A Deep Learning-based COVID-19 Automatic Diagnostic Framework using Chest X-ray Images

ABSTRACT

The lethal novel coronavirus disease 2019 (COVID-19) pandemic is affecting the health of the global population severely and a huge number of people may have to be screened in the future. There is a need for effective and reliable systems that perform automatic detection and mass screening of COVID-19 as a quick alternative diagnostic option to control its spread. A deep learning-based robust system is proposed to detect the COVID-19 using chest X-ray images. Infected patient's chest X-ray images reveal numerous opacities (denser, confluent, and more profuse) in comparison to healthy lungs images which is used by a deep learning algorithm to generate a model to facilitate an accurate diagnostics for multi-class classification (COVID vs. normal vs. bacterial pneumonia vs. viral pneumonia) and binary classification (COVID-19 vs. non-COVID). COVID-19 positive images have been used for training and model performance assessment from several hospitals of India and also from countries like Australia, Belgium, Canada, China, Egypt, Germany, Iran, Israel, Italy, Korea, Spain, Taiwan, USA, and Vietnam. The data were divided into training, validation and test sets. The test accuracy of 97.11 ±2.71% was achieved for multi-class and 99.81 ± 0.00% for binary classification. The proposed model performs real-time disease detection in 0.137 seconds per image in a system equipped with a GPU device and can reduce the workload of radiologists by classifying thousands of images on a single click to generate a probabilistic report in real-time.

Keywords: Chest X-ray Radiographs; Coronavirus; Deep Learning; Image Processing; Pneumonia

1. INTRODUCTION

An eruption of novel coronavirus disease or COVID-19 (previously known as 2019-nCoV) started in China in December 2019. As of 16th September 2020, more than 29.5 million cases have been reported in more than 188 countries and it has more than 930000 deaths [1]. COVID-19 caused by severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) is a disease that can be severe in patients with comorbidities and has a fatality rate of 2% [2]. There is an urgent need to take an effective step for the containment of COVID-19 by performing screening tests on a suspected fellow so that the infected person can receive immediate care and more specific treatment and quarantine of the patient can be ensured to limit the spread of the virus.

The SARS-CoV-2 infection has a wide range of clinical manifestations ranging from asymptomatic infection and mild upper respiratory tract illness to severe viral pneumonia that may culminate in failure of the respiratory system and sometimes death [3]. Real-time reverse transcriptase-polymerase chain reaction (RT-PCR) tests are performed for the qualitative detection of nucleic acid from upper and lower respiratory tract specimens (i.e. nasal, lower respiratory tract aspirates, sputum, nasopharyngeal or oropharyngeal swabs, nasal aspirate) of infected person [4]. Performing RT-PCR testing for COVID-19 will most probably remain main detection method, however it is expensive, complicated, and time-consuming for countless patients with a lack of time and also other

methods are required to detect infected patients. Because of the shortage of kits for RT-PCR and still also relatively high false-negative rate, the examination of chest X-rays (CXR) can be an alternative method of screening and early identification of lung involvement. It may be noted that the detection of lung involvement may predict a potentially life-threatening outcome in patients with COVID-19 [5] [6].

CXR images are non-invasive and X-rays of the chest are usually done in either anteroposterior (AP view) or Posterior anterior (PA view) of a suspected patient's chest to generate cross-sectional images [7]. These X-ray images are examined by expert radiologists to find abnormal features suggestive of COVID-19 based on extent and type of lesions. Imaging features of the X-ray image of coronavirus affected persons varies as these depend on the stage of infection. The spectrum of radiological findings varies from normal (18% of cases) to 'whiteout Lung'. The usual abnormality seen is bilateral peripheral sub-pleural ground glass opacities (GGO) and consolidations. "Crazy-paving" pattern and reversed halo sign may be seen [5] [6]. There may be a rapid progression in extent of lesion in 24 to 48 hours to multilobar to total lung involvement in severe disease [8]. With an increase in the number of patients with COVID-19 disease, the medical community may have to depend on portable CXR images because of its extensive accessibility and reduced infection controlling issues which presently limit the utilization of computed tomography (CT) services. With an increase in patient numbers, the workload on radiologists for this diagnostic process is also increasing and lack of availability of radiologists in certain places is also a challenge. Thus, there is an urgent requirement of a device or system which identifies the disease with an acceptable level of accuracy, even without a radiologist's help to save time as well as to preserve the effort for the neediest in these time-constrained settings. Analysing medical imaging for disease classification is one of the highest priority research areas. With the help of an expert radiologists and based on the aforementioned features of CXR images, a computer-aided diagnostic system can be generated to correctly interpret COVID-19 cases from the input X-ray image.

Several artificial intelligence systems using deep learning [9] as a pre-screening test for COVID-19 detection using CXR images are proposed in [10] [11]. Narin et al. [12] and Zhang et al. [13] used a similar approach with ResNet 50 and basic ResNet respectively, as a base neural network to classify normal and COVID-19 patients. A range of fine-tuned deep convolutional neural network (DCNN) based COVID-19 detection proposed by Khalid et al. [14] classify the result into normal and pneumonia. Khan et al [15] proposed coroNet for detection of COVID-19 in which CXR images are trained on Xception deep neural architecture for COVID-19 classification. Wang et al. [16] used X-ray images of four categories- COVID-19, Bacterial Pneumonia, Viral Pneumonia, and Normal. Fine-tuned SqueezeNet is proposed by Ferhat et al. [17] for COVID-19 diagnosis with Bayesian optimization additive giving the accuracy of 98.3% on CXR images. Ghosal et al. [18] used CXR images for computer-aided diagnostic of COVID-19 and normal patients. The CNN model is trained with 70 COVID-19 and other images of normal subjects and claimed 92% accuracy with that dataset. Shashank et al. [19] used public dataset of 181 COVID-19 Images, 364 healthy Images to detect COVID-19 using deep transfer learning. The model achieved the accuracy of 96.3% and loss of 0.151. Luca et al. [20] proposed a COVID-19 detection using deep learning on the dataset of 250 COVID-19 CXR images. Random forest classifier-based screening system to differentiate between COVID-19 patients and community-acquired pneumonia was implemented by Shi et al. [21] with 87.9% accuracy. Khobahi et al. [22] proposed a novel semi-supervised deep neural network architecture that can distinguish between healthy, non-COVID pneumonia, COVID-19 infection based on the CXR manifestation of these classes while taking very few numbers of parameters. It comprised of Task-Based Feature Extraction Network (TFEN), and COVID-19 Identification Network (CIN). Ozturk et al. [23] implemented 17 convolutional layers where each layer has different filters for each layer using DarkNet as a feature extractor layer. Abbas et al. [24] proposed a Decompose, Transfer, and Compose (DeTraC) method for COVID-19 classification. The author trained the chest computes tomography (CT) dataset with re-trained models on ImageNet and for identification use ResNet. For all previous works dataset for COVID-19 patients is too small or limited in size to make the training model robust, as compared in Table 10. The screening systems should be developed in the way to diagnose the CXR of the person and classify that image according to the probability of the diagnosed disease. It creates a need for a user-friendly diagnosis system where there is no need for trained manpower. The whole framework should be standalone and for practical reasons, there is often requirement not to depend on internet connectivity. The system should work fast to reduce workload and give results much faster than human experts. The dataset collected from multiple places and multiple conditions to train the deep learning model can be helpful to develop a diagnostic system which will be less prone to errors with universal acceptance.

The main contribution of the work is to develop a deep learning-based system that can automatically identify the COVID-19 disease in CXR images. For this purpose, we collected so far the largest dataset of COVID-19 patients and examined several different architectures where the most accurate was identified. The used dataset contains CXR database of 659 COVID-19, 1660 healthy and 4265 non-COVID (viral and bacterial pneumonia)

samples which was also extended by 300 abnormal samples. Those samples were collected from three local hospitals of India and other countries like China, Italy, Australia, Iran, Spain, Germany, Vietnam, Israel Belgium, Canada, USA, Egypt, Korea and Taiwan making the dataset comprised of the large variety that may train the model for high robustness. The dataset was split into training, validation (5-fold cross-validation) and test datasets. Different approaches were examined including binary classification (COVID-19 or non-COVID), three class classification (COVID-19, pneumonia or non-COVID), and four-class classification model (COVID-19, normal, bacterial pneumonia, viral pneumonia). The results outperform the previous works in terms of accuracy, speed, and other parameters.

Unlike many existing works that only consider a classification task on COVID-19 and non-COVID classes, the trained deep-learning network on comprehensive dataset can extract the best region in the X-Ray images to be further fed into the succeeding classifier network. This is unlike many existing works that naively feed the X-ray image to the classification network.

The variety of the sample images was collected from various pubic data sources and globally from several countries and extended from data collected from several hospitals. Thanks to this, we expect to achieve high comprehensiveness of the train model and robustness among variety of different imaging devices with various settings. We suppose we achieved higher acceptance among wide range of countries. The image test takes approximately 137 milliseconds per image (NVIDIA Quadro P600) thus making the model suitable for online screening of COVID-19 patients on any system equipped with modern GPU device. The source code and datasets will be released as an open-source and is free for download so anyone can benefit from the work or can also extend the work in future with other sources. The experiment is fully reproducible.

The rest of the paper is structured as follows. Section II describes the datasets used in the experiment and how it was created. It also discusses clinical aspects of the problem and architectures used for detection of COVID-19 cases including the methodology used for evaluation. Section III includes experiments and results for COVID-19 detection and comparison to other works. Finally, section IV concludes the paper and discusses possibilities regarding future work.

2. MATERIALS AND METHODOLOGY

2.1 Dataset – Chest X-Ray Images

For this study, datasets of CXR images are taken from two publicly available databases which were supplemented by data collected from hospitals in India. Indeed, public database geographical and X-ray image acquisition variance brings diversity and richness in the configuration and performance assessment phase.

a) Dataset A is from the open-source repository [25] that has 237 COVID-19 CXR images from various parts of the world like Australia, Belgium, Canada, China, Egypt, Germany, Iran, Israel, Italy, Korea, Spain, Taiwan, USA and Vietnam on (12 May, 2020). This open repository contains a database of chest images of COVID-19, acute respiratory distress syndrome (ARDS), severe acute respiratory syndrome (SARS) 1, SARS 2, Middle East respiratory syndrome (MERS) patients.

b) Dataset B consists of chest X-ray images of pneumonia infected and normal people of 5848 images from open source repository [26]. It is a combination of 1583 normal images, 2772 bacterial pneumonia images, and 1493 viral pneumonia CXR images.

c) Dataset C[1] is collected by the authors from 3 different hospitals from Uttar Pradesh and Rajasthan, India.
   - 188 images (28 COVID-19, 83 non-COVID, 77 healthy images were collected from King George's Medical University (K.G.M.U.), Lucknow, Uttar Pradesh, India.
   - 68 images of COVID-19 patients were collected from Uttar Pradesh University of Medical Sciences (U.P.U.M.S.), Saifai, Etawah, Uttar Pradesh, India.
   - 543 X-ray images (326 COVID-19, 217 Non-COVID) from Government Medical College, Kota, Rajasthan, India.

CXR images from all the databases are divided into training, validation, and test sets. Training and validation sets were split in ratio of 7:3. Test dataset contains 3112 samples for multi-class classification in 4 classes (194 COVID vs. 583 normal vs. 1772 bacterial pneumonia vs. 493 viral pneumonia cases) whereas 3042 samples for binary classification and COVID-19 detection (194 positive and 2848 non-COVID samples). Sample CXR images of COVID-19, healthy, viral pneumonia, bacterial pneumonia is shown in Fig. 1-4.

[1]https://drive.google.com/drive/folders/1TILc4dLrpZdfuPUmvG0RzUvsMQ-PKgAj?usp=sharing ()

Fig. 1: Chest X-ray example images of a healthy person (Dataset B and C)

Fig. 2: Chest X-ray example images of COVID-19 patient (Dataset A and C)

Fig. 3: Chest X-ray example images of viral pneumonia patient (Dataset B)

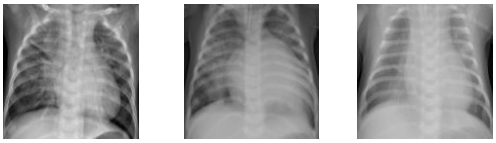
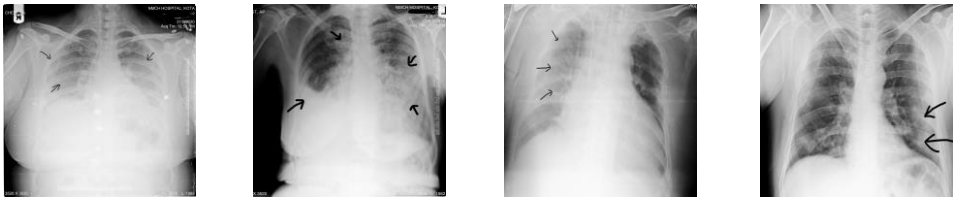


Fig. 4: Chest X-ray example images of Bacterial pneumonia patient (Dataset B)

CXR images of COVID-19 are patched or with opacities which almost look the same as viral pneumonia images. At the initial stage of the COVID-19 infection, images do not indicate any kind of abnormalities. Though with the increase of viruses the images gradually become unilateral. The lower zone and the mid-zone of the lung started transforming into patchy and smudged.

2.2 Clinical perspective of X-ray images for COVID-19 Detection

Bilateral and peripheral opacities (areas of hazy opacity) are the common trademark features of COVID-19 affected patients X-ray report [27] with consolidations of the lungs (compressible lung tissue filled with fluid instead of air). The presence of air space opacities in more than one lobe is unlikely bacterial pneumonia since bacterial pneumonia is likely to be unilateral and involves a single lobe [28]. Other significant signs for COVID-19 pneumonia are consolidation, peripheral, and diffused air space opacities. Initially, the researcher of COVID-19 found the air-space disease likely to have a lower lung distribution and is most commonly bilateral and peripheral [29]. These kinds of peripheral lung opacities have also characteristics to be confluent, either patchy or, multifocal, and can be easily recognized on CXR images. Diffused lung opacities in COVID-19 patients have a similar pattern of CXR as other prevalent inflammatory or infectious processes such as in ARDS. Some other rare findings in COVID-19 affected patients are pneumothorax, lung cavitation, and pleural effusion (water in pleural spaces of the lung) [30]. It is mostly, if at all found at the later stage of the disease. Some of the COVID-19 CXR features are depicted in Fig. 5.



(a) (b) (c) (d)

Fig. 5: Chest X-ray images of COVID-19 infected patients (a) diffuse ill-defined hazy opacities (black arrows) (b) diffuse lung disease and right pleural effusion (black arrows) (c) subtle ill-defined hazy opacities in right side (black arrows) (d) patchy peripheral left mid to lower lung opacities (black arrows)

The conceptual schematic diagram of the proposed work is given in Fig. 6. Once the model is trained using a deep learning algorithm it can be utilized for rapid screening in health care centres. Mobile van-based screening can be performed in hot spot areas and public places. For pre-screening, a digital X-ray machine is required to get CXR.

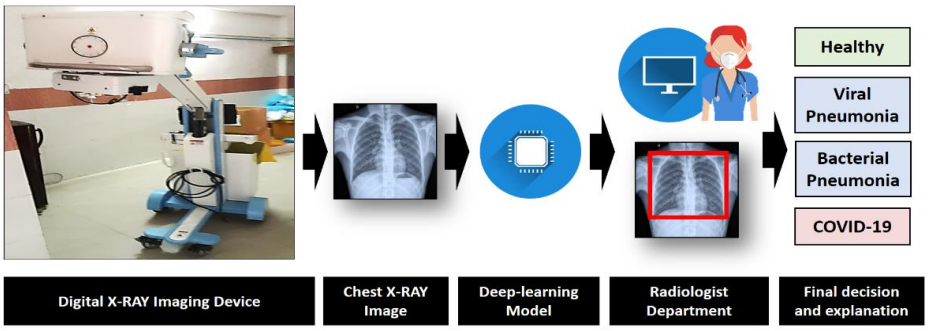


Fig. 6: Conceptual schematic representation of the proposed COVID-19 screening framework

Thereafter, the image can be tested on any computing device which contains the proposed model. The model can classify the image in 0.137 seconds. It can classify thousands of images on a single click and generate a report.

2.3 Methodology

Deep learning (DL) is a part of machine learning which is utilized to solve complex problems with the state-of-the-art performance on computer vision and image processing [31]. DL methods are widely used for medical imaging giving a high performance in segmentation, classification, and detection tasks including breast cancer detection, tumour detection and skin cancer detection [32]. The block diagram for dataset preparation, training and analysing using the deep learning model in the proposed work is depicted in Fig. 7. All the collected images from various sources of different countries were merged into one large dataset. The most of the collected samples were in Digital Imaging and Communications in Medicine (DICOM) format with extension ".dcm". All digital X-ray files were converted in one common image format. The samples were then pre-processed where CXR images were cropped to remove redundant portions and resized to fit better to dimensions of used artificial neural networks. Augmentation was carried out which not only increases the dataset but gives robustness to the trained model and mitigates the occurrence of overfitting problems. Rotation, shear, scaling, flips, and shifts are few of the augmentation techniques which were used to prepare the model to increase the efficiency of test images in a different orientation. Dataset was labelled in different classes as per the opinion of the medical experts which are annotated and categorized accordingly. CXR images are annotated manually for proper training and bounding boxes made around the targeted area. The respective information about the labels and area is saved. After dividing the dataset of CXR in training, test, and validation set, deep learning models are trained for multiple iterations with the prepared dataset. The validation set is used for tuning the parameters and to escape overfitting of the training model. After sufficient training, the model adjusts its weights and the final trained model is tested on the new set of CXR images of various categories which were analysed for performance evaluation of the trained model.

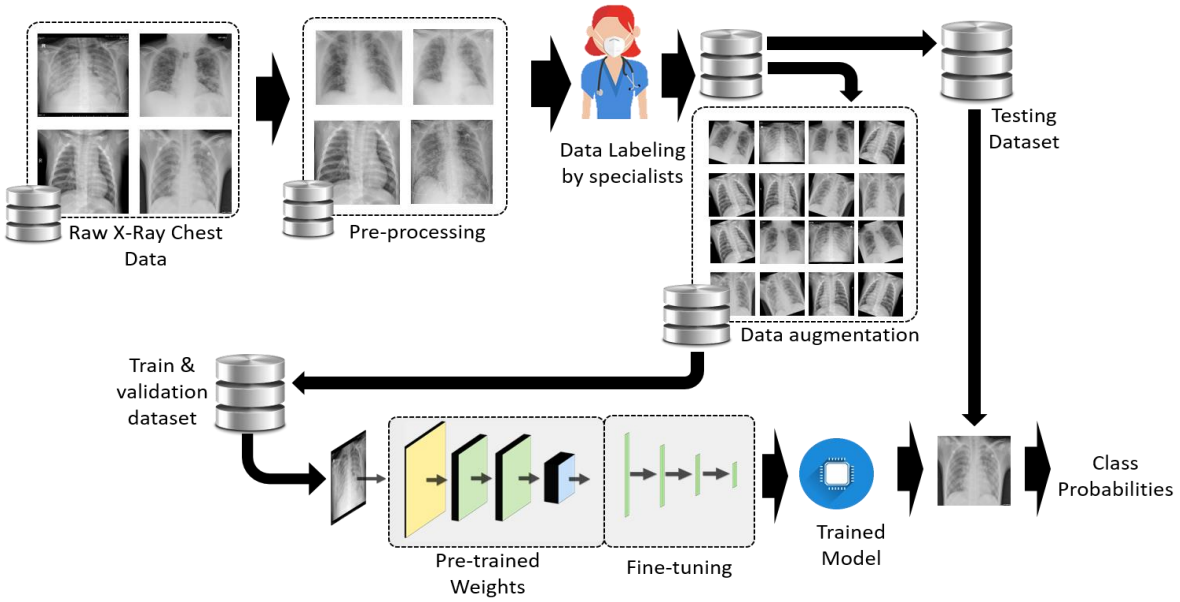


Fig. 7: Methodology of training and testing of the deep learning based COVID-19 detection algorithm

For the image recognition and classification tasks, various architectures of convolutional neural networks or CNNs have proven their accuracy and are used widely. CNNs are commonly composed of multiple building blocks of layers consisting of convolution layer, activation layers, pooling layers, and fully connected (FC) layers which are designed to learn spatial hierarchies of features automatically and adaptively through backpropagation to perform vision task. The convolutional layer is an important part of the deep learning neural network, which extracts the common features from the input images. Input images are convolved with a filter or kernel to generate a convolved feature matrix using different strides. After convolution, the output is passed through an activation function (ReLU, Tanh, or Sigmoid). The activation layer is used to increase non-linearity without effecting its receptive field. Convolution layers are interleaved with pooling layers that is used to decrease the spatial size of the convolved feature matrix. It looks for a larger area of the input image matrix and takes aggregate information (maximum, average, and sum). FC layer is a dense layer which is the final learning phase of CNN architecture performing classification tasks.

Training Model

Architecture selection for backbone network plays an important role in feature extraction in object detection tasks. Stronger the backbone network stronger the detection speed and accuracy of the detection result. DarkNet-53 is used as a backbone network which consists of 53 layers pre-trained on ImageNet [33]. Instead of random weights for initialization of training of the model, pre-trained weights using transfer learning is used in the proposed work which reduces the training time and make more efficient training. The DarkNet-53 network composed of 3 × 3 and 1 × 1 filters with shortcut connections. To perform detection tasks, 53 additional layers are merged with DarkNet layers resulting in a total of 106 network layers. The considered YOLO-v3 based-architecture [34] for the proposed work with processed data to train with different CXR images of various classes, is shown in Fig. 8. This neural network architecture provides high speed of detection and desired precision. Due to multiscale search, it can detect large or as well as smaller objects. DarkNet-53 reaches the highest measured floating-point operations per second resulting in higher utilization of GPU by the structure of the network, which offers higher performance.

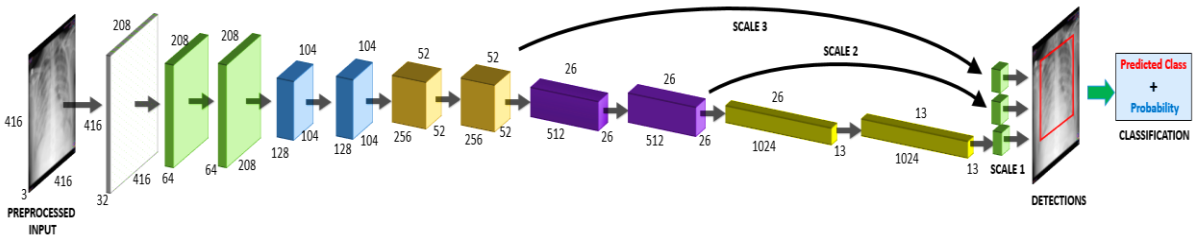


Fig. 8: Architecture of proposed deep learning model for COVID-19 detection with processed dataset

While training, input CXR images are converted into multi-level perceptions after passing through various CNN layers and these are flattened into a column vector and transferred into the FC layer to detect and classify different diseases. For training the proposed model with CXR images in a computationally efficient manner, an advanced form of ReLU activation function, leaky ReLU (LReLU) is used. LReLU activation function saves the value of gradients from getting saturated in case of constant negative bias alike in ReLU. Instead of pruning the negative part to completely zero (as ReLU does), the negative part is multiplied by α which is a small constant value and non-zero number, usually taken as 0.01. The output of the LReLU activation function used in the trained model can be represented as:

$$R(x) = \begin{cases} x, & if\ x > 0 \\ \alpha x, & if\ x \leq 0 \end{cases} \quad (1)$$

In the proposed methodology, max-pooling is utilized together with convolutional layers for extraction of sharp features such as edges from input CXRs. In max-pooling, the maximum value from the rectified feature map is selected. For a CNN architecture, where s is the pooling size and f is pooling function, the output feature on j^{th} local receptive for i^{th} pooling layer is:

$$X_j^i = f(X_j^{i-1}, s) \quad (2)$$

Binary cross-entropy loss is used during training for the class predictions of input CXR images. The input CXR images are divided into N × N grids. In the proposed work, predictions were made on three different scales as shown in Fig. 8. Thus, an input CXR image of 416 x 416 dimension is divided into grids of 13 x 13, 26 x 26 and 52 x 52 for the respective stride values of 32, 16, and 8. The grid cells are responsible for detecting the objects

if the centres of the objects lie in those grid cells. The grid cells predict bounding boxes and determine the confidence score associated with those boxes. The confidence score describes the confidence of the model that the object lies in the box and the accuracy of the box is predicted. Each grid in the input CXR image predicts B bounding boxes with confidence scores, as well as C class conditional probabilities. Confidence score formula is given in equation 3:

$$Confidence = pred(Obj) * IoU_{pred}^{truth} \quad (3)$$

where, IoU_{pred}^{truth} represents the common value in between predicted and reference bounding box and $pred(Obj) = 1$, if the target is in the grids otherwise it would be 0.

Detection of the targeted object for input CXR image in the proposed methodology is shown in Fig. 9. For a CXR image of size 416 × 416 pixels and stride of 32, the input image is divided into 13 × 13 cells. The cell which contains the centre of ground truth box is responsible for predicting the trained object class of CXR. The red grid cell in Fig. 9 is depicting the center of ground truth which responsible for detecting COVID-19 related features.

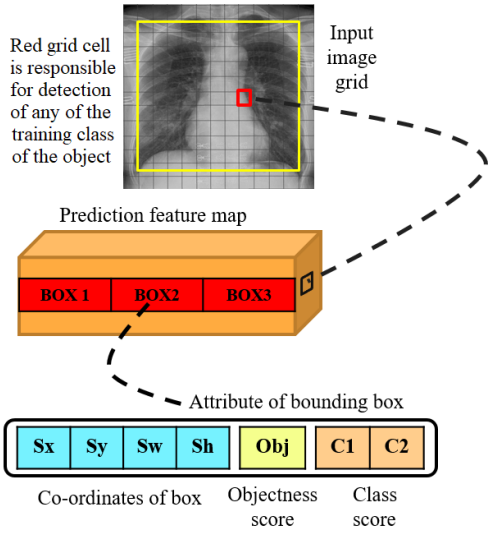


Fig. 9: Detection task through trained model in an image

Detection takes place at three different scales like feature pyramid network (FPN) [35] which is done by downsampling the input image dimensions by 32, 16, and 8. Detection at three different scales makes the deep learning model detect the smallest objects. The last layer of the training network performs bounding box and class prediction for an input CXR image. Attributes of the bounding box contain co-ordinate points of the bounding box, objectness score, and target classes (COVID-19 and non-COVID). Objectness score is defined as the likelihood of containing the targeted object in a given bounding box. Objectness score is calculated by logistic regression for each bounding box and it should be one if ground truth object has more overlapping of bounding box prior as compared to others. The best bounding box is selected out of multiple bounding boxes with the help of non-maximum suppression (NMS). It suppresses less likely bounding box and keep the best bounding box. NMS considers objectness score and intersection over union (IoU) parameters of the bounding box where IoU is the ratio between area of overlap and area of union of the predicted bounding box and true bounding box. NMS chooses the box with highest score for multiple iterations and eliminate higher overlapping bounding boxes after the computation of overlap with other boxes. The deep neural network computes four coordinate points for each bounding box, S_x, S_y, S_w, S_h. Then, corresponding predictions for respective x-coordinate, y-coordinate, width and height of bounding box represented by B_x, B_y, B_w, and B_h, respectively calculated as-

$$B_x = \sigma(S_x) + C_{Ox} \quad (4)$$

$$B_y = \sigma(S_y) + C_{Oy} \quad (5)$$

$$B_w = B_{pw}\, e^{Sw} \quad (6)$$

$$B_h = B_{ph}\, e^{Sh} \quad (7)$$

where, the cell is offset from the top left corner of the image by (C_{Ox}, C_{Oy}). B_{pw} and B_{ph} are bounding box width and height prior, respectively. For training, the dataset sum of squared error loss is used. Logistic regression is used for predicting each class score and threshold for multiple labels for multi-labels prediction on CXR images. Objects which has higher class score value than the defined threshold value are assigned to the respective bounding box.

For an input CXR image for testing the trained model predicts single or multiple detection results based on the values of threshold probability (0.5 in the proposed work), the best one is chosen as output to make screening processes rapid for a large number of testing samples.

3. EXPERIMENTAL RESULTS AND DISCUSSION

For performing all the training and testing of the model python language is used Intel Xenon processor with graphics processing unit (GPU). Considering the memory limitations of the server, the batch size of the training model is taken as sixteen. Other factors such as momentum to accelerate network training, an initial learning rate to affect the speed at which the algorithm reaches the optimal weights, weight decay to regularize the training model with complete software and hardware specifications for training different CXR images are given in Table 1.

TABLE 1
PARAMETERS FOR TRAINING A DEEP LEARING MODEL

Name	Parameters
Development Environment	Anaconda, Jupyter Notebook, Tensorflow, Keras, OpenCV
Processor	Intel Xenon Gold 5218 CPU @ 2.30GHz, 2.29GHz
Installed RAM	64 GB
Operating System	Windows 10, 64 bit
Graphics	NVIDIA, Quadro P600
Graphics Memory	24 GB
Programming Language	Python
Input	Image Dataset
Input dimension	416*416
Batch Size	16
Decay	0.0001
Initial Learning Rate	0.001 (will reduced to 10^{-2} times after every 50000 steps)
Momentum	0.9
Epochs	250
Optimization algorithm	Stochastic Gradient Descent (SGD)

Experiments are performed on the dataset collected from different sources for binary classification (COVID-19 vs. Non COVID) and multi-classification of diseases. The statistics of the number of images in the datasets are given in Table 2. Apart from keeping images in the test set, the remaining dataset from all the sources is divided for training and validation in the ratio of 7:3 respectively. Test dataset was kept updating with new CXR images during the experiments while training and validation sets were kept constant. The deep learning model is trained with CXR images of different categories for several iterations until the loss gets saturated. Generated trained models are analysed with multiple images in test dataset to get overall performance.
The following classes are subjected to classification:

1. **Bacterial Pneumonia**
2. **Viral Pneumonia**
3. **COVID-19**
4. **Normal** (Healthy)
5. **Non - COVID** (Combination of Normal, Bacterial Pneumonia, Viral Pneumonia and Abnormal)

TABLE 2
CHEST X-RAY IMAGES IN DIFFERENT DATASETS

Dataset	COVID-19			NON-COVID												Total
				Normal (Healthy)			Viral Pneumonia			Bacterial Pneumonia			Abnormal (Used only for Binary Classification)			
	Train	Valid	Test	Train	Valid	Test	Train	Valid	Test	Train	Valid	Test	Train	Valid	Test	
Dataset A	166	71	0	0	0	0	0	0	0	0	0	0	0	0	0	**237**
Dataset B	0	0	0	700	300	583	700	300	493	700	300	1772	0	0	0	**5848**
Dataset C	159	69	194	54	23	0	0	0	0	0	0	0	161	69	70	**799**
Total	**325**	**140**	**194**	**754**	**323**	**583**	**700**	**300**	**493**	**700**	**300**	**1772**	**161**	**69**	**70**	**6884**

Performance metrics used for the calculation of experiment is:

$$Sensitivity\ or\ Recall = \frac{TP}{TP+FN} \tag{8}$$

$$Specificity = \frac{TN}{TN+FP} \tag{9}$$

$$Precision = \frac{TP}{TP+FP} \tag{10}$$

$$F1\ Score = 2 * \frac{Precision*Recall}{Precision+Recall} \tag{11}$$

$$Accuracy = \frac{TP+TN}{Positive+Negative} \tag{12}$$

where, true positive (TP) is the case where model correctly predicts the positive labelled image, false positive (FP) is the case where the model predicts as COVID-19 although the image is labelled as non-COVID-19. True negative (TN) is when the model correctly predicts negative image and false-negative (FN) is the case where the model incorrectly predicts a positive labelled image. Positive includes true and false positive images where negative is true negative and false negative images. The confusion matrix is used for measuring the performance of the machine learning classification problem. It is a combination of actual and predicted classes. Abnormal cases are those cases which do not belong to COVID-19 and normal category of CXR images. As abnormal images are limited in number in its category so these are considered only for binary (COVID-19 vs. non-COVID) classification and not considered for multi-classification in 3 and 4 classes. First, image data is analysed without augmentation and later the proposed methodology is tested with application of augmentation techniques.

3.1 Binary Classification

The combined database of datasets A, B, and C is grouped to perform binary classification i.e. COVID-19 or non-COVID, which contains 465 (237 images from dataset A and 228 from dataset C) images of COVID-19, 3307 (3000 images from dataset B and 307 from dataset C) CXR of non-COVID. The dataset is divided in the ratio of 7:3 for images of different categories from the collected sources in respective classes. The confusion matrix is shown in Fig. 10.

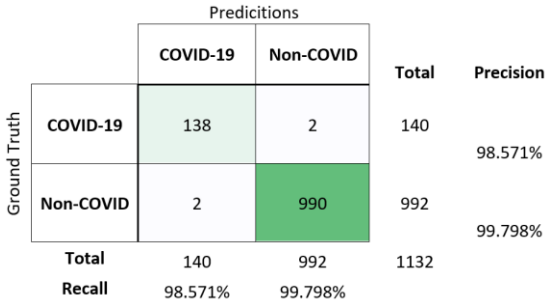


Fig. 10: Confusion matrix for Combined Dataset A, B, and C

After having 5-fold cross validation, overall performance evaluation of detected outputs in binary classification is achieved as in Table 3. The values of TP, TN, FP and FN are averaged for 5-fold cross validation and that mean value is used to calculate those parameters. An accuracy in terms of confidence interval (95%) is achieved as 99.61 ± 0.00%, which shows very less false positives and false negatives cases.

TABLE 3
5-FOLD CROSS VALIDATION RESULT FOR 2-CLASS CLASSIFICATION: COVID-19 VS. NON-COVID ON DATASET (A +B +C)

Samples of other category	Accuracy (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	Precision (95% CI)	F1-Score (95% CI)
Standard Deviation- COVID-19 and non-COVID	0.00	0.84	0.84	1.06	0.96
Overall Results- COVID-19 and non-COVID	99.61 ± 0.00	99.17 ± 1.16	99.17± 1.16	99.05± 1.47	99.10 ± 1.33
Standard Deviation - COVID-19	0.19	2.08	0.11	0.77	0.81
Overall Results for COVID-19	99.61 ± 0.17	98.57 ± 1.83	99.76 ± 0.10	98.30 ± 0.68	98.42 ± 0.71

Tests are performed on new 3112 number of test CXR images belonging to different classes for each of the 5 models received after cross validation. Results of the binary classification on test images are shown in Table 4. The values of TP, TN, FP, and FN are averaged and other parameters values are calculated. Overall results are represented in terms of confidence interval 95%.

TABLE 4
TEST RESULT AFTER CROSS VALIDATION FOR 2-CLASS CLASSIFICATION: COVID-19 VS. NON-COVID ON DATASET (A +B +C)

Class	Samples of testing class	Samples of other classes	TP	TN	FP	FN	Accuracy (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	Precision (95% CI)	F1-Score (95% CI)
COVID-19	194	2918	188.8	2916.8	1.2	5.2	99.79	97.32	99.96	99.37	98.33
Non-COVID	2918	194	2916.8	188.8	5.2	1.2	99.79	99.96	97.32	99.82	99.89
Standard Deviation-All Classes							0.00	1.87	1.87	0.32	1.11
Overall Results- All classes							99.79 ± 0.00	98.64 ± 2.59	98.64 ± 2.59	99.6 ± 0.44	99.11 ± 0.53
Standard Deviation for COVID-19							0.12	2.04	0.02	0.22	1.00
Overall Results for COVID-19							99.79 ± 0.10	97.32 ± 1.79	99.96 ± 0.02	99.37 ± 0.20	98.32 ± 0.88

The confusion matrix for binary classification using averaged values of results obtained from 5 different weights of trained model is shown in Fig. 11.

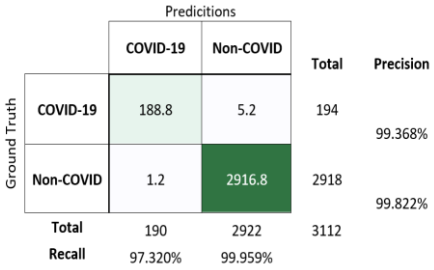


Fig. 11: Confusion matrix for binary classification (COVID-19 vs non-COVID) on test dataset

The testing results for new images gives an accuracy of 99.79 ± 0.00% which is signifying the robustness of the proposed model. Hence this model can be utilized for performing detection and classification of COVID-19 and non-COVID X-ray images of chest.

3.2 Multi-class Classification

The combined database i.e. Dataset A, B, and C which contains 465 images of COVID-19, 1077 normal CXR images, 1000 bacterial pneumonia, and 1000 viral pneumonia images are trained. In this set, CXR images of 77 normal classified people and 228 COVID-19 diagnosed patient data from local hospitals are also involved along with images of the dataset A and B which is divided into training and validation set in the ratio of 7:3. The confusion matrix for multi-class classification is shown in Fig. 12.

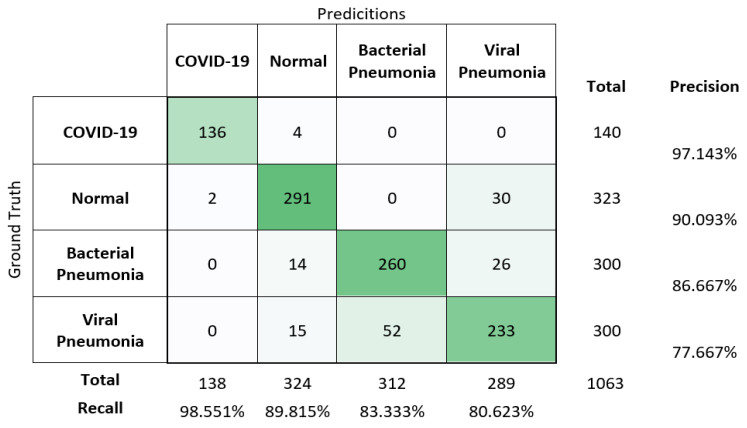


Fig. 12: Confusion matrix of COVID-19 for Combined Dataset of A, B and C for multi-class classification

To better examine the classification model generated by a deep learning algorithm, 5-fold cross-validation is used. The complete CXR image dataset is divided into five different parts and trained for five iterations. The model is trained with the four-fifth part and validated with remaining one-fifth part of CXR image dataset.

After checking the datasets for all Test configurations and evaluating the validation performance, all datasets are chosen for cross-validation to get the actual performance of the model. Confidence interval (CI) is used to analyse the results which gives more information than point estimates. It measures the degree of certainty and uncertainty in a sampling method and gives the range of values which likely to contain the unknown parameter. 95% confidence interval is most commonly used criteria for such estimations. For calculations of confidence intervals, mean and standard deviation for different folds of cross validation are calculated as in equation 13-

$$Confidence\ Interval = \overline{x} \pm \frac{(z* \sigma)}{\sqrt{N}} \quad (13)$$

where, $\overline{x}$ is the mean, σ is standard deviation and N is the sample size. The constant z =1.96 is confidence level value for 95% confidence interval.

After performing 5-fold cross validation for overall performance evaluation for detected outputs in multi-classification are achieved as in Table 5. The accuracy for 95% confidence interval is achieved as 94.79 ±3.81 % whereas the results concerning only the COVID-19 patients the achieved accuracy is for 99.70 ± 0.23%, which shows very less false positives and false negatives cases.

TABLE 5
5-FOLD CROSS VALIDATION FOR 4-CLASS CLASSIFICATION: NORMAL VS. VIRAL PNEUMONIA VS. BACTERIAL PNEUMONIA VS. COVID-19

Samples of other category	Accuracy (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	Precision (95% CI)	F1-Score (95% CI)
Standard Deviation- 4 classes	3.89	9.50	2.52	5.82	6.69
Overall Results- 4 classes (95% CI)	94.79 ±3.81	90.82 ±9.31	96.48 ±2.47	91.35 ±5.70	90.88 ±6.56
Standard Deviation- COVID-19	0.26	1.72	0.09	0.61	0.91
Overall Results for COVID-19	99.70 ± 0.23	98.14 ± 1.51	99.91 ± 0.08	99.42 ± 0.53	98.77 ± 0.80

Testing of trained models is done after different cross-validation folds to authenticate the performance. Trained model after each cross-validation has been tested on new test images and the mean values of different parameters for all 5 trained models are shown in Table 6.

TABLE 6
TEST RESULT AFTER CROSS VALIDATION FOR 4-CLASS CLASSIFICATION: NORMAL VS. VIRAL PNEUMONIA VS. BACTERIAL PNEUMONIA VS. COVID-19 ON DATASET (A +B +C)

Class	Samples of testing category	Samples of other classes	TP	TN	FP	FN	Accuracy (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	Precision (95% CI)	F1-Score (95% CI)
COVID-19	194	2848	187.8	2845.8	2.2	6.2	99.72	96.80	99.92	98.85	97.81
Normal	583	2459	556.4	2365	94	26.6	96.04	95.44	96.18	85.61	90.24
Bacterial Pneumonia	1772	1270	1004.2	1172.8	97.2	767.8	71.56	56.67	92.35	91.18	69.89
Viral Pneumonia	493	2549	356.6	1773.8	743.6	136.4	70.03	72.33	70.45	32.43	44.77
Standard Deviation- 4 classes							15.72	19.35	13.22	30.22	23.74
Overall Results- 4 classes (95% CI)							84.34 ± 15.41	80.31 ± 18.96	89.72 ± 12.95	77.02 ± 29.61	75.68 ± 23.27
Standard Deviation- COVID-19							0.07	0.99	0.06	0.84	0.54
Overall Results for COVID-19							99.72 ± 0.06	96.80 ± 0.87	99.92 ± 0.05	98.85 ± 0.74	97.81 ± 0.47

The confusion matrix for multi-classification using averaged values of results obtained from 5 different weights of trained model is shown in Fig. 13.

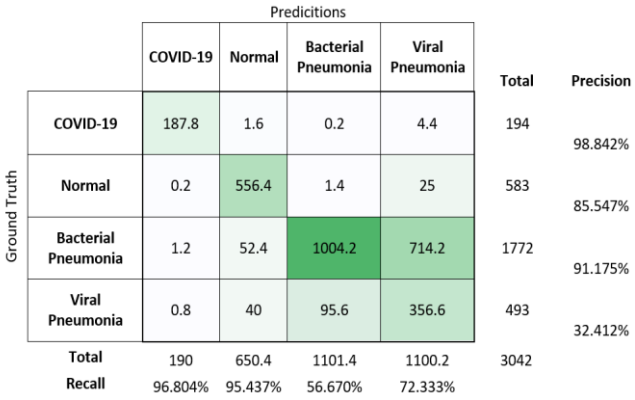


Fig. 13: Confusion matrix for multi-classification on test dataset

If only the COVID-19 result is taken into consideration, the results in a confidence interval (95%) achieved as accuracy of 99.72 ± 0.06%, the sensitivity of 96.80 ± 0.87 %, specificity if 99.92 ± 0.05%, the precision value of 98.85 ± 0.74% and F1-score value of 97.81 ± 0.47%. While considering all four classes the accuracy is achieved as 84.34 ±15.41% in terms of 95% CI. After analysing the testing results, classification of bacterial pneumonia and viral pneumonia is giving lower classification results when compared with other classes, as shown in Table 4. Chaos occurs for a model to classify more precisely the CXR images of viral and bacterial pneumonia. So, the results of 4-class classification are interpreted in the form of 3 class classification where the results of bacterial and viral pneumonia are considered as a single class of pneumonia. After combining both pneumonia classes in one output class, results are significantly improved. Table 7 represents the test result after 5-fold cross-validation for performing three classifications.

TABLE 7
TEST RESULT AFTER CROSS VALIDATION FOR 3-CLASS CLASSIFICATION: NORMAL VS. COVID-19 VS. PNEUMONIA (VIRAL PNEUMONIA + BACTERIAL PNEUMONIA) ON DATASET (A +B +C)

Class	Samples of testing category	Samples of other classes	TP	TN	FP	FN	Accuracy (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	Precision (95% CI)	F1-Score (95% CI)
COVID-19	194	2848	187.8	2845.8	2.2	6.2	99.72	96.80	99.92	98.84	97.81
Normal	583	2459	556.4	2365	94	26.6	96.04	95.44	96.18	85.55	90.22
Pneumonia	2265	777	2170.6	746	31	94.4	95.88	95.83	96.01	98.59	97.19
Standard Deviation- All Classes							2.18	0.70	2.21	7.57	4.21
Overall Results- All classes							97.21 ± 2.46	96.02 ± 0.80	97.37 ± 2.50	94.35 ± 8.57	95.08 ± 4.76

The testing results show the increase in the overall classification results from 84.34 ±15.41% to 97.21 ± 2.46%, signifying high accuracy results. The values of other parameters are also substantially increased. Thus, this consideration can be used for 3 class classification for more surety in case of rapid large scale testing.

3.3 Dataset Augmentation

After performing the augmentation on the training and validation set with rotation (15 degree clockwise and anti-clockwise and scaling (half and double) of images, the dataset of 3772 CXR images is increased to 18860 images, 5 times that of original dataset. After augmentation, CXR images are used for training which were tested on the test dataset of 3112 images for binary and 3042 images for multi-class classification. The results are shown in Table 8, where 97.11 ± 2.71% accuracy is achieved for multi-class (COVID-19 vs. Normal vs. Pneumonia) and 99.81 ± 0.00% accuracy is achieved for multi-class (COVID-19 vs. non-COVID).

TABLE 8
TEST RESULTS FOR BINARY AND MULTI-CLASS CLASSIFICATION AFTER AUGMENTATION

Classification	Class	Samples of testing category	Samples of other categories	TP	TN	FP	FN	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F1-Score (%)
Multi-class	Covid-19	194	2848	191	2847	1	3	99.87	98.45	99.96	99.48	98.96
	Normal	583	2459	555	2359	100	28	95.79	95.20	95.93	84.73	89.66
	Pneumonia	2265	777	2164	746	31	101	95.66	95.54	96.01	98.59	97.04
	CI (95%)							**97.11 ± 2.71**	**96.40± 2.02**	**97.30 ± 2.61**	**94.27 ± 9.36**	**95.22 ± 5.56**
	Standard Deviation							**2.39**	**1.79**	**2.30**	**8.27**	**4.91**
Binary (COVID-19 vs. Non-COVID)	COVID-19	194	2918	191	2915	3	3	99.81	98.45	99.90	98.45	98.45
	Non-COVID	2918	194	2915	191	3	3	99.81	99.90	98.45	99.90	99.90
	CI (95%)							**99.81 ± 0.00**	**99.18 ± 1.42**	**99.18 ±1.42**	**99.18 ± 1.42**	**99.18 ± 1.42**
	Standard Deviation							**0**	**1.03**	**1.03**	**1.03**	**1.03**

The performance comparison table for all kind of classification with and without augmentation is given in Table 9. The dataset augmentation enhanced the performance of the proposed methodology. Binary classification is giving the best output among other.

TABLE 9
COMPARISION OF THE OBTAINED TEST RESULTS USING THE PROPOSED METHODOLOGY

Augmentation	Classification	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F1-Score (%)
Without Augmentation	Multi-class (4 classes)	84.30 ± 6.15	84.34 ± 15.41	80.31 ± 18.96	89.72 ± 12.95	77.02 ± 29.61
	Multi-class (3 classes)	97.21 ± 2.46	96.02 ± 0.80	97.37 ± 2.50	94.35 ± 8.57	95.08 ± 4.76
	Binary (COVID-19 vs. Non-COVID)	99.79 ± 0.00	98.64 ± 2.59	98.64 ± 2.59	99.6 ± 0.44	99.11 ± 0.53
With Augmentation	Multi-class (3 classes)	97.11 ± 2.71	96.40± 2.02	97.30 ± 2.61	94.27 ± 9.36	95.22 ± 5.56
	Binary (COVID-19 vs. Non-COVID)	99.81 ± 0.00	99.18 ± 1.42	99.18 ±1.42	99.18 ± 1.42	99.18 ± 1.42

All these generated models can be used in primary health care centres for performing the screening test on CXR images respectively. It can be utilized at places where there is a lack of availability of expert radiologists and also can assist them to make an accurate diagnosis whenever there are more patients. The model can be utilized as a real-time screening device that has an average detection time of 0.137 seconds per image for detection (NVIDIA Quadro P600) and its classification from input CXR images with dimensions of 416*416 pixels in a system with 6 GB GPU. Some of the detected CXR images are shown in Fig. 14.

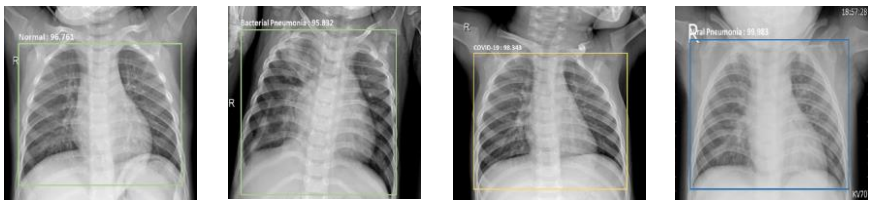


Fig. 14. Prediction results of trained model on augmented dataset for multi-classification

Table 10 compares different existing works for diagnosis of COVID-19 detection using CXR images which gives a reference of some similar existing and reported methods. The proposed method achieved good accuracy with low time complexity for detection of COVID-19 using CXR images, which is encouraging.

TABLE 10
COMPARISON WITH STATE-OF-THE-ART METHODS

Work	Dataset	Methodology	Classification	Time (in seconds)	Performance Metrics (%)
[13]	70 COVID-19, 1008 Pneumonia	ResNet-18	Binary class	NA	Sen=96 Spe=70.65
[16]	266 COVID-19, 8,066 Normal, 5,538 Pneumonia	COVID-Net	3-class	NA	Acc=93.3 Sen=91
[15]	284 Covid-19, 310 Normal, 330 Bacterial Pneumonia, 327 Viral Pneumonia Images	CoroNet	4-class	NA	Acc=89.6 Pre=97 F1-Score=98
			3-class		Acc=99 Pre=95 F1-Score=95.6
[23]	127 COVID-19, 500 Normal, 500 Pneumonia Images	DarkCovidNet (CNN)	Binary class, 3-class	< 1 seconds	2-classes: Acc=98.08 Spe=95.3 Sen=95.13 Pre=98.03 F1-Score=96.51
					3 classes: Acc=87.02 Spe=92.18 Sen=85.35 Pre=89.96 F1-Score=87.37
[36]	455 COVID-19, 2109 Non-COVID Images	MobileNet V2	Binary class	NA	Acc=99.18 Sen=97.36 Spe=99.42
[37]	231 Covid19, 1583 Normal, 2780 Bacterial Pneumonia, 1493 Viral Pneumonia	Inception ResNetV2	3-class	0.1599	Acc=92.18 Sen=92.11 Spec=96.06 Pre=92.38 F1-Score=92.07
[38]	224 Covid-19, 504 Normal, 400 Bacteria Pneumonia, 314 Viral Pneumonia	MobileNet	Binary class	NA	Acc=96.78 Sen=98.66 Spe=96.46
[39]	180 COVID-19, 8851 Normal, 6054 Pneumonia	Concatenation of Xception and ResNet50V2	3-class	NA	Acc=91.4
[40]	250 COVID-19, 3520 Normal, 2753 Other Pulmonary Diseases	VGG-16	Binary class	2.5	Acc=97 Sen=87 Spe=94

[41]	305 COVID-19, 1888 Normal, 3085 Bacterial Pneumonia, 1798 Viral Pneumonia	Stacked MultiResolution CovXNet	Binary class	NA	Acc=97.4 Spe=94.7 F1-score=97.1 Recall=97.8 Pre=96.3 AUC=96.9
Proposed Approach	659 COVID-19, 1660 Normal, 1493 Viral Pneumonia, 2772 Bacterial Pneumonia, 300 Abnormal	YOLO-v3, DarknNet-53	Binary class	0.137	Acc= 99.81 ± 0.00, Sen=99.18 Spe= 99.18 Precision = 99.18 F1-score=99.18
			3-class		Acc= 97.11± 2.71, Sen=96.40 Spe= 97.30 Precision = 94.27 F1-Score=95.22
			4-class		Acc= 84.34 ± 15.41, Sen=80.31, Spe= 89.72, Precision = 77.02, F1-Score=75.68

Authors are aware that the datasets used in these various studies are not the same and that confrontation of the different methods should be done on the same database. Since there is no access to these data, this task remains difficult so dataset is made open for further research.

3.4 Misclassified images

On the analysis of the experimental results, it was found that most of the misclassified images were low-quality images or have some artefacts. Table 9 includes some of the images classified either as false-positives or false-negatives and has a clinical input given by radiologists which may be a reason for misclassification.

TABLE 11
CLINICAL INPUT BY RADIOLOGIST FOR MISCLASSIFIED IMAGES

IMAGES	GROUND TRUTH	PREDICTION	CLINICAL INPUT
	COVID-19	Normal	X-ray image of pediatric patient has less filed of lung as compare with mediastinum so the soft ware learning algorithm picks up as normal.
	COVID-19	Normal	No explanation has to correlate with chest auscultation findings
	Normal	COVID-19	X-ray image have an area of retro cardiac opacity and cardiac silhouettes deviation so the software learning algorithm may have picked up as COVID-19.
	Bacterial Pneumonia	COVID-19	X-ray image have hilar lymph nodes and peripheral opacity so the software learning algorithm may have picked up as COVID-19.
	COVID-19	Normal	No explanation has to correlate with chest auscultation findings

In Table 9, three COVID-19 images were classified as normal whereas one bacterial image is misclassified as COVID-19, and one normal image is classified as COVID-19. Clinically it was found that since the lungs of children are not fully developed, it is difficult to predict the diseases using their CXR image. Short inputs from the clinical point of view of doctors are included in the Table as a possible reason of misclassification.

4. CONCLUSION AND FUTURE WORK

The 2019 novel coronavirus (COVID-19) pandemic appeared in Wuhan, China in December 2019 and has become a serious public health problem worldwide. In the proposed work a deep learning algorithm-based model is proposed for pre-screening of COVID-19 with CXR images. To make the system robust, the model is trained

with a dataset of chest X-ray images collected from local hospitals of India and also from countries like Australia, Belgium, Canada, China, Egypt, Germany, Iran, Israel, Italy, Korea, Spain, Taiwan, USA, and Vietnam. The database is been manually processed and trained with a deep convolutional neural network. In order to detect COVID-19 at an early stage, this study uses transfer learning methods. The performance of the developed convolutional neural network model after 5 fold cross validation was giving the accuracy of 99.61 ± 0.11% for binary classification (*is or is not COVID-19 disease*) using 1132 CXR image samples and accuracy of 94.79 ± 1.59% for multi-class classification of COVID-19, normal, bacterial pneumonia, and viral pneumonia using 1063 CXR image samples. The test accuracy for the augmented dataset is achieved as 97.11 ± 2.71% on 3112 CXR images samples for 3-class classification and 99.81 ± 0.00% for binary (COVID-19 vs. non-COVID) classification on 3042 different CXR image samples. It shows the proposed model has a high accuracy to identify COVID-19 cases from other categories. Since in the current scenario identification of the COVID 19 is the most important task, the experimental results of the proposed model support the COVID 19 identification with very high accuracy. For the future, the model can be trained with images of more diseases to make an automatic prediction for those diseases.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this research paper.

References

[1] COVID-19 Dashboard by the Centre for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU)". ArcGIS. Johns Hopkins University. Retrieved 16th September 2020. (Web Link: https://gisanddata.maps.arcgis.com/apps/opsdashboard/index.html#/bda7594740fd40299423467b48e9ecf6)

[2] Wu F, Zhao S, Yu B et al. "A new coronavirus associated with human respiratory disease in China". Nature. 2020; (published online Feb 3.) DOI: 10.1038/s41586-020-2008-3 Crossref Scopus (124) Google Scholar.

[3] F. Zhou *et al.*, "Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study," *Lancet*, vol. 395, no. 10229, pp. 1054–1062, Mar. 2020, doi: 10.1016/S0140-6736(20)30566-3.

[4] Bernheim A, Mei X, Huang M, et al. Chest CT findings in coronavirus disease-19 (COVID-19): relationship to duration of infection. Radiology 2020:200463

[5] Rodrigues JC, Hare SS, Edey A, Devaraj A, Jacob J, Johnstone A, McStay R, Nair A, Robinson G. An update on COVID-19 for the radiologist-A British society of Thoracic Imaging statement. Clinical radiology. 2020 May 1;75(5):323-5.

[6] Bai Y, Yao L, Wei T, et al. Presumed Asymptomatic Carrier Transmission of COVID-19. JAMA. 2020; 323(14):1406–1407. doi:10.1001/jama.2020.2565

[7] Whiting, P.; Singatullina, N.; Rosser, J. H. Computed Tomography of the Chest: I. Basic Principles. Contin Educ Anaesth Crit Care Pain 2015, 15 (6), 299−304.

[8] Li M, Lei P, Zeng B, et al. Coronavirus Disease (COVID-19): Spectrum of CT Findings and Temporal Progression of the Disease. Acad Radiol. 2020;27(5):603-608. doi:10.1016/j.acra.2020.03.003

[9] LeCun et al. Deep learning. Science, 2015

[10] A Ulhaq, A Khan, D Gomes, M Paul, "Computer vision for COVID-19 control: a survey," engrXiv, 1-24, 2020

[11] F. Shi et al., "Review of Artificial Intelligence Techniques in Imaging Data Acquisition, Segmentation and Diagnosis for COVID-19," in IEEE Reviews in Biomedical Engineering, doi: 10.1109/RBME.2020.2987975.

[12] A. Narin, C. Kaya, and Z. Pamuk, "Automatic detection of coronavirus disease (COVID-19) using X-ray images and deep convolutional neural networks," arXiv:2003.10849, 2020 (Zonguldak Bulent Ecevit University, Turkey)

[13] J. Zhang, Y. Xie, Y. Li, C. Shen, and Y. Xia, "COVID-19 screening on Chest X-ray images using deep learning based anomaly detection," arXiv:2003.12338, 2020. (China)

[14] Khalid et al, "Automated Methods for Detection and Classification Pneumonia based on X-Ray Images Using Deep Learning" https://arxiv.org/abs/2003.14363v1

[15] Asif Iqbal Khan, Junaid Latief Shah, Mohammad Mudasir Bhat, CoroNet: A Deep Neural Network for Detection and Diagnosis of COVID-19 from Chest X-ray Images, Computer Methods and Programs in Biomedicine, 2020, 105581, ISSN 0169-2607, https://doi.org/10.1016/j.cmpb.2020.105581.

[16] Linda Wang, Zhong Qiu Lin, and Alexander Wong, "COVID-Net: A Tailored Deep Convolutional Neural Network Design for Detection of COVID-19 Cases from Chest X-Ray Images" arXiv:2003.09871v3 [eess.IV] 15 Apr 2020

[17] F. Ucar and D. Korkmaz, "Covidiagnosis-net: Deep bayes-squeezenet based diagnostic of the coronavirus disease 2019 (covid-19) from x-ray images," Medical Hypotheses, p. 109761, 2020.

[18] B. Ghoshal and A. Tucker, "Estimating uncertainty and interpretability in deep learning for coronavirus (COVID-19) detection," arXiv: 2003.10769, 2020. (Brunel University, London, United Kingdom)

[19] S. Vaid, R. Kalantar, and M. Bhandari, "Deep learning covid-19 detection bias: accuracy through artificial intelligence," International Orthopaedics, p. 1, 2020.

[20] L. Brunese, F. Mercaldo, A. Reginelli, and A. Santone, "Explainable deep learning for pulmonary disease and coronavirus covid-19 detection from x-rays," Computer Methods and Programs in Biomedicine, p. 105608, 2020.

[21] F. Shi, L. Xia, F. Shan, D. Wu, Y. Wei, H. Yuan, et al., "Large-scale screening of COVID-19 from community acquired pneumonia using infection size-aware classification," arXiv:2003.09860, 2020. (China)

[22] S. Khobahi, C. Agarwal, and M. Soltanalian, "CoroNet: A Deep Network Architecture for Semi-Supervised Task-Based Identification of COVID-19 from Chest X-ray Images," *medRxiv*, p. 2020.04.14.20065722, Jan. 2020, doi: 10.1101/2020.04.14.20065722. (University of Illinois, Chicago)

[23] T. Ozturk, M. Talo, E.A. Yildirim, U.B. Baloglu, O. Yildirim, U.R. AcharyaAutomated detection of COVID-19 cases using deep neural networks with x-ray images Comput. Biol. Med. (2020), p. 103792.
[24] Asmaa Abbas, Mohammed M Abdelsamea, and Mohamed Medhat Gaber. Classification of COVID-19 in chest x-ray images using detrac deep convolutional neural network. arXiv preprint arXiv:2003.13815, 2020.
[25] Joseph Paul Cohen and Paul Morrison and Lan Dao COVID-19 image data collection, arXiv: 2003.11597, 2020 https://github.com/ieee8023/COVID-chestxray-dataset.
[26] Chest X-Ray Images (Pneumonia) https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia.
[27] Kobayashi, Y.; Mitsudomi, T. Management of Ground-Glass Opacities: Should All Pulmonary Lesions with Ground-Glass Opacity Be Surgically Resected? Transl. Lung Cancer Res. 2013, 2 (5), 354−363.
[28] Vilar J, Domingo ML, Soto C et al. Radiology of Bacterial Pneumonia. J.Eur J Radiol. 2004 Aug;51(2):102 -13.
[29] Wong HYF, Lam HYS, Fong AH et al. Frequency and Distribution of Chest Radiographic Findings in COVID -19 Positive Patients. Radiology. 2019 Mar 27:201160. doi: 10.1148/radiol.2020201160.
[30] Salehi S, Abedi A, Balakrishnan S et al. Coronavirus Disease 2019 (COVID -19): A Systematic Review of Imaging Findings in 919 Patients. AJR Am J Roentgenol . 2020 Mar 14:1 -7. doi: 10.2214/AJR.20.23034.
[31] M.Z. Alom, T.M. Taha, C. Yakopcic, S. Westberg, SM. Hasan, B.C. Van Esesn, A.A.S. Awwal and V.K. Asari, The History Began from AlexNet: A Comprehensive Survey on Deep Learning Approaches. arXiv preprintarXiv:1803.01164, 2018.
[32] G. Litjens, T. Kooi, B.E. Bejnordi, A.A.A. Setio, F. Ciompi, M. Ghafoorian, J.A.W.M. Van der Laak B. Van Ginneken and C.I. Sánchez, A survey on deep learning in medical image analysis. Medical image analysis, 42, 60-88, 2017.
[33] K rizhevsky, A.; Sutskever, I.; Hinton, G. ImageNet classification with deep convolutional neural networks. In Proceedings of the 25th International Conference on Neural Information Processing Systems, Lake Tahoe, NV, USA, 3 – 6 December 2012; pp. 1097 –1105.
[34] J. Redmon and A. Farhadi. Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767, 2018.
[35] Lin, T.-Y., Dollar, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature Pyramid Networks for Object Detection. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). doi:10.1109/cvpr.2017.106
[36] I. D. Apostolopoulos, S. I. Aznaouridis, and M. A. Tzani, “Extracting possibly representative covid-19 biomarkers from x-ray images with deep learning approach and image data related to pulmonary diseases,” Journal of Medical and Biological Engineering, p. 1, 2020.
[37] K. Elasnaoui and Y. Chawki, “Using x-ray images and deep learning for automated detection of coronavirus disease,” Journal of Biomolecular Structure and Dynamics, no. just-accepted, pp. 1–22, 2020.
[38] I. D. Apostolopoulos and T. A. Mpesiana, “Covid-19: automatic detection from x-ray images utilizing transfer learning with convolutional neural networks,” Physical and Engineering Sciences in Medicine, p. 1, 2020.
[39] M. Rahimzadeh and A. Attar, “A modified deep convolutional neural network for detecting covid-19 and pneumonia from chest x-ray images based on the concatenation of xception and resnet50v2,” Informatics in Medicine Unlocked, p. 100360, 2020.
[40] L. Brunese, F. Mercaldo, A. Reginelli, and A. Santone, “Explainable deep learning for pulmonary disease and coronavirus covid-19 detection from x-rays,” Computer Methods and Programs in Biomedicine, p. 105608, 2020.
[41] T. Mahmud, M. A. Rahman, and S. A. Fattah, “Covxnet: A multidilation convolutional neural network for automatic covid-19 and other pneumonia detection from chest x-ray images with transferable multireceptive feature optimization,” Computers in Biology and Medicine, p. 103869, 2020.

---------- End of Main Manuscript ----------

Appendix

(Not a part of Main Manuscript)

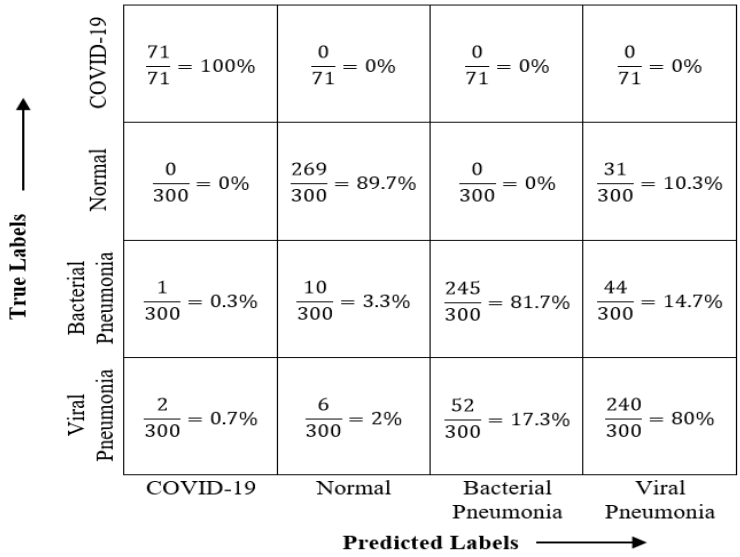
These experimental results will not be a part of the Main Manuscript – Included here for the reference of Reviewers only.

Authors have carried out various experiment's in many possible combinations and permutations – like Classification in only Dataset C, Dataset A+B, Combination of A+B+C, Various combinations like ***COVID vs. Non-COVID****,* ***COVID vs. Pneumonia (Bacterial + Virus Combined) vs. Healthy****,* ***COVID vs. Viral Pneumonia vs. Bacterial Pneumonia vs. Healthy*** *– and the results are included in the Appendix.*

---

1. Performance Evaluation for 4-Class Classification: Normal vs. Viral Pneumonia vs. Bacterial Pneumonia vs. COVID-19 on Dataset (A +B)

Class	Samples of validating class	Samples of other classes	TP	TN	FP	FN	Accuracy (%)	Sensitivity (%)	Specification (%)	Precision (%)	F1 Score (%)
COVID-19	71	900	71	897	3	0	99.69	100.00	99.67	95.95	97.93
Normal	300	671	269	655	16	31	95.16	89.67	97.62	94.39	91.97
Bacterial Pneumonia	300	671	245	619	52	55	88.98	81.67	92.25	82.49	82.08
Viral Pneumonia	300	671	240	596	75	60	86.10	80.00	88.82	90.94	90.66
Mean							**92.48**	**87.83**	**94.59**	**95.95**	**97.93**



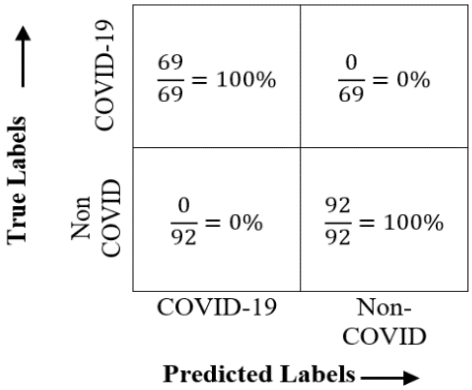
2. Performance Evaluation for 2 Class Classification: COVID-19 vs. Non-COVID on Dataset C

Class	Samples of validating class	Samples of other classes	TP	TN	FP	FN	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F1 Score (%)
COVID-19	69	92	69	92	0	0	100.00	100.00	100.00	100.00	100.00
Non-COVID	92	69	92	69	0	0	100.00	100.00	100.00	100.00	100.00
Mean							100.00	100.00	100.00	100.00	100.00

Appendix

(Not a part of Main Manuscript)

These experimental results will not be a part of the Main Manuscript – Included here for the reference of Reviewers only.



3. Cross Validation for 4-Class Classification: Normal vs. Viral Pneumonia vs. Bacterial Pneumonia vs. COVID-19 on Dataset (A +B +C)

Fold	Class	Samples of validating class	Samples of other classes	TP	TN	FP	FN	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F1-Score (%)
1	COVID-19	140	923	136	921	2	4	99.44	97.14	99.78	98.55	97.84
	Normal	323	740	291	707	33	32	93.89	90.09	95.54	89.81	89.95
	Bacterial Pneumonia	300	763	260	711	52	40	91.35	86.67	93.18	83.33	84.97
	Viral Pneumonia	300	763	233	707	56	67	88.43	77.67	92.66	80.62	79.12
	Mean							**93.27**	**87.89**	**95.29**	**88.08**	**87.97**
2	COVID-19	140	923	138	922	1	2	99.72	98.57	99.89	99.28	98.92
	Normal	323	740	321	706	34	2	96.61	99.38	95.41	90.42	94.69
	Bacterial Pneumonia	300	763	270	725	38	30	93.60	90.00	95.02	87.66	88.82
	Viral Pneumonia	300	763	238	741	22	62	92.10	79.33	97.12	91.54	85.00
	Mean							**95.51**	**91.82**	**96.86**	**92.23**	**91.86**
3	COVID-19	140	923	134	923	0	6	99.44	95.71	100.00	100.00	97.81
	Normal	323	740	322	696	44	1	95.77	99.69	94.05	87.98	93.47
	Bacterial Pneumonia	300	763	269	706	57	31	91.72	89.67	92.53	82.52	85.94
	Viral Pneumonia	300	763	216	742	21	84	90.12	72.00	97.25	91.14	80.45
	Mean							**94.26**	**89.27**	**95.96**	**90.41**	**89.42**
4	COVID-19	140	923	140	923	1	0	100.00	100.00	99.89	99.29	99.64
	Normal	323	740	323	699	41	0	96.14	100.00	94.46	88.74	94.03
	Bacterial Pneumonia	300	763	265	716	47	35	92.29	88.33	93.84	84.94	86.60
	Viral Pneumonia	300	763	228	745	18	72	91.53	76.00	97.64	92.68	83.52
	Mean							**94.99**	**91.08**	**96.46**	**91.41**	**90.95**
5	COVID-19	140	923	139	923	0	1	99.91	99.29	100.00	100.00	99.64
	Normal	323	740	323	715	25	0	97.65	100.00	96.62	92.82	96.27
	Bacterial Pneumonia	300	763	260	732	31	22	93.32	92.20	95.94	89.35	90.75
	Viral Pneumonia	300	763	233	754	9	42	92.85	84.73	98.82	96.28	90.14
	Mean							**95.93**	**94.05**	**97.84**	**94.61**	**94.20**
Overall Results- All classes								**94.79**	**90.82**	**96.48**	**91.35**	**90.88**
Overall Results for COVID-19								**99.70**	**98.14**	**99.91**	**99.42**	**98.77**

Appendix

(Not a part of Main Manuscript)

These experimental results will not be a part of the Main Manuscript – Included here for the reference of Reviewers only.

---

4. Test Result After Cross Validation for 4-Class Classification: Normal vs. Viral Pneumonia vs. Bacterial Pneumonia vs. COVID-19 on Dataset (A +B +C)

Fold	Class	Samples of testing category	Samples of other category	TP	TN	FP	FN	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F1-Score (%)
1	COVID-19	194	2848	190	2846	2	4	99.80	97.94	99.93	98.96	98.45
	Normal	583	2459	548	2379	80	35	96.22	94.00	96.75	87.26	90.50
	Bacterial Pneumonia	1772	1270	1013	1168	102	759	71.70	57.17	91.97	90.85	70.18
	Viral Pneumonia	493	2549	353	1795	754	140	70.61	71.60	70.42	31.89	44.13
	Mean							**84.58**	**80.18**	**89.77**	**77.24**	**75.81**
2	COVID-19	194	2848	188	2843	5	6	99.64	96.91	99.82	97.41	97.16
	Normal	583	2459	560	2390	69	23	96.98	96.05	97.19	89.03	92.41
	Bacterial Pneumonia	1772	1270	960	1178	92	812	70.28	54.18	92.76	91.25	67.99
	Viral Pneumonia	493	2549	371	1752	797	122	69.79	75.25	68.73	31.76	44.67
	Mean							**84.17**	**80.60**	**89.63**	**77.36**	**75.56**
3	COVID-19	194	2848	187	2847	1	7	99.74	96.39	99.96	99.47	97.91
	Normal	583	2459	556	2357	102	27	95.76	95.37	95.85	84.50	89.61
	Bacterial Pneumonia	1772	1270	1014	1163	107	758	71.56	57.22	91.57	90.45	70.10
	Viral Pneumonia	493	2549	345	1661	730	148	65.94	69.98	69.47	32.09	44.01
	Mean							**83.25**	**79.74**	**89.22**	**76.63**	**75.40**
4	COVID-19	194	2848	189	2846	2	5	99.77	97.42	99.93	98.95	98.18
	Normal	583	2459	558	2352	107	25	95.66	95.71	95.65	83.91	89.42
	Bacterial Pneumonia	1772	1270	1024	1170	100	748	72.12	57.79	92.13	91.10	70.72
	Viral Pneumonia	493	2549	352	1839	710	141	72.02	71.40	72.15	33.15	45.27
	Mean							**84.89**	**80.58**	**89.96**	**76.78**	**75.90**
5	COVID-19	194	2848	185	2847	1	9	99.67	95.36	99.96	99.46	97.37
	Normal	583	2459	560	2347	112	23	95.56	96.05	95.45	83.33	89.24
	Bacterial Pneumonia	1772	1270	1010	1185	85	762	72.16	57.00	93.31	92.24	70.46
	Viral Pneumonia	493	2549	362	1822	727	131	71.79	73.43	71.48	33.24	45.76
	Mean							**84.80**	**80.46**	**90.05**	**77.07**	**75.71**
Overall Results- ALL classes								**84.34**	**80.31**	**89.72**	**77.02**	**75.68**
Overall Results for COVID-19								**99.72**	**96.8**	**99.92**	**98.85**	**97.81**

5. Test Result After Cross Validation for 3-Class Classification: Normal vs. COVID-19 vs. Pneumonia (Viral Pneumonia + Bacterial Pneumonia) on Dataset (A +B +C)

Fold	Class	Samples of testing category	Samples of other category	TP	TN	FP	FN	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F1-Score (%)
1	COVID-19	194	2848	190	2846	2	4	99.80	97.94	99.93	98.96	98.45
	Normal	583	2459	548	2379	80	35	96.22	94.00	96.75	87.26	90.50
	Pneumonia	2265	777	2184	739	38	81	96.09	96.42	95.11	98.29	97.35
	Mean							97.37	96.12	97.26	94.84	95.43
2	COVID-19	194	2848	188	2843	5	6	99.64	96.91	99.82	97.41	97.16
	Normal	583	2459	560	2390	69	23	96.98	96.05	97.19	89.03	92.41
	Pneumonia	2265	777	2191	748	29	74	96.61	96.73	96.27	98.69	97.70
	Mean							97.74	96.56	97.76	95.04	95.76
3	COVID-19	194	2848	187	2847	1	7	99.74	96.39	99.96	99.47	97.91
	Normal	583	2459	556	2357	102	27	95.76	95.37	95.85	84.50	89.61
	Pneumonia	2265	777	2166	747	30	99	95.76	95.63	96.14	98.63	97.11
	Mean							97.09	95.80	97.32	94.20	94.87
4	COVID-19	194	2848	189	2846	2	5	99.77	97.42	99.93	98.95	98.18
	Normal	583	2459	558	2352	107	25	95.66	95.71	95.65	83.91	89.42
	Pneumonia	2265	777	2158	749	28	107	95.56	95.28	96.40	98.72	96.97
	Mean							97.00	96.14	97.32	93.86	94.86
5	COVID-19	194	2848	185	2847	1	9	99.67	95.36	99.96	99.46	97.37
	Normal	583	2459	560	2347	112	23	95.56	96.05	95.45	83.33	89.24
	Pneumonia	2265	777	2154	747	30	111	95.36	95.10	96.14	98.63	96.83
	Mean							96.87	95.51	97.18	93.81	94.48
Overall Results- ALL classes								**97.21**	**96.02**	**97.37**	**94.35**	**95.08**
Overall Results for COVID-19								**99.72**	**96.8**	**99.92**	**98.85**	**97.81**

Appendix

(Not a part of Main Manuscript)

These experimental results will not be a part of the Main Manuscript – Included here for the reference of Reviewers only.

---

6. Cross Validation Fold-1 Result for 2-Class Classification: COVID-19 vs. Non-COVID on Dataset (A +B +C)

Fold	Class	Samples of validating class	Samples of other classes	TP	TN	FP	FN	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F1-Score (%)
1	COVID-19	140	992	138	990	2	2	99.65	98.57	99.80	98.57	98.57
	Non-COVID	992	140	990	138	2	2	99.65	99.80	98.57	99.80	99.80
	Mean							**99.65**	**99.18**	**99.18**	**99.18**	**99.18**
2	COVID-19	140	992	140	988	4	0	99.65	100.00	99.60	97.22	98.59
	Non-COVID	992	140	988	140	0	4	99.65	99.60	100.00	100.00	99.80
	Mean							**99.65**	**99.80**	**99.80**	**98.61**	**99.19**
3	COVID-19	140	992	133	991	1	7	99.29	95.00	99.90	99.25	97.08
	Non-COVID	992	140	991	133	7	1	99.29	99.90	95.00	99.30	99.60
	Mean							**99.29**	**97.45**	**97.45**	**99.28**	**98.34**
4	COVID-19	140	992	140	990	2	0	99.82	100.00	99.80	98.59	99.29
	Non-COVID	992	140	990	140	0	2	99.82	99.80	100.00	100.00	99.90
	Mean							**99.82**	**99.90**	**99.90**	**99.30**	**99.59**
5	COVID-19	140	992	139	989	3	1	99.65	99.29	99.70	97.89	98.58
	Non-COVID	992	140	989	139	1	3	99.65	99.70	99.29	99.90	99.80
	Mean							**99.65**	**99.49**	**99.49**	**98.89**	**99.19**
Overall Results- All classes								**99.61**	**99.16**	**99.16**	**99.05**	**99.10**
Overall Results for COVID-19								**99.61**	**98.57**	**99.76**	**98.30**	**98.42**

7. Test Result After Cross Validation for 2-Class Classification: COVID-19 vs. Non-COVID on Dataset (A +B +C)

Fold	Class	Samples of testing category	Samples of other category	TP	TN	FP	FN	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F1-Score (%)
1	COVID-19	194	2918	185	2917	1	9	99.68	95.36	99.97	99.46	97.37
	Non-COVID	2918	194	2917	185	9	1	99.68	99.97	95.36	99.69	99.83
	Mean							**99.68**	**97.66**	**97.66**	**99.58**	**98.6**
2	COVID-19	194	2918	191	2917	1	3	99.87	98.45	99.97	99.48	98.96
	Non-COVID	2918	194	2917	191	3	1	99.87	99.97	98.45	99.9	99.93
	Mean							**99.87**	**99.21**	**99.21**	**99.69**	**99.45**
3	COVID-19	194	2918	184	2917	1	10	99.65	94.85	99.97	99.46	97.1
	Non-COVID	2918	194	2917	184	10	1	99.65	99.97	94.85	99.66	99.81
	Mean							**99.65**	**97.41**	**97.41**	**99.56**	**98.45**
4	COVID-19	194	2918	192	2917	1	2	99.9	98.97	99.97	99.48	99.22
	Non-COVID	2918	194	2917	192	2	1	99.9	99.97	98.97	99.93	99.95
	Mean							**99.9**	**99.47**	**99.47**	**99.71**	**99.59**
5	COVID-19	194	2918	192	2916	2	2	99.87	98.97	99.93	98.97	98.97
	Non-COVID	2918	194	2916	192	2	2	99.87	99.93	98.97	99.93	99.93
	Mean							**99.87**	**99.45**	**99.45**	**99.45**	**99.45**
Overall Results- ALL classes								**99.79**	**98.64**	**98.64**	**99.6**	**99.11**
Overall Results for COVID-19								**99.79**	**97.32**	**99.96**	**99.37**	**98.32**

Appendix: *Authors have carried out various experiment's in many possible combinations and permutations – like Classification in only Dataset C, Dataset A+B, Combination of A+B+C, Various combinations like* ***COVID vs. Non-COVID****,* ***COVID vs. Pneumonia (Bacterial + Virus Combined) vs. Healthy****,* ***COVID vs. Viral Pneumonia vs. Bacterial Pneumonia vs. Healthy*** *– and the results are included in the Appendix.*