# Detection of COVID-19 Using X-ray Image Classification

Kavya Garlapati, Naveena Kota, Yasaswini Swarna Mondreti, Preethi Gutha, and AswathyK Nair

Department of Electronics and Communication Engineering, Amrita Vishwa Vidyapeetham Amritapuri, India

garlapatikavya@am.students.amrita.edu
kotanaveena@am.students.amrita.edu
myasaswiniswarna@am.students.amrita.edu
guthapreethi@am.students.amrita.edu
aswathykn@am.amrita.edu

**Abstract:** COVID 19 disease rooted in China, spread across other parts of the world and became a devastating pandemic. The detection of COVID-19 has become a crucial task in the medical sector because of the soaring cases and the paucity of pharmaceutical supplies for detection. Considering the urgency, an immediate auxiliary automatic detection system is required for early diagnosis of the disease and helps the affected patients to be under immediate care. In this work, we aimed to propose an automatic detection system based on lung X-ray images, as radiography modalities is a promising way of faster diagnosis. In this work, we built a machine learning model considering X-ray images taken from publicly available data sets of 2000 images. The relevant features from the images were taken for building the model, prior that proper segmentation was applied to the X-ray images. The X-ray images are prone to noise and spatial aliasing which leads the boundary to be indistinguishable, so proper image segmentation is required. Comprehensive validation has been performed on different segmentation techniques, among those, Sobel demonstrated an accurate result, which is not only effective in detecting edges but also good in removing noises within the image. Further, the preprocessed image is fed to a support vector machine (SVM) model, which accomplished the maximum classification accuracy of 99.17%, also SVM achieved precision, recall, and F1 score of 99.24%,98.13%,98.68% respectively in predicting the COVID-19 versus other pulmonary diseases. Taking the advantage. the model can be helpful to medical persons that can be used as an initial screening of individuals.

**Keywords:** Covid Detection, data augmentation, pre-processing, edge detection, gradient methods, filtering, Support Vector Machine.

## 1 Introduction

The outbreak of COVID-19 virus emerged reportedly from China has infected millions of people around the world and many countries are in lockdown. At the time of writing this report (May 2021), the total affected cases worldwide have been more than 158,334,441 and death reported were 3,297,034 which is considered as a disaster to the humankind [17]. The symptomatic or asymptomatic nature of COVID 19 has been changing over some time, based on age, personal immune, and this leads to delay in finding an effective vaccine to fight against this virus.

The unavailability of therapeutic equipment and the cost of purchasing those for the diagnosis of COVID –19 disease put the medical sectors in a major crisis especially in developing and underdeveloped countries as the daily COVID-19

affected cases are rapidly rising. There are several prevailing methods for COVID detection viz. NAAT tests (Nucleic Acid Amplification Tests), Serological Tests, Antigen tests. The standard prevailing confirmatory medical test for COVID-19 is polymerase chain reaction (RT-PCR) or sequencing, which requires manual intervention and also time consuming. In an emergency situation, where the daily cases are significantly rising, these tests might not be sufficient because of the unavailability of domain experts and medical kits. Computed tomography (CT) is one of the available screening methods for the diagnosis of pneumonia which can be considered as a potential complication of COVID-19 in this prevailing epidemic context [1][14].

The restricted availability of viral testing kits and the time-consuming nature of the tests such as RT-PCR leads radiology come to the front line of diagnosis. CT scans are mostly used for patients with severe symptoms related to lungs and image based diagnosis significantly help to assess the seriousness of such diseases[15]. Typical CT findings included bilateral ground-glass opacity, pulmonary consolidation, and prominent distribution in the posterior and peripheral parts of the lungs [2]. Chest X-Rays (CXR) helps to differentiate a patient affected by COVID-19 from other lung affected illness[16]. Like other types of pneumonia, COVID-19 pneumonia increases the density of lungs. This may be seen as white patches in the lungs on radiography known as Ground-glass opacities (GGOs) [4] that can be considered as a sign of abnormality. It causes inflammation in air spaces of the lungs increasing fluid build-up and makes difficult the transfer of oxygen into the bloodstream. According to Fleischner study, CXR images gives a visual indexes of such abnormalities, and medical practitioners use CXR for the diagnosis of COVID-19 as a preliminary modality [18].

COVID-19 affected time period became a significant phase for doctors to rapidly screen and diagnose patients and isolate them from other individuals to prevent further spreading. Also in a short time, it is a challenge for medical persons to extract the features of coronavirus and discriminate it from other viruses or pulmonary disease. Many articles and studies came up with machine learning algorithms to detect corona and its features. Such works with Artificial Intelligence (AI) can be considered as a global solution to tackle COVID from lung x-rays and an early diagnosis. Ramsey W etal. [6] conducted study on COVID using lung X-ray. The authors study validated that machine learning algorithms were able to detect COVID-19 faster about ten times through X-ray images and more accurately than thoracic radiologists. Various edge detection methods were also carried out to find the abnormalities related to lungs in X-ray images. O. R. Vincent et al. had demonstrated their work based on X-ray image edge detection using Sobel which concluded that Sobel performs a two-dimensional spatial gradient method on an image and also justified with reasons for the superiority of Sobel over other edge detection techniques [7] . In [8], the authors had explored various edge detection algorithms for the analysis of X-ray images.

The trends in Machine Learning (ML) and artificial intelligence (AI) aided us to classify whether a patient is diagnosed COVID or other pulmonary disease. From the detailed literature study on Machine learning algorithms related to COVID and lung X-ray, Support Vector Machine (SVM) was found to be an efficient and simplest technique for classification of COVID techniques using CXR images over

other existing techniques [3][6]. Compared to other techniques the model using SVM is much easier to implement as well as efficient too.

While performing image based analysis, it is necessary to differentiate Region of Interest from other background noises in the image. The CXR image taken for the study contains many background noises and the connecting wires are also evident in the CXR images. For identifying the infectious lesions associated with Coronavirus on CXR, the aforementioned noises need to be removed while conducting AI based studies. This work focuses on the comparison of various image analysis techniques based on X-Ray images and to find the appropriate technique for the classification of COVID-19. The X-rays are pre-processed using image processing techniques such as feature extraction, filtering, noise reduction and edge detection. The tool, OpenCV was used for pre-processing the images, Scikit-learn was used to import required libraries, and models that helped to recognize X-rays of COVID. SVM was used for classification and the performance of the model was evaluated using the performance metric: precision, recall, f1-score, and the confusion matrix and the accuracy of the model was predicted.

The entire work is organized as follows: Section II demonstrates System Architecture, Section III focuses on Methodology, Section IV demonstrates Results and Analysis and section V concludes the entire work.

## 2 System Architecture

### 2.1 Process

As aforementioned in the paper, we have exploited Chest X-ray (CXR) images for COVID-19 among other pulmonary diseases as radiographic based studies are much extensive and cost efficient compared to conventional diagnosis. The image dataset was taken from publicly available repository which consists of COVID affected patients, non COVID and pneumonia cases and a model was built based on that. The CXR images were preprocessed to increase the accuracy of the system. The CXR images are affected by noise, this has to be removed effectively before being fed to the ML model. So suitable preprocessing was performed on the images which further increased the accuracy and reliability of the model and the preprocessed data were given to the ML algorithm for the classification of COVID.

In a nutshell, the entire work can be illustrated as follows: starting from data collection which contains COVID and Non-COVID CXR images, proceeded to get reliable classification data preprocessing such as data resizing, augmentation and noise removal filtering were performed. As a first step to the AI process, the data was split into appropriate ratios for training and testing. The split data was passed through an SVM classifier, undergone the classification and finally arrived to analyze the performance of the system by examining different parameters such as confusion matrix, accuracy, precision, recall, F1-score.

For a better understanding of the model and to get an idea of the entire workflow, the process is demonstrated with the help of the flowchart as in Fig.1.

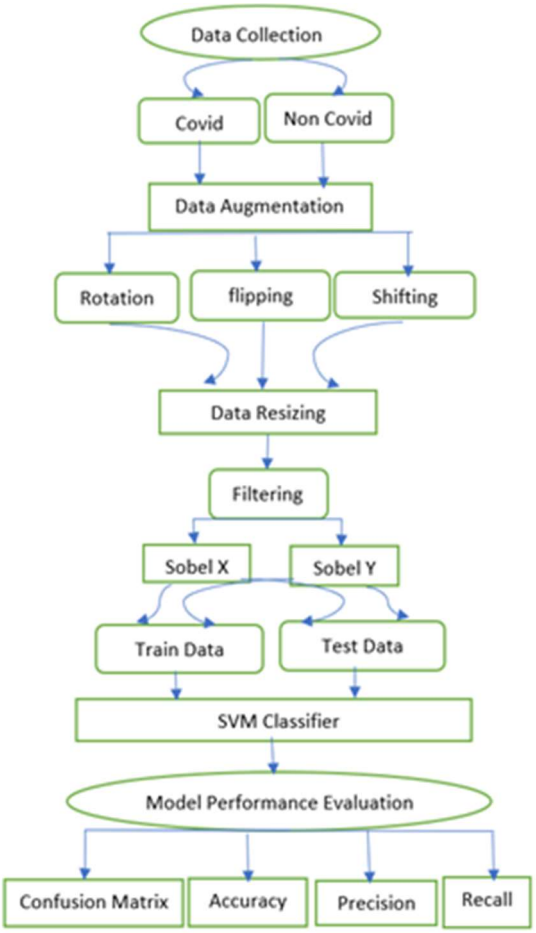


**Fig. 1.** Flow Chart of process

## 2.2 Data Collection

Data collection is chosen based on the required factors for training our model. A dataset that contains images of COVID affected along with other non-COVID images is considered and classified the whole data set into two categories. These two categories are named COVID and Non-COVID. CXR of COVID affected is seen as shown in Fig.2(a), Fig.2(b) is a CXR of Non-COVID cases of other pulmonary disease and normal person. A total of 3000 images were taken for classification and among those, augmented images were also included to avoid overfitting. The training, validation and test data were taken as a ratio of 70:10:20. The dataset of two categories were trained to get the output of interest. A trained dataset in were also taken for reference and validation and added more sets of images for accurate observation. The image that is affected by noise is removed using suitable filters.


**Fig.2** Lung x-rays of (2a) COVID, (2b) Non COVID

# 3 Methodology

## 3.1 Data Preprocessing

**Data Augumentation**. In this work, we have adopted the Data augmentation technique to increase the dataset size, as there are limited CXR images available from public repositories. Lack of sufficient data will result in wrong predictions, when given to AI based classifiers. Data augmentation is a basic tool used to create datasets from existing data sets. It creates new and different images from the existing image data sets. This process is done by applying various transformation techniques such as rotation of the existing image by appropriate degrees, zooming, cropping, shearing, and flipping off the existing set of images in different directions. The original CXR image is given in Fig. 3a. The image that is flipped is shown in Fig. 3c. Fig.3d shows the rotated images of given dataset. After all the processing the image data was increased to 3648, which includes the CXR images with medical connecting wires and clips. We have intentionally taken those images with wires for training the model, considering the fact that in real scenario those factors exists in CXR images. Using our model we, could effectively remove those wires during preprocessing.

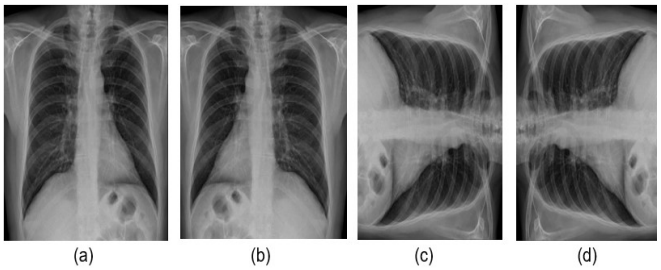


**Fig. 3** Preprocessed images of (3a) original sample, (3b) shows the Flipped Image, figure (3c), (3d) shows the rotated images.

**Data Reshaping.** The dataset for the classification was taken from different publicly available datasets and the images have a dynamic range of resolutions that need to be scaled down to a common acceptable size for the design developed. Based on the design and trial experiment, an acceptable image resolution was chosen for the proposed model as 250 × 250. After reshaping, the images were split into the training set and testing set. All the processed images were then stored into a common repository for further investigation.

### 3.2 Image Filtering

Image Filtering. Removing the unwanted data is necessary for every model to accomplish greater accuracy that can be achieved through some of the filtering techniques- smoothening and enhancing. Enhancing the image plays a major role in image pre-processing techniques because it provides us a better image for further examination [9].

In image filtering, the first task was to select the best filter that removes noise in the image and enhances the quality of the image. Taking all these into consideration, the images of lung X-rays are filtered using different image processing techniques such as Gaussian, Sobel X, and Sobel Y as shown in Fig.4. Other filters such as Laplacian and HOG were also taken and implemented for the study, but it showed poor results. In the CXR image data set, many connecting wires and clips used for medical examinations were seen. The inclusion of these in the images for processing may result in wrong prediction, so to remove these unwanted effects, the appropriate filter has to be selected. By examining these filtered images using the mentioned techniques and by running edge detection algorithms, Sobel X and Sobel Y gave accurate results and was chosen for further classification. Our proposed model in the preprocessing stage was able to figure out the pulmonary lesions in the image.

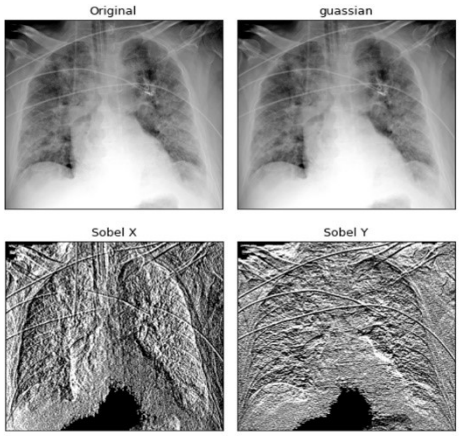


**Fig.4** Images with filtering

### 3.3 Edge Detection

In order to classify an image whether a COVID or not, the important factors that need to be considered are edges which help to get an accurate outcome. It is also a basic feature of an image that provides valuable information for image perception [11]. Edge detection considerably reduces the processing time by reducing the amount of data while preserving the structural boundary of the region of interest (ROI). The proposed method preserves the structural information of the boundary where pulmonary lesions were present and removed the background.

Edges preserves the contour features of the ROI in the image. The detection of edges mainly involves the measurement and localization of gradient change in gray scale image as edge is a sharp discontinuity in intensity over the boundary in an image

[11]. Smoothing was the first step performed in edge detection as it suppresses the noise without disturbing the edges. While smoothing the image, especially X-ray images the image quality may degrade. So to enhance the quality of the edges, image sharpening was performed. Then, to determine which edge pixels must be removed as noise and which must be retained and finally exact localization of the edges corresponding to ROI need to be evaluated. All these factors were considered and our model gave accurate results. There are numerous ways to perform edge detection, out of which, the gradient method was chosen that calculates the gradient change in the pixel values of the image in a given direction. In this work, the Sobel filter which comes under the family of edge detection filters based on gradient method were used where we have done iterative approaches to obtain optimal results for edge detection.

**Sobel Filtering - A Gradient method of filtering.** Sobel is an orthogonal gradient operator which corresponds to the first derivative. It uses two slicing windows in which, one is column and other is row [13]. These slices use 3x3 matrices that take each pixel value one by one by shifting one unit to the right. In this method, the convolution of two kernels is performed to calculate the gradient Gx and gradient Gy, along x-axis and y-axis respectively.

Absolute gradient in Sobel is calculated as :

$$G = \sqrt{(Gx^2 + Gy^2)} \qquad (1)$$

And it is approximated as

$$|G| = |Gx| + |Gy| \qquad (2)$$

The output of a gradient edge detector is its magnitude (1) (2). After calculating the magnitude of the first order derivative, identify the pixels corresponding to an edge. For that, thresholding the gradient image is necessary. So, the edge is recognized by the pixels with gradient value greater than that of the threshold. One of the problems with edge detection is it prone to noise. The noise is increased as a result of spatial domain differentiation strengthening high frequencies [12].

**Kernel Calculation .** Kernels are used to calculate the pixel orientation with respect to location denoted as Gx and Gy its gradient is given by

$$\boldsymbol{\nabla f}\,(\boldsymbol{x},\boldsymbol{y}) = [\boldsymbol{Gx}\ \boldsymbol{Gy}]^{\boldsymbol{T}} = [\tfrac{\boldsymbol{\partial f}}{\boldsymbol{\partial x}}\ ,\tfrac{\boldsymbol{\partial f}}{\boldsymbol{\partial x}}] \qquad (3)$$

The gradient magnitude and angle is given in (4) and (5):

$$mag\ (\nabla f) = |\nabla f(2)| = [\,Gx^2\ \ Gy^2]^{½} \qquad (4)$$

$$\phi\ (x, y) = arctan\ (\ Gx\ /Gy\ ) \qquad (5)$$

These derivatives have to be estimated for every pixel region. We use two 3x3 convoluted templates with weighte indexes, one for Gx and the other for Gy. These templates are shown below :

| | | |
|---|---|---|
| -1 | 0 | 1 |
| -2 | 0 | 2 |
| -1 | 0 | 1 |

Gx

| | | |
|---|---|---|
| 1 | 2 | 1 |
| 0 | 0 | 0 |
| -1 | -2 | -1 |

Gy

The above two kernels are used to perform convolution with every point in the image. Also, kernel coefficents can be varied with respect to user choice without violiating its properties. İnitially the image is convolved in the direction of X, later in direction of Y. As a result, one kernel has its maximum approximation of the derivative to the column edge and the other to the row edge. The edge magnitude is the maximum values of the two convolutions. In genral sobel operator is a combinaion smoothning and differencing functions, here it is notated as s(k,l). Their convolutions are as follows (6)(7)(8).

$$g1(x,y) = \sum_{k=1}^{Row} \sum_{I=-1}^{List} s_1(k,l)\, f(x+k, y+l) \quad (6)$$

$$g2(x,y) = \sum_{k=1}^{Row} \sum_{I=-1}^{List} s_2(k,l)\, f(x+k, y+l) \quad (7)$$

$$g\,(x,y) = g_1^2(x,y) + g_2^2(x,y) \quad (8)$$

Here $g1(x,\ y) > g2(x,\ y)$ indicates that the edge is passing through the vertical coordinates. If f(x,y) satisfies the below conditions, then f(x,y) of a point (x,y) is considered as an edge

$$g(x,y) > 4 * \sum_{i=1}^{Row} \sum_{j=-1}^{List} \frac{g_{(i,j)}^2}{Row} * List$$

$$g1(x,y) > g2(x,y)$$

$$g(x,y-1) <= g(x,y)$$

$$g(x,y) >= g(x,y+1) \quad (9)$$

Simillarly, when the edge is passing through the horizantal coordinates. If f(x,y) satisfies the below conditions, then f(x,y) of a point (x,y) is considered as an edge

$$g(x,y) > 4 * \sum_{i=1}^{Row} \sum_{j=-1}^{List} \frac{g_{(i,j)}^2}{Row} * List$$

$$g1(x,y) > g2(x,y)$$

$$g(x-1,y) \leq g(x,y)$$

$$g(x,y) >= g(x+1,y) \qquad (10)$$

### 3.4 Support Vector Machine

Here, SVM was preferred because It can facilitate image classification problems easily and results in the best classification measures by transforming it into a problem of squared optimization. This helps in the reduction of steps in the process of learning and also a faster solution compared to other algorithms such as KNN, Random forest and so on. SVM is a powerful and convenient classification and regression algorithm used in supervised Machine Learning by producing efficient results by optimal computation.

In classification, it can separate features into data classes and tries to get the maximum marginal hyperplane. It can be done by generating hyperplanes iteratively and providing the best class division and choosing the best plane among them. It selects a point far from the features of the data classes to get a line. The distance between each data class is calculated for the best-chosen hyperplane by the SVM model.

## 4 Results and Analysis

### 4.1 Performance Measures of SVM model

In this division, we provided the common methodology for the experiments to measure the performance using a validation set. Classification performance metrics are used for the prediction of diseases. A confusion matrix is used to understand and evaluate the SVM model by using standard metrics. For binary classification, the confusion metrics use true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN). Accuracy is obtained as

$$Accuracy = \frac{TP+TN}{TP+TN+FP+F}$$

The ratio of truly predicted positive values to the total number of predicted positive values gives the Precision. Higher value of precision results in a lower false-positive value

$$precision = \frac{TP}{(TP+FP)}$$

Recall is known as sensitivity; it is the ratio of truly predicted values to the total values in all original sets. It results in the positive values that are correctly identified.

$$Recall = \frac{TP}{(TP+FN)}$$

The weighted mean of Precision and Recall results in F1-Score. Particularly, if the data is non-uniform distribution, the precision and recall values sometimes are erratic, and F-Measure (F-Score) tends to be intuitive.

$$F1-Score = 2 * \frac{((Precision * Recall)}{((Precision + Recall)}$$

### 4.2 Experimental Results

The experimental results provide information on the validation process. Through the proposed SVM model, the accuracy achieved was Accuracy:0.98. The data set division is as follows as shown in Table I. Confusion Matrix of the model is shown in Fig.5, where the coloured diagonal elements represent the correctly predicted samples and the black diagonal elements shows the false classification. We can observe that the out of 1095 tested images, 556 are truly predicted as COVID cases and 525 are predicted as truly Non COVID cases.

The performance measures such as Recall and precision is obtained from the confusion matrix and F1-score is also calculated from precision and recall.

**Table 1.** Data set for train and test division

| Type of Data | COVID | Non COVID |
|---|---|---|
| Original data | 125 | 200 |
| Augmented data | 912 | 912 |
| Sobel Filtering | 1824 | 1824 |
| Total Images | 3648 | |
| Train Set | 2553 | |
| Test set | 1095 | |

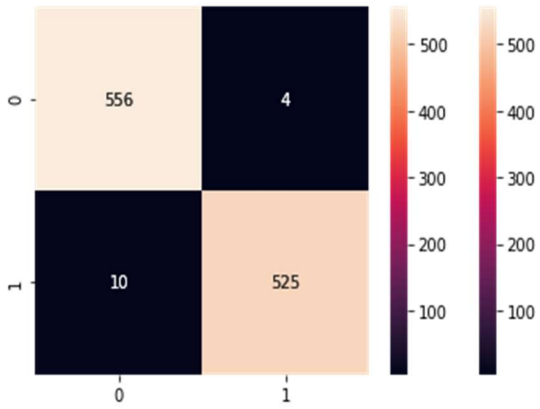


Fig.5 Confusion Matrix

Precision, Recall, and F1-Score from proposed model are tabulated as shown in Table II.

**Table 2.** Performance Mesures of the model

| | |
|---|---|
| Precision value | 0.99243 |
| Recall / Sensitivity | 0.9813 |
| F1-Score | 0.9868 |

## 5 Conclusions

The foremost objective of this work is to study SVM machine learning model for the identification and detection of COVID-19. The dataset taken for this study has 2000 X-rays images that included augmented images also, this helped to further achieve the training efficiency of 99 %. The Sobel filtering helped in intensifying the image features and increases the performance of the model. The suggested system provides good accuracy of 98.72. Moreover, the model obtains good precision, recall, F1-Score. It is proved that the values obtained from this model are quite enough to predict the COVID-19 as a preliminary diagnosis. This concludes that the implemented model can be useful for the physicians to classify and analyse the COVID 19 as a supporting system. In future works, the real time medical X-ray examination using AI model along with proper reference to medication can be used.

## References

[1] Hani C, Trieu NH, Saab I, Dangeard S, Bennani S, Chassagnon G, Revel MP. COVID-19 pneumonia: A review of typical CT findings and differential diagnosis. Diagn Interv Imaging.;101(5):263-268. doi: 10.1016/j.diii.2020.03.014. Epub 2020 Apr 3. PMID: 32291197; PMCID: (2020).

[2] Li B, Li X, Wang Y, et al: Diagnostic value and key features of computed tomography in Coronavirus Disease 2019. Emerg Microbes Infect. 2020;9(1):787-793. doi:10.1080/22221751.2020.1750307(2019).

[3] Syeda HB, Syed M, Sexton KW, et al. Role of Machine Learning Techniques to Tackle the COVID-19 Crisis: Systematic Review. JMIR Med Inform. 2021;9(1):e23811. Published 2021 Jan 11. doi:10.2196/23811.

[4] Li Y, Xia L. Coronavirus Disease (COVID-19): Role of Chest CT in Diagnosis and Management. AJR Am J Roentgenol. 2020;214(6):1280-1286. doi:10.2214/AJR.20.22954 (2019)

[5] Miah, Md Badrul & Yousuf, Mohammad: Detection of lung cancer from CT image using image processing and neural network. 10.1109/ICEEICT.2015.7307530. (2015)

[6] Ramsey M. Wehbe, Jiayue Sheng, Shinjan Dutta, Siyuan Chai, Amil Dravid, et al. DeepCOVID-XR: An Artificial Intelligence Algorithm to Detect COVID-19 on Chest Radiographs Trained and Tested on a Large U.S. Clinical Data Set, -Radiology 2021 299:1, E167-E176(2021).

[7] Olufunke Vincent, Olusegun Folorunso: " A Descriptive Algorithm for Sobel Image Edge Detection" - Proceedings of Informing Science & IT Education Conference (InSITE) (2009)

[8]Mahendran, S.k: “A comparative study on edge detection algorithms for computer aided fracture detection systems".international Journal of Engineering and Innovative Technology, 2(5), 190-193 (2012).

[9] Ali Mohammad Alqudah,Shoroq Qazan “Augmented COVID-19 X-ray Images Dataset”, Mendeley Dataset , V4, DOI:10.17632/2fxz4px6d8.4 (2020).

[ 10] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, Proc. IEEE Comput. Soc. Conf.Comput. Vis. Pattern Recognit. 770–778 (2016).

[11] Mohammad Elha, Jawadkadhim Mohammed, et al: ”Study Sobel Edge Detection Effect on the ImageEdges Using MATLAB”.International Journal of Innovative Research in Science, Engineering and Technology (An ISO 3297: 2007 Certified Organization) Vol. 3, Issue 3, March 2014. ISSN: 2319-8753

[12] R. Fisher, S. Perkins, A. Walker and E. Wolfart: "Edge Detectors", (2003).

[13] T. A. Al-Aish: "Edge Detection in Sensor Networks using Image Processing", Diala , Jour., vol. 31 , Iraq, (2008).

[14]K. Roy et al: "A Comparative study of Lung Cancer detection using supervised neural network," International Conference on Opto-Electronics and Applied Optics (Optronix), Kolkata, India, 2019, pp. 1-5, doi: 10.1109/OPTRONIX.2019.8862326. (2019)

[15] Adithya J, Nair B, Aishwarya S, Nath LR. The Plausible role of Indian Traditional Medicine in combating Corona Virus (SARS-CoV 2): a mini-review. Curr Pharm Biotechnol. doi: 10.2174/138920102 1666200807111359. Epub ahead of print. PMID: 32767920.( 2020 Aug 6).

[16] Gopal K.,Varma P.K.: Department of Cardiovascular and Thoracic Surgery, Amrita Institute of Medical Sciences, Amrita Vishwa Vidyapeetham (Amrita University), Kochi, Kerala, India,Cardiac surgery during the times of COVID-19, Indian J Thorac Cardiovasc Surg, p.1-2 (2020).

[17]https://www.worldometers.info/coronavirus/?utm_campaign=homeAdvegas1

.

[18] Rubin GD, Ryerson CJ, Haramati LB, Sverzellati N, Kanne JP, Raoof S, Schluger NW, Volpi A, Yim JJ, Martin IBK, Anderson DJ, Kong C, Altes T, Bush A, Desai SR, Goldin J, Goo JM, Humbert M, Inoue Y, Kauczor HU, Luo F, Mazzone PJ, Prokop M, Remy-Jardin M, Richeldi L, Schaefer-Prokop CM, Tomiyama N, Wells AU, Leung AN: The Role of Chest Imaging in Patient Management During the COVID-19 Pandemic: A Multinational Consensus Statement From the Fleischner Society. Chest. Jul;158(1):106-116. doi: 10.1016/j.chest.2020.04.003. Epub 2020 Apr 7. PMID: 32275978; PMCID: PMC7138384.( 2020).