A Hybrid Model for Prediction and Progression of COVID-19 Using Clinical Text Data and Chest X-rays

Swetha V. Devan and K. S. Lakshmi

Abstract COVID-19 is an infectious disease caused by a virus known as novel corona virus or severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Since December 2019, the world is dealing with a pandemic as a result of this sickness. In addition to the medical field, technologies such as deep learning and machine learning aid in the fight against COVID-19. With the use of image or textual data, these technologies can anticipate the existence of disease. The suggested method is a hybrid model that predicts COVID-19 along with illness development using both machine learning and deep learning. Patients' entire medical records (clinical text data) and chest X-rays are regarded significant data for the suggested method. Because this disease mostly affects the respiratory system, X-rays of the chest are utilized to determine how far the sickness has progressed. A logistic regression model is trained with the clinical text data to classify them as COVID or not. Since the classification is binary, logistic regression is an efficient and easy to implement method. Then a VGG-16 model, which is considered as one of the best vision models, is trained by using several chest X-Rays. The trained model is then used to predict the patient's status and the progression of disease. A GUI is also developed for a user-friendly experience so that user can directly input the data. The overall output of the proposed model includes the COVID-19 status, the percentage of progression and the masked image of chest X-ray. The combined classifications using clinical data and chest X-Ray improve the effective disease progression detection of COVID-19.

Keywords COVID-19 · Machine learning · Deep learning · Clinical text · Chest X-Rays

S. V. Devan (✉) · K. S. Lakshmi
Department of Information Technology, Rajagiri School of Engineering and Technology, Ernakulam, Kerala, India

K. S. Lakshmi
e-mail: lakshmiks@rajagiritech.edu.in

S. Smys et al. (eds.), *Inventive Computation and Information Technologies*, Lecture Notes in Networks and Systems 336, https://doi.org/10.1007/978-981-16-6723-7_47

1 Introduction

In December 2019, the first case of COVID-19 was reported in China. Since then, the disease has posed numerous hazards and challenges. According to WHO, there are about 1.63 billion reported cases till May 2021 and about 3.3 million deaths were reported. This disease affects various persons in different ways. Some people may not need hospitalization to recover and experience only mild symptoms. Others, on the other hand, have suffered from serious health problems. COVID-19 is characterized by a dry cough, fever and lethargy. In addition, the majority of persons have had mild to moderate respiratory disease [1]. The virus that causes COVID-19 can be detected using the real-time reverse transcription-polymerase chain reaction (RT-PCR) test which is the most preferred method of illness detection. But there are several false negative and false positive results for RT-PCR. It also takes 2–4 h to get the results. Chest X-rays (CXRs) and computed tomography (CT) scans can also be used to screen for COVID-19 infection and evaluate disease progression in hospital admitted cases. Despite the fact that they are not officially suggested [2] as primary diagnostic procedures, they can be utilized as a secondary test to confirm the existence of disease and track disease development.

The application of artificial intelligence to medical diagnostics has a number of benefits for the evolution of the healthcare business. Artificial intelligence-based software [3] can tell if a patient is sick even before symptoms appear. Software's like this can assist doctors in making decisions. These systems work with digital data like texts or photos to deliver results in seconds. Among the textual information is clinical text, which covers the patient's medical data, comprises patient history and assessments, as well as a wealth of information for clinical decision making. Image includes CT scans or X-Rays. Although CT has a higher sensitivity [4] for pulmonary illness, it comes with several drawbacks. It includes the difficulties of sanitizing the room and equipment after each patient, as well as the time it will take to do so after each person. On the other hand, X-Rays of the patients can be collected with less complications since the sanitization of equipment for collecting X-Ray is not a big deal when compared to CT.

The proposed model is a system which uses both machine learning and deep learning technologies for predicting COVID-19 and its progression. The system receives text data as well as chest X-Rays as input and outputs. The logistic regression technique is utilized to handle text data, while the VGG-16 architecture is used to process CXRs. A logistic regression model can effectively handle the binary classification problems. In this case, the logistic regression model will analyse the clinical data and will return yes or no based on the findings. The VGG-16 model is used to detect COVID-19 patients in the case of CXR. The probability value used to calculate the progression is returned by VGG16. The CXR may contain the whole thoracic region and our region of interest is just the lung areas. So, to segment out the lung areas, the lung masks are used. Lung masks are generated by using a pre-trained UNet model. The developed masks are then merged with the X-rays before they are introduced into VGG-16 model. The masked X-rays highlight the lung areas. The

proposed model intents to be a helping hand to medical practitioners to confirm the COVID-19 case and to show how much it has progressed within the lung.

The rest of the sections are organized as follows. Section 2 will give the brief idea about past related works. Section 3 describes about the architecture and techniques involved in the proposed methodology. Then, Sect. 4 deals with results and discussion. Finally, Sect. 5 will conclude the study.

2 Related Works

The literature review reveals more about the previous works in detecting COVID-19 which uses deep learning or machine learning algorithms.

The author of [4] proposed a deep learning model for detecting pulmonary manifestations of COVID-19 with chest X-rays. This method includes a convolutional neural network (CNN) and several pre-trained ImageNet model. The model is trained to learn the features while using each ImageNet with the CNN. The learned knowledge is then used to find the relative performance and the best performing models are iteratively pruned. Author collected several CXR datasets and segmentation is performed at the pre-processing stage. The process of segmenting separates the region of interest from the rest of the picture. In [4] for segmenting the lung regions, the author used a pre-trained UNet model. The size of the dataset and its inherent uncertainty, as well as the computing resources required for effective implementation and usage, are two major determinants of this approach's performance.

In [5], the author explains a neural network architecture that can be trained with a small amount of data while still producing radiologically interpretable results in finding COVID-19. The author has used FC-DenseNet103 to segment the lung areas from the Chest X-Rays. Then a patch-by-patch training approach is used for classifying the CXR as normal, bacterial pneumonia, tuberculosis (TB) or viral pneumonia which includes the pneumonia caused by COVID-19 infection. The segmented images were cropped randomly with a size of 224×224 in the classification network, and the resulting patches were used as network inputs. The centres of patches were randomly selected within the lung areas to avoid cropping the patch from the empty area of the segmented image. Several patches were randomly acquired during the inference, with the number of patches chosen to cover all lung pixels several times, allowing each image to represent the entire attribute of the entire image. The patches were then fed into the network to generate the required output. The classification algorithm's backbone is ResNet-18 model. When a model is overly complex for a limited set of data, overfitting may occur. The ResNet architecture would aid in the prevention of overfitting. The performance of this model slightly affected when reducing the patch size, and there was also no benefit with increasing the patch size. So, the patch size must be maintained as 224×224.

The author of [6] proposed a weakly supervised deep learning framework to detect COVID-19 infected regions fully automatically using chest CT data acquired from multiple centres and multiple scanners. Based on the CT radiological features, the

disease classifies COVID-19 cases from community-acquired pneumonia (CAP) and non-pneumonia (NP) scans using the developed deep neural networks. The author used TCIA dataset for the proposed model. They trained a multi-view UNet model for the segmentation task. For the classification, they developed a network which was inspired by VGG-16 architecture. In the architecture, configuration of CNN depth increased using small convolution filters stacked with non-linearity injected in between them. All convolution layers consisted of 3×3 kernels, batch normalization and rectified linear units. The proposed CNN was fully convolutional, consisting of five convolutional blocks. Then, a multi-scale learning scheme is adapted to cope with variations of the size and location of the lesions. To implement this, the intermediate CNN representations, i.e. feature maps, at third, fourth and fifth convolution layers were fed into the weakly supervised classification layers. A 1×1 convolution was applied to mapping the feature maps down to the class score maps. Though this model is not discriminative enough when it comes to separate the community-acquired pneumonia from COVID-19, this model can pinpoint the regions of inflammation or lesions within the lung effectively.

Babukarthik et al. [7] is about the COVID-19 detection from CT scans and CXRs. The author of [7] proposed a model which consist of CNN models such as VGG16, ResNet50, DenseNet121, InceptionResNetV2 and several machine learning methods. The CNNS are used for extracting features from the CXRs and CT scans. Then, COVID-19 is identified from the extracted features by using various machine learning algorithms and statistical modelling techniques. In the feature extraction phase, each model is implemented in a hierarchical fashion so that to ensure obtained features are finely refined. The classification logic of the proposed model uses several algorithms such as k-nearest neighbors (kNNs), support vector machine (SVM), Gaussian process (GP), random forest (RF), multilayer perceptron (NN) and Adaboost to process the features. Each algorithm is implemented for different purposes. This method is too complex since it incorporates multiple deep learning and machine learning algorithms.

Shamsi et al. [3] explains a model which helps in classifying the lungs as COVID-19 affected and healthy lungs (normal person) using CXR images. This method proposes an independent and continuous learning algorithm for generating a DCNN architecture spontaneously. The process includes the operations of partitioning DCNN into numerous weighted fully connected and meta-convolutional block. Each block possesses the operations like pooling, convolution, batch normalization, dropout, fully connection and activation operation. The genetic operations such as selection, crossover and mutation process are performed to evolve the population for DCNN architectures. The fitness value will be generated after these processes which will be the prediction results. The model is trained by using about 5000 CXRs. Due to storing and evaluating a huge amount of DCNN structure, GDCNN has high computation and space complexity.

3 Methodology

The proposed methodology consists of two phases such as text processing and image processing. The first phase is detecting COVID-19 from clinical text data using logistic regression model. Second phase is finding COVID-19 from the chest X-Rays and there by calculating the progression, by using pre-trained ImageNet model. The backbone of classification architecture is VGG-16 model. The system architecture is shown in Fig. 1. The proposed methodology consists of the following steps within the upcoming sections. Section 3.1 describes about the datasets that are used; Sect. 3.2 describes about the procedures involved in text data pre-processing. Section 3.3 is the machine learning classification; Sect. 3.4 describes about pre-processing of images; and finally, Sect. 3.5 is the section which deals with the classification of CXRs and progression detection.

3.1 Data Collection

The proposed system requires two types of data, clinical text and chest X-Rays. The clinical text data [8] can be collected from Kaggle repository. The available dataset included 1057 entries of details of patients having symptoms of COVID-19 or other viral diseases. The dataset consists of several attributes including the patient's gender, age, label which specifies the disease, other details of symptoms and tests of the patients.

For the image processing part, multiple CXR datasets were used which were also collected from Kaggle repository; this includes COVID-19 chest X-Ray Data and CoronaHack—Chest X-Ray dataset [9]. These datasets include chest X-Rays of normal people, COVID-19 patients and people with viral pneumonia. Then, lung segmentation from chest X-Ray dataset [10] is also used which consist of CXRs of

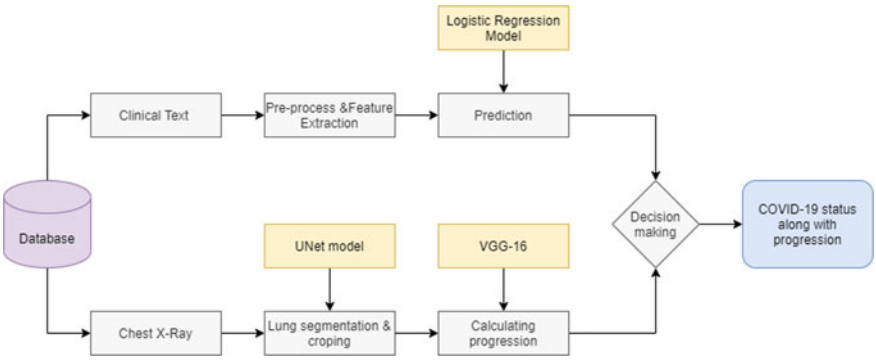

Fig. 1 Architecture of the proposed methodology

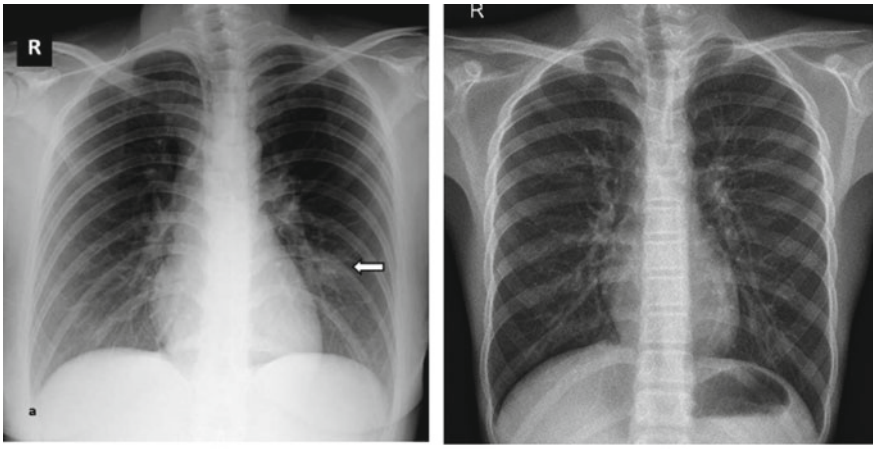

Fig. 2 Images from CXR dataset **a** COVID-19 X-Ray. **b** Normal X-Ray

normal people and CXRs showing pneumonia and non-pneumonic related abnormalities [11]. In Fig. 2a is the X-Ray of COVID patient and b is that of a normal patient.

3.2 Clinical Text Pre-processing

The first step is to prepare the clinical texts to perform machine learning algorithm [3]. The raw data were subjected for several pre-processing methods. Data cleaning is done to remove unwanted texts which includes special characters, white spaces, etc. Then, from the pre-processed data the features can be extracted. For feature extraction, the TF-IDF technique is used. The TF-IDF [12] stands for **term frequency-inverse document frequency**. The TF-IDF technique will identify the relevant features and will convert into the vectorized form. The TF-IDF technique helps in identifying how important a particular word or phrase is to a given document in the process of feature extraction. The basic working of this technique is based on two statistical concepts such as term frequency and inverse term frequency. The term frequency refers to the number of times a term t appears in the document. Then, inverse document frequency will measure the importance or relevance of a particular word in the overall document. TF-IDF value can be simply computed by multiplying both or can be found out by using Eq. (1).

$$W_{i,j} = tf_{i,j} \times \log \frac{N}{\mathrm{d}fi} \tag{1}$$

where $tf_{i,j}$ referred to the number of occurrence of term i in document j. df_i is the number of documents containing term i and N is the total number of documents. Here, each record will be considered as different documents. Then, by applying weight function to the extracted features, the vectorized input will be given into the logistic regression model.

3.3 Clinical Text Classification

For classifying the clinical data as COVID positive or negative, a classification algorithm is required. Since the data is text data, it can be classified by using a machine learning algorithm. Here, the classification is a binary classification, so the logistic regression [13] is one of the best candidates for this job. The supervised machine learning technique logistic regression can accurately predict the tags for a binary classification task. The algorithm will return 1 if the test case is of a COVID-19 patient or else it will return 0. The class membership probability can be calculated by the Eq. (2). In which P stands for probability which can have values from 0 to 1; a and b are independent variables, which will vary according with the extracted features.

$$p = \frac{e^{a+bx}}{1 + e^{a+bx}} \tag{2}$$

Primarily, our main focus is on the clinical notes section of the data set which consist of description about the character of patient regarding the symptoms or likely causes, etc. Basically, the primary observations about the patient is included in the clinical notes. Then, there is a column named findings which consist of class label. If the finding is COVID, it will be set as 1 and all other findings such that other viral diseases are set as 0. Logistic regression is used for classification [14]. In the training phase, the logistic regression model is set to train with the vectorized tokens and the findings which will be either 0 or 1 so that the machine can learn what to return when a text is given. In the testing phase, the pre-processed vectorized clinical texts can be classified by the trained model.

3.4 Chest X-Ray Pre-processing

The second phase of the proposed methodology is to identify the CXRs with COVID-19 and to compute the progression. Before implementing the classification algorithm, the raw X-Ray images need to be pre-processed [4]. The fully connected layers in convolutional neural networks, for example, demanded that all images be of the same size arrays. Image pre-processing can also speed up model inference and reduce

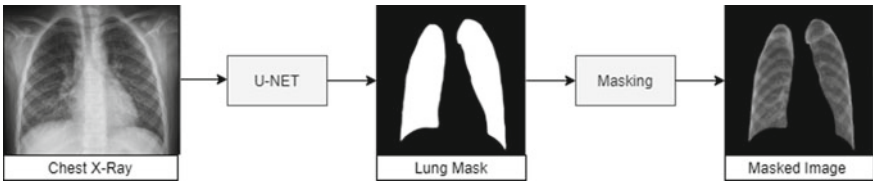

Fig. 3 The CXR segmentation

model training time. If the input photographs are very huge, shrinking them will significantly reduce model training time without compromising model performance.

One of the most important pre-processing is the image segmentation. This is highly applicable in case of using deep learning techniques in medical imaging processes. Separating the foreground from the background, or clustering pixels based on colour or shape similarity are examples of image segmentation. Here, image segmentation is performed on the CXR data sets to visualize the required regions of the data. For this purpose, lung masks will be generated. A pre-trained UNet model is used for segmenting lung areas. The dataset in [10] is a collection of CXRs along with lung masks, so this dataset is used to retrain the UNet architecture so that it could return masks of size 512 $\times$ 512 pixels. These masks are then placed on the CXR images to create a bounding box that contains the lung pixels. Figure 3 shows the mask generation process.

After segmentation the image, further pre-processing techniques includes pixel rescaling and edge preservation are done to maintain the picture quality. The masked images are given into the classification unit for further proceedings.

3.5 CXR Classification

The masked CXR images are fed into the classification network along with the label and progression. The classification network is a VGG-16 model. The VGG-16 [15] model is one of the best architectures in classifying images. There are only two class labels here, Yes and No. Yes, if COVID-19 is identified, else No. A data frame has been created which consist of the class label and progression for each image in training set. This data frame is used to train the network. The network takes images of size 256 $\times$ 256. So, the segmented CXR are reshaped into 256 $\times$ 256 pixels. The number of neurons in hidden layers is set in to 512. The classification logic is based on the two activation functions, rectified linear activation function (ReLU) and sigmoid function. ReLU activation is applied to the hidden layers, and Softmax activation is applied to the output layer. The ReLU activation will output the value if it is positive, and output zero if the value is negative. Whereas sigmoid activation function will always keep the values between 0 and 1. The sigmoid function can be calculated by the Eq. 3.

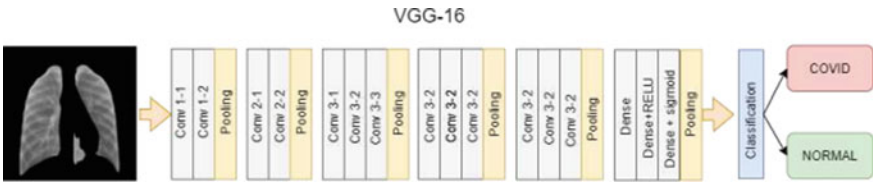

Fig. 4 Classification using VGG-16

$$f(x) = \frac{1}{1 + e^{-(x)}} \tag{3}$$

After setting the values and activation function, the model is trained with the segmented images and the model will be saved. The VGG-6 architecture is shown in Fig. 4. In the architecture, Maxpool layer of 2×2 filter of stride 2 and convolution layers of 3×3 filter with stride 1 and always used identical padding. Throughout the architecture, the convolution and max pool layers are arranged in the same way. It has two FC (completely connected layers) in the end, followed by a Softmax for output. The 16 in VGG16 alludes to the fact that it contains 16 layers with different weights. In the testing phase, the saved model can be used for classifying the CXR. If the CXR is identified as of a COVID-19 patient, then the percentage of progression will be returned.

4 Results and Discussion

As the final step, a graphical user interface has been built with which can join the two phases as well as a user-friendly approach. In the GUI, we can upload clinical texts and once the COVID is detected from text, we can upload CXR. By uploading these data there, we can see the steps by step processes. The output of the model includes the masked image, the two messages such that findings from both text and CXR, then the percentage value of progression which shows how severe is the case. Figures 5 and 6 show the screenshot from the GUI. The first page in the application is a login window where the user can enter userID and password. The login credentials will be checked in the background and if it is matched, the home page will appear. In the home page, there provided three buttons “upload image”, “Start Processing” and “Cancel”. The user has to upload the X-Ray of the patient and the clinical text data, which will be a clinical note consisting the details, symptoms etc. When the processing begins, the system will call the past modules and step by step process will be carried out. The steps will be listed in the window itself. After the processing, the output will be printed.

The proposed system is developed in a windows system having a 4 GB RAM and 2.20 GHz processor. Several tools and libraries are used throughout the system. The text processing phase is implemented with the help of the natural language toolkit

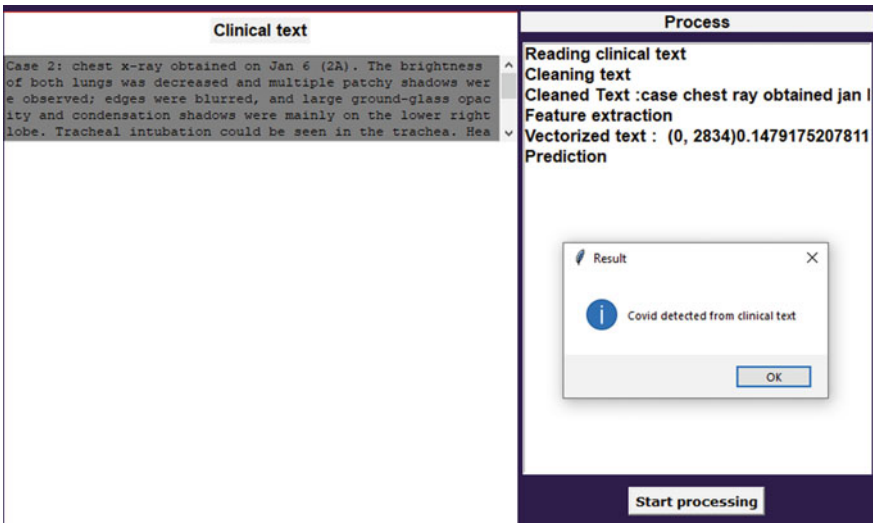

Fig. 5 Confirmation from clinical text

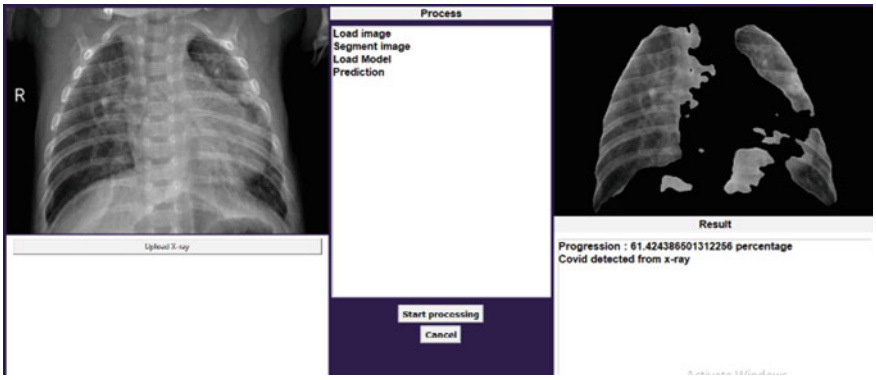

Fig. 6 Progression of COVID-19

(NLTK) which consists a group of libraries and packages that can help in processing natural language. In the second phase, Keras library functions are used which is a well-known deep learning framework. At the end, a GUI has been developed with the help of Tkinder toolkit. The required dataset includes both text and image data. The text data is split into a 70:30 ratio, with 70% of the data being used to train the model, whereas the remaining 30% being used to test the model. The clinical data for COVID-19 is less available. There is a probability for improving the performance if more data is available [16]. The system can be updated according to the availability of more data. Even though this system can aid support in analysing clinical data. For

Table 1 Accuracy

	Phase 1	Phase 2
Accuracy	91	97.7%

X-Ray classification, the text to train ratio is as 80:20, i.e. 80% of data were used for training and 20% being used for testing.

The accuracy of the model can be calculated by the Eq. (4). In the proposed model, we use two algorithms such as logistic regression and VGG-16; the accuracy in each case is given in the Table 1.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FN} + \text{FP}} \tag{4}$$

TP is the number of positive classes predicted as positive. TN is the number of negative classes predicted as negative. FN is the number of positive classes predicted as negative. FP is the number of negative classes predicted as positive.

The phase 1 of the model has been tested in two steps to determine its true accuracy. It employed 75% of the available data that were taken manually in the first stage, which results in lower accuracy than the stage where the entire data were used for experimentation. As a result, it can deduce that if more data is provided to these algorithms, performance may improve. The graph plotted with the obtained value is shown in Fig. 7.

Then in phase 2, the epoch value is given as 7 since the data set is images; in this case, the size of the data set is greater and the time taken for training is more. The least number of topically relevant images are used though; it is recognized that the training time and memory limits are required for practical deployment using computer resources. The model achieved 97.2% precision and $F1$ score is 97%. The graph plotted with the obtained value is shown in Fig. 8.

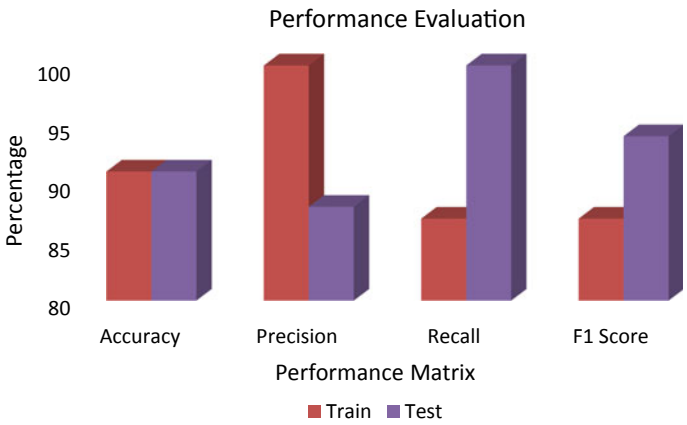

Fig. 7 Graph of phase 1

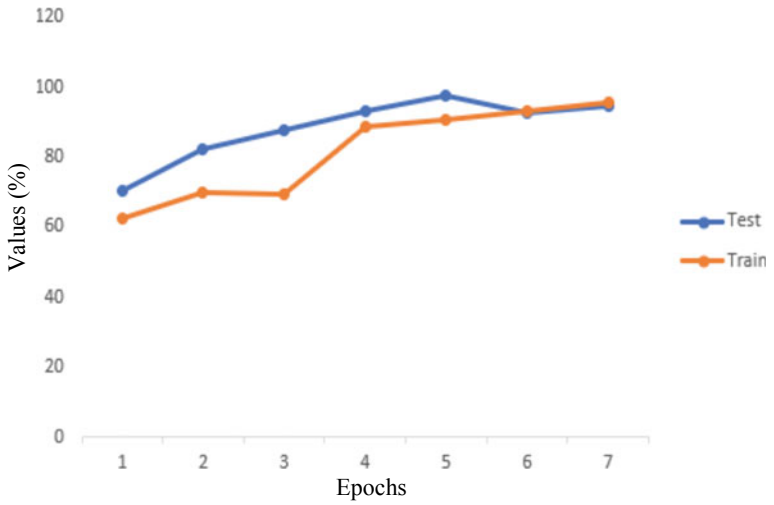

Fig. 8 Graph showing accuracy of phase 2

Table 2 Comparative study

	Type of data	Architectures used	Accuracy obtained (%)
Proposed model	• Clinical text • CXR	• Logistic regression • UNet • VGG-16	95
[4]	• CXR	• UNet • VGG-16 • VGG19 • InceptionV3	90.01
[5]	• CXR	• FC DenseNet103 • ResNet-18	91.9
[6]	• CT scans	• Multiview UNet • CNN	89.2
[7]	• CXR	• CNN	94.84

A comparative study can be done with other similar works Table 2 shows the comparative analysis of the proposed model with other works.

5 Conclusion and Future Scope

The COVID-19 disease had a negative impact on almost every industry. Not only the ones infected, but also all the people had to face a lot difficulties due to lockdown and all. Many researches are carrying out in defending against this disease. Here, a system is proposed to identify the disease and progression of COVID-19 disease. The system uses clinical texts as well as chest X-Rays for finding the disease. A dataset consists of about 1057 entries of clinical texts and a group of chest X-ray datasets which contain X-Rays of normal lungs and the lungs showing abnormalities such as pneumonic related or COVID-19-related abnormalities. These were used to train the proposed system. While testing the system effectively identifies and return the image as either normal or as COVID-19 along with the progression. Machine learning algorithm is applied for classifying clinical text data and deep learning algorithm in CXR processing. The system resulted in an overall accuracy of 95%. Increasing the quantity and quality of data can improve the efficiency of the model. The dataset for clinical text is a growing data repository, so the dataset can be updated eventually according with the availability of data. For CXRs, more X-Ray images can be included for training the VGG-16 architecture. In the future, we would like to add a gradient feature and heat maps so that we can pinpoint and visualize the disease-affected regions within the X-Ray.

References

1. World Health Organization Corona Virus Informations [Online]. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/
2. COVID-19 and imaging: an article on the limited role for CT and CXR in diagnosis of COVID-19 [Online]. Available: https://blog.radiology.virginia.edu/covid-19-and-imaging
3. A. Shamsi, H. Asgharnezhad, S.S. Jokandan, A. Khosravi, P.M. Kebria, D. Nahavandi, S. Nahavandi, D. Srinivasan, An uncertainty—aware transfer learning-based framework for COVID-19 diagnosis. IEEE Trans. Neural Networks Learn. Syst. **32** (2021)
4. S. Rajaraman, J. Siegelman, P.O. Alderson, L.S. Folio, L.R. Folio, S.K. Antani, Iteratively pruned deep learning ensembles for COVID-19 detection in chest X-rays. IEEE Access **8** (2020)
5. Y. Oh, S. Park, J. Chul Ye, Deep learning COVID-19 features on CXR using limited training data sets. IEEE Trans. Med. Imaging **39** (2020)
6. S. Hu, Y. Gao, Z. Niu, Y. Jiang, L. Li, X. Xiao, M. Wang, E.F. Fang, W. Menpes-Smith, J. Xia, H. Ye, G. Yang, Weakly supervised deep learning for COVID-19 infection detection and classification from CT image. IEEE Access **8** (2020)
7. R.G. Babukarthik, V. Ananth Krishna Adiga, G. Sambasivam, D. Chandramohan, J. Amudhavel, Prediction of COVID-19 using genetic deep learning convolutional neural network (GDCNN). IEEE Access (2020)
8. COVID-19 clinical text data [Online]. Available: https://www.kaggle.com/bachrr/covid-chest-xray/metadata.csv
9. COVID-19 X-ray dataset—corona hack dataset [Online]. Available: https://www.kaggle.com/praveengovi/coronahack-chest-xraydataset
10. Lung segmentation from chest X-ray dataset. Available: https://www.kaggle.com/nikhilpandey360/lung-segmentation-from-chest-x-ray-dataset

11. T. Zebin, S. Rezvy, COVID-19 detection and disease progression visualization: deep learning on chest X-rays for classification and coarse localization. Appl. Intell. **51**, 1010–1021 (2021)
12. Description for TF/IDF technique [Online]. available: https://www.geeksforgeeks.org/sklearn-feature-extraction-with-tf-idf/
13. S.S. Aljameel, I.U. Khan, N. Aslam, M. Aljabri, E.S. Alsulmi, Machine learning-based model to predict the disease severity and outcome in COVID-19 patients—5587188 2021/04/20
14. A.M.U.D. Khanday, S.T. Rabani, Q.R. Khan, N. Rouf, M.M.U. Din, Machine learning based approaches for detecting COVID-19 using clinical text data. Int. J. Inf. Technol. (2020)
15. VGG-16 architecture explanation [Online]. Available: https://towardsdatascience.com/step-by-step-vgg16-implementation-in-keras-for-beginners-a833c686ae6c
16. E. Sogancioglu, E. Çallı, B. Ginneken, K. Leeuwen, K. Murphy, Deep learning for chest X-ray analysis: a survey (2021)