fig 2.

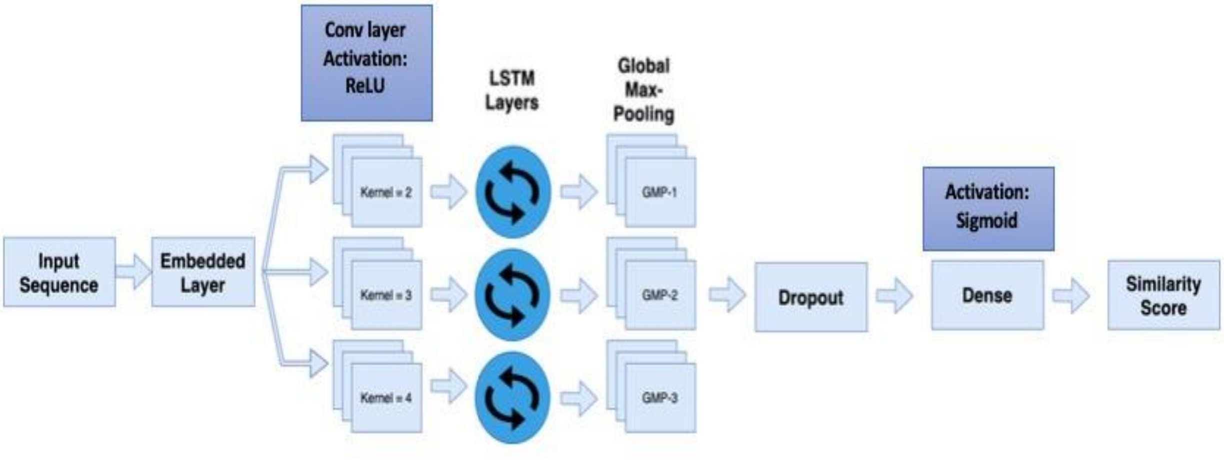

Fig 2: Architecture of 'GenomeSimilarityPredictor'.

4.2 Training Parameters

The model proposed in this manuscript makes effective use of Adam optimizer [22]. It includes the following training parameters.

I. **Alpha (α).** It denotes the learning rate of the neural network. The proportion of weights assigned to the networks are updated dynamically. The learning rate is directly dependent on the values of weights. Higher the values, the faster will be the learning rate of the neural network.

II. **Beta 1 (β_1) and Beta 2 (β_2):** These are the exponential decay rates. On performing the set of experiments, the authors determine the optimum value of β_1 as 0.9 and β_2 as 0.999. The decrease in the values of $\beta_1$and β_2 below 0.5 and above 1.2 yields high values of the loss function. This degrades the accuracy of the model. A low impact on the value of accuracy is observed when the values of β_1 and β_2 are decreased from 0.9 to 0.5 and increased from 0.9 to 1.2.

III. **Epsilon (ε):** It is a small number to prevent the division by zero error. In this model, the authors used the default value as1e-0.8 as discussed in [24].

IV. **Binary Cross-Entropy (BCE) loss** [25]: It is effective in dealing with the binary classification. It measures the performance of a classification model which yields the probability score between 0 and 1. The value of BCE loss increases with an increase in deviation of the predicted probability from its actual label. Equation 1 gives the formula to calculate its value. In this equation $H_p(\text{q})$ is the BCE loss, N is the number of points for classification, Y_i represents the label of the class. The value 0 of Y_i indicates the genome of the virus other than 'SARS-CoV-2'. The value 1 indicates the genome of 'SARS-CoV-2'. $p(y_i)$ is the probability of occurrence of a genome in the class label 1.

$$H_p(\text{q}) = -\frac{1}{N}\sum_{i=1}^{N} y_i \,.\log(p(y_i) + (1-y_i).\log(1-p(y_i)) \tag{1}$$

V. **Activation Function:** The activation function is applied to introduce the non-linearity into the output of a neuron. This improves the learning of the model. In the proposed model, the authors use the ReLU activation function for the CNN layers and the Sigmoid activation function for the Dense Layer. For the positive value of the input, the ReLU function does no modification. On the contrary, it returns '0' for the negative or '0' values given as input. For example, x is the input given to the ReLU function then it returns x for x>0 and return '0' for x<=0 [25]. The Sigmoid function is effective in predicting the probability of occurrence of an input sequence in a labeled class. In this research work, the authors aim to predict the probability of occurrence of a genomic sequence in the 'SARS-CoV-2' class. Thus, they chose the sigmoid activation function for the dense layer of the model.

4.3 Working of the Model

The three convolutional layers of the model C_1, C_2, and C_3 receive the genomic sequences of 'SARS-CoV-2', 'SARS-CoV', 'MERS-CoV', HIV, HTLV, and Bat SARS-like virus as inputs and gives the degree of similarity of 'SARS-CoV-2' with the remaining input genomic sequences. Algorithm 1 shows the sequence of steps followed by the model. The model also detects the genomes of the 'SARS-CoV-2' in the human genome.

The publicly available dataset [26-30] containing the genomic sequences is embedded by the embedding layer. This layer performs the pre-processing and computes the score for each sequence based on the position and frequency of the nitrogenous bases: Adenine (A), Cytosine (C), Guanine (G) and Thymine (T) (for DNA) or Uracil (U) (for RNA). The score is useful in deriving the context of the genomic sequence. Now, the output obtained from the embedding layer is passed to the convolutional layers C1, C2, and C3. These layers extract the features based on the filters and kernel size. Now, the bidirectional LSTM layers perform training of the model on the original input sequence as well as its reverse copy. The LSTM layers pass the outputs to the GMP layers. Each GMP layer extracts the maximum value from each feature map. Now, all three GMP layers are concatenated together and connected to the dropout layer. The dropout layer acts as an intermediate between the GMP layers and the dense layer. It allows only the selected neurons to establish connections to the dense layer. This is important to reduce the problem of overfitting. It also reduces the size of connections and hence, useful in reducing the time complexity. Here, the authors set the probability score as 0.2. It means that the dropout layer randomly removes 20 neurons from every 100 neurons. The remaining 80 neurons are passed to the next layer for connection. Finally, the dense layer employs the sigmoid as an activation function to calculate the probability score. The three convolutional layers of the model C1, C2, and C3 are trained on 'SARS-CoV-2' genomic sequences. Thus, the model becomes efficient in understanding the pattern and structure of its genome. When the model receives the input sequence of other viruses, then it generates the similarity score of 'SARS-CoV-2' with the input sequence. This score demonstrates the similarity index of the genome of 'SARS-CoV-2' with other input genomes. Score 1 indicates that the genome is the same as the 'SARS-CoV-2'. Score 0 indicates no matching of input genome with 'SARS-CoV-2'. The probability score between 0 and 1 indicates the degree of similarity of a genome with the 'SARS-CoV-2'. Algorithm 1 shows the sequence of steps the model follows to complete the prediction.

Algorithm 1: Genomic Similarity Prediction.

Input: GenomeSimilarityPredictor (genomic sequences of SARS-CoV-2, human beings, SARS-CoV, MERS-CoV, HIV, HTLV, Bat SARS Like Virus).

Output: Similarity Score among input genomic sequences.

```
1. $Embedding(genomic\ sequences) \rightarrow$
   $Score\ for\ each\ genomic\ sequence$
2. $Convolution(Scores\ calculated\ in\ step\ 1) \rightarrow$
   $Extracted\ Features$
3. $B\text{idirectional LSTM}(\text{Extracted Features}) \rightarrow$
   $\text{FeatureMap}$
4. $Global\ Max\ Pooling(Feature\ Maps) \rightarrow$
   $MaxValue\ Feature\ Maps$
5. $Dropout\ layer(Neurons) \rightarrow$
   $ReducedNumber\ of\ Neurons$
```

6. $Dropout\ Layer \rightarrow Sigmoid\ Activation \rightarrow Dense\ Layer$
7. $Probability\ score \rightarrow (0\ to 1)$
8. $Interpret(Probability\ Score) \rightarrow SimilarityScore\ among\ input\ genomes$

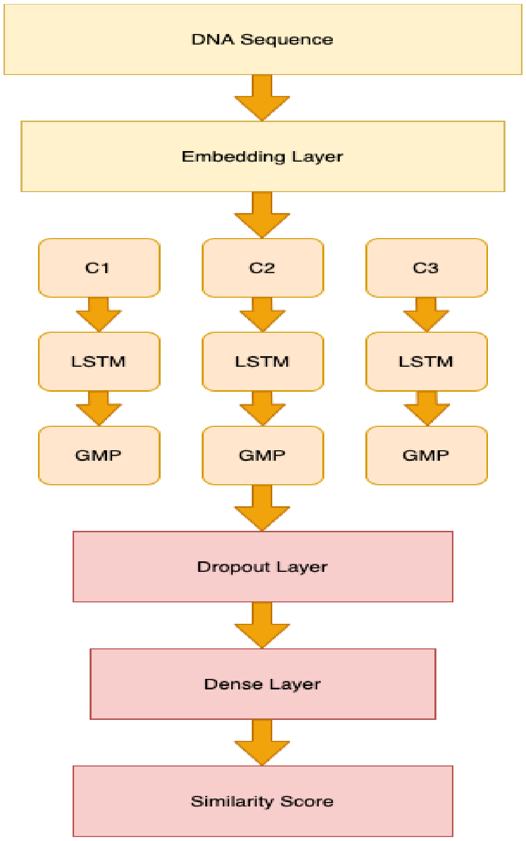

Fig 3: Working of 'GenomeSimilarityPredictor'.

5. Experiments and Results Analysis

The authors use Google Colab [31], the freely available online training platform for implementing the model. It has Tesla K80 GPU and 12 GB RAM. The Google Colab provides the facility of continuous execution for 12 hours.

For experiments, the authors used the dataset publicly available at [26]. The dataset contains 852 genomic sequences. The 100 sequences are genomes of 'SARS-CoV-2' in the host human beings and the remaining 752 sequences are genomes of non-infected human beings. These genomic sequences have been extracted from blood samples, oronasopharynx, and lungs of the people from different geolocations such as China, Spain, USA, Vietnam, and Thailand.

As a part of the preprocessing step, the authors assigned unique numbers 0, 1, 2, and 3 to the nitrogenous bases: A, G, C, and T respectively. They sliced genomic sequence is at the interval of the length 300. The authors divided the dataset of 852 genomic sequences into a batch size of 128. They obtained 7 batches each of size 128 for training the model. The batch size is selected based on the experiments conducted on different batch sizes such as 32, 64, 128, and 256. The model reported the best performance on the batch size of 128. This batch size initializes the training of the model before familiarizing it with the complete dataset. This is effective in dealing with the problem of generalization [23]. The authors also used 172 genomic sequences of 'SARS-CoV' [27] 'MERS-CoV' [28], 'HIV' [29] and 'Bat-SARS like virus' [30] for testing the efficacy of the model in calculating the degree of similarity of these sequences with the genome of 'SARS-CoV-2'. The genomic sequences of 'SARS-CoV-2' are labeled as 1 and the remaining sequences are labeled as 0.

The authors used 80% of the total dataset for training, 10% for the validation, and 20% for testing the model. They set the value of the threshold as 0.75 for the sigmoid function at the dense layer. This value indicates that the model will give the probability score 1 if the genomic sequence resembles 75% or higher to the genome of 'SARS-CoV-2'. Score 0 is obtained, if the genomic sequence shows the similarity lower than the pre-set value of the threshold. The model uses 2,69,937 parameters for the training. The number of parameters is dependent on the number of filters and the kernel size of the model. The model is executed to detect the presence of the genome of 'SARS-CoV-2' in the input genomic sequence of human beings. It also predicts the similarity score of the above-mentioned viruses with the genome of 'SARS-CoV-2'.

5.1 Performance of the Model

To evaluate the efficacy of the proposed model, the authors used the following nine evaluation metrics.

I. **BCE loss:** The value of loss function indicates the error in the prediction. The value decreases with the training of the model. Fig 4 demonstrates the trends for the values of BCE loss with an increase in the number of epochs. It is clear from Fig 4 that the values of BCE loss decrease with an increase in the number of epochs. This demonstrates that the proposed model continuously learns from the values of the loss function and updates the weights to minimize it. This leads to a decrease in the value of the loss. The minimum value of loss after 30 epochs, indicates that the model has become effective in classifying the input genomic sequences.

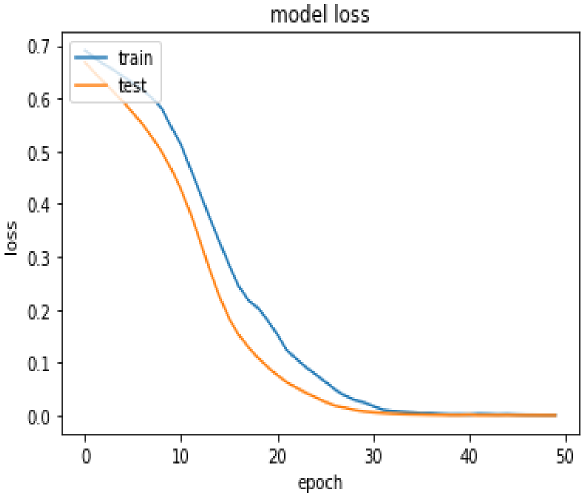

Fig 4: Variation in BCE loss with Number of Epochs.

II. **Confusion Matrix:** This is used to display the number of correctly and incorrectly classified instances from the test dataset. Table 1, shows the confusion matrix for the 172 genomic sequences used in the testing set of the ‘GenomeSimilarityPredictor’. Here, the True Positive (TP) represents the number of instances correctly predicted in the positive class. True Negative (TN) shows the number of instances correctly predicted in the negative class. False Positive (FP) gives the number of instances incorrectly predicted in the positive class. False Negative (FN) displays the number of instances incorrectly predicted in the negative class.

Table 1: Confusion Matrix.

	Positive	Negative
Positive	(TP) 149	(FP) 0
Negative	(FN) 1	(TN) 22

III. **Accuracy:** This is the ratio of correctly classified genomic sequences to that of total dataset size. Equation (2), gives the formula to calculate the value of accuracy (ACC). The ‘GenomeSimilarityPredictor’ misclassifies only 1 instance as the false negative. Thus, it achieved the highest **accuracy** of **99.27%**. Fig 5 demonstrates the variation in the accuracy of ‘GenomeSimilarityPredictor’ with a change in the number of epochs. There is a random increase and decrease in the value of accuracy when it is executed from the 0th to 30th epoch. This reveals that the model is continuously learning and updating the weights. After 30 epochs the **accuracy** achieves its maximum value and becomes **99.27%.** On further increasing the number of epochs, no significant change is observed in the accuracy of the model. This shows that the model is trained on all the parameters when executed for 30 epochs.

$$\text{Accuracy} = \frac{TP+FP}{TP+FP+TN+FN} \qquad (2)$$

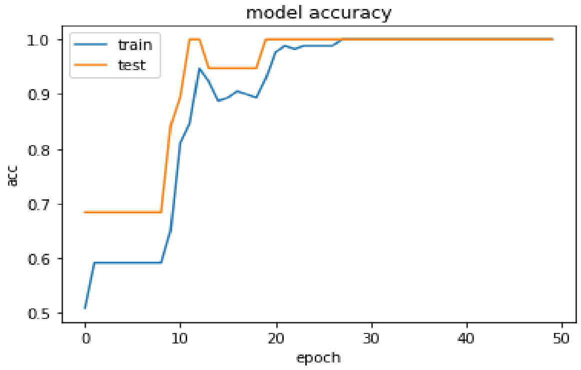

Fig 5: Variation in Accuracy with Number of Epochs.

IV. **Receiver Operating Curve (ROC):** This is the graphical representation of the TP Rate versus FP Rate at different values of the threshold. The confusion matrix shown in Table 1, displays the number of TP, FP, TN, and FN genomic sequences obtained on testing the model with 172 genomic sequences. The ROC as shown in Fig 6, illustrates the efficacy of the classifier ‘GenomeSimilarityPredictor’. The high value **0.9783** of Area Under Curve **(AUC)** below the ROC demonstrates the effectiveness of the model to categorize the test sequences of ‘SARS-CoV-2’ from genomic sequences of human beings.

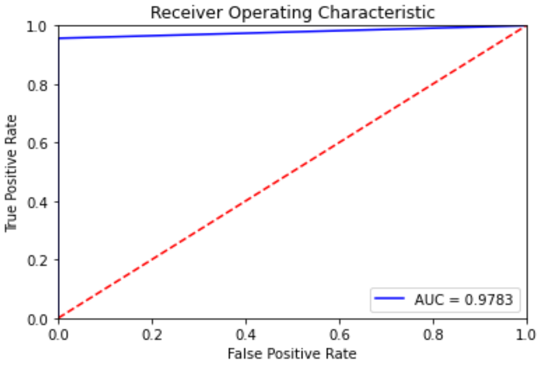

Fig 6: ROC and AUC

V. **Precision (P):** This is the ratio of correctly classified genomic sequences in the expected class to that of the total number of genomic sequences classified correctly to both classes. Equation 3 gives the formula to calculate the precision of the model. The model ‘GenomeSimilarityPredictor’ gives a **precision of 99%.** The high value of precision proves that the model is effective in extracting the relevant instances *i.e.* genome of ‘SARS-CoV-2’ from the total number of extracted instances.

$$P = \frac{TP}{TP+FP} \qquad (3)$$

VI. **Recall (R):** This is the proportion of correctly classified genomic sequences to the expected class to that of the total number of genomic sequences classified to the expected class. Equation (4) gives the formula for calculating the value of recall. The model 'GenomeSimilarityPredictor' achieves the highest **recall of 100%.** This proves that the model is effective in extracting all the relevant instances from the given instances. The value of TP=149 and TN=0 shows that all the 149 genomes are correctly classified into its relevant class.

$$R = \frac{TP}{TP+FN} \qquad (4)$$

VII. **F1 Score:** This is the harmonic mean of precision and recall. The formula for calculating the value of the F1 score is given in equation (5). The 'GenomeSimilarityPredictor' achieves the highest **F1 score of 100%.** This proves the efficacy of the model in classifying the genomes of 'SARS-CoV-2' and Homo sapiens.

$$F1\ Score = 2 * \frac{P*R}{P+R} \qquad (5)$$

VIII. **Degree of Similarity:** The model 'GenomeSimilarityPredictor' predicts the degree of genomic similarity of four viruses: 'HTLV', 'HIV', 'MERS-CoV', and 'SARS-CoV' with the genome of 'SARS-CoV-2'. Table 2 shows the degree of similarity obtained. Its first column contains the name of the virus and the second column displays the degree of similarity with 'SARS-CoV-2' in %. It is clear from the similarity score shown in Table 2 that 'SARS-CoV-2' shows the highest genomic similarity of 98.11% with the 'SARS-CoV'.

Table 2: Degree of Genomic Similarity.

Name of Virus	Degree of Similarity with SARS-CoV-2 in %
HTLV	86.5
HIV	89.7
MERS-CoV	95.3
SARS-CoV	98.11

IX. **K-Fold Cross Validation:** To validate the reliability of the 'GenomeSimilarityPredictor', the authors applied the 10-fold cross-validation. They divided the dataset into ten equal-sized subsets. The nine subsets are used for training the model and the remaining one subset is used for testing. The process is repeated until each subset becomes the testing subset. The 10-fold cross-validation minimizes the problem of overfitting and underfitting of the model.

The trend in the value of the loss function at each fold is shown in Fig 7. The values obtained at different iterations of the 10-fold cross-validation are shown in Table 3. The model gives the 0.005904 as the average value of loss function. The low value of the loss function proves its efficacy in predicting the similarity score of 'SARS-CoV-2' with 'HTLV', 'HIV', 'MERS', and 'SARS' viruses.

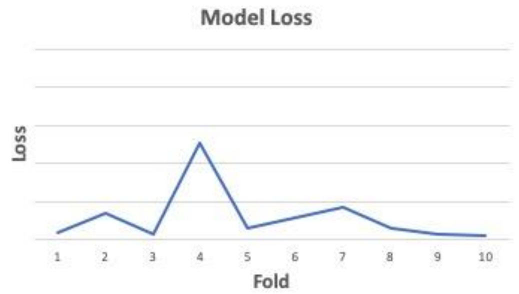

Fig 7: Trend in the Loss Function of 'GenomeSimilarityPredictor'.

Table 3: Loss at Each Fold.

Fold_Number	Model Loss
1	0.00194
2	0.00714
3	0.00139
4	0.02560
5	0.00312
6	0.00582
7	0.00850
8	0.00318
9	0.00127
10	0.00108

5.2 Discussion

As per the discussion given in [2], less information is available about the response from the human immune system to the genomic sequence of 'SARS-CoV-2'. The knowledge about the genomes, genes, protein sequences, protein structures of pathogens, and the response from the host cell are the key requirements for the discovery of vaccine or medicine effective in the treatment of any pathogen. The resemblance of the genome or protein structure of 'SARS-CoV-2' with previously discovered viruses can give useful insights for the treatment of 'COVID-19'.

The genome matching of a virus is possible with high-end computing devices and AI [14]. The deep learning models proposed from [14-21] worked for the detection of viral genome in the host cell. Among all the models proposed in the state of the art, the Bi-path Convolutional Neural Network proposed in [20] achieved the accuracy of **97.05%** in detecting the infection due to 2019-nCoV, SARS, and SARS-CoV. But, its high computation complexity is a limitation for its use. Moreover, it does not find the similarity among the different genomic sequences of viruses. Thus, it does not give insight into drug discovery. The deep learning model proposed in [21] reported the better accuracy than the model proposed in [20]. The model achieved an accuracy of 98.17% for the classification of 'SARS-CoV-2' from the given genomic sequences.

The 'GenomeSimilarityPredictor' proposed in this manuscript is a hybrid of CNN and LSTM. It improved the accuracy of prediction even for the multi-class classification. The model also worked for identifying the virus showing the maximum similarity with 'SARS-CoV-2'. The authors optimized the number of convolution layers in their model to minimize the problem of overfitting and underfitting on the dataset used for training and testing. The optimized combination of CNN and LSTM in 'GenomeSimilarityPredictor' improved the accuracy reported in [21] by 1.10% and achieved the highest accuracy of **99.27%** in detecting the presence of 'SARS-CoV-2' in the genome of human beings. The model achieved the **recall** of **100%** on the test dataset. In addition, it finds the similarity score of a given genomic sequence with other genomic sequences.

The comparison of the experimental results with the techniques proposed in [14], [21-23] prove that the model outperforms the existing models with an accuracy of **99.27%.** The authors present the comparison of the accuracy of the existing techniques and the 'GenomeSimilarityPredictor' in Fig 8.

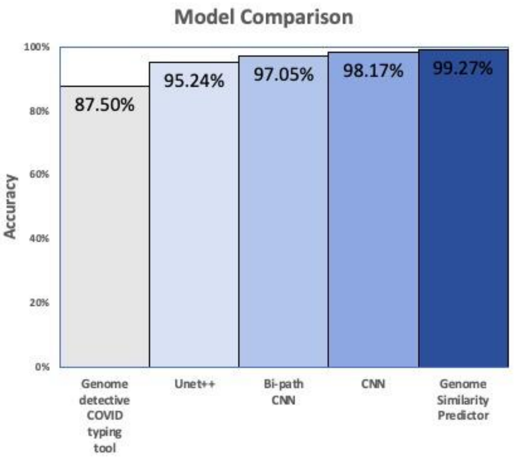

Fig 8: Performance Comparison of 'GenomeSimilarityPredictor'.

6. Conclusion and Future Work

Fighting with 'COVID-19' is the prime objective of the world at present. The scientists, health experts, and research community are contributing in their best ways to meet this objective.

In this manuscript, the authors proposed a technological solution to deal with 'COVID-19'. Their model 'GenomeSimilarityPredictor' is an optimum CNN model with LSTM. It is efficient in detecting the genome of 'SARS-CoV-2' in human beings with an **accuracy of 99.27%.** It gives the low value of **loss function 0.005904** on applying k-fold cross-validation. The model also determines the degree of similarity in the genomes of 'SARS-CoV', 'MERS-CoV', 'HTLV, and 'HIV' with 'SARS-CoV-2'. The experimental results shown in Table 2 demonstrates that the genome of 'SARS-CoV-2' is the most similar to the genome of 'SARS-CoV'. This can be a useful clue for the clinicians to find the most effective vaccine or drug for the treatment of 'COVID-19'.

The comparison of the proposed CNN model 'GenomeSimilarityPredictor' with models proposed in the literature [10-21] shows that model has reported higher accuracy and outperforms the existing techniques as shown in Fig 8. Its effectiveness in dealing with noisy data, low time complexity makes it applicable for the screening of infected genomes in the present situation of 'Global Pandemic'. The zero instance in the FP and only 1 instance in the FN increase the acceptability of this model. Thus, it can be used for mass screening of patients infected with 'SARS-CoV-2'. It may prove a quick and reliable tool for the doctors.

Future Scope: The application of this model can be extended to quickly find the degree of similarity between genomes of any two microbes. The model can be trained with the labelled genomic sequences of infected and healthy human beings for mass screening of patients. It can quickly detect the mutation in the human genome. It can be used for the multiclass classification and may prove useful in screening of infected patients with more than one virus.

Compliance with ethical standards

Funding: This work is partially funded by FCT/MCTES through national funds and when applicable co-funded EU funds under the Project UIDB/EEA/50008/2020; and by Brazilian National Council for Scientific and Technological Development (CNPq) under Grant No. 309335/2017-5.

Conflict of Interest: The authors declare that they have no conflict of interest.

References

[1] WHO (2020) Naming the coronavirus disease (COVID-19) and the virus that causes it. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/naming-the-coronavirus-disease-(covid-2019)-and-the-virus-that-causes-it. Accessed 21 June 2020.

[2] WHO (2020) Covid-19 Coronavirus Pandemic. https://www.worldometers.info/coronavirus/. Accessed 9August 2020.

[3] Gerald Choon-Huat KOH, Helen HOENIG (2020) How Should the Rehabilitation Community Prepare for 2019-nCoV. Archives of Physical Medicine and Rehabilitation. https:// doi.org/10.1016/j.apmr.2020.03.003.

[4] World Health Organization (2020) Laboratory testing for Coronavirus disease 2019 (COVID-19) in suspected human cases. Interim guidance 2 March 2020.https://apps.who.int/iris/bitstream/handle/10665/331329/WHO-COVID-19-laboratory-2020.4-eng.pdf?sequence=1&isAllowed=y. Accessed 5 July 2020.

[5] C. Long, H. Xu, et al (2020) Diagnosis of the Coronavirus disease (COVID-19): RRT-PCR or CT?

Eur. J. Radiol. 126:108961. doi:10.1016/j.ejrad.2020.108961

[6] Ushmani, Azhar. (2019). Machine Learning Pattern Matching. 10.13140/RG.2.2.16276.96649.

[7] Robson B. (2020). Computers and viral diseases. Preliminary bioinformatics studies on the design of a synthetic vaccine and a preventative peptidomimetic antagonist against the SARS-CoV-2 (2019-nCoV, COVID-19) coronavirus. Computers in biology and medicine. 119: 103670. https://doi.org/10.1016/j.compbiomed.2020.103670S

[8] M. B. Jamshidi et al (2020) Artificial Intelligence and COVID-19: Deep Learning Approaches for Diagnosis and Treatment. IEEE Access 8: 109581-109595. doi: 10.1109/ACCESS.2020.3001973.

[9] Kumar, R., Nagpal, S., Kaushik, S. et al (2020) COVID-19 diagnostic approaches: different roads to the same destination. VirusDis. 31: 97–105 https://doi.org/10.1007/s13337-020-00599-7

[10] Skewes-Cox, P., Sharpton, T. J., Pollard, K. S., & DeRisi, J. L. (2014). Profile hidden Markov models for the detection of viruses within metagenomic sequence data. PloS one. 9(8): e105067. https://doi.org/10.1371/journal.pone.0105067

[11] Cleemput S, Dumon W, Fonseca V, et al (2020) Genome Detective Coronavirus Typing Tool for rapid identification and characterization of novel coronavirus genomes. Bioinformatics. 36(11):3552-3555. doi:10.1093/bioinformatics/btaa145

[12] Aly O. Abdelkareem, Mahmoud I. Khalil, Ali H. A. Elbehery, Hazem M. Abbas (2020) bioRxiv 2020.01.10.892158. https://doi.org/10.1101/2020.01.10.892158

[13] Ren J, Ahlgren NA, Lu YY, Fuhrman JA, Sun F (2017) VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. Microbiome. 5(1):69. doi:10.1186/s40168-017-0283-5

[14] Zhu H, Guo Q, Li M, Wang C, Fang Z, Wang P, Tan J, Wu S, Xiao Y (2020) Host and infectivity prediction of Wuhan 2019 novel coronavirus using deep learning algorithm. bioRxiv 2020.01.21.914044; doi: https://doi.org/10.1101/2020.01.21.914044

[15] Bzhalava Z, Tampuu A, Bała P, Vicente R, Dillner J (2018) Machine Learning for detection of viral sequences in human metagenomic datasets. BMC Bioinformatics. 19(1):336. doi:10.1186/s12859-018-2340-x

[16] Kelley DR, Snoek J, Rinn JL (2016) Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. Genome Res. 26(7):990-999. doi:10.1101/gr.200535.115

[17] Tampuu, A., Bzhalava, Z., Dillner, J., & Vicente, R. (2019). ViraMiner: Deep learning on raw DNA sequences for identifying viral genomes in human samples. PloS one, 14(9): e0222271. https://doi.org/10.1371/journal.pone.0222271

[18] Un Chen, Lianlian Wu, Jun Zhang, et al (2019) Deep learning-based model for detecting 2019 novel coronavirus pneumonia on high-resolution computed tomography: a prospective study. medRxiv 2020.02.25.20021568. doi: https://doi.org/10.1101/2020.02.25.20021568.

[19] Ren J, Song K, Deng C, Ahlgren NA, Fuhrman JA, Li Y, Xie X, Sun F (2018) Identifying viruses from metagenomic data by deep learning. arXiv preprint arXiv:1806.07810

[20] Le, Ho, Lee, & Jung. (2019) Application of Long Short-Term Memory (LSTM) Neural Network for Flood Forecasting. Water, 11(7): 1387. doi:10.3390/w11071387

[21] Alejandro Lopez-Rincon, Alberto Tonda, Lucero Mendoza-Maldonado et. Al (2020) Accurate Identification of SARS-CoV-2 from Viral Genome Sequences using Deep Learning. bioRxiv 2020.03.13.990242; doi: https://doi.org/10.1101/2020.03.13.990242

[22] Gu J, Wang Z, Kuen J, Ma L, Shahroudy A, Shuai B, Liu T, Wang X, Wang G, Cai J, Chen T (2018) Recent advances in convolutional neural networks. Pattern Recognition. arXiv:1512.07108v6

[23] Ruder S. An overview of gradient descent optimization algorithms (2016) arXiv preprint arXiv:1609.04747

[24] Pradhan, N., Dhaka, V.S., Rani, G, Choudhary H (2019) Transforming view of medical images using deep learning. Neural Comput & Applic. https://doi.org/10.1007/s00521-020-04857-z.

[25] Creswell A, Arulkumaran K, Bharath AA (2017) On denoising autoencoders trained to minimize binary cross-entropy. arXiv preprint arXiv:1708.08487

[26] National Center for Biotechnology Information (2020) https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide&VirusLineage_ss=Severe%20acute%20respiratory%20syndrome-related%20coronavirus,%20taxid:694009. Accessed 26 March 2020.

[27] National Center for Biotechnology Information (2020) https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide&VirusLineage_ss=SARS%20coronavirus%20ExoN1,%20taxid:627440. Accessed 26 March 2020.

[28] National Center for Biotechnology Information (2020) https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide&VirusLineage_ss=Middle%20East%20respiratory%20syndrome-related%20coronavirus%20(MERS-CoV),%20taxid:1335626. Accessed 26 March 2020.

[29] National Center for Biotechnology Information (2020) https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide&VirusLineage_ss=Human%20immunodeficiency%20virus%201%20(HIV-1),%20taxid:11676. Accessed 26 March 2020.

[30] National Center for Biotechnology Information (2020) https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide&VirusLineage_ss=Bat%20S

ARS-like%20coronavirus,%20taxid:1508227. Accessed 26 March 2020.

[31] Carneiro T, Da Nóbrega RV, Nepomuceno T, Bian GB, de Albuquerque VH, Reboucas Filho PP (2018). Performance analysis of google collaboratory as a tool for accelerating deep learning applications. IEEE Access. 6:61677-85.