

# Artificial Intelligence Based Transfer Learning Approach in Identifying and Detecting Covid-19 Virus from CT-Scan Images



Soubraylu Sivakumar, D. Haritha, Ratnavel Rajalakshmi, S. Shanmugan, and J. Nagaraj

**Abstract** A virus that influences the world since late 2019 is Coronavirus (Covid-19). The WHO has given preliminary guidelines to be followed by every individual in this pandemic situation. Even then many countries are facing a lot of consequences and human loss. Many countries tried to produce different vaccines in the market for this covid-19. But, none of these vaccines is 100% immune against this virus. Early detection of this virus from a victim will provide a faster recovery and good health. One such solution can be provided through the determination of covid-19 from Computerized Tomography (CT) scanned images. A strong and effective machine learning approach is needed in this pandemic situation to locate, stop, and control the spread of COVID-19. To extract the complex features and to provide a better classifier from these high-dimensional CT scan images an effective feature extractor and optimal classifier are required. A transfer learning

---

S. Sivakumar (✉)

Computing Technologies, SRM Institute of Science and Technology, Kattankulathur, Chennai, India

e-mail: [sivas.postbox@gmail.com](mailto:sivas.postbox@gmail.com)

D. Haritha

Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Guntur, Andhra Pradesh, India

e-mail: [haritha\\_donavalli@kluniversity.in](mailto:haritha_donavalli@kluniversity.in)

R. Rajalakshmi

School of Computing, Vellore Institute of Technology, Chennai, India

e-mail: [rajalakshmi.r@vit.ac.in](mailto:rajalakshmi.r@vit.ac.in)

S. Shanmugan

Research Centre for Solar Energy, Department of Engineering Physics, Koneru Lakshmaiah Education Foundation, Guntur, Andhra Pradesh, India

e-mail: [s.shanmugam1982@gmail.com](mailto:s.shanmugam1982@gmail.com)

J. Nagaraj

Computer Science and Engineering, Madanapalle Institute of Technology and Science, Madanapalle, Andhra Pradesh, India

e-mail: [nagrajan31@gmail.com](mailto:nagrajan31@gmail.com)

based machine learning method is proposed named as VGG16-SVM. At the top of this proposed model, a multiple stacked smaller size kernel named VGG16 is used to select the complex features that are hidden from the CT-Scan images. These reduced features are fed into a kernel-based Support Vector Machine (SVM) for analysis. Three distinct experiments are explored to determine the best model in the identification of Covid-19 which includes VGG16-SVM, VGG16-Random Forest and VGG16-XGBOOST. The experiments are conducted on 348 CT-Scan images. Image data augmentation strategies are employed to enhance the proposed model's accomplishment and prevent overfitting. The metrics used in our research work are training and validation accuracy. From the experimental analysis, we found that applying VGG16-SVM on CT scanned images obtained a best performance result in recognizing the Coronavirus.

**Keywords** VGG16 · Support Vector Machine · Deep convolutional neural network · Covid-19 · CT-Scan images · Transfer learning · Random Forest · XGBoost

## 1 Introduction

A novel virus named Covid-19 has out broken from china in December 2019 from Wuhan city. It has infected globally a large unit of populations regardless of their community, race and location. People affected with a disease like cancer, diabetes, chronic respiratory and cardiovascular are affected severely by this SARS-CoV-2. It badly strikes the respiratory canal, makes a lesion and affects the regular functionality of the lungs. The symptoms of this epidemic can be identified as tiredness, dry cough, respiratory illness, fever and loss of sensation of taste.

There are 591,683,619 COVID-19 proved cases around world. This epidemic caused 6,443,306 human losses in the world on 23 August 2022 [30]. To reduce human loss, more tests are to be performed per million populations to identify this disease and to know the spreading rate of the epidemic. Covid-19 infects and creates an impact on the respiratory organ of humans. So, it will be easy to detect Covid-19 rapidly through medical imaging of the chest. A Computer Tomography Scan of a Coronavirus victim is the simplest and easy method to quickly detect and automatically [16] diagnose Covid-19.

The motivation of this study is to automatically detect Covid-19 from patient CT scan images [11]. A collection of 348 CT-Scan images from a senior radiologist is used in this experiment. These scanned images are taken from 216 COVID-19 patients. A data augmentation method that magnifies a particular image from the dataset is added to optimise the model's performance. In the augmentation process, the 300 training and 100 validation images are expanded into 600 training and 200 validation images.

Two methods are analyzed in the various experiments to find an effective model for detecting Covid-19 namely, (i) Cross Validation and (ii) the Learning Curve

method. Both these methods are applied to three sets of experiments viz., (i) VGG16 with Support Vector Machine (VGG16-SVM), (ii) VGG16 with Random Forest (VGG16-RF) and (iii) VGG16 with eXtreme Gradient Boost (VGG16-XGBoost). In the folding method, each fold is represents an experiment. So, a total of 10 experiments are conducted in 10 fold method. In this method, VGG16-SVM have achieved a mean Accuracy of 87.75% compared with the other two models. In the learning curve method, the no. of training samples is increased to the multiple of 27. A total of 10 experiments are conducted with a training size of 270. The VGG16-SVM has obtained an Accuracy of 84.66% while compared with VGG16-RF and VGG16-XGBoost.

Each block of VGG16 is composed of 2 convolution layers with a receptive size of  $3 \times 3$  along with a ReLu activation and a max pooling layer of  $2 \times 2$ -pixel window. This neural network is composed of five such blocks in a series. This complex non-linear layer helps to extract a latent rich feature from the high dimensional space. These features are fed into the SVM. The quadratic programming of SVM used to perform iterative calculation on the input until a feasible solution is obtained. It scales well for sparse data like images, especially for binary classification problems. It has attained good results for both 10-fold cross-validation and linear curve method while compared with VGG16-RF or VGG16-XGBoost models. The contribution in this article comes in three folds.

1. The complex chronological features from the high dimension CT-Scan images are extracted with a sequence of multiple non-linear layers from VGG-16,
2. The rich extracted features from CT-Scan are gave into the SVM an iterative computation sequential optimizer classifier for classification and
3. Using a Transfer Learning (TL) approach, the deep CNN and a conventional classifier are integrated to provide a VGG16-SVM model to classify and detect Covid-19 patients.

The entire research work structure is as follows. The existing work limitations are discussed in Sect. 2. In Sect. 3, the proposed methodology, the feature selection method, the data augmentation method and the different image classifiers are discussed. In Sect. 4, the dataset, metrics, experimental setup and experimental results are explained. The experiment results are concluded with future enhancement in Sect. 5.

## 2 Related Work

Covid-19 [1] is asserted as a pandemic disease by WHO. The Covid-19 disease affects the lungs of humans and lung abnormality is the well-known diseases among all ages of human being. Due to lung infection, the respiratory system is also damaged. Image modeling detects the infection severity. The scanned images of the patient are helpful in detecting the covid-19 through a three phase model and the steps involved is as follows,

**Phase 1:** Data accession uses stagnant wavelets

**Phase 2:** CNN model is used for detection of covid-19

**Phase 3:** This phase uses CT scan images. It uses well known pre-trained architecture [18] of ResNet Series such as 18, 50 and 101 including SqueezeNet for assessment.

In this classification process, 70 and 30% of images are used in training and validation. The performance is calculated by computing the common performance measures. The ResNet18 uses TL approach is a pre-trained model used in evaluating the experiments.

For image analysis, the deep learning methodology [2] is greatly expanded. Similar to how deep learning applications (DL) to medical images of the respiratory system have emerged, these trials are displaying great potential. The development of DL applications for medical image analysis with a focus on lung image analysis is discussed in order to better understand the achievements to COVID-19. Deep learning tasks like segmentation, categorization, and recognition [19] as well as abnormal lung pathologies like respiratory illnesses, lung cancer, COVID-19, as well as other infections are emphasised in this survey that has 160 contributions. It also provides an overview of the current pandemic situation.

Before universal testing (CT imaging and PCR) prior to surgery of 72–120 h blood tests and anesthesiology related to imaging and surgical procedure respectively are scheduled [3, 4]. Of the 211 emblematic victims who have been tested ahead surgery, six trusted positive PCR findings. One Hundred Four patients who underwent elective surgery as part of the current study were healthy both before and after the procedure. Out of 336 patients who were cancelled, only 12% agreed to a new appointment time right away. This led to a 70% decrease in elective surgery and a 50% decrease in arthroplasties following lockdown. The post-COVID period did not see an increase in the complication rate. Patients struggle to comprehend the new guidelines founded by health organizations and have unclear ideas about screening.

In two orthopaedic surgery centres, 1397 patients underwent chest CT scans between March 1 and May 10, 2020, during the peak of the epidemic [3, 4]. We chose 118 patients out of 1397 for orthopaedic surgical treatment who displayed trauma symptoms. Thirty nine patients out of 118 were subjected to PCR testing to investigate COVID-19 infection. The Chest CT scan provides useful information for orthopaedic surgeons based on clinical status (symptomatic or non-symptomatic) and is graded from 0 (no value) to 3 (high value) (high value). Specific pathways and exposure to radiation were analyzed and discussed with existing possibilities. The evaluation of the increase in CT scanning during the COVID-19 pandemic led to the development of this result, which was based on prior treatment performed during the same time period (1st March 2018–15th April 2018). Total 118 patients are used for evaluation with 102 patients with negative and 16 patients with positive chest CT scan. Regarding the PCR results, the chest value is reported as having a sensitivity of 81%, a specificity of 93%, a positive predictive value of 86%, and a positive predictive value of chest CT ( $p = 0.001$ ). The following are the grades (1 for 71

cases, 2 for 5 cases, 3 for 11 cases). Only 2% of the CT scans performed during the pandemic period were for orthopaedic or trauma patients, which accounted for 20% of the total. This outcome was ten times greater when compared to the preceding control period.

It is difficult to identify Covid-19 while looking at clinical photographs of patients [5]. The TL method has been employed in the clinical imaging of a variety of lung conditions, which include covid-19 [10]. Similar condition that predated covid-19 is pneumonia lung disease. We must put forth a new model to stop the disease from spreading in order to anticipate COVID-19. We can learn that COVID-19 and viral pneumonia are the same thanks to the TL approach. Due to noise burden outside of tissues and lesions, it is difficult to see aberrant features in images. Texture characteristics are extracted employing haralick functionalities, which target solely on the area of attraction.

The deep learning method was trained using 120 chest CT scans with pulmonary infiltrates [6]. The regions and arteries of the lungs are segmented using this technique. Using 72 consecutive scans from 24 COVID-19 individuals, this method determines the existence and evolution of infiltrates connected to COVID-19. The components of this method are as follows: (1) Computerized lung borders and artery edge detection; (2) Lung frontiers validation among scans; (3) Digital recognition of the Digitized Pneumonitis Zones; and (4) Evaluation of Cancer Progression. The understanding among the regions that the radiologist manually outlined and the regions that the computer detected was evaluated using the Dice coefficient. A heat map illustrating the difference between scans was produced using serial scans that were enrolled and recorded. Using a five-point Likert scale, two radiologists evaluated the accuracy of the heat map used to indicate progression. They had a Dice coefficient of 81% (CI 76–86%), which indicated understanding among computer identification and human delineation of pneumonic regions. The algorithm had an 84% (CI 81–86%) specificity and a 95% (CI 94–97%) sensitivity. In order to detect significant pneumonia zones, radiologists assessed 95% (CI 72–99) of heat maps as at least “adequate” for portraying disease progression.

The world is anticipating what the novel coronavirus [7] can do to humanity because it has a number of distinctive properties. It is essential to identify and isolate the ill patients as soon as practical because a coronavirus-specific vaccination is not yet available. There aren't as many testing facilities and tools available as we had anticipated. This study paper discusses the use of machine learning techniques for gaining crucial insights. A lung CT scan should be the first obtainable test for real-time reverse transcriptase-polymerase chain reaction (RT-PCR). Patients with COVID-19 pneumonia seem to be difficult to differentiate from those who have other viral pneumonias on a global scale [9]. For training and testing, Microsoft Azure's custom vision software is built on machine learning methods. With the suggested method, COVID-19 classification accuracy is 91%.

Our goal [8] is to compare the diagnostic efficacy of a 30-mAs chest CT protocol with a 150-mAs standard-dose regular procedure using imaging of COVID-19 pneumonia. After the IRB approved the trial, COVID-19 participants 50 years of age or older were sent for chest CT scans and received normal CXRs the same

day. A standard dosage (150 mAs) chest CT scan was performed first. Low-dose CT (30 mAs) for COVID-19 identified individuals was carried out right away. The accuracy of CTs with low and high doses was distinguished. The 20 patients in this study had an average age of  $64.20 \pm 13.8$ . The results were: low-dose  $1.80 \pm 0.42$  mSv and standard-dose  $6.60 \pm 1.47$  mSv. Absolute cancer risk per mean cumulative effective dose values as follows standard-dose  $2.71 \times 10^{-4}$  and low-dose  $0.74 \times 10^{-4}$ .

## 2.1 Existing Limitations

A major virus that influences the world from the year 2020 is named Coronavirus (COVID-19). Identifying the Covid-19 patient immediately is a great challenge in these pandemic situations. The various approaches and restrictions for identifying Covid-19 employing CT-Scan images are presented in Table 1. Some of the inferences we come across in this literature review limitations are:

- (i) Not able to learn deep features,
- (ii) It causes overfitting and overemphasizes outliers.
- (iii) It needs feature scaling and
- (iv) It is time-consuming and requires more computational power.

To address all the above issues, a robust model is required to provide a better result. The proposed VGG16-SVM model addresses some of the limitations listed above and its performance is discussed in the below section.

**Table 1** List the existing methodology and limitations of identification of Covid-19 from CT-Scan images

S.No.	Author	Methodology	Limitations
1.	M. Loey et al. [32]	ALexnet	It struggles to scan for all features and its performance is poor in most of the use cases
2.	V. Perumal et al. [5]	Inception V3	It is time-consuming and requires more computational power
3.	G. Muscas et al. [25]	Gradient boosting machine	It causes overfitting and overemphasize for outliers
4.	Xiangjun Wu et al. [26]	Multi-View deep fusion model	It is time-consuming and requires more computational power
5.	H. Yasar and Ceylan [29]	Local Binary Pattern-KNN 10 fold cross validation	Does not scale well with high dimensions and needs feature scaling
6.	V. Shah et al. [33]	CTnet-10	Not able to learn deep features

### 3 Materials and Methods

The methodology used, image pre-processing method, data augmentation method, feature weighting method, proposed architecture, machine learning methods and multivariate analysis are discussed in this section.

#### 3.1 Methodology

The CT scan images from the datasets are applied to this deep neural network design by Ganesan et al. [12]. There are 348 Covid and 348 Non-Covid CT Scan pictures in the original dataset. In order to increase the quantity of photos and prevent the results from being overfit, a data augmentation approach is used in this process. The quantity of photographs in each category has increased to 52 after the augmentation process (i.e., 400 images). Three categories of augmented images—training, validation, and testing—are used. The training, validation, and testing sets each comprise 200, 100, and 100 photos after the images have been divided. With the help of the VGG-16 deep neural network, the characteristics are extracted. It is a 16 layer neural network that assists in extracting significant essential features from the Covid and Non-Covid CT-Scan dataset [15].

The SVM, RF, and XG Boost are three machine learning models whose hyperparameters are selected using the grid search approach. The classifiers’ ideal parameters are set with the use of grid search. In order to forecast the testing set, many classifiers are learned. To forecast the model’s performance outcome, two techniques are used. When predicting CT-Scan images using various classifiers, K-Fold cross-validation and learning curve approach are utilised. The effectiveness of the various classifiers is evaluated using accuracy as a criterion. Figure 1 displays the process used to predict Covid CT-Scan results.

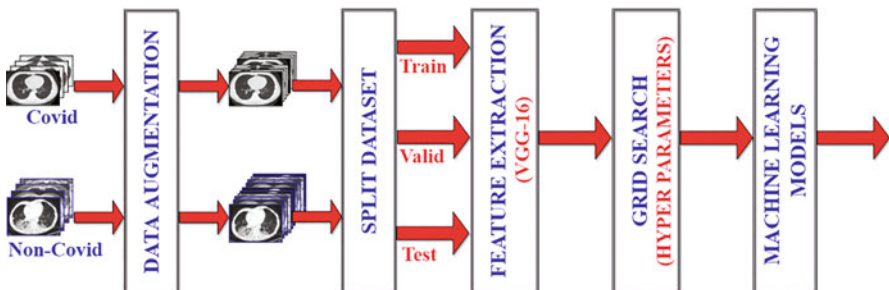


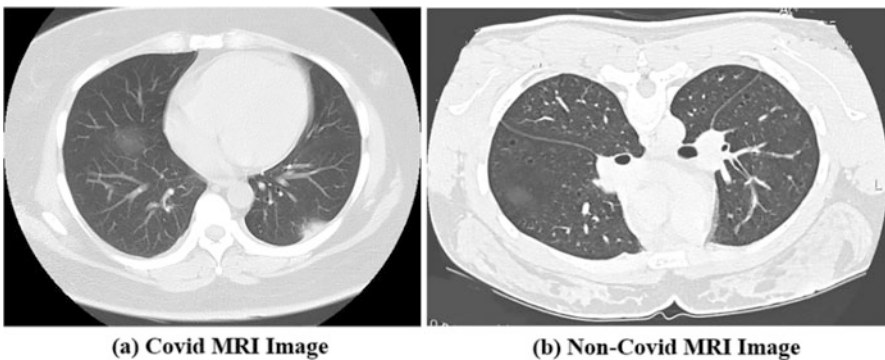
Fig. 1 Methodology used in the prediction of Covid CT-Scan images

### 3.2 Preparing Dataset

The *path* variable holds the folder path of all the files to be renamed. The function *listdir()* returns the list of all files, while *enumerate()* function helps to iterate over each file in the folder. In every loop iteration, the *count* takes a value from 0 to  $n - 1$ , where  $n$  is the number of files in the folder. The *dst* variable holds the file name as *covid0.png* and it helps to rename the file. The source and destination path names of the file are stored in the variables *src* and *dst*. The function *rename()* helps to rename the file from source name to destination name. Base directory and train folder are combined with the *join()* method and stored in *train\_dir* variable. The function *mkdir()* creates a new folder for training images to be placed. In similar way, a Covid and Non-Covid folder will be created inside the *train* folder. The same operations are continued for creating *validation* and a *test* folder for storing the images. *Shutil* is a library that copies a file from the original folder to the newly created folder. The *src* represents the source of the file to be copied and the *dst* represents the destination of the file to be placed.

### 3.3 Data Augmentation

A sample instance of Covid and Non-Covid CT-Scan image is shown in Fig. 2. There are 348 images in each category of this dataset. To increase the performance and avoid overfitting of the model, we have included an image data augmentation step before feature extraction [23]. Figure 2a, b shows the Covid and Non-covid MRI Image respectively. *ImageDataGenerator* is a class responsible to define an object that is required for augmentation. There are several methods available for augmentation and a few of them are flipping, rotating, shearing, cropping and zooming. A zoom operation is applied to the images as an augmentation step in



**Fig. 2** Displays a sample instance from Covid and Non-Covid CT-Scan images before augmentation



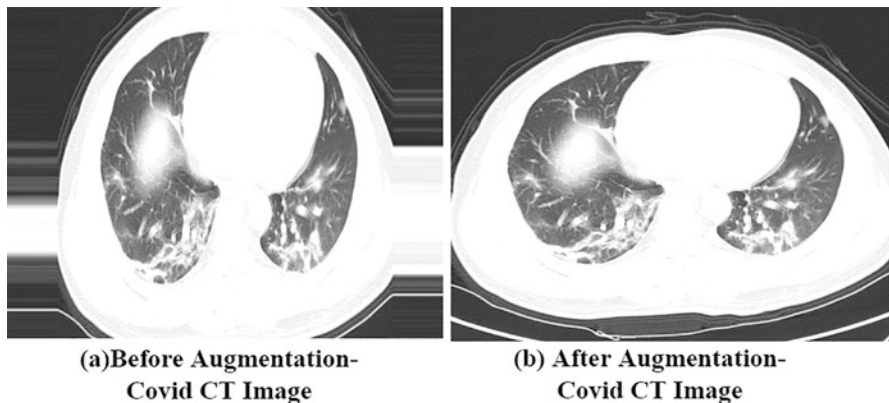


Fig. 3 Sample instance from Covid CT-Scan images before augmentation and after augmentation

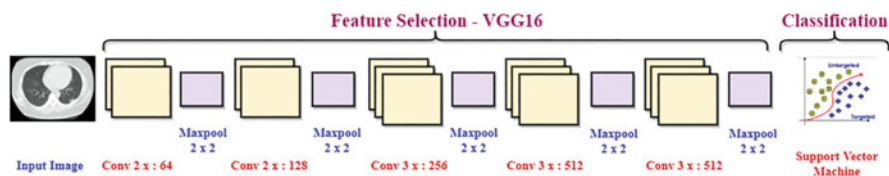


Fig. 4 Proposed architecture VGG16 with Support Vector Machine (VGG16-SVM)

our preprocessing. Images are loaded from the folder using `load_img( )` function and the image are converted into 3D Numpy array using `img_to_array( )` function. The `reshape( )` method changes the dimension of the Numpy array.

The method `flow( )` is responsible for applying the zoom effect to the image, saving the image in a specific folder and specific format with an added prefix file name as “*aug*”. Figure 3a, b shows an sample instance from Covid CT-Scan images before augmentation and after augmentation respectively. After augmentation, the CT scan image dataset contains 400 images in each category of Covid and Non-Covid set. Fifty two images are added by the augmentation process to each category of the dataset.

### 3.4 Proposed Architecture

VGG16 is used at the top of the proposed architecture, and SVM is used at the bottom. It is depicted in Fig. 4. The VGG16 extracts the complex feature from the high dimensional space and it is fed into the SVM for classification. The SVM is a more suitable classifier for the high dimensional vector without any preprocessing and provides good results.

### 3.4.1 VGG16

VGG16 is employed to extract the features from the images [21]. The deep neural network is made up of the first five blocks in the architectural diagram. VGG16 is composed of 2 two layer and 3 three layer convolution layer. In between the convolution layer a maxpooling layer is used to choose the prominent features. These five blocks constitutes a 16-layer VGG, with an input shape (width, height, channel) as (224, 224, 3). In VGG-16, the top three fully connected are made as false, to convert this deep neural network to act as a feature extractor instead of an image classifier. The output of the fifth block is provided as an input to well-known machine learning techniques like SVM, RF, and XG Boost rather than a fully linked layer. A method from the *Keras* library named *summary()* is used to display the entire structure of the VGG16 neural network and the definition of VGG16.

### 3.4.2 Support Vector Machine

Given a set of data  $D = \{x_i, y_i\}_{i=1}^n$  with  $x_i \in \mathbb{R}^d$  and  $y_i \in \{-1, +1\}$ , let us take for instant the data points are linearly separable. Formerly we can define, that all data points marked  $y_i = -1$  lie on one side ( $h(x) < 0$ ) and other data points marked as  $y_i = +1$  lie on another side ( $h(x) > 0$ ). The basic idea of SVMs is to take the canonical hyperplane, defined by the weight vector ‘w’ and the bias ‘b’, that gives the maximum margin among all potential classifying hyperplanes  $h(x) \equiv w^T x + b = 0$ . If  $\delta_h^*$  presents the margin for hyperplane  $h(x) = 0$ , then the goal is to determine the optimum hyperplane  $h^*$  and it is given in the Eq. (1).

$$h^* = \arg \max_h \{ \delta_h^* \} = \arg \max_{w,b} \left\{ \frac{1}{\|w\|} \right\} \quad (1)$$

The SVM task is to determine the hyperplane that enlarges the margin  $\frac{1}{\|w\|}$ , subject to the ‘n’ constraints, namely,  $y_i (w^T x_i + b) \geq 1$  for all data points  $x_i \in D$ .

Rather than enlarging the margin  $\frac{1}{\|w\|}$ , we obtain an equivalent expression if we minimize  $\|w\|$ . In fact, we can obtain an equivalent minimization expression given as follows in Eq. (2)

$$\begin{aligned} \text{Objective Function : } & \min_{w,b} \left\{ \frac{\|w\|^2}{2} \right\} \\ \text{Linear Constraints : } & y_i (w^T x_i + b) \geq 1, \forall x_i \in D \end{aligned} \quad (2)$$

It is a supervised machine algorithm [13] used for both regression and classification purposes. The wide functionality of SVM suits this method for classification problems. The features of each data point in this algorithm are plotted in n-dimensional space. Each coordinate point represents a feature in the problem space. Then, classification is done by distinguishing between the two classes with a hyperplane. This classifier is best suitable for discriminating the two classes. It

works well in separating the two classes with a margin. This method is suitable for images, videos and audios which are in high dimensional nature. In the decision function of SVM, a subset of training points is used that effectively uses the memory during the classification.

### **3.5 *Machine Learning Methods***

The machine learning methods used in the classification of the CT-Scan images are discussed below:

#### **3.5.1 XGBoost**

It is otherwise called Extreme Gradient Boost. It is a decision tree based model [20] designed to increase in speed of training and improve the model performance. The algorithm is designed to optimize using the memory for situations like distributed computing, cache optimization, out-of-core optimization and parallelization. In general, the gradient boosting model corrects the error of the new model from features learned from the prior model. This method supports the prediction problems through classification and regression. It can handle automatically the missing values in any instance of the dataset. This constructs the tree parallel and increases the computation speed. In the training phase, the model is designed to fit any new data.

#### **3.5.2 Random Forest**

The important feature of the Random Forest is the ease of use, robustness and accuracy. Two important feature selection methods are included in this algorithm mean decrease accuracy and mean decrease impurity. In the mean decrease accuracy method, feature values are permuted to decrease the model accuracy. Important variables have significant influences in decreasing the accuracy of the model [14], while unimportant variables do no effect on accuracy. In another method, feature that decreases weighted impurity of the tree are considered for the training. For classifications, Information Gain or Gini impurity and for regression, variances are used in feature computations. Decreases in each feature impurity are averaged and this measure is used to rank those features.

### **3.6 *Multivariate Analysis***

The multivariate analysis deals with images that have more than one measurement per pixel. High dimensional data will be analyzed using principle component

analysis (PCA) to simplify the process while conserving patterns and trends. We have adopted the image reconstruction technique using PCA to capture more information and variance from the data. For that purpose, we have taken a Covid CT scan image from the dataset and the image reconstruction method using PCA is applied. Figure 5a, b shows an original instance from Covid CT-Scan image and reconstructed Covid CT-Scan image using PCA with  $k = 25$  respectively. A line chart is shown in the Fig. 6 with a number of components and cumulative explained variance in 'x' and 'y' respectively. From the chart, it is clear that 25 'k' components are required with a cumulative variance of 95% for successful image reconstruction.

From the chart, we have observed that the number of principal components makes a difference in the reconstruction of the image. We also plotted different subplots to compare the relative difference between the number of 'k' components and as shown in the Fig. 7. Although the image clarity has modified, the suggested model's

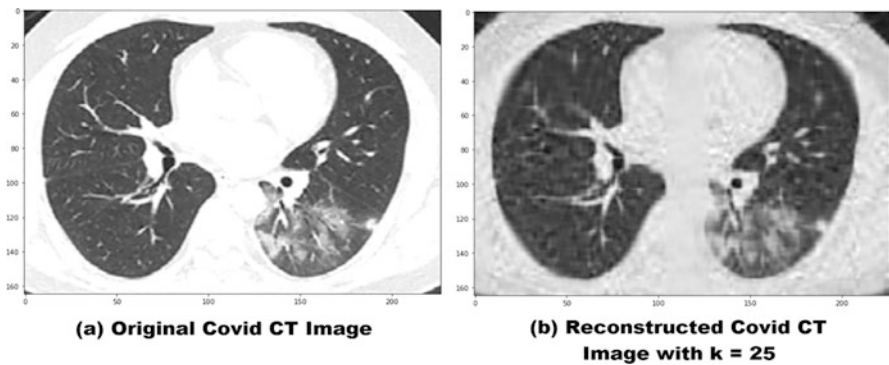


Fig. 5 An original instance from Covid CT-Scan image and reconstructed Covid CT-Scan image

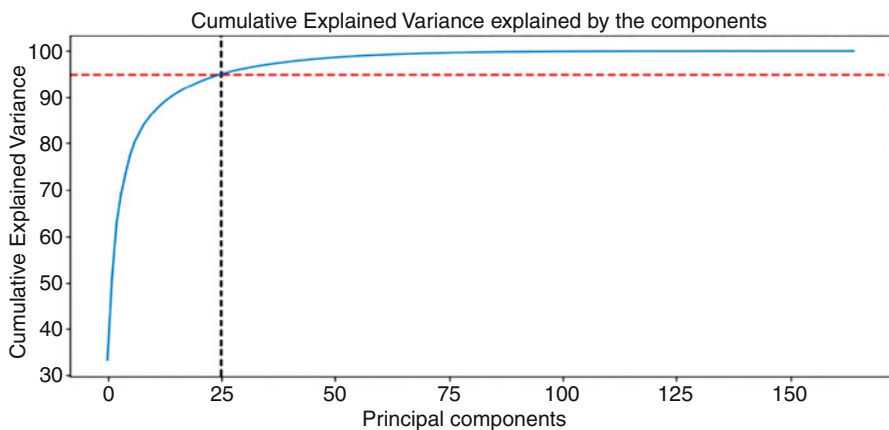
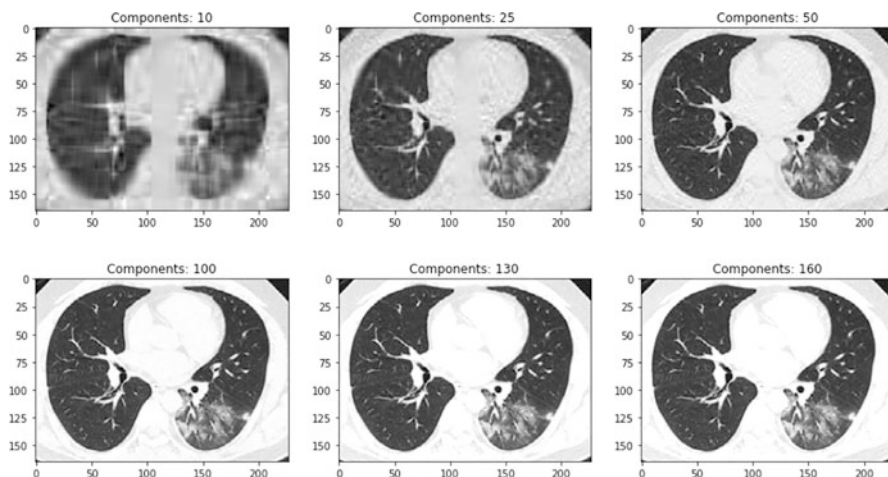


Fig. 6 A line chart with cumulative variance and 'k' principal components



**Fig. 7** Subplots of Covid CT-Scan Image for different 'k' principle components using reconstruction approach

overall accuracy has not improved as a result of the multivariate approach. However, the increase in image reconstruction pixel size results in an increase in computational time.

## 4 Results and Discussion

This section deals with the software and hardware specification, the dataset used for the analysis of Covid-19, metrics used for evaluation, hyperparameter tuning of the model, experimental results and comparative analysis of existing with proposed models are discussed.

### 4.1 Experimental Setup

On the CT-Scan X-ray images from the Covid dataset, a comparison analysis of the three TL methods is done [24]. For the experimental configuration, we used a computer with an Intel Core i3 8th generation, 12 GB of RAM, and a 1 TB hard drive running Windows 10. A top radiologist at Tongji Hospital in Wuhan, China provided this dataset. There are 349 Covid-19 CT scan images in this collection. Tensorflow 1.14, Keras 2.2.4, and Python 3.7 are all used to carry out the tests. The features are extracted from the images by employing the VGG-16 DL model. The VGG-16 is a predefined model loaded from the Keras application. Three experiments are carried on the Covid dataset. It includes viz., SVM, RF and XG Boost [31]. A TL approach is applied to the extracted feature with the above given traditional machine learning methods.

### 4.2 Dataset and Metrics

The dataset is collected from Tongji Hospital, Wuhan, China between the period January 2020 and April 2020. It contains a collection of 348 CT images aggregated from 216 patients. The dataset is taken from Github. CT scan images are otherwise called Computerized Tomography or CAT scan images. These images are aggregated from Covid victims by the senior radiologist. These images are helpful in the diagnosis of Covid patients. Quick treatment can be given with the help of a better and more efficient image diagnosis procedure through deep learning methods.

Each Covid and normal victims have 348 CT scan images. It is divided into 250, 50 and 48 images for training, validation and testing sets respectively and it is shown in Fig. 8. Image augmentation is a method or technique [17] that expands the capacity of the dataset to prevent overfitting issues. It also helps to increase or ameliorates the functioning of the model. After image augmentation, the dataset consists of 300 and 100 images for training and validation set in each category of Covid and Non-Covid dataset and it is shown in Fig. 9. The augmentation dataset is prepared by zooming on the original dataset with the help of *ImageDataGenerator()*. The accuracy measures the goodness of a model in predicting the result and it is given in the Eq. (3). In this equation, the TP, FP, TN and FN represent True Positive, False Positive, True Negative and False Negative respectively.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

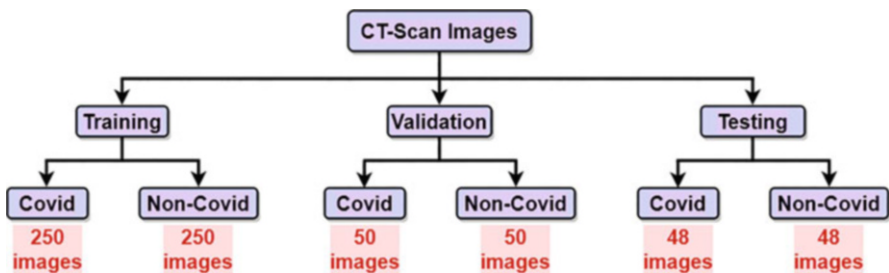
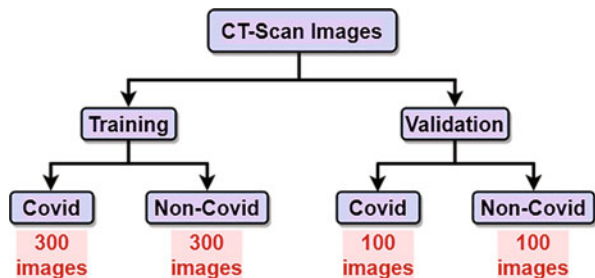


Fig. 8 Categorization of covid and non-covid CT-Scan images before augmentation

Fig. 9 Categorization of covid and non-covid CT-Scan images after augmentation



### 4.3 *Hyper Parameter Selection*

Grid search is a technique that helps to choose/compute parameters of a model that enables us to produce accurate predictions in the classification. It is also called an estimator that helps to perform optimal hyperparameter selection for a given model. This search helps us to maximize the prediction. The hyperparameters of the Random Forest classifier for the parameters random state, max. features, no. of estimators, max. depth and criterion are 42, 'log2', 200, 8, and 'entropy' respectively. The parameters that are set for the XGBoost classifier are small fractions of columns (0.7), learning rate (0.05), max. depth (6), min. sum of instance weight (11), missing input value (-999), no. of estimator (5) and no. of thread (4). To obtain optimal results in the prediction of Covid-19 victims, the SVM classifier hyperparameters are tuned as penalty (12), c (0.01), loss (squared\_hinge) and kernel (linear).

### 4.4 *Experimental Results*

To find the effectualness of the proposed architecture in finding the Covid-19 patients from CT-Scan figures two methods are employed namely (i) Cross Validation method and (ii) Learning Curve method.

#### 4.4.1 **Experiment 1: Cross Validation**

##### **Conceptual Representation**

The various types of cross-validation approaches are used in machine learning to validate the dataset. A few important types of cross-validation methods are (i) K fold Cross Validation, (ii) Stratified K fold Cross Validation, (iii) Leave P out cross validation and (iv) Time Series Cross Validation. A 10-fold cross-validation method is chosen to validate the CT scan images. The total size of the dataset is 400 images. The test set size used in each experiment of this cross-validation method is  $400/10 = 40$  images. A Total of 10 experiments are conducted in this cross-validation. Each experiment [22] is representing one fold of the cross-validation. In experiment one, 1–40 images are assigned as the test set, while the remaining 41–400 images i.e., 360 images for the training set. Likewise, in experiment two, the training set is assigned with a range of 1–40 and 81–400 with a total of 360 images. The test set in experiments 1, 2, 3, ... 10 ranges from 1–40, 41–80, 81–120, ... , 361–400 respectively and it is shown in Fig. 10 with a yellow color. The sky blue color represents the training set used in the 10-fold cross-validation.

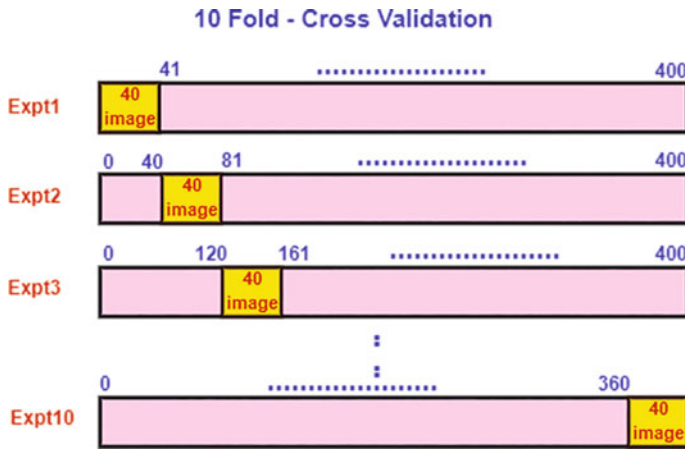


Fig. 10 Illustrates the conceptual representation of 10-Fold cross validation in CT scan images

Table 2 Experimental results of three TL models in terms of accuracy

	Expt1	Expt2	Expt3	Expt4	Expt5	Expt6	Expt7	Expt8	Expt9	Expt10
VGG16-SVM	92.5	90.0	90.0	82.5	90.0	85.0	92.5	87.5	82.5	85.0
VGG16-RF	75.0	77.5	87.5	77.5	87.5	87.5	82.5	82.5	80.0	82.5
VGG16-XGB	67.5	55.0	60.0	75.0	85.0	77.5	52.5	60.0	75.0	80.0

**Results Discussion**

Table 2 lists the accuracy results of three TL models in the order VGG16-SVM, VGG16-RF and VGG16-XGB. Ten experiments are conducted for 10-fold cross-validation in each model. The SVM obtained a maximum accuracy of 92.5% in experiment one, while XGBoost obtained a minimum value of 52.5% in the 7th fold cross-validation. The Mean accuracy of SVM, Random Forest and XGBoost models for 10-fold cross-validation is 87.75%, 82.00% and 68.75% respectively. SVM, Random Forest and XGBoost achieved a standard deviation of 3.61%, 4.30% and 10.73% respectively. A comparison of the three models is displayed in Fig. 11. Blue, Red and Green lines represent the SVM, Random Forest and XGBoost accuracy for cross-validation. SVM lies top in the experiment results of CT Scan images of the Covid19 dataset. The SVM classifier is composed of a function that is maximized in compliance with linear constraints on its variable using minimal consecutive optimization. This approach allows the identification of the viable solution by iterative computation, which generated an accuracy of 87.75%.

**4.4.2 Experiment 2: (Learning Curve Method)**

This method helps to monitor the various aspects of machine learning model performance and identify the lacking areas of the model. Thereby it helps to see the



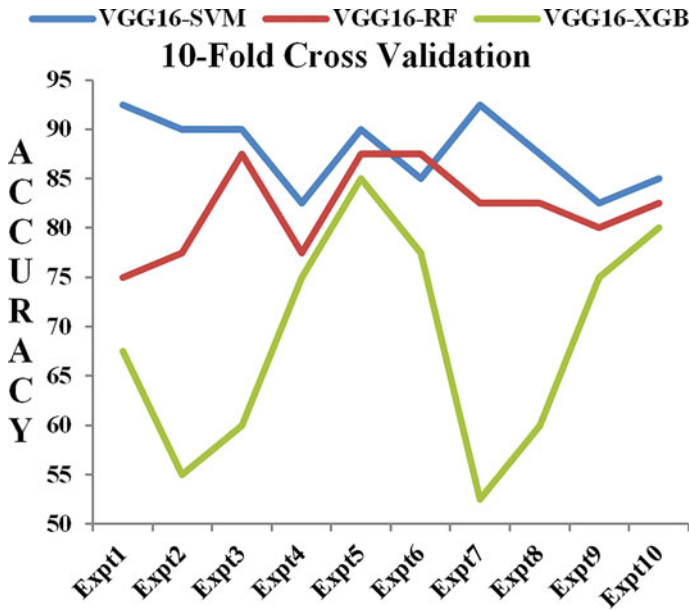
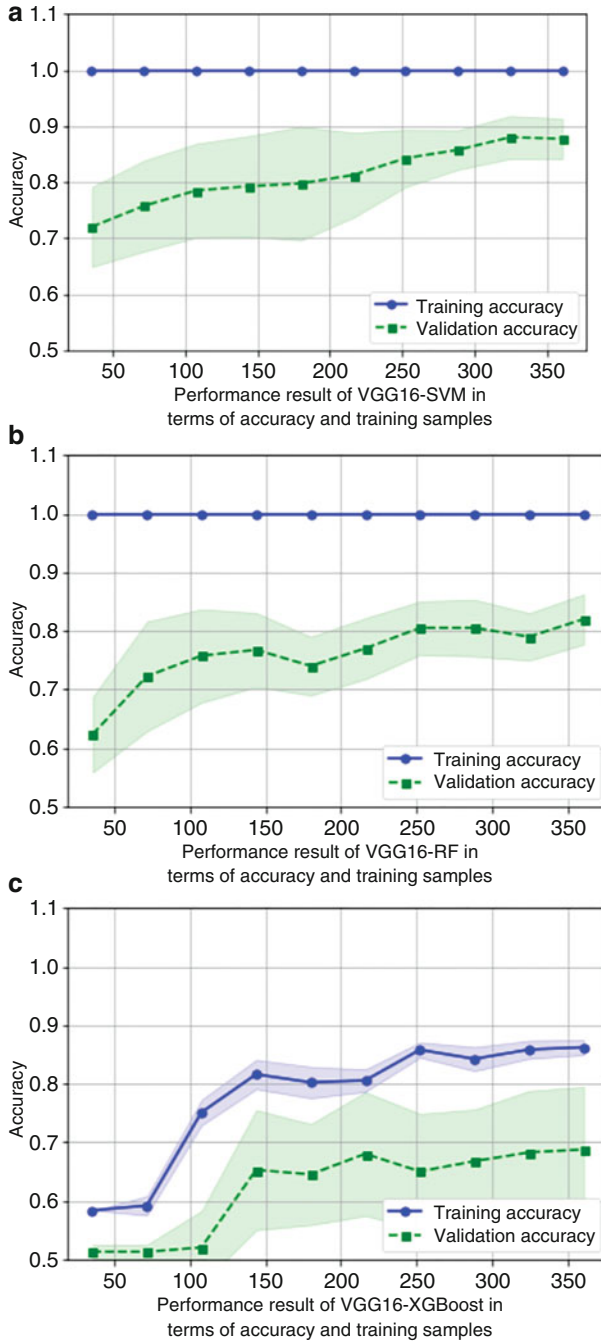


Fig. 11 Comparison of the three TL models in terms of accuracy

progress of the model, helps to build a plan to improve the performance of the model and helps to forecast the decision making process to be easy. The experimental results of the three TL models are compared between accuracy and the number of training samples as shown in Fig. 12. In the figure, (a), (b) and (c) represents the performance result of SVM, Random Forest and XGBoost respectively.

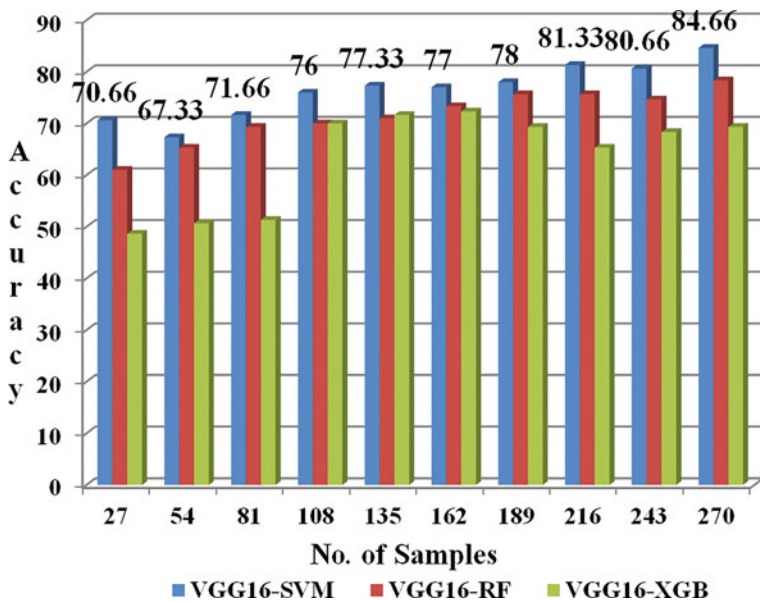
The validation accuracy of the model gradually increases as the training set size of the Covid dataset increases. Support Vector Machine obtained 70.66% accuracy at 27 training samples and 84.66% accuracy at 270 training images. SVM achieves the highest accuracy of 84.66% which is 6.33% and 15.33% greater than the Random Forest and XGBoost classifier respectively. The performance result of the SVM classifier is higher than the Random Forest and XG Boost for training samples. Table 3 shows the output of the learning curve method on the different classifiers. The Fig. 12a–c displays the performance of the learning curve method of SVM, Random Forest and XG Boost classifiers respectively. Random Forest and XGBoost are tree based classifiers, they need a special preprocessing step to handle. In this tree based method, output at each stage depends on the correlation between two trees in the forest and the strength of an individual tree. This enables us to overfit the two models for high-dimensional data.



**Fig. 12** Comparison of the three TL models in terms of accuracy with no. of training samples

**Table 3** Experimental results of three TL models for learning curve method

	No. of samples									
	27	54	81	108	135	162	189	216	243	270
VGG16-SVM	70.66	67.33	71.66	76.00	77.33	77.00	78.00	81.33	80.66	84.66
VGG16-RF	61.00	65.33	69.33	70.00	71.00	73.33	75.66	75.66	74.66	78.33
VGG16-XGB	48.66	50.66	51.33	70.00	71.66	72.33	69.33	65.33	68.33	69.33



**Fig. 13** Performance results of three TL models for learning curve method

### 4.5 Comparative Analysis

The author (G. Muscas et al. [25]) used Gradient Boost Machine in the classification of Covid-19 X-ray and obtained a result of 67.00%. A ResNet 101 CNN is employed to detect abnormalities in X-ray images by Che Azemin MZ et al. [27]. They have achieved an Accuracy of 71.9% and it is shown in the Table 4. A multi-view fusion DL model is proposed to improve the efficacy in the diagnosis of CT chest images. It has resulted in an increase in Accuracy of 4.1% than the Che Azemin MZ et al. [27] approach. S. Wang et al. [28] and H. Yasar and Ceylan [29] proposed Inception CNN and Local Binary Pattern-KNN to obtain 79.30% and 79.73% of Accuracy respectively. The author has proposed TL based VGG16-SVM to classify the Covid-19 CT Scan images which resulted in an Accuracy of 84.66%.

**Table 4** Comparative analysis of proposed method with existing methods

S.No.	Reference	Methodology used	Accuracy (%)
1.	G. Muscas et al. [25]	Gradient boosting machine	67.00
2.	Che Azemin MZ [27]	ResNet 101- CNN	71.90
3.	Xiangjun Wu et al. [26]	Multi-View deep fusion model	76.00
4.	S. Wang et al. [28]	Inception CNN	79.30
5.	H. Yasar and Ceylan [29]	Local Binary Pattern-KNN 10 fold cross validation	79.73
6.	Proposed method	VGG16-SVM	84.66

## 4.6 Limitations

The VGG16 model suffers from a vanishing gradient problem. This can be solved through a ResNet architecture. The ResNet uses residual learning which might be used in the architecture to overcome the gradient problem. When the number of images in the dataset increases, the SVM performance will be reduced due to the requirement for high computational power. A dimensional reduction method like PCA or t-Distributed Stochastic Neighbor Embedding (t-SNE) might be used to scale down the aspect of the data for better classification with less computation.

## 5 Conclusion

A major virus that influences the world in the year 2020 is named as Coronavirus (COVID-19). This novel virus has out breaks the epidemic situation into pandemic one. In order to identify, prevent and control the spreading of coronavirus an effective machine learning technique is required in this situation. The work carried out helps to detect and diagnose Covid-19 automatically through medical imaging of CT-Scan images. In this study, CT-Scan images are used to classify the Covid-19 patients. Three experiments are carried out in the order SVM, RF and XGBoost to identify the proposed method. The multiple stacked convoluted neural network VGG16 in the proposed architecture is capable of extracting the complex feature from the Covid19 CT-Scan images. The extracted chronological features are fed into the SVM for classification. The SVM classifier uses a minimal sequential optimizer that iterates to find a feasible solution and it provides a better solution for spare data. The VGG16-SVM TL method obtained an accuracy of 84.66% which is greater than VGG16-RF and VGG16-XGB by 6.33% and 15.33% respectively. Thus, it can be used as a tool to identify and diagnose Covid-19 without any delay in processing. In the future, the traditional machine model can be replaced with a DL method for classifying the CT-Scan images which can obtain a better result.

## References

1. Ahuja, S., Panigrahi, B. K., Dey, N., et al. (2021). Deep transfer learning-based automated detection of COVID-19 from lung CT scan slices. *Applied Intelligence*, 51(1), 1–15. <https://doi.org/10.1007/s10489-020-01826-w>
2. Farhat, H., Sakr, G. E., & Kilany, R. (2020). Deep learning applications in pulmonary medical imaging: Recent updates and insights on COVID-19. *Machine Vision and Applications*, 31(53), 1–42. <https://doi.org/10.1007/s00138-020-01101-5>
3. Hernigou, J., Valcarenghi, J., Safar, A., et al. (2020a). Post-COVID-19 return to elective orthopaedic surgery—Is rescheduling just a reboot process? Which timing for tests? Is chest CT scan still useful? Safety of the first hundred elective cases? How to explain the “new normality health organization” to patients? *International Orthopaedics (SICOT)*, 44, 1905–1913. <https://doi.org/10.1007/s00264-020-04728-1>
4. Hernigou, J., Cornil, F., Poignard, A., El Bouchaibi, S., Mani, J., Naouri, J. F., Younes, P., & Hernigou, P. (2020b). Thoracic computerised tomography scans in one hundred eighteen orthopaedic patients during the COVID-19 pandemic: Identification of chest lesions; added values; help in managing patients; burden on the computerised tomography scan department. *International Orthopaedics*, 44, 1571–1580. <https://doi.org/10.1007/s00264-020-04651-5>
5. Perumal, V., Narayanan, V., & Rajasekar, S. J. S. (2020). Detection of COVID-19 using CXR and CT images using Transfer Learning and Haralick features. *Applied Intelligence*, 51, 1–18. <https://doi.org/10.1007/s10489-020-01831-z>
6. Pu, J., Leader, J. K., Bandos, A., et al. (2020). Automated quantification of COVID-19 severity and progression using chest CT images. *European Radiology*, 31, 436–446. <https://doi.org/10.1007/s00330-020-07156-2>
7. Sharma, S. (2020). Drawing insights from COVID-19-infected patients using CT scan images and machine learning techniques: A study on 200 patients. *Environmental Science and Pollution Research*, 27, 37155–37163. <https://doi.org/10.1007/s11356-020-10133-3>
8. Tabatabaei, S. M. H., Talari, H., Gholamrezanezhad, A., Farhood, B., Rahimi, H., Razzaghi, R., Mehri, N., & Rajebi, H. (2020). A low-dose chest CT protocol for the diagnosis of COVID-19 pneumonia: A prospective study. *Emergency Radiology*, 27, 607–615. <https://doi.org/10.1007/s10140-020-01838-6>
9. Balaji, P., Nagaraju, O., & Haritha, D. (2017). Levels of sentiment analysis and its challenges: A literature review. In *2017 International Conference on Big Data Analytics and Computational Intelligence (ICBDAC)*. <https://doi.org/10.1109/ICBDACI.2017.8070879>
10. Nagaraju, O., Balaji, P., & Haritha, D. (2018). An overview on opinion mining techniques and sentiment analysis. *International Journal of Pure and Applied Mathematics*, 118(19), 61–69.
11. Penubakabalaji, P., Haritha, D., & Nagaraju, O. (2018). Feature based summarization system for e-commerce based products by using customers’ reviews. In *2018 IADS International Conference on Computing, Communications & Data Engineering (CCODE)*, pp. 1–7. <https://doi.org/10.2139/ssrn.3168342>
12. Ganesan, T., Sivakumar, S., Zeelan Basha, C. M. A. K., & Haritha, D. (2018). Classification of mining techniques in multiclass data sets using wavelets. *International Journal of Pure and Applied Mathematics*, 118(10), 217–222. <https://doi.org/10.12732/ijpam.v118i10.26>
13. Nimmagadda, S., Sivakumar, S., Kumar, N., & Haritha, D. (2020). Predicting airline crash due to birds strike using machine learning. In *2020 7th International Conference on Smart Structures and Systems (ICSSS), Chennai, India*, pp. 1–4. <https://doi.org/10.1109/ICSSS49621.2020.9202137>
14. Rajesh Kumar, T., Videla, L. S., SivaKumar, S., Gupta, A. G., & Haritha, D. (2020). Murmured speech recognition using Hidden Markov model. In *2020 7th International Conference on Smart Structures and Systems (ICSSS), Chennai, India*, pp. 1–5. <https://doi.org/10.1109/ICSSS49621.2020.9202163>

15. Sivakumar, S., Videla, L. S., Rajesh Kumar, T., Nagaraj, J., Itnal, S., & Haritha, D. (2020). Review on Word2Vec Word Embedding Neural Net. In *2020 International Conference on Smart Electronics and Communication (ICOSEC), Trichy, India*, pp. 282–290. <https://doi.org/10.1109/ICOSEC49089.2020.9215319>
16. Soniya, V., Swetha Sri, R., Swetha Titty, K., Ramakrishnan, R., & Sivakumar, S. (2020). Attendance automation using face recognition biometric authentication. In *IEEE 2017 International Conference on Power and Embedded Drive Control (ICPEDC), 16th–18th March 2017*, pp. 122–127. <https://doi.org/10.1109/ICPEDC.2017.8081072>
17. Vidya Sagar, S., Ragav Kumar, G., Xavier, L. X. T., Sivakumar, S., & Durai, R. B. (2020). SISFAT: Smart Irrigation System With Flood Avoidance Technique. In *IEEE 2017 Third International Conference on Science Technology Engineering & Management (ICONSTEM), 23th–24th March 2017*, pp. 28–33. <https://doi.org/10.1109/ICONSTEM.2017.8261252>
18. Sivakumar, S., Rajalakshmi, R., Prakash, K. B., Kanna B. R., & Karthikeyan, C. (2019, July). Virtual vision architecture for VIP in ubiquitous computing. In S. Paiva (Ed), *Technological trends in improved mobility of the visually impaired* (EAI/Springer innovations in communication and computing, pp. 145–179). Springer. [https://doi.org/10.1007/978-3-030-16450-8\\_7](https://doi.org/10.1007/978-3-030-16450-8_7)
19. Videla, L. S., Rao, M. R. N., Anand, D., Vankayalapati, H. D., & Razia, S. (2019). Deformable facial fitting using active appearance model for emotion recognition. In S. Satapathy, V. Bhateja, & S. Das (Eds.), *Smart intelligent computing and applications* (Smart innovation, systems and technologies) (Vol. 104). Springer. [https://doi.org/10.1007/978-981-13-1921-1\\_13](https://doi.org/10.1007/978-981-13-1921-1_13)
20. Videla, L. S., et al. (2018). Modified feature extraction using Viola Jones algorithm. *Journal of Advanced Research in Dynamical and Control Systems*, 10(3 Special Issue), 528–538.
21. Videla, L. S., & Ashok Kumar, P. M. (2020). Fatigue monitoring for drivers in advanced driver-assistance system. In S. R. Nayak & J. Mishra (Eds.), *Examining fractal image processing and analysis* (pp. 170–187). IGI Global. <https://doi.org/10.4018/978-1-7998-0066-8.ch008>
22. Shanmugan, S., & Essa, F. A. (2020). Experimental study on single slope single basin solar still using TiO<sub>2</sub> nanolayer for natural clean water invention. *Journal of Energy Storage*, 30, 101522. <https://doi.org/10.1016/j.est.2020.101522>
23. Essa, F. A., Elsheik, A. H., Sathyamurthy, R., Muthu Manokar, A., Kandeal, A. W., Shanmugan, S., Kabeel, A. E., Sharshir, S. W., & HiteshPanchal, M. M. (2020). Younes, extracting water content from the ambient air in a double-slope half-cylindrical basin solar still using silica gel under Egyptian conditions. *Sustainable Energy Technologies and Assessments*, 39, 100712. <https://doi.org/10.1016/j.seta.2020.100712>
24. Panchal, H., Mevada, D., Sadasivuni, K. K., Essa, F. A., Shanmugan, S., & Khalid, M. (2020). Experimental and water quality analysis of solar stills with vertical and inclined fins. *Ground-water for Sustainable Development*, 11, 100410. <https://doi.org/10.1016/j.gsd.2020.100410>
25. Muscas, G., Matteuzzi, T., Becattini, E., et al. (2020). Development of machine learning models to prognosticate chronic shunt-dependent hydrocephalus after aneurysmal subarachnoid hemorrhage. *Acta Neurochirurgica*, 162, 3093–3105. <https://doi.org/10.1007/s00701-020-04484-6>
26. Wu, X., Hui, H., Niu, M., Liang, L., Wang, L., He, B., Yang, X., Li, L., Li, H., Tian, J., & Zha, Y. (2020). Deep learning-based multi-view fusion model for screening 2019 novel coronavirus pneumonia: A multicentre study. *European Journal of Radiology*, 128, 109041, ISSN 0720-048X. <https://doi.org/10.1016/j.ejrad.2020.109041>
27. Che Azemin, M. Z., Hassan, R., Mohd Tamrin, M. I., & Md Ali, M. A. (2020, August 18). COVID-19 deep learning prediction model using publicly available radiologist-adjudicated chest X-ray images as training data: Preliminary findings. *International Journal of Biomedical Imaging*, 2020, 8828855. <https://doi.org/10.1155/2020/8828855>. PMID: 32849861; PMCID: PMC7439162.
28. Wang, S., Kang, B., Ma, J., Zeng, X., Xiao, M., Guo, J., Cai, M., Yang, J., Li, Y., Meng, X., & Xu, B. (2021). A deep learning algorithm using CT images to screen for Corona virus disease (COVID-19). *European Radiology*, 31(8), 6096–6104. <https://doi.org/10.1007/s00330-021-07715-1>

29. Yasar, H., & Ceylan, M. (2021). A novel comparative study for detection of Covid-19 on CT lung images using texture analysis, machine learning, and deep learning methods. *Multimedia Tools and Applications*, 80, 5423–5447. <https://doi.org/10.1007/s11042-020-09894-3>
30. World Health Organization. (2022, August 23). *Coronavirus (COVID-19) dashboard*. <https://covid19.who.int/>
31. Transfer Learning. (2020, October 18). *Source code for proposed VGG16-SVM*. <https://github.com/SIVAKUMAR-SOUBRAYLU/VGG16-SVM-RF-XGB>
32. Loey, M., Smarandache, F., & Khalifa, N. E. M. (2020). Within the lack of chest COVID-19 X-ray dataset: A novel detection model based on GAN and deep transfer learning. *Symmetry*, 12(4), 651. <https://doi.org/10.3390/sym12040651>
33. Shah, V., Keniya, R., Shridharani, A., Punjabi, M., Shah, J., & Mehendale, N. (2021). Diagnosis of COVID-19 using CT scan images and deep learning techniques. *Emergency Radiology*, 28(3), 497–505. <https://doi.org/10.1007/s10140-020-01886-y>