Multimedia Systems (2022) 28:1251–1262
https://doi.org/10.1007/s00530-021-00824-3

SPECIAL ISSUE PAPER

Applying deep learning-based multi-modal for 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]

Received: 21 September 2020 / Accepted: 20 June 2021 / Published online: 21 July 2021

Abstract

Amidst the global pandemic and catastrophe created by 'COVID-19', every research institution and scientist 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-based multi-modal for the screening of COVID-19 using chest radiographs and genomic sequences. The modal is also effective in finding the degree of genomic similarity among the Severe Acute Respiratory Syndrome-Coronavirus 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. The experimental results on the datasets available at National Centre for Biotechnology Information, GitHub, and Kaggle repositories show that it is successful in detecting the genome of 'SARS-CoV-2' in the host genome with an accuracy of 99.27% and screening of chest radiographs into COVID-19, non-COVID pneumonia and healthy with a sensitivity of 95.47%. Thus, 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 a serious health pandemic since the last 14 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 '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 114,944,666 people and caused 2,550,287 deaths till 2 March 2021, across 219 countries [2]. The majority of patients has shown 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

✉ Vijaypal Singh Dhaka
vijaypalsingh.dhaka@jaipur.manipal.edu

Geeta Rani
Geeta.rani@jaipur.manipal.edu

Meet Ganpatlal Oza
meetoza08@gmail.com

Nitesh Pradhan
nitesh.pradhan@jaipur.manipal.edu

Sahil Verma
sahilverma@ieee.org

Joel J. P. C. Rodrigues
joeljr@ieee.org

[1] Department of Computer and Communication Engineering, Manipal University Jaipur, Jaipur, Rajasthan, India

[2] Department of Computer Science and Engineering, Manipal University Jaipur, Jaipur, Rajasthan, India

[3] Department of Computer Science and Engineering, Chandigarh University, Mohali 140413, India

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

[5] Instituto de Telecomunicações, Aveiro, Portugal

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 a blood sample for detection of the presence of ribonucleic 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]. The 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 image processing, object detection, pattern recognition, learning, and matching can be used for mass screening of COVID-19 and the detection of mutation caused by 'SARS-CoV-2' in the human genome [6]. This motivated the authors to provide a deep learning-based solution for quick and mass screening of 'COVID-19' using multiple modalities.

The main objective of this research is to develop a quick, reliable, and multi-modal 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-term memory (LSTM) for improving the accuracy of classification and similarity score prediction. Thus, the authors propose a multi-model system for screening of COVID-19. It is a unique architecture comprising deep learning models 'GenomeSimilarityPredictor' and 'COVID-Screen-Net' for the screening of COVID-19 using genomic data or chest radiographs. The 'GenomeSimilarityPredictor' is effective in analyzing the textual data. It detects 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'. Whereas, the tailored CNN model 'COVID-Screen-Net' is efficient in dealing with imagery data. It classifies the chest X-rays into COVID-19, non-Covid pneumonia, and healthy classes. The interface of the multi-model system is shown in Fig. 1.

The main contributions of this research are as follows.

- To provide a multi-modal technological solution for the screening of COVID-19 using chest radiographs and genomic sequences.
- Precisely detecting the presence of 'SARS-CoV-2' in the human genome.
- Quickly classifying the host genomes and chest radiographs into infected and non-infected categories.
- Finding the genomic similarity of 'SARS-CoV-2' with other viruses.
- Providing insights into finding the effective drug/vaccine for the treatment of 'COVID-19' from the available drugs or vaccines.

The remaining paper is organized as follows. Section 2 presents a brief description of the background and related works. Section 3 provides a detailed discussion of the methodology used to design and develop the model. Section 4 demonstrates the experimental results and discussion. In Sect. 5, the authors conclude the work and give directions for future work.

2 Background and related works

Since the onset of 2020, the research community is contributing to find the solutions for an early diagnosis and treatment of 'COVID-19'. The researchers in [7] focused 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 convolutional neural networks (CNNs),

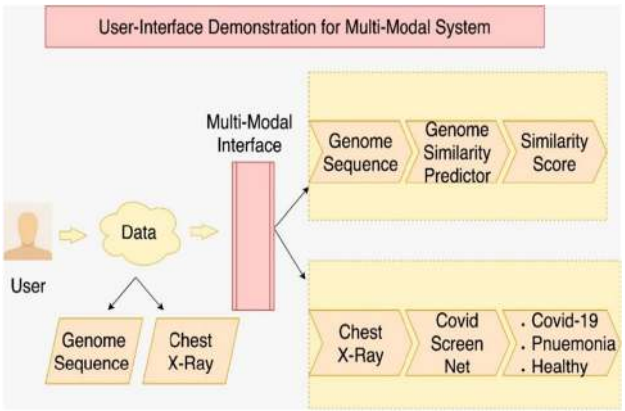

Fig. 1 User interface of multi-model system

LSTM to resolve the identified challenges in detecting the genome of 'SARS-CoV-2' in the human genome and screening of COVID-19.

A CNN works well for identifying the simple patterns present in the dataset. Further, it uses these patterns to form the complex patterns at higher layers of the network. Thus, the networks become very effective in deriving the interesting features from small and fixed-length segments of the dataset [9]. To improve the performance of the CNN models, the authors employ optimizers such as Root Mean Square Propagation (RMSProp), Adaptive Gradient (AdaGrad), Stochastic Gradient Descent (SGD), and Adaptive Moment Estimation (Adam). The optimizer is selected based on the values of gradients and types of parameters. The RMSProp [9] works well when there is a large variation in the values of gradients. The AdaGrad [10] 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 [11]. It computes the individual weights for each parameter and determines the individual learning rate for each gradient. It is also effective in dealing with sparse parameters.

The LSTM works on the current instance as well as the previously passed input instances. It uses its hidden state and preserves the selective information from the previously received inputs [12]. It is useful in teaching the additional context to the network. Moreover, it takes a short time duration to give the results.

The researchers in [13] discussed different strategies for the diagnosis of 'COVID-19'. They claimed that genomic sequence detection is a fast and reliable technique for the diagnosis of the disease.

The authors in [14] take clues from the work proposed in [13] and employed the hidden Markov's model. Their model identifies the viral genome from the host cell with an accuracy of 87%. Also, the model is effective in identifying the elusive genomic sequences. The authors claimed that the model overcomes the challenges of polymerase chain reaction (PCR) and reports higher accuracy. Another group of researchers developed a web-based software named 'Coronavirus Typing Tool' [15] for rapid detection of the Coronavirus 'SARS-CoV-2' from the given genomic sequences. It uses the phylogenetic and genotypic information to recognize the virus with an accuracy of 87.5%. However, this improved the accuracy reported in [14] by 0.5%. But, the inaccuracy of 12.5% is still a point of concern for its use in real-time.

The researchers in [16] applied the random forest and artificial neural network (ANN) based model for the classification of viral genomes present in the metagenomics sequences. They achieved an area under curve (AUC) score of 79%. For providing a more reliable model, the authors in [17] applied natural language processing for the detection of the viral genome. They reported the AUC of 85%. They 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. In line with the ongoing research, the authors in [18] developed the deep learning model 'Basset' for the multi-class classification of viral genomes. They improved the correctness of the model reported in [17], and achieved the highest AUC of 89%. To further improve the performance of the models, the researchers in [19] applied the K-tuple word frequencies for the detection of a viral genome from the metagenomics sequence. Their model outperformed the model proposed in [18] and achieved the AUC score of 91.4%. The researchers in [20] proposed the applied CNN-based k-mer classification and achieved an accuracy of 93%. The authors in [21] applied UNet for detecting the 2019 novel Coronavirus. Their model achieved the highest accuracy of 95.24% and reported the improvement of 2.24% over the work proposed in [20].

The researchers in [22] applied the Bi-path CNN and improved the accuracy of detection of the novel Coronavirus 'SARS-CoV-2'. They claimed that their model is effective in identifying the genomes of the 2019-nCoV or 'SARS-CoV-2', and 'SARS-CoV' with an accuracy of 97.05%. But, the high computation complexity of Bi-path CNN is a limitation in its use. Another group of researchers proposed the deep learning model [23] that distinguishes the 'SARS-CoV-2' from other Coronaviruses with an accuracy of 98.17%. They reported an improvement of 1.12% over the work proposed in [22]

Tulin Ozturk et al [24] further enhanced the applicability of CNN in identifying the simple as well as complex patterns present in the dataset. They developed CNN models for the detection of COVID-19 from the imagery dataset. They employed the Darknet19 model for the diagnosis of COVID-19 using X-ray images obtained from two different sources [25, 26]. The model achieved the classification accuracy of 98.08% for binary classification of the dataset into Covid, and Normal classes. But, it could achieve an accuracy of 87.02% for multi-class classification of the dataset into Covid, Normal, and Pneumonia classes. The low accuracy in multi-class classification leaves scope for improving and optimizing the model. Similarly, the authors in [27] employed the VGG-19, MobileNet-V2, Xception, Inception, and Inception ResNet models for automatic detection of COVID-19 from X-ray images. They collected the datasets from GitHub repository [25], Radiological Society of North America (RSNA), Radiopaedia, and the Italian Society of Medical and Interventional Radiology (SIRM). The dataset comprises 224 X-ray images of confirmed

patients of COVID-19, 504 cases of healthy instances, 400 patients infected with bacterial pneumonia, and 314 people infected with viral pneumonia. The authors harnessed the potential of transfer learning. They claimed that MobileNet-V2 reports the highest accuracy of 96.78%, a sensitivity of 98.66%, and specificity of 96.46% for the binary classification. However, their work reported a higher accuracy than the model presented in [24]. But, they did not focus on the multi-class classification.

The authors in [28] proposed a deep convolutional neural network-based architecture 'CovXNet' for the classification of chest X-ray images into classes namely normal, non-Covid viral pneumonia, and bacterial pneumonia. They evaluated the performance of their model on the dataset comprising 1583 X-ray images of healthy persons, 1493 images of the non-Covid viral pneumonia, 2,780 X-ray images of the persons infected with bacterial pneumonia, and 305 X-ray images of the patients infected with COVID-19. They collected the dataset from Medical Center, China, and Sylhet Medical College, Bangladesh. The 'CovXNet' achieved an accuracy of 90.2% for the multi-class classification of the dataset.

The analysis of related research works shows that 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. Also, limited techniques are available for the screening of 'COVID-19' using multiple modalities, viz., chest radiographs and the genomic sequences. Moreover, the techniques based on single modality either chest radiographs or genomic sequences report low accuracy for multi-class classification. Further, these techniques have high computation complexity. Thus, there is a requirement to provide the multi-modal solution for the screening of COVID-19. Also, there is a scope to improve the accuracy and reduce the time complexity of the existing models.

3 Methodology

In this section, the authors present the architectures, training parameters, and working of the CNN and LSTM based multi-modal. The multi-modal comprises the 'GenomeSimilarityPredictor' and COVID-Screen-Net [29]. The 'GenomeSimilarityPredictor' is applied for the prediction of the genomic similarity of 'SARS-CoV-2' with other viruses and the classification of genomes into infected and non-infected. Whereas, the architecture of COVID-Screen-Net is adopted from the work presented in [29] to classify the chest radiographs into healthy, bacterial pneumonia, and COVID-19.

3.1 The architecture of 'GenomeSimilarityPredictor'

The architecture of 'GenomeSimilarityPredictor' as shown in Fig. 2, 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 that is further connected to the dense layer. The dense layer employs the sigmoid activation function for the prediction of similarity scores.

3.2 Training parameters

The model proposed in this manuscript makes effective use of the Adam optimizer [10]. It includes the following training parameters.

1. *Alpha (α)* It denotes the learning rate of the neural network. It is directly dependent on the values of dynami-

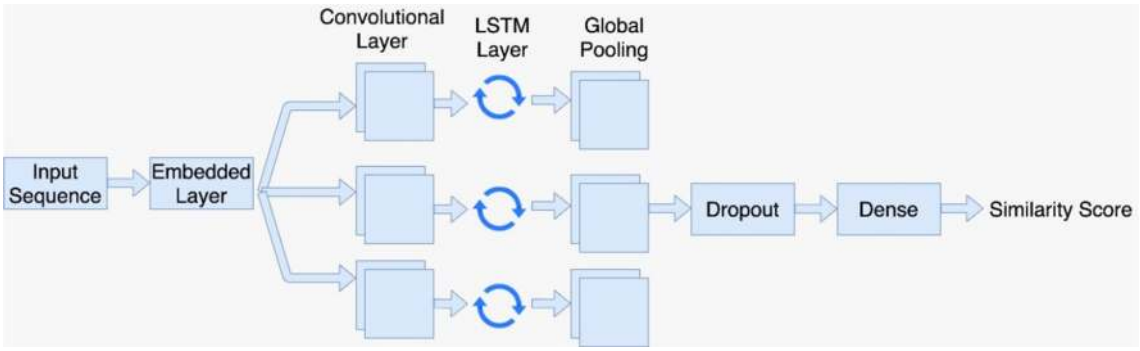

Fig. 2 Architecture of 'GenomeSimilarityPredictor'

cally updated weights of neurons. Higher values of weights favor the faster learning of the model.

2. *Beta 1 (β_1) and Beta 2 (β_2)* These are the exponential decay rates. The authors pre-set the optimum value of β_1 as 0.9 and β_2 as 0.999 based on the set of experiments conducted. The decrease in the values of β_1 and β_2 below 0.5 and above 1.2 yields high values of the loss function and hence 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 decreased from 0.9 to 0.5 and increased from 0.9 to 1.2.
3. *Epsilon (ε)* It is a small number to prevent the division by zero error. In this model, the authors used the default value as 1e−0.8 as discussed in [10].
4. *Binary Cross-Entropy (BCE) loss* It is effective in dealing with binary classification. It measures the performance of a classification model which yields a 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 $H_p(q)$ 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.

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

3.3 Activation function

The activation function is applied to introduce 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' input values. For example, x is the input given to the ReLU function then it returns x for $x > 0$ and return '0' for $x \leq 0$ [11]. The authors adopted ReLU activation function to randomly activate the neurons rather than activating all the neurons at the same time. It also prevents the problem of gradient saturation and leads to faster convergence of the model than Tanh activation function. The expected value of the Tanh activation function is zero. It is useful in the quick learning of deep neural networks. 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.

3.4 Working of the model

The neural network models, viz., 'COVID-Screen-Net' and 'GenomeSimilarityPredictor' work as parallel networks of the proposed multi-modal. The 'COVID-Screen-Net' is invoked if the chest radiographs are given as input. It classifies the chest radiographs into COVID-19, bacterial pneumonia, and healthy, whereas, another neural network the 'GenomeSimilarityPredictor' is invoked for the genomic sequences. Initially, three convolutional layers of this 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. Figure 3 shows the sequence of steps followed by the model. The 'GenomeSimilarityPredictor' also detects the genomes of the 'SARS-CoV-2' in the human genome.

The publicly available datasets [31–35] containing the genomic sequences are 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

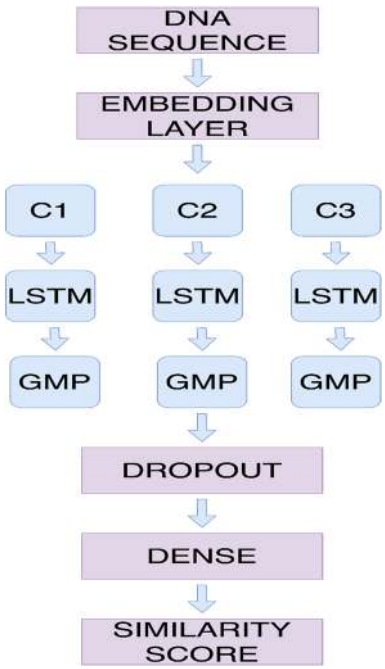

Fig. 3 Working of 'GenomeSimilarityPredictor'

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'.

4 Results and discussion

The authors use Google Colab [36], 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 h.

4.1 Dataset

For experiments, the authors used the datasets publicly available at [25, 26]. It contains 300 chest radiographs comprising of 100 images of non-infected persons, 100 images of patients infected with 'COVID-19', and 100 images of patients infected with 'Bacterial Pneumonia'. The authors used 75% of the total dataset for training and 25% for testing the performance of the model. The model is trained on 202 images, validated on 35 images, and tested on a set of 63 images. The authors used the batch size of 32 images. Therefore, they input eight batches of the dataset for training the model and two batches for testing the model. All batches are given as input in each epoch during training and testing the model. The authors selected the above-mentioned batch size based on the set of experiments performed. This batch size facilitates the model to learn gradually about the dataset without making it familiar with the whole dataset. Thus, it reduces the problem of generalization. Another dataset available at [31] 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 people from different geolocations such as China, Spain, the USA, Vietnam, and Thailand. As a part of the pre-processing step, the authors assigned unique numbers 0, 1, 2, and 3 to the nitrogenous bases: A, G, C, and T respectively. They sliced a genomic sequence at the interval of the length 300. The authors divided the dataset of 852 genomic sequences into a batch size of 128. They obtained seven 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. The authors also used 172 genomic sequences of 'SARS-CoV' [32] 'MERS-CoV' [33], 'HIV' [34] and 'Bat SARS-like virus' [35] 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 70% 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 of 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'.

4.2 Performance of the model

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

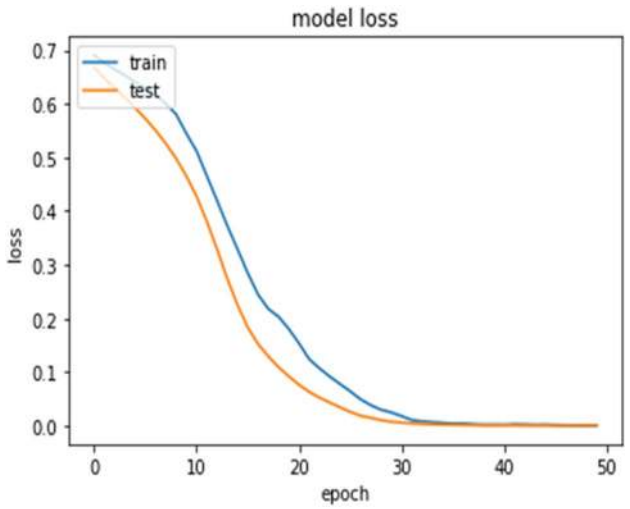

Fig. 4 Variation in BCE loss with the number of epochs

1. *BCE loss* The value of the loss function indicates the error in the prediction [11]. The value decreases with the training of the model. Figure 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.
2. *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', and Table 2 shows the confusion matrix obtained by applying 'COVID-Screen-Net' on the test dataset of 63 chest radiographs. In Table 1, 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. In Table 2, the class labels have been used to demonstrate the correctly and incorrectly classified instances.
3. *Accuracy of 'GenomeSimilarityPredictor'* 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%**

Figure 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.

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ size\ of\ test\ dataset} \tag{2}$$

Accuracy of 'COVID-Screen-Net' Its accuracy randomly varies when it is executed from the 0th to 120th epoch. Fig. 6 demonstrates the trends of the accuracy of the model. It is clear from Fig. 6 that the accuracy sharply increases in executing the model from 0th to 10th epochs. From the 10th to 118th epoch, there is a random increase and decrease in the accuracy. At the 118th epoch, the model achieves a maximum **accuracy of 95.47%.** The trends of increase and decrease in accuracy, give the information that the model gradually learns about the dataset and the training parameters. It completes its training in 118 epochs.

The accuracy of the model is also dependent on the batch size. It achieves an **accuracy of 94%** when a **batch size of 16 images** is given as input. On the contrary, the accuracy decreases and becomes **90.5%** when the **batch**

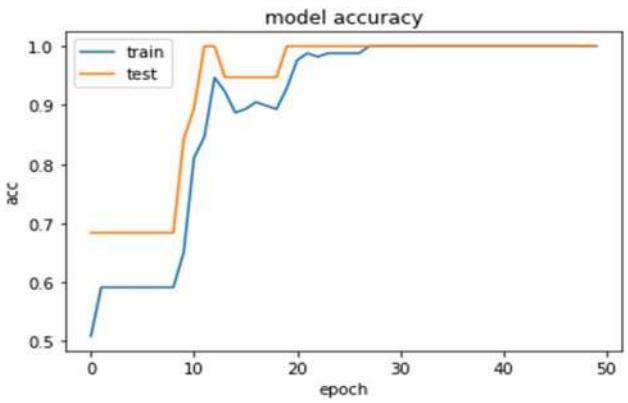

Fig. 5 Variation in accuracy of 'GenomeSimilarityPredictor' with number of epochs

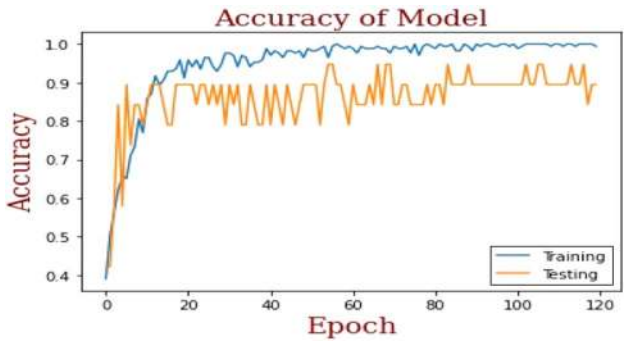

Fig. 6 Trends of accuracy of 'COVID-Screen-Net' with change in the number of epochs

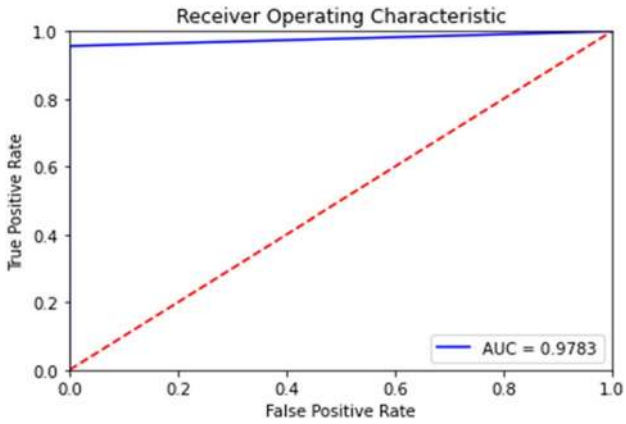

Fig. 7 ROC and AUC

Table 1 Confusion matrix 'GenomeSimilarityPredictor'

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

size is increased to 128 images. The maximum **accuracy of 95.47%** is achieved for the optimum batch size of 32 images.

4. *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. 7, 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.
5. *Precision (P)* It is the ratio of the number of correct predictions in a particular class to that of the total number of correct predictions made in all the classes. The model 'GenomeSimilarityPredictor' gives a precision of 99%. Whereas, the model 'COVID-Screen-Net' achieves the average precision of 95.4649% on the test dataset. The high values of precision prove that the models are effective in extracting the relevant instances of each class label from the total number of extracted instances.
6. *Recall (R)* It is the ratio of the number of correct predictions to a particular class to that of the total number of predictions made in that class. 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. Whereas, the model 'COVID-Screen-Net' gives an average **recall of 95.2381%**. The high values of recall prove that the model extracts all the relevant instances from the given instances. It makes the correct predictions for the 'COVID-19'.
7. *F1 score* This is the harmonic mean of precision and recall. The formula for calculating the value of the F1 score is given in Eq. (3).

$$F1\ \text{Score} = 2 \times \frac{P \times R}{P + R}. \tag{3}$$

Table 2 Confusion matrix 'COVID-Screen-Net'

		Predicted		
		Non-Infected	**Bacterial Pneumonia**	**COVID-19**
Actual	**Non-Infected**	20	0	0
	Bacterial Pneumonia	1	19	0
	COVID-19	0	2	21

The 'GenomeSimilarityPredictor' achieves the highest **F1 score of 100%** which proves the efficacy of the model in classifying the genomes of 'SARS-CoV-2' and Homo sapiens. The model 'COVID-Screen-Net' yields an average **F1 score of 95.2434 %** which prove, its efficacy in classifying the test images correctly into their actual classes.

8. *Degree of Genomic 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 3 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 3 that 'SARS-CoV-2' shows the highest genomic similarity of 98.11% with the 'SARS-CoV'.

9. *K-Fold Cross Validation* To validate the reliability of the 'GenomeSimilarityPredictor', the authors applied the tenfold 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 tenfold cross-validation minimizes the problem of overfitting and under fitting of the model.

The trend in the value of the loss function at each fold is shown in Fig 8. The values obtained at different iterations of the tenfold cross-validation are shown in Table 4. The model gives 0.005904 as the average value of the 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.

Table 3 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

5 Discussion

The state-of-art models as discussed in Sect. 2 [15–23] have been applied for the detection of viral genome in the host cell. Among all the machine learning models, the bi-path convolutional neural network model proposed in [22] achieved the highest accuracy of **97.05%.** However, the model is efficient 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. Further, the deep learning model proposed in [23] reported the higher accuracy than machine learning models and achieved an accuracy of 98.17% for the classification of 'SARS-CoV-2' from the given genomic sequences. Although, the model proves its efficacy in classification. But, it does not focus on finding the similarity among different genomic sequences.

Furthermore, the research works proposed in [24–28] extended the applications of deep learning models for screening of COVID-19 using chest radiographs. But, they reported the highest **accuracy of 87.02%** for multi-class classification of chest radiographs into non-infected, infected with COVID-19 and bacterial pneumonia [24]. But, these

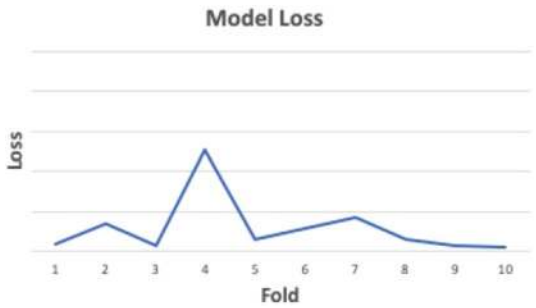

Fig. 8 Trend in the loss function of 'GenomeSimilarityPredictor'

Table 4 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

models employed the deep neural networks even though the dataset is small. Therefore, these encountered the problem of overfitting. Also, the biased training of models reduces their reliability.

Further, the use of multiple modalities viz. chest radiographs and genomic sequences together for screening of COVID-19 is an unaddressed research problem.

The multi-modal comprising two parallel networks viz. 'COVID-Screen-Net' and 'GenomeSimilarityPredictor' proposed in this manuscript overcomes the above-stated challenges. The 'COVID-Screen-Net' performs the multi-class classification with an accuracy of 95.47%. It is effective in automatic screening of COVID-19. The authors used an equal number of images for each class to avoid the problem of biased training. They used 100 X-ray images of each category namely COVID-19 confirmed cases, non-infected, and bacterial pneumonia. Moreover, the authors restricted the number of convolution layers to five to avoid the problem of overfitting on the small dataset.

Another model the 'GenomeSimilarityPredictor' 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 under fitting on the dataset used for training and testing. The optimized combination of CNN and LSTM in 'GenomeSimilarityPredictor' improved the accuracy reported in [23] 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. Besides, it finds the similarity score of a given genomic sequence with other genomic sequences. Its comparison with the techniques proposed in [15, 21–23] as shown in Fig. 9 proves

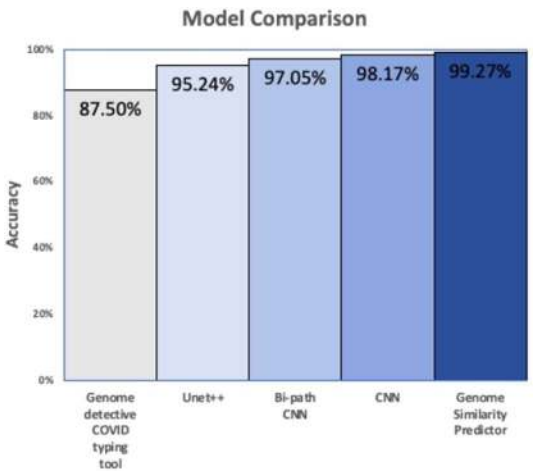

Fig. 9 Performance comparison of 'GenomeSimilarityPredictor'

the supremacy of the proposed model over the existing models. But, the performance of the model may degrade if the genomic sequences have been damaged during extraction. The authors trained the proposed model with different mutants of the same virus to improve its robustness.

6 Conclusion and future work

Screening and treatment of 'COVID-19' have become the prime objective of the globe. 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 multi-modal comprising the 'GenomeSimilarityPredictor' and 'COVID-Screen-Net' is effective in the screening of COVID-19 using multiple modalities viz chest radiographs and genomic sequences. The model 'GenomeSimilarityPredictor' 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 3 demonstrate 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'.

On the other hand, the model 'COVID-Screen-Net' distinguishes the X-ray images of non-infected, patients infected with 'Bacterial Pneumonia' and patients infected with 'COVID-19' with an accuracy of **95.47%.**

The comparison with the prior art shows that there is a lack of research works that are efficient in the screening of COVID-19 using multiple modalities. Also, both the models employed in the multi-modal proposed in this manuscript report higher accuracy than the models proposed in the literature [15–24]. Its effectiveness in dealing with noisy data, low time complexity makes it applicable for the screening of infected genomes as well as chest radiographs. 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 doctors.

Future scope: The application of this model can be extended for the screening of multiple classes of lung infections. It can also be used to quickly find the degree of similarity between genomes of any two microbes. The model can be trained with the labeled genomic sequences and chest radiographs of infected and healthy human beings for mass screening of patients. It can quickly detect the mutation in

the human genome. Thus, it may prove useful in the situation when the virus has a high tendency to generate its mutants.

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.

Declarations

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 9 Aug 2020
3. Koh, G.C.H., Hoenig, H.: How should the rehabilitation community prepare for 2019-nCoV. Arch. Phys. Med. Rehabil. **101**(6), 1068–1071 (2020). 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. Long, C., Xu, H., Shen, Q., Zhang, X., Fan, B., Wang, C., Li, H.: Diagnosis of the Coronavirus disease (COVID-19): rRT-PCR or CT? Eur. J. Radiol. **126**, 8961 (2020)
6. Ushmani, A.: Machine learning pattern matching. https://doi.org/10.13140/RG.2.2.16276.96649 (2019)
7. Robson, B.: 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. Comput. Biol. Med. **119**, 103670 (2020)
8. Jamshidi, M.B., et al.: Artificial intelligence and COVID-19: deep learning approaches for diagnosis and treatment. IEEE Access **8**, 109581–109595 (2020). https://doi.org/10.1109/ACCESS.2020.3001973
9. Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G., Cai, J., Chen, T.: Recent advances in convolutional neural networks. Pattern Recognition. rXiv:1512.07108v6 (2018)
10. Ruder, S.: An overview of gradient descent optimization algorithms. arXiv preprint arXiv preprint arXiv:1609.04747 (2016)
11. Pradhan, N., Dhaka, V.S., Rani, G., Choudhary, H.: Transforming view of medical images using deep learning. Neural. Comput. & Appli. (2019). https://doi.org/10.1007/s00521-020-04857-z
12. Le, H., Ho, H.V., Jung, S.: Application of long short-term memory (LSTM) neural network for flood forecasting. Water **11**(7), 1387 (2019). https://doi.org/10.3390/w11071387
13. Kumar, R., Nagpal, S., Kaushik, S., Mendiratta, S.: COVID-19 diagnostic approaches: different roads to the same destination. Virusdisease **31**(2), 97–105 (2020)
14. Skewes-Cox, P., Sharpton, T.J., Pollard, K.S., DeRisi, J.L.: Profile hidden Markov models for the detection of viruses within metagenomic sequence data. PLoS ONE **9**(8), e105067 (2014). https://doi.org/10.1371/journal.pone.0105067
15. Cleemput, S., Dumon, W., Fonseca, V., et al.: Genome Detective Coronavirus Typing Tool for rapid identification and characterization of novel coronavirus genomes. Bioinformatics **36**(11), 3552–3555 (2020). https://doi.org/10.1093/bioinformatics/btaa145
16. Bzhalava, Z., Tampuu, A., Bała, P., Vicente, R., Dillner, J.: Machine Learning for detection of viral sequences in human metagenomic datasets. BMC Bioinform. **19**(1), 1–11 (2018)
17. Abdelkareem, A.O., Khalil, M.I., Elbehery, A.H.A., Abbas, H.M.: Viral sequence identification in metagenomes using natural language processing techniques. 1–13(2020). bioRxiv 2020.01.10.892158
18. Kelley, D.R., Snoek, J., Rinn, J.L.: Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. Genome Res. **26**(7), 990–999 (2016). https://doi.org/10.1101/gr.200535.115
19. Ren, J., Ahlgren, N.A., Lu, Y.Y., Fuhrman, J.A., Sun, F.: VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. Microbiome. **5**(1), 69 (2017). https://doi.org/10.1186/s40168-017-0283-5
20. Tampuu, A., Bzhalava, Z., Dillner, J., Vicente, R.: ViraMiner: Deep learning on raw DNA sequences for identifying viral genomes in human samples. PLoS ONE **14**(9), e0222271 (2019). https://doi.org/10.1371/journal.pone.0222271
21. Ren, J., Song, K., Deng, C., Ahlgren, N.A., Fuhrman, J.A., Li, Y, Xie, X., Sun, F.: Identifying viruses from metagenomic data by deep learning. arXiv preprint arXiv:1806.07810 (2018)
22. Lopez-Rincon, A., Tonda, A., Mendoza-Maldonado, L., et al.: Accurate Identification of SARS-CoV-2 from viral genome sequences using deep learning. bioRxiv (2020). https://doi.org/10.1101/2020.03.13.990242
23. Chen, U., Wu, L., Zhang, J., et al.: Deep learning-based model for detecting 2019 novel coronavirus pneumonia on high-resolution computed tomography: a prospective study. medRxiv (2019). https://doi.org/10.1101/2020.02.25.20021568
24. Ozturk, T., Talo, M., Yildirim, E.A., Baloglu, U.B., Yildirim, O., Acharya, U.R.: Automated detection of COVID-19 cases using deep neural networks with X-ray images. Comput. Biol. Med. **121**, 103792 (2020)
25. Cohen. Covid chest x-ray dataset. https://github.com/ieee8023/covid-chestxray-dataset, 2020. Accessed 3 April 2020
26. Mooney." Kaggle chest x-ray images (pneumonia) dataset. https://github.com/ieee8023/covid-chestxray-dataset, 2020. Accessed 3 April 2020
27. Hemdan, E. E. D., Shouman, M. A., Karar, M. E.: Covidx-net: A framework of deep learning classifiers to diagnose covid-19 in x-ray images. arXiv preprint arXiv:2003.11055 (2020)
28. Mahmud, T., Rahman, M.A., Fattah, S.A.: CovXNet: A multi-dilation convolutional neural network for automatic COVID-19 and other pneumonia detection from chest X-ray images with transferable multi-receptive feature optimization. Comput. Biol. Med. **122**, 869 (2020)
29. Dhaka, V., Rani, G., et al.: A deep learning model for mass screening of COVID-19. Int. J. Imaging Technol. (2021). https://doi.org/10.1002/ima.22544
30. Creswell, A., Arulkumaran, K., Bharath, A.A.: On denoising autoencoders trained to minimize binary cross-entropy. arXiv preprint arXiv:1708.08487 (2017)
31. National Center for Biotechnology Information: 2020. https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide&VirusLineage_ss=Severe%20acute%20respiratory%20syn

drome-related%20coronavirus,%20taxid:694009. Accessed 26 Mar 2020
32. 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 Mar 2020
33. 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 Mar 2020
34. 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 Mar 2020
35. National Center for Biotechnology Information: 2020. https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide&VirusLineage_ss=Bat%20SARS-like%20coronavirus,%20taxid:1508227. Accessed 26 Mar 2020
36. Carneiro, T., Da Nóbrega, R.V., Nepomuceno, T., Bian, G.B., de Albuquerque, V.H., Reboucas Filho, P.P.: Performance analysis of google collaboratory as a tool for accelerating deep learning applications. IEEE Access. **6**, 61677–61685 (2018)