



A multi-modal bone suppression, lung segmentation, and classification approach for accurate COVID-19 detection using chest radiographs

Geeta Rani^a, Ankit Misra^a, Vijaypal Singh Dhaka^{a,*}, Deepak Buddhi^c,
Ravindra Kumar Sharma^c, Ester Zumpano^b, Eugenio Vocaturo^b

^a Manipal University Jaipur, India

^b University of Calabria, Italy

^c R.G. Stone Urology and Laparoscopy Hospital, India

ARTICLE INFO

Keywords:

SARS-CoV-2

COVID-19

Deep learning

Biomedical

Medical imaging

ABSTRACT

The high transmission rate of COVID-19 and the lack of quick, robust, and intelligent systems for its detection have become a point of concern for the public, Government, and health experts worldwide. The study of radiological images is one of the fastest ways to comprehend the infectious spread and diagnose a patient. However, it is difficult to differentiate COVID-19 from other pneumonic infections. The purpose of this research is to provide an automatic, precise, reliable, robust, and intelligent assisting system 'Covid Scanner' for mass screening of COVID-19, Non-COVID Viral Pneumonia, and Bacterial Pneumonia from healthy chest radiographs. To train the proposed system, the authors of this research prepared novel a dataset called, "COVID-Pneumonia CXR". The system is a coherent integration of bone suppression, lung segmentation, and the proposed classifier, 'EXP-Net'. The system reported an AUC of 96.58% on the validation dataset and 96.48% on the testing dataset comprising chest radiographs. The results from the ablation study prove the efficacy and generalizability of the proposed integrated pipeline of models. To prove the system's reliability, the feature heatmaps visualized in the lung region were validated by radiology experts. Moreover, a comparison with the state-of-the-art models and existing approaches shows that the proposed system finds clearer demarcation between the highly similar chest radiographs of COVID-19 and Non-COVID viral pneumonia. The copyright of "Covid Scanner" is protected with registration number SW-13625/2020. The code for the models used in this research is publicly available at: https://github.com/Ankit-Misra/multi_modal_covid_detection/.

1. Introduction

The world has been facing a global pandemic since January 2020 due to the COVID-19 outbreak. This is a type of viral pneumonia caused by the Coronavirus 'SARS-CoV-2'. The viral load can be detected in the trachea, bronchi, and lungs. It attacks the respiratory system of a person by latching its spiky surface proteins to the Angiotensin-Converting Enzyme2 (ACE2) receptors. It destroys the healthy cells and multiplies its viral proteins to replicate the virus genome. The virus has a high transmission rate and currently has no anti-viral drug against it (Abdulrahma & Salem, 2020, Saddik et al., 2021). Also, its potential to mutate challenges the efficacy of the vaccines developed till date. According to the data presented by the World Health Organization (WHO), 491,090,990

positive cases of COVID-19 and 6,174,488 deaths have been reported till 3 March 2022. Early diagnosis and isolation of the patients are one of the most effective ways to control the spread of this Coronavirus.

Reverse Transcription-Polymerase Chain Reaction (RT-PCR) is the most widely adopted screening test for COVID-19. The test has a detection sensitivity of 65% to 75%, which is highly dependent on external factors such as temperature maintained for storage and proper transportation of collected samples. Also, only a few pathology labs have the facilities to perform this test; thus it cannot be performed in remote locations where there is a lack of medical facilities. Therefore, there is a need for finding alternative ways of testing. As per the reports and results published in (Harmon et al., 2020), Computer Tomography (CT) scans can be used for the early diagnosis of COVID-19. But, the high

* Corresponding author.

E-mail addresses: geetachhikara@gmail.com (G. Rani), jaimishra2345@gmail.com (A. Misra), vijaypalsingh.dhaka@jaipur.manipal.edu (V.S. Dhaka), e.zumpano@dimes.unical.it (D. Buddhi), e.vocaturo@dimes.unical.it (R.K. Sharma), drdeepdejau@mail.com (E. Zumpano), drravisharma2812@gmail.com (E. Vocaturo).

<https://doi.org/10.1016/j.iswa.2022.200148>

Received 30 June 2022; Received in revised form 8 October 2022; Accepted 3 November 2022

Available online 7 November 2022

2667-3053/© 2022 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

cost of the machine, spatial complexity in 3-Dimensional (3-D) volume, and the need for the sanitization after each screening poses real-time challenges. CT scan delivers about 70 times higher doses of harmful radiation than a digital X-ray. Moreover, CT scan requires to move the patient which is infeasible for intubated patients. Thus, we shift our focus on chest x-rays (CXR), as a viable diagnostic tool. The possibility of scanning via portable X-ray machines does not require moving the intubated patients. Its low cost, high availability, and non-contact mechanism overcome the drawbacks of CT scans.

Although many pioneering works, mentioned in Section - 2, have been published for screening of COVID-19 using CXRs, most of these works focus on the binary classification of CXRs into COVID-19 (CV) and healthy class (N). These works ignore the class interference between pneumonic infections of COVID-19, Bacterial Pneumonia, and Non-COVID Viral Pneumonia. Only a few research works, such as in (Afshar et al., 2020, Mangal et al., 2020), address the problem of screening of COVID-19 from Bacterial pneumonia (BP) and Non-COVID Viral Pneumonia (VP). In a recent work (Zumpano et al., 2021), the authors discriminate cases of Non-COVID Viral Pneumonia from Bacterial Pneumonia and healthy cases using multiple instance learning approach. However, the small size of testing data and lack of clinical validation from radiologists undermines the reliability of their results. Also, many research works put minimum emphasis on data preparation and preprocessing. Bony structures, such as the ribs, collar bones, shoulders, etc. captured in a CXR affect the model training and classification performance. Thus, it is necessary to extract the lung region and reduce the visibility of bones in the soft tissues. Moreover, in the early stages, the nature of infection spread of COVID-19 is similar to Non-COVID viral pneumonia. Thus, it is challenging to distinguish it from the VP class. Also, VP and BP can co-exist. Therefore, it is quite difficult to distinguish these diseases using a CXR.

To address the aforementioned challenges, we developed a COVID-19 screening system that focuses on differentiating COVID-19, Non-COVID Viral Pneumonia, and Bacterial Pneumonia from Normal CXRs in two stages - preprocessing (samples shown in Fig. 2) and classification. The preprocessing stage comprises bone suppression (BS) and lung segmentation (LS) models. The BS model reduces the visibility of ribs and collar bones in the lung region, whereas the LS model extracts the lung region. The classification stage comprises the proposed architecture, 'EXP-Net' that predicts the diseased category. To train this classification model, we built a benchmarking dataset "COVID-Pneumonia CXR", comprising bone suppressed and lung segmented CXRs of COVID-19, Non-COVID Viral Pneumonia, Bacterial Pneumonia, and healthy patients. The dataset is validated by radiological experts. To justify the impact of bone suppression and lung segmentation models on the classification performance, we provide an ablation study as well. We also present a comparison of the proposed classifier 'EXP-Net' with other state-of-the-art classifiers viz DenseNet121, ResNet50, InceptionV3, ResNet101, EfficientNetB2, and Xception net by employing them on the same dataset. Furthermore, we also showcase the difference in performance between our proposed methodology and the existing approaches. The major objectives of this manuscript are as follows.

- Provide a validated and benchmarking dataset for the detection of COVID-19.
- Analyze the impact of preprocessing on the model's classification performance.
- Identify the loss function useful for resolving the problem of class imbalance.
- Compare the performance of the proposed model with state-of-the-art models.
- Validate the reliability of the model by generating feature heatmaps using GradCam ++.

The paper has been structured as follows. In Section - 2, we present the review of related works. In Section - 3, we provide details of the ma-

terials and methods. This section entails details on the data preparation, architectures of the deep learning models, problem formulation, and training details of the proposed system. In Section - 4, we demonstrate the results obtained by employing the preprocessing, and classification models on the prepared dataset. We also show an exhaustive comparative analysis of the performance of the proposed system and the state-of-the-art methods. In Section 5, we present the discussion of the significance of the proposed work. We also elaborate on the advantages and drawbacks of proposal in reference to the state-of-the-art. Finally, the future developments of our research are presented in Section - 6.

2. Related works

The study of related literature shows that many pioneering works have been proposed for the screening of COVID-19. In this section, we present a brief discussion on the most related works on COVID-19 detection.

Ohata et al. (2020) proposed a binary classifier to distinguish between chest X-rays of healthy people and COVID-19 patients. The authors collected a small dataset comprising 194 X-ray images of both COVID-19 and healthy classes. To improve feature representation learning on their collected small dataset, they utilized the potential of transfer learning (Hussain et al., 2018). For classification, they combined the feature extractor with ML techniques - Support Vector Machine (SVM), k-Nearest Neighbor, Random Forest, and Multilayer Perceptron. Based on the experimental results, they found that the integration of MobileNet as a feature extractor and SVM as a classifier achieves the highest accuracy of 98.5%. However, their model extracts the features from the CXR showing lungs and bony structures such as rib cage, collar bone, etc. Also, the authors do not focus on the visualization of the extracted features. Moreover, they do not justify the need for integrating DL and ML techniques rather than employing only DL models for feature extraction and classification. A similar ensemble approach has been utilized by Roy and Kumar (2022) where they have utilized different DL models - ResNet50 (He et al., 2016), DenseNet201 (Huang et al., 2017), InceptionV3 (Szegedy et al., 2016), VGG-16 (Simonyan & Zisserman, 2014), VGG-19 (Simonyan & Zisserman, 2014), Xception (Chollet, 2017), and MobileNetV2 (Sandler et al., 2018) to differentiate between Covid-19 and Normal CXRs.

In another research, Chandra et al. (2021) proposed an automatic COVID screening system. They employed an ensemble network consisting of five supervised classification algorithms viz. Decision Tree, KNN, Naive Bayes, ANN, and SVM. Their system first extracts radiomic texture descriptors from CXRs and classifies them into Normal and Infected classes. Then, the CXRs classified to infected class are further distinguished as Non-COVID pneumonia or COVID-19 pneumonia. The authors trained their model on 696 images of each class. First, they evaluated their system to show the impact of augmentation on the robustness of the model. Second, they demonstrated the generalization potential of SVM in identifying the radiological characteristics specific to COVID-19. Lastly, they also discussed the role of ensemble networks in reducing the chances of false diagnosis. The authors claimed that their system achieved an accuracy of 91.329%. Moreover, they observed that the system is useful even when a small annotated dataset is available for training. But, they do not elaborate on why two-phase classification is employed rather than using single-phase multi-class classification. Also, their system is not capable of distinguishing Bacterial and Non-Covid Viral Pneumonia. It also lacks visualization of features used for classification.

Kusakunniran et al. (2021) employed the ResNet-101 architecture for the detection of COVID-19 from the dataset containing chest X-rays of healthy lungs and lungs infected with diseases other than COVID-19. They applied a pre-trained U-Net model for the segmentation of lungs. They also visualized the region of interest using the heatmaps. They reported a sensitivity of 97% for the detection of COVID-19. The authors calculated the confidence score of their model and claimed that

it is efficient in the primary screening of COVID-19. Thus, it reduces the number of chest X-rays to be examined manually. But, this model lacks in distinguishing COVID-19 from Non-COVID Viral Pneumonia and Bacterial Pneumonia.

Zhong et al. (2021) proposed a deep metric learning-based model that uses the optimized embedding space created by a collection of images with the same labels for this task. First, they applied preprocessing techniques such as anonymization, image cropping, resizing, windowing, and lung segmentation. Then, they applied a pre-trained model for feature extraction. These features are integrated with the information derived from lab tests, and medical histories of patients of COVID-19, for decision making. The authors reported an accuracy of 76.7% for the detection of the COVID-19 class. However, the authors prepared the training dataset using CXRs collected from various websites and validated by two senior radiologists, but the problem of class imbalance is still unresolved. Furthermore, the image dimensions of 256×256 may lead to the loss of features from CXR. As per the discussion given in (Rajpurkar et al., 2017), 512×512 is suggested as the most suitable dimension for medical images.

Jain et al. (2021) presented a comparison in the performance of Xception, Inception V3, and ResNet models. They trained these models on 5,467 chest x-ray images and validated on 965 images. The authors claimed that the Xception model outperforms all three aforementioned models. It achieved the highest accuracy of 97.97% for the classification of chest X-rays into healthy and COVID-19. The authors applied rotation, zoom, and shearing for data augmentation. But, rotation changes the direction of the heart in chest X-rays that may lead the model to decide based on the wrong features. A similar augmentation technique was used by Minaee et al. (2020) to compare the performance of ResNet18, ResNet50, SqueezeNet, and DenseNet-121, in the detection of COVID-19 from CXR. In both papers, the authors did not focus on the spatial relationship between different components of the chest X-rays. This decreases the reliability of the model.

Another group of researchers Sekeroglu and Ozsahin (2020) compared the performance of non-pre-trained CNN models, ML models, and pre-trained CNN models for the detection of COVID-19. They used eightfold cross-validation and reported the mean accuracy of 98.50% for binary classification of CXRs into Normal and COVID-19. They also calculated the macro-averaged F1 score as 94.10% for classifying the samples into three classes viz COVID-19, Pneumonia, and Normal. The authors claimed that the shallow CNN model without preprocessing is efficient in detecting COVID-19 from a small and imbalanced dataset. The authors also claimed that reducing the dimension of images reduces the efficacy of the CNN model.

A team of researchers, Abbas et al. (2021) proposed a DL-based self-supervised model that assigns pseudo labels to an unlabeled dataset of CXRs. The model extracts visual features and applies DBSCAN clustering algorithms for clustering of extracted features. Based on the features, a label is assigned to each cluster. To further optimize the clusters, the authors used the k-nearest-neighbor (kNN) algorithm. Now, they employed transfer learning for better image recognition. The authors reported an accuracy of 97.54% on the test dataset comprising 529 chest X-ray images and an accuracy of 99.8% on 283 samples of COVID-19. The experimental results indicate that the model is effective in detecting COVID-19 using chest radiographs even if a small labeled dataset is available. However, their system fails to address the class interference between Non-COVID Viral Pneumonia, and Bacterial Pneumonia.

Afshar et al. (2020) extended the applicability of CNN models and proposed a capsule network to classify Normal, Bacterial Pneumonia, Non-COVID Viral Pneumonia, and COVID-19 infections using chest radiographs. This model captures the spatial relationship between different sample images and multiple objects enclosed in an image. This improves the robustness of the model. The model becomes effective in recognizing a pattern and an object even after flipping or rotating the chest radiograph. They pre-trained their model on the NIH dataset (Rajpurkar et al., 2017) containing chest X-rays of five categories viz. No

findings, tumor, pleural disease, lung infection, and other diseases (Afshar et al., 2020). The authors claimed that their network is equally efficient even when the training dataset is small. The model achieved an accuracy of 98.3%. Similarly, Mangal et al. (2020) proposed the CNN model 'CovidAID' for the classification of chest X-rays into Normal, Bacterial Pneumonia, Viral Pneumonia, and COVID-19. Their model comprises a pre-trained CheXNet (Rajpurkar et al., 2017) model and Dense Convolutional Network (DenseNet) (Nguyen et al., 2020) that are further connected to a fully connected layer. They trained their model on 1,341; 2,530; 1,337; and 115 samples of Normal, Bacterial Pneumonia, Viral Pneumonia, and COVID-19 classes, respectively. The authors in (Mangal et al., 2020) resolved the problem of class imbalance by using the weighted binary cross-entropy loss function. In which the total loss is calculated as a weighted sum of cost from each class. The cost from each class is multiplied by a unique scalar value, calculated differently for every class, depending upon the number of training samples in each class. They also visualized the features using saliency maps. Their model gave an accuracy of 90.5%.

Loey et al. (2020) extended the scope of the COVID-19 detection systems. They proposed the classification and regression-based dual approach that predicts the estimated time for reaching a patient to the severe stage of illness. The authors reported an accuracy of 85.91% on the test dataset. But, the dataset contains only 86 samples of severely infected patients in a total of 408 trainable samples. This may cause the problem of class imbalance. Also, the authors did not determine a threshold for deciding the non-severe, mild and severe stages of illness. Further, Greenspan et al. (2020) presented a discussion on the role of AI techniques in early disease detection, resource management, and developing patient-specific diagnostic models. Based on the review of related literature, the authors claimed that sufficient AI-based tools are available for the diagnosis of COVID-19. But, there is an immediate requirement to address the challenge of collecting a huge and labeled dataset. Especially, for COVID-19 where the virus is changing its behavior unexpectedly and the rate of spread is very fast. The authors recommended the development of tools that not only work on imaging data but also consider the patient-level clinical information for decision making.

Another research group, Michael Roberts et al. (2021) presented a review of ML and DL techniques proposed for the detection of COVID using CT scans and chest X-ray images. They claimed that most researchers performed binary classification of the chest radiographs into COVID and Normal classes, or ternary classification into COVID, Pneumonia, and Normal. Only in two research articles, the chest radiographs were divided into four classes viz. COVID, Bacterial Pneumonia, Viral Pneumonia, and Normal. Based on the review, they also claimed that there is a lack of external validation, robustness, or sensitivity analysis of the proposed models. Moreover, there is little attention given to the preparation of the dataset. The dataset available so far consists of complete chest radiographs rather than bone suppressed and lung segmented images. Also, there is a lack of validation of the dataset and sensitivity of DL models by the radiology experts. This is highly required for the practical implementation of the model in healthcare. Further, there is a scope to improve the classification accuracy and reduce the training and response time. Also, there is a need to focus on the visualization of features involved in prediction. This is useful to convince clinicians about the reliability of the prediction models.

3. Materials and methods

In this section, we discuss the preparation, handling, and splitting of the dataset, architectures of models employed for preprocessing of the prepared dataset, and multiclass classification. Next, we present the training details of the proposed model. An overview of the proposed methodology is demonstrated in Fig. 1.

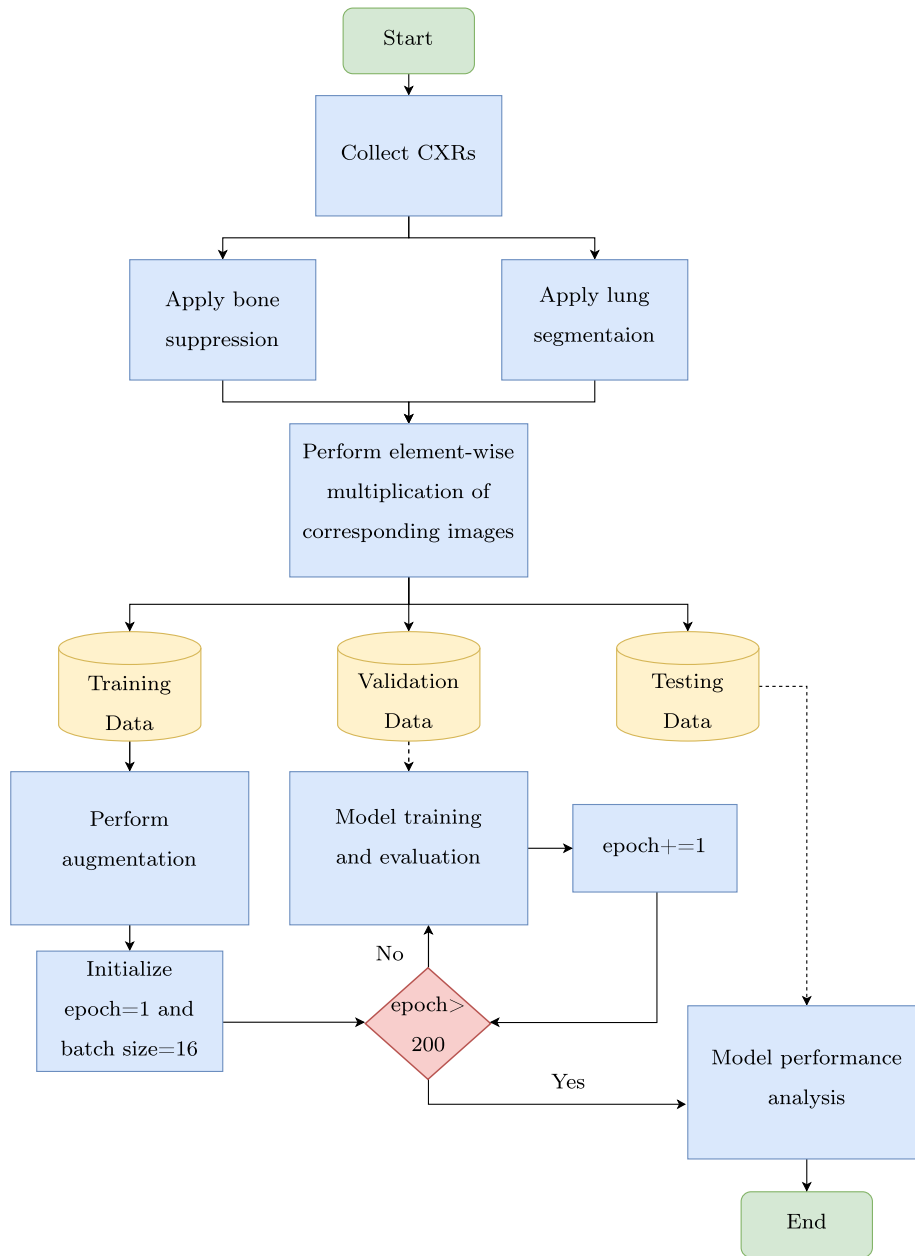


Fig. 1. Methodology overview.

3.1. Dataset preparation

Acquisition and preparation of anonymous, labeled, good quality, and preprocessed dataset of chest radiographs is a challenging task.

The online dataset providers claim that the data available at their source is medically approved. But, it is observed that the images available at many sources are randomly scraped from Google search. Radiologists involved in this research assessed the available online dataset and found many images under the wrong class labels. Such images may lead to erroneous results if used for training the deep learning models. Thus, we collected CXRs only from Radiopaedia (<https://radiopaedia.org/articles/covid-19-4>), Eurorad (<https://www.eurorad.org/advanced-search?search=COVID>), IEEE (<https://github.com/ieee8023/covid-chestxray-dataset>), Kaggle (<https://www.kaggle.com/c/pneumoniabacteriavirus>), and from the R.G. Stone Urology and Laparoscopy hospital in India. The radiology experts validated the quality and labels of the dataset. Finally, we obtained 6,367 CXRs, containing 412, 1,273, 2,406, and 2,205 chest radiographs from CV,

N, VP, and BP classes, respectively. The labels of the collected dataset are validated by the radiology experts involved in this research. Further, gathering the dataset from validated sources does not guarantee the good quality of CXRs. The quality of CXRs may be degraded by the artifacts, movement of a patient while capturing the CXRs, improper handling of films, fault in the x-ray machines, wearing of accessories by the patients, etc. Moreover, it is not feasible to capture good quality CXRs when patients are facing difficulty in breathing. So, the severity of the disease also affects the quality of a radiograph. For precise training of the DL model, it is important to discard poor quality CXRs. Thus, our associated team of radiologists discarded the CXRs with ornaments, artifacts, incomplete and poor visibility of lungs through a meticulous manual examination. The final dataset prepared comprises 5,188 CXRs.

We divided this dataset into training, and validation sets in the ratio of 70% and 30% respectively. The training dataset comprising 219, 1255, 764, and 1395 images in the CV, N, VP, and BP classes respectively. The validation dataset contains 93, 538, 327, and 597 images in the CV, N, VP, and BP classes respectively. Further, we prepared a test

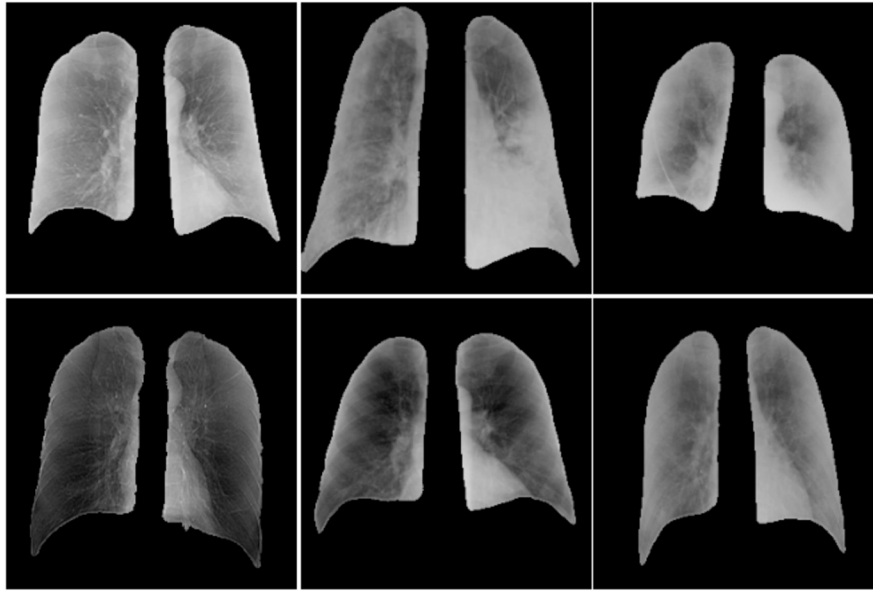


Fig. 2. Sample images of COVID-Pneumonia CXR Dataset.

dataset including 401 images comprising 52, 140, 89, and 120 images from the CV, N, VP, and BP respectively. To prevent data leakage, the CXRs of one patient were kept a part of either the training or validation or testing set.

Since, the performance of the DL models is directly dependent on the size of the training data, the robustness of the model is determined by the variety, statistical characteristics, and demographics of the dataset. Therefore, we applied the augmentation techniques such as Contrast Limited Adaptive Histogram Equalization (CLAHE) (Yadav et al., 2014) and blurring filters viz. mean and median only on the training dataset. These techniques increased the training size from 3,633 to 14,532 radiographs comprising 876, 3,056, 5,580, and 5,020 images from CV, N, VP, and BP classes respectively. The training dataset is used to train the model while the validation dataset is used to find the best fit model during training. Lastly, the test set is used to evaluate and compare the performance of different model architectures.

After preparing the dataset, we observed an imbalance in the data due to the difference in the number of images in each class. Training the classifier with an imbalanced dataset influences its decision towards the class with a larger number of images. Thus, reducing the overall prediction performance and reliability of the model. This issue was resolved by experimenting with the weighted sigmoid loss (Rajpurkar et al., 2017) and focal-loss (Lin et al., 2017) function. The weighted sigmoid loss calculates weight values per class which are inversely proportional to the number of samples. Classes with a lower number of samples are assigned higher weights and contribute more to the loss function, whereas classes with a higher number of samples contribute less to the loss function (Rajpurkar et al., 2017). In the case of focal-loss, samples classified with high probability of occurrence to one of the classes contribute less to the loss value. On the other hand, the samples with an equivalent probability of occurrence in all the classes are hard to classify. Such samples contribute more to the loss value (Lin et al., 2017). Therefore, to further minimize both the loss functions, the deep learning model has to improve its overall feature extraction capability and make unbiased decisions. Thus, they overcome the problem of class imbalance.

Further, the prepared dataset comprising CXR images contains bony structures such as the neck, arms, ribs, collar bones, etc. These structures interfere in feature extraction from the lung region. Thus, increase the training time and decrease the precision of the DL model in the detection, localization, and visualization of the infection. Also, features from solid tissues outside the lung region adversely affect the

decision-making of a DL model. Therefore, we applied a bone shadow suppression model for reducing the visibility of bony structures and a lung segmentation model for extracting the lung region from the CXR. We label this bone suppressed and lung segmented dataset as 'COVID-Pneumonia CXR Dataset' (CP-CXR). Sample images of this prepared dataset are shown in Fig. 2. We use this dataset for training the classification models.

3.2. Proposed multi-modal classification system

The proposed classification system, "Covid Scanner", as shown in Fig. 3, is divided into two stages: processing stage (PS)-A, and processing stage (PS)-B. The input CXR is passed through these two stages to get the output label(s).

3.2.1. Processing stage - A

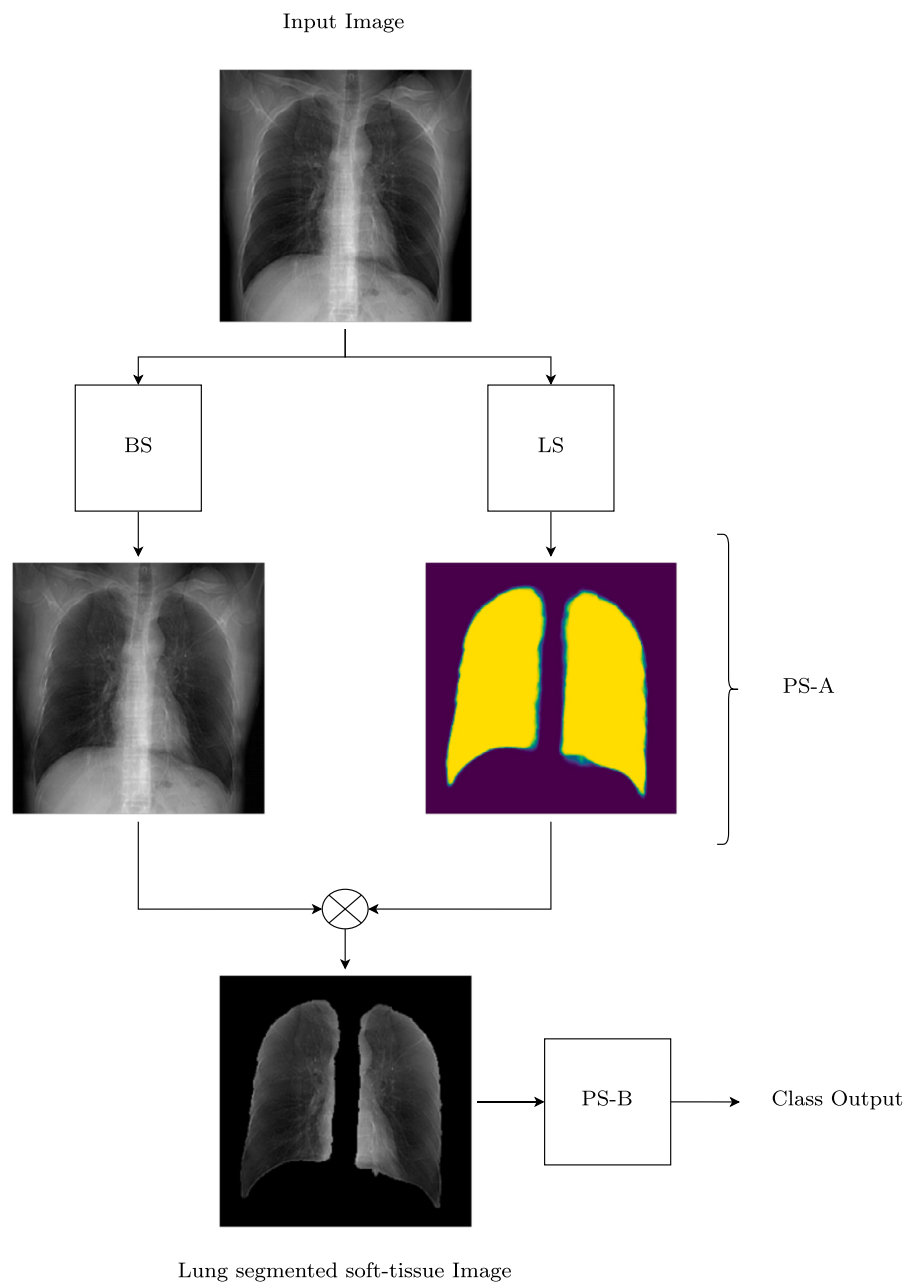
The first stage (PS-A) works as a preprocessing unit for reducing the feature complexity. This stage includes a conditional Generative Adversarial Network (cGAN) model for Bone Suppression (BS) (Isola et al., 2017), and a semantic segmentation based custom U-Net model for Lung Segmentation (LS) (Ronneberger et al., 2015). The architecture of these models is described below.


For BS, we employed the Pix2Pix (Isola et al., 2017) model comprising a Generator (G), and a Discriminator (D). The Generator of the model is a U-Net-based auto-encoding network. Its encoder is designed by combining eight Convolution-BatchNorm-ReLU layers (Ck) with k filters as shown below.

Encoder : C64 – C128 – C256 – C512 – C512 – C512 – C512 – C512

At each layer, the network performs a convolution operation for feature extraction. The number of filters is 64, 128, 256 for the first, second, and third layers respectively. For the next five layers, the number of filters is 512. Now, batch normalization is applied to each batch of features obtained after the convolution operation. It improves the learning stability of the model even at higher learning rate. Moreover, it helps in reducing the number of epochs required for training the model. Next, the model employs the ReLU activation function on the output received after batch normalization.

The Discriminator (D) of the BS model follows the Markovian discriminator architecture (Isola et al., 2017) which divides an image into $N \times N$ patch. It returns one feature map for real or fake predictions. Now,



 Element-wise multiplication

LS : Lung Segmentation

BS : Bone Suppression

PS-A : Processing Stage-A

PS-B : Processing Stage-B

Fig. 3. Covid Scanner system architecture.

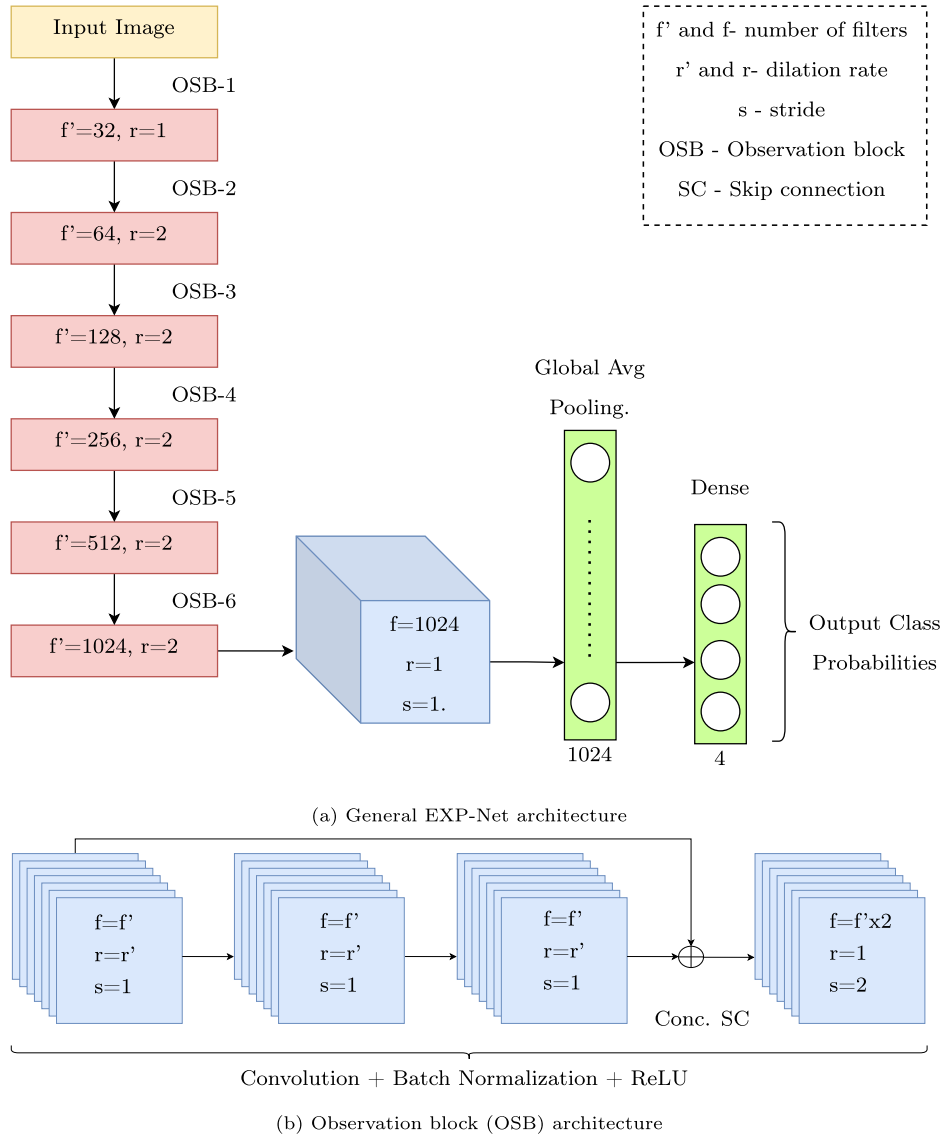


Fig. 4. EXP-Net architecture.

it finds an average of the obtained feature maps to give a single score as an output of the discriminator.

The architecture of the generator of the decoder is a combination of eight Convolution-BatchNorm-Dropout-ReLU layers (CDk) with a dropout rate of 50%. Its symbolic representation is shown below.

Decoder : $CD512 - CD512 - CD512 - C512 - C256 - C128 - C64$

In this architecture, the convolution operation is followed by batch normalization. Now, the dropout is applied to randomly reduce the feature map obtained after batch normalization. Next, the ReLU activation function is employed on the output feature map obtained.

In each of the first four layers there are 512 filters. Whereas, the fifth, sixth, and seventh layers have 256, 128, and 64 filters respectively. At the end of second last layer (CD128), a final convolutional operation with tanh non-linearity is applied to get the output. All convolution filters employed are of size 4×4 with a stride of 2.

The discriminator (D) of the decoder of the BS model follows the Markovian architectural configuration (Isola et al., 2017). After the fourth layer, it applies a convolution operation with a sigmoid function to identify convolutional patches as real or fake. Further, it avoids batch normalization in the first layer. The architecture employs leaky ReLU functions at all the layers with a slope of 0.2.

For LS, we employed a semantic segmentation-based custom U-Net model. The architectural design for the contraction and expansion paths of the LS model is similar to the aforementioned encoder-decoder architecture of the Generator (G) of the BS, rather it uses the sigmoid function in place of the tanh activation function in the last layer.

The BS model is executed to produce a dual-energy subtracted soft-tissue image by removing the rib cage and collar bones. Simultaneously, the LS model is executed to generate a binary mask of the lung region. The independent, and parallel execution of BS and LS models minimizes the compounding of error while generating the preprocessed CXR. Now, the individual outputs obtained from these models undergo element-wise multiplication to give the resultant preprocessed image, as shown in Fig. 3. This resultant image is passed onto the next stage (PS-B).

3.2.2. Processing stage - B

In this section, we present a detailed discussion on the proposed classifier 'EXP-Net' developed for multiclass Classification.

The proposed classifier 'EXP-Net' is an integration of six observation blocks (OSB), as shown in Fig. 4. While progressing from OSB-1 to OSB-6, the number of filters is increased to 32, 64, 128, 256, 512, and 1024, respectively. Each observation block, from OSB-1 to OSB-6, consists of four convolution blocks. Every convolutional block is a

combination of the Convolution-BatchNorm-ReLU layers, with a varying number of filters and sizes. In all observation blocks, the first three convolution blocks have a spatial filter size of 3×3 and stride of 1, followed by a concatenated skip-connection made between the output of the first and third convolutional block. We did not use dropout in these convolutional blocks. The purpose of the first three convolutional blocks is to locate areas where there is a maximum probability of finding the features. Now, we used the fourth convolutional block with filter size 1×1 and stride of 2. This block reduces the spatial size of the feature map and works similar to a pooling strategy. The objective of this last convolutional block is to utilize the mid-level feature representations to decide the relevance of the extracted features. This allows the propagation of the most relevant information for decision-making.

We have used a concatenated skip-connection over an additive one to provide better reusability of the feature representations and preservation of information. In every observation block, except OSB-1, the first three convolutional blocks utilize dilated convolutional operation, with a rate of 2. This increases the receptive strength of the convolutional operation (Yu & Koltun, 2015). Further, the leaky Relu activation function with a slope of 0.2 is employed to avoid the problem of dying Relu (Lu et al., 2019). The convolution layer is followed by a global average pooling layer and a fully connected layer. The sigmoid activation function is employed for predicting probabilities of CXRs into CV, NP, BP, and N classes.

3.3. Problem formulation

Let X denote the set of input feature matrices comprising Antero-posterior (AP) view chest radiographs, and let Y denote the multiset of corresponding output probability vectors. Both X and Y have a cardinality of n . (X_i, Y_i) denotes the i^{th} instance of the input matrix and its corresponding output vector. Each instance vector in Y is of size $1 \times k$; where k is the number of different output classes under observation. In this manuscript, k is 4 which represents the four classes viz. CV, NP, BP, and N. Further, we assume that k classes are independent events. Therefore, each index value in an instance vector in Y denotes the independent probability of the occurrence of a particular output class in 0 or 1 for a given X . The summation of conditional probabilities in an instance vector of Y can be written as reported in Eq. (1). Y is the set of predicted output vectors on X , and $J^* : Y' \times Y \rightarrow R$ is the cost function. A and B are instances of X and Y respectively.

$$\sum_{j=1}^k P(B_j|A) \geq 1 : (A, B) \in \{(X_i, Y_i)_{i=1}^n\}. \quad (1)$$

This type of setting was done to consider that both Non-COVID Viral Pneumonia and Bacterial Pneumonia can co-exist in a community. The system aims to minimize J^* by reducing the overall probabilistic error calculated over the multiset: $\{(Y'_i, Y_i)_{i=1}^n\}$, thereby leading to an effective feature distribution mapping.

3.4. Objective analysis at processing stage

In this section, we elaborate on how the objective functions are achieved by the BS, LS, and classification model.

3.4.1. Analysis at PS-A

BS is trained on a separate set of feature matrix pairs $\{(X_{B(i)}, Y_{B(i)})_{i=1}^n\}$. Here, X_B and Y_B are two sets of the same cardinality m . X_B is the set of input chest radiographs and Y_B is the set of output dual-energy subtracted chest radiographs with $|X_B| = |Y_B| = m$. We employed the L1 loss function to define the objective functions (JG) for the generator and the minimax strategy for the objective function (JD) for the discriminator. Further, we followed the experiments conducted in reference (Isola et al., 2017) and pre-set the values of hyperparameters β and γ as 0.5 and 1000, respectively. The definitions of both the objective functions JG and JD are given in Eq. (2) and (3) respectively.

$$J_D = \beta[\mathbb{E}_{X_B, Y_B} \log D(X_B, Y_B) + \mathbb{E}_{X_B} \log(1 - D(X_B, G(X_B)))] \quad (2)$$

$$J_G = \mathbb{E}_{X_B, Y_B} \log D(X_B, G(X_B)) + \gamma \mathbb{E}_{X_B, Y_B} \|Y_B - G(X_B)\|_1 \quad (3)$$

In this adversarial training, the generator tries to generate plausible images $G(X_B)$ to fool its adversary discriminator, by minimizing its objective function J_G . In contrast, the discriminator tries to better distinguish between the ground truth image Y_B and generated image $G(X_B)$ by minimizing its objective function J_D . As the discriminator continues to improve, the generator tends to generate images similar to Y_B . As soon as the cGAN approaches the Nash equilibrium (Nash, 1950), we observe that the generated image is highly similar to the ground truth image $G(X_B) \sim Y_B$. After successful training the cGAN, its generator learns effective feature translation mapping from $X_B \rightarrow Y_B$. Thus, it is used to represent BS and translate $X \rightarrow X_{BS}$. Now, in parallel to the BS, LS is trained on a set of feature matrix pairs $\{(X_{L(i)}, Y_{L(i)})_{i=1}^m\}$. Here, $X_{L(i)}$ and $Y_{L(i)}$ are two sets of same cardinality m . $X_{L(i)}$ is the set of input CXRs and $X_{L(i)}$ is the set of output lung mask images. The objective function (J_{LS}) for the LS is defined in Eq. (4).

$$J_{LS} = \min(Y_L \log U(X_L) + (1 - Y_L) \log(1 - U(X_L))) \quad (4)$$

In this equation, U represents the custom U-Net model and $U(X_L)$ denotes the set of predicted output masks. The objective of the model is to generate effective lung binary masks by determining the area of the lung region at a pixel level. Pixel values of the lung region are emphasized to shift towards 1 while the remaining ones are to 0. Once, the LS model is trained, it generates the masks of the lung region.

In this way, PS-A utilizes the potential of BS and LS to translate X into a new set of reduced feature complexity matrices; X' . Both BS and LS process the input X in parallel for computing individual output sets X_{BS} and X_{LS} respectively. Then, an element-wise multiplication, as shown in Eq. (5), is performed between the corresponding feature matrices in X_{BS} and X_{LS} to obtain X' .

$$X' = \{p \odot q : (p, q) \in \{(X_{BS(i)}, X_{LS(i)})_{i=1}^n\}\} \quad (5)$$

3.4.2. Analysis at PS-B

PS-B maps the preprocessed input set $X' \rightarrow Y$ and computes J^* , given in Eq. (6).

$$J^* = \min(\mathbb{E}_{Y', Y} L(Y', Y)) \quad (6)$$

where $L(Y', Y)$ is defined in Eq. (7):

$$L(Y', Y) = \sum_{i=1}^n -\alpha(1 - t_i)^\gamma \log t_i \quad (7)$$

We chose focal-loss (Lin et al., 2017) as our baseline loss function (L) so that easier training samples contribute less and harder samples contribute more to the loss function. For example, the visual symptoms of Bacterial Pneumonia are clearly distinguishable from other classes. Therefore, the deep learning model is able to classify these samples with high inter-class differences in probability. On the other hand, the samples with an equivalent probability of occurrence in all classes are hard to classify. Such samples are considered harder samples. For example, the CXRs of Non-COVID Viral Pneumonia and COVID-19 have a high degree of similarity in visual symptoms. Thus, the deep learning model may report less difference in the probability of occurrence between classes of such samples.

In Eq. (7), α is the balancing factor and is set to 0.25, and γ is the modulating factor and is set to 2, and t_i is the modified output vector of Y'_i , given in Eq. (8):

$$t_{ij} = \begin{cases} Y'_{ij}, & Y_{ij} = 1 \\ 1 - Y'_{ij}, & Y_{ij} = 0 \end{cases} \quad (8)$$

Here, Y'_{ij} , Y_{ij} , and t_{ij} correspond to the j^{th} (ranging from 0 to k) index value of the vector Y'_i , Y_i and t_i respectively. This constitutes a

new set of modified output vectors $T(T = \{t_1, t_2, t_3, \dots, t_n\})$. Now, from Eq. (7) and (8), we obtain our new cost function changes as described in Eq. (9).

$$J^* = \min(E_T - \alpha(1 - T)^r \log T) \quad (9)$$

3.5. Training details

In this section, we present the training details of the models employed for bone suppression, lung segmentation, and multiclass classification.

3.5.1. Training of bone suppression model

The Pix2Pix (Isola et al., 2017) model used for bone suppression is trained on the BSE-JSRT dataset (<http://db.jsrt.or.jp/eng.php>) containing 247 chest radiographs and their corresponding dual-energy subtracted images. The dataset is divided into the training, and testing sets in the ratio of 60%, and 40%, respectively. The training dataset comprises 147 images whereas, the testing dataset contains 100 images. The augmentation techniques such as rotation at angles of $\pm 5^\circ$, $\pm 10^\circ$, and $\pm 15^\circ$, and Histogram Equalization (HE) of both input and rotated combinations are applied to the training dataset. Further, Contrast Limited Adaptive Histogram Equalization (CLAHE) (Yadav et al., 2014) is applied for histogram equalization. Applying HE improves the robustness of the model by familiarizing the model with both low and high contrast chest radiographs. Finally, the training dataset comprising 1,911 CXRs is prepared.

We employed the Adam (Kingma & Ba, 2014) optimizer to train the Pix2Pix model. The value of α is 0.0002, β_1 is 0.5, and β_2 is 0.999 as per the experimentation and observations in (Isola et al., 2017). Further, the discriminator is updated twice before updating the generator. The initial updation is done over the actual image and then a second updation is done over the generated image. The model is trained for 100 epochs with a batch size of 2 and rate decay of 25% per 20 epochs.

3.5.2. Training of lung segmentation model

The U-Net (Ronneberger et al., 2015) based model employed for lung segmentation was trained on 6,395 images from the repository maintained by 'v7 labs' (<https://github.com/v7labs/covid-19-xray-dataset>). The repository contains lung region masks of both Normal and pathology affected radiographs, including COVID-19. All the masking images available in the dataset were annotated and validated by medical experts and radiologists. The dataset was divided into training, validation, and testing containing 4,476, 959, and 960 CXRs, respectively. To train the model, Adam optimizer (Kingma & Ba, 2014) was used with default values of hyperparameters for 100 epochs with a batch size of 128.

3.5.3. Training of classification model

The training of bone suppression and lung segmentation models is followed by training of the proposed classification model, 'EXP-Net'. We trained this model on the benchmarking dataset 'COVID-Pneumonia CXR', comprising bone suppressed and lung segmented CXRs of Normal, Bacterial Pneumonia, Non-COVID Viral Pneumonia, and COVID-19. Similar to the BS and LS models, we used the Adam optimizer with default hyperparameters for training the model. The model was trained for 200 epochs with a batch size of 16 and a learning rate decay of 25% per 50 epochs.

4. Results

In this section, we demonstrate impacts of augmentation, loss functions, BS, and LS on the performance of the proposed classifier. We also discuss the visualization of features extracted by the classifier using

Table 1

Comparison in performance of bone suppression models.

Model	PSNR (dB)	SSIM	Entropy (bits/pixel)
CNN	29.75	0.9379	6.8936
VAE	31.05	0.9622	6.7702
Pix2Pix	33.64	0.9776	6.6470

GradCAM++. Further, we give a comparative analysis of the proposed system with the state-of-the-art classification models and the systems proposed in the related literature. To evaluate the model efficacy and reliability, we have used quantitative measurements such as Accuracy, Precision, Recall, AUC, F1 score, and Cohen-Kappa score (Vieira et al., 2010).

4.1. Impact of augmentation

In this sub-section, we demonstrate the impact of augmentation on the classifier's performance. We present the confusion matrix obtained on the testing data when using the augmented and non-augmented training data. It is clear from the confusion matrix shown in the Fig. 5a and 5b that there is an increase of 3, 4, 15, and 4 correct classifications in CV, VP, BP, and N classes, respectively. The improvement of 6% in the accuracy of classification on testing dataset, comprising 401 CXRs, proves the significance of augmentation.

We also showcase the trends of AUC on the validation data when Weighted Sigmoid Loss (WSL) and Focal Loss (FL) functions are employed in the classification model. The impacts on AUC when using the augmented, and non-augmented datasets for training are shown in Fig. 6a and 6b respectively. The AUC is independent of the number of samples; therefore, it does not reflect the impact of class imbalance on the experimental results.

4.2. Impact of bone suppression

We trained a Convolutional Neural Network (CNN) (Gusarev et al., 2017), a Variational Auto Encoder (Gusarev et al., 2017), and a Pix2Pix model for bone suppression on the dataset discussed in subsection 3.5.1. We evaluated the performance of these models by calculating the values of Peak Signal to Noise Ratio (PSNR) (Gusarev et al., 2017), Structural Similarity Index (SSIM) (Gusarev et al., 2017), and entropy (Pradhan et al., 2020). This is evident from the results shown in Table 1 that Pix2Pix based bone suppression model reports the highest values of PSNR, and SSIM. Moreover, it reports the minimum difference of 0.1138 bits/pixel in the mean entropy of the generated and actual images. Mean entropy of actual images is reported as 6.5332 bits/pixel.

Therefore, we employed Pix2Pix based BS model to minimize the interference of the features extracted from the bony structures. It is evident from the confusion matrix obtained on the clinical testing dataset (shown in Fig. 5b and 5c) that there is an increase of 4, 3, 5, and 4 correct classifications for the CV, VP, BP, and N classes, respectively, when BS is employed on the augmented training dataset. There is an improvement of 4% in the overall accuracy of the classification. Further, the impact of the BS model on the performance of the proposed classifier on the validation dataset is visible when comparing the graphs shown in Fig. 6a and 7, respectively. It is apparent from the trends of loss functions obtained on the bone suppressed images of the validation dataset and confusion matrix of the testing data that BS improves the overall classification performance.

4.3. Impact of lung segmentation

The authors trained the U-Net (Ronneberger et al., 2015), Pyramid Scene Parsing Network (PSP-Net) (Zhao et al., 2017), Fully Convolutional Networks (FCN) (Long et al., 2015), and Custom U-Net models on the dataset discussed in subsection 3.5.2 for segmentation of the

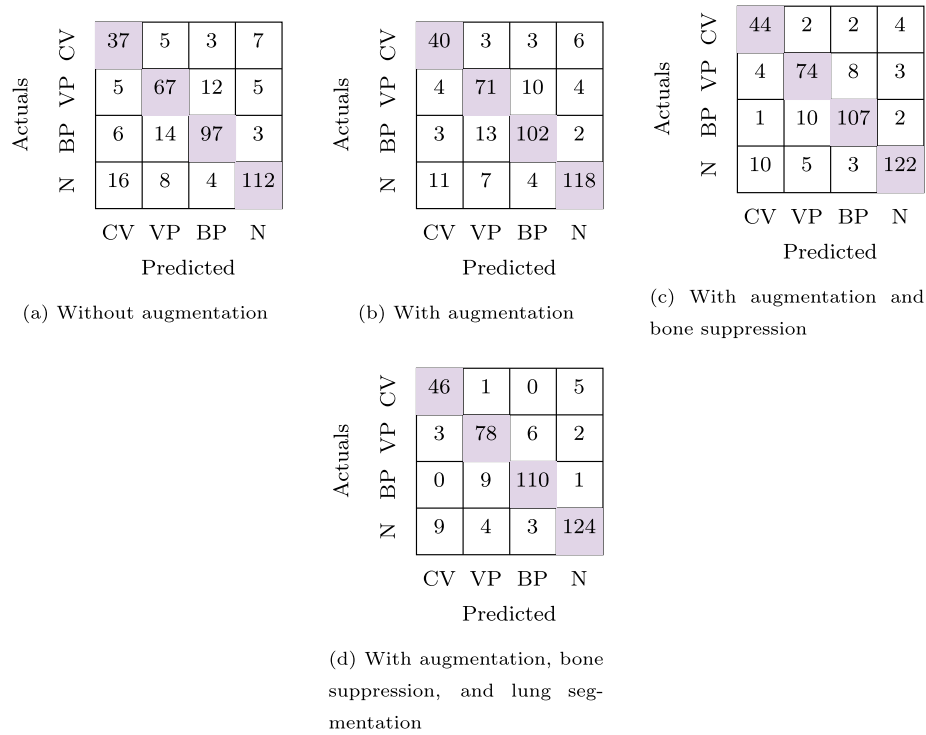


Fig. 5. Impact of augmentation on performance of classifier.

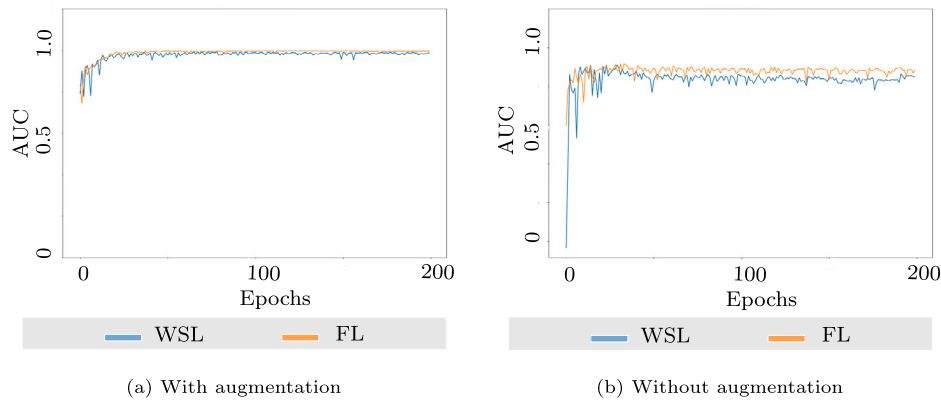


Fig. 6. Impact of augmentation on validation data when using different loss functions; Weighted Sigmoid Loss (WSL) and FL (Focal Loss).

lung region from the CXRs. They evaluated the performance of these models by calculating the Intersection Over Union (IoU) score, and binary accuracy on the validation and testing datasets. It is apparent from the results shown in Table 2, that Custom U-Net model reports the highest value of IoU as 91.65%, and 91.87% on the validation and testing datasets respectively. Also, it reports the highest binary accuracy of 97.07%, and 97.09% on the validation, and testing dataset, respectively.

Therefore, the authors employed the Custom U-Net for the LS model to extract the Region of Interest (i.e. lungs) from the CXRs. They record the AUC and confusion matrix to demonstrate the impact of employing LS with BS and augmentation. It is evident from the confusion matrix shown in Fig. 5c and 5d that there is an increase of 2, 4, 3, and 2 correct classifications in CV, VP, BP, and N classes, respectively. It results in an improvement of 2.7% in the overall accuracy of classification. Further, it is noticeable from the graphs shown in Fig. 7 and 8 that employing the LS model improves the AUC values. Hence, it ensures better training of the model.

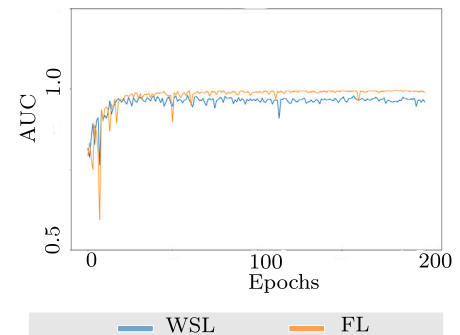


Fig. 7. Impact of bone suppression on validation data when using different loss functions; Weighted Sigmoid Loss (WSL) and Focal Loss (FL).

4.4. Feature visualization using GradCAM++

The health experts are naïve to the technical details of the classification model. Therefore, it is difficult for them to rely on the screening

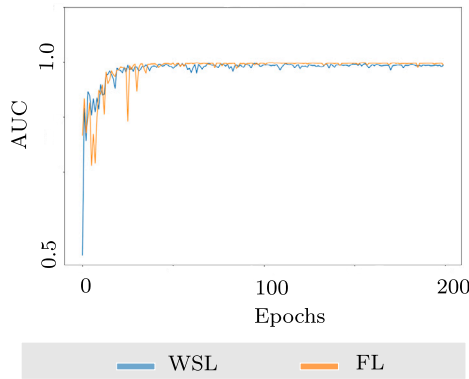


Fig. 8. Impact of lung segmentation on the Loss function; Weighted Sigmoid Loss (WSL) and FL (Focal Loss).

Table 2

Performance of different lung segmentation models on validation and test datasets.

Model	IoU		Binary Accuracy	
	Validation	Test	Validation	Test
U-Net	0.9163	0.9141	0.9703	0.9697
PSP-Net	0.9146	0.9161	0.9699	0.9704
FCN	0.9073	0.9081	0.9594	0.9643
Custom U-Net	0.9165	0.9187	0.9707	0.9709

Table 3

Comparison in performance of classification models on validation dataset.

Model	Loss	Accuracy	Precision	Recall	AUC
DenseNet121	0.1631	0.9084	0.9315	0.8184	0.9614
ResNet50	0.1825	0.8863	0.9332	0.7770	0.9544
InceptionV3	0.1702	0.9135	0.9343	0.7841	0.9576
EXP-Net	0.1249	0.9407	0.9411	0.8533	0.9658
Xception	0.1547	0.9088	0.9346	0.8192	0.9617
ResNet101	0.1755	0.9022	0.9318	0.7926	0.9574
EfficientNetB2	0.1417	0.9342	0.9374	0.8388	0.9631

performed by the DL-based classification model. To build their trust in the classification model, we plotted feature heatmaps using Grad-CAM++ (Chattopadhyay et al., 2018). The GradCAM++ shows the regions where the most prominent features involved in classification have been extracted. The change in color from blue to red demonstrates the increase in the relevancy of features picked from a region. We used these plots on top of the actual finding annotations of our radiologists and found that our model covers the maximum area where abnormalities are present. The sample images of plotted GradCam++ are shown in Fig. 9.

4.5. Comparative analysis

To validate the efficacy of the proposed classifier ‘EXP-Net’, we compared its performance with the state-of-the-art models viz. ResNet50 (He et al., 2016), ResNet101 (He et al., 2016), DenseNet121 (Huang et al., 2017), InceptionV3 (Szegedy et al., 2016), Xception (Chollet, 2017), and EfficientNetB2 (Tan & Le, 2021). It is evident from the confusion matrices shown in Fig. 10 that the performance of the above-stated models is comparable. Here, the EfficientNetB2 model reports the highest sensitivity for the CV and N classes. Its sensitivity is similar to the DenseNet121 model for the BP class. The Xception model reports the highest sensitivity for the VP class.

Further, we recorded the values of loss function, accuracy, precision, recall, and Area Under Curve (AUC) for all the above-mentioned models, and the proposed classifier ‘EXP-Net’ as shown in Table 3. It is clear from the values shown in column 2 that the ResNet50 model

Table 4

Model performance comparison on test data.

Model	Cohen-Kappa	AUC	F1
DenseNet121	0.8317	0.9582	0.8513
ResNet50	0.7765	0.9323	0.7948
InceptionV3	0.8114	0.9561	0.8305
EXP-Net	0.8535	0.9648	0.8726
Xception	0.8353	0.9501	0.8505
ResNet101	0.8064	0.9460	0.8151
EfficientNetB2	0.8398	0.9528	0.8596

reported the highest value of 0.1825, whereas the ‘EXP-Net’ reported the lowest value of loss function as 0.1249 on the validation dataset. Thus, it has better convergence among all the models. There is a slight variation in the values of loss functions achieved by DenseNet121, InceptionV3, Xception, ResNet101, and EfficientNetB2. Moreover, it is evident from column 3 of Table 3 that the ‘EXP-Net’ reported the highest accuracy of 94.07%. The ResNet50 model reported the lowest accuracy of 88.63%. There are slight differences in the accuracy reported by the other models viz. DenseNet121, InceptionV3, Xception, and EfficientNetB2 models. Also, these values are lower than the accuracy reported by the proposed classifier. Similarly, it is evident from column 4 that the ‘EXP-Net’ reports the highest value of precision as 94.11% among all the above-stated classifiers. Additionally, it is clear from the results given in the columns 5 and 6 of Table 3 that the ‘EXP-Net’ outperforms other classifiers in terms of recall and AUC.

Next, we also compared the performance of all the above-stated classifiers on the testing dataset. It is obvious from the values of Cohen-Kappa, AUC, and F1 scores shown in Table 4 that there is a minor variation in these values for all the above-mentioned state-of-the-art models. Simultaneously, it is also evident from the results shown in Table 4, that the ‘EXP-Net’ gave the highest value of 85.35%, 96.48%, and 87.26% for Cohen-Kappa, AUC, and F1 score, respectively. These results indicate that ‘EXP-Net’ is more capable of drawing decision boundaries and has a higher likelihood of generating matching probability distributions for all four categories. This validates the efficacy and reliability of the proposed classifier.

Apart from comparing the performance of our proposed classifier, ‘EXP-Net’ with other models, we also showcase comparison between our complete ‘Covid Scanner’ system and methodologies presented in Section - 2. For this task, we selected the following approaches - combination of MobileNetv2 and Support Vector Machine (SVM) (Ohata et al., 2020), integration of lung segmentation (LS) and ResNet101 (Kusakuniran et al., 2021), and a pre-trained DenseNet121 model (Mangal et al., 2020) (‘CovidAID’). We trained and evaluated these models on our collected CXRs with the same training, validation, and testing strategy. For this comparative analysis, we pre-trained our EXP-Net model on the same NIH data (available at - <https://www.kaggle.com/datasets/nih-chest-xrays/data>) (Rajpurkar et al., 2017) as used for the CovidAID model (Mangal et al., 2020).

It is evident from the confusion matrices shown in Fig. 11 that ‘Covid Scanner’ with a pre-trained ‘EXP-Net’ performs the maximum number of correct classifications for CV, VP, and BP classes. Although it performs a slightly higher number of wrong classification from the N class in contrast to the ‘CovidAID’ model, it reports a total of merely 19 misclassified images from the testing dataset of 401 CXRs. This proves the efficacy of the model.

5. Discussion

In this manuscript, we achieved the purpose of developing an effective and intelligent system for quick mass screening of COVID-19 by using chest X-ray images. The system proposed in this paper is a coherent integration of DL-based preprocessing and classification models. The preprocessing unit consists of bone suppression and lung segmentation modules. These modules are employed in parallel to avoid the

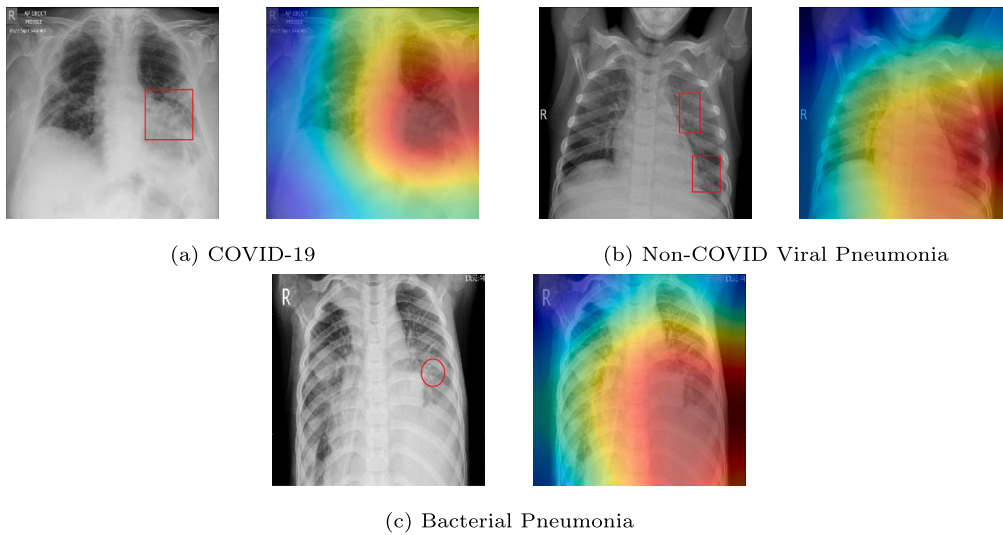


Fig. 9. Feature heatmap visualization using GradCAM++.

Actuals	CV	VP	BP	N
	40	2	3	7
	5	73	9	2
	3	13	97	7
	13	10	7	110
Predicted				
(a) ResNet50				
Actuals	CV	VP	BP	N
	42	4	2	4
	6	75	8	0
	1	10	101	8
	14	10	1	115
Predicted				
(b) ResNet101				
Actuals	CV	VP	BP	N
	44	2	1	5
	3	75	5	6
	0	10	105	5
	7	7	6	120
Predicted				
(c) DenseNet121				
Actuals	CV	VP	BP	N
	43	1	1	7
	1	77	9	2
	6	6	102	6
	20	2	1	117
Predicted				
(d) InceptionV3				
Actuals	CV	VP	BP	N
	44	2	2	4
	0	78	7	4
	2	10	103	5
	10	6	5	119
Predicted				
(e) Xception				
Actuals	CV	VP	BP	N
	45	1	3	3
	1	77	10	1
	4	9	106	1
	5	7	6	122
Predicted				
(f) EfficientNetB2				

Fig. 10. Comparison in performance of classification models on the testing dataset.

compounding of errors. The bone suppression module minimizes the possibility of feature extraction from the collar bone and rib cage visible in the lung region. Then, the lung segmentation model extracts only the soft tissues i.e. lungs and heart, from the CXR. These bone suppressed and lung segmented CXRs are validated by radiology experts and then used for training the classification model. Both BS and LS ensure that the features from the ROI are involved in decision-making. Thus, it improves the reliability of the system. Also, the selection of loss function was done by conducting a series of experiments. The results shown in Fig. 6a, 6b, 7, and 8 show the comparison in the AUC score of the 'EXP-Net' by employing weighted sigmoid loss and focal loss, respectively in various stages. It was observed that 'EXP-Net' gave better performance by employing the focal loss rather than the weighted sigmoid loss function. Both the loss functions help to resolve the problem of class imbalance in the training dataset.

Apart from the performance of the system on the validation, and testing datasets, visualization of the features using GradCam++ proved the reliability of the model. The team of radiologists verified the re-

sults shown by GradCam++, and validated that the features involved in decision-making were picked from the regions of infection.

Further, the comparison of the proposed system with the works proposed in the literature shows that the systems proposed so far lack in determining, and enhancing the quality of the dataset. None of the existing works focus on preprocessing the dataset by employing both the BS and LS. Also, there is a lack of data validation from radiology experts. In many cases, either the dataset used for training and testing is small or collected from a single source. This may fail to develop a robust model for screening COVID-19. Furthermore, most of the systems proposed in the literature focus on binary class classification. These systems fail to distinguish the lesions caused by Bacterial Pneumonia, Non-COVID Viral Pneumonia and COVID-19. As per discussion with radiology experts, it has been found that there is a class interference between Non-COVID Viral Pneumonia and COVID-19. Thus, it is hard for a DL model to distinguish between the samples of these classes. Also, there is a high degree of similarity in the samples of Bacterial and Non-COVID Viral Pneumonia. These samples are hard to classify. Although

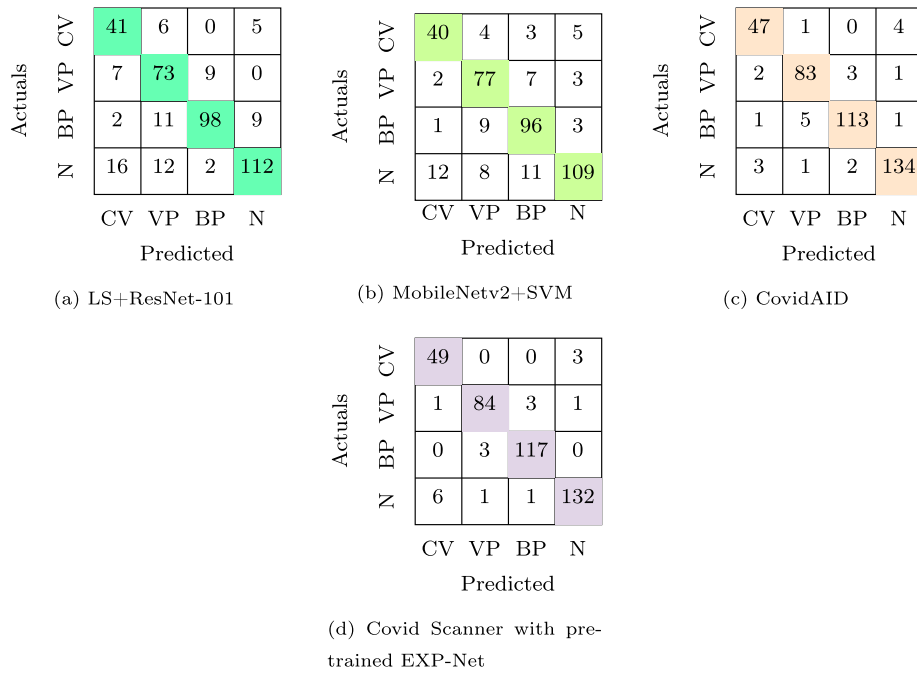


Fig. 11. Comparison in performance of pre-trained 'Exp-Net' and models proposed in literature.

Mangal et al. (2020) and Afshar et al. (2020) proposed their systems to address this problem, their systems have drawbacks. First, in both the researches the small test dataset undermines the reliability of the system. The value of 100% sensitivity reported for COVID-19 is based on 30 images only (Mangal et al., 2020). Similarly, the accuracy, sensitivity, and specificity reported in (Afshar et al., 2020) are based on 25 images of COVID-19 only. Moreover, both the works lack the evaluation of their model on real-life clinical datasets and validation by radiology experts. To further validate the supremacy of our 'Covid Scanner', we pre-train our "EXP-Net" on the NIH data and compare its performance with models proposed in (Kusakunniran et al., 2021), (Afshar et al., 2020), and (Mangal et al., 2020). It is evident from the results shown in Fig. 11 that the 'Covid Scanner' system performs the highest number of 382 correct classifications from the testing dataset comprising 401 CXRs.

Moreover, the performance of the 'EXP-Net' was compared with the state-of-the-art models viz. ResNet50, ResNet101, DenseNet121, InceptionV3, Xception, and EfficientNetB2. The supremacy of 'EXP-Net' is evident from the confusion matrices shown in Fig. 5d and 10. The improvement in the classification performance is due to the optimum selection of loss function, tailored architecture, and integration of preprocessing methods such as bone suppression and lung segmentation.

6. Conclusions

The research work proposed in this manuscript prepared a labeled dataset 'COVID-Pneumonia CXR Dataset' (CP-CXR) validated by radiology experts. The dataset comprises Anteroposterior (AP) view chest radiographs of COVID-19, Non-COVID Viral Pneumonia, Bacterial Pneumonia, and Normal classes. The AI-based screening system proposed in this manuscript is an integration of bone suppression, lung segmentation, and classification models. The proposed system successfully achieved its objective of correctly classifying the chest radiographs into CV, VP, BP, and N classes. It is evident from the confusion matrix shown in Fig. 5d and 11d that the system is efficient in classifying the BP from the VP. Also, the system is equally efficient in classifying the hard to distinguish samples of Non-COVID Viral Pneumonia and COVID-19.

Employing augmentation techniques increases the size of the dataset for training the proposed models and improves the classification accuracy by approximately 4.48%. It also improves the robustness, and reliability of the system. Further, applying the bone suppression and lung segmentation models ensure the feature extraction from the lung region only. It improves the classification sensitivity by 12.71%. This is proved by evaluating the system on the validation and testing datasets. Although there was a difference in the quality of the dataset, and degree of severity of infection, the system achieved the highest classification accuracy of 94.07% on the validation dataset. Further, it achieved the highest AUC of 96.58% on the validation, and 96.48% on the testing dataset. Next, employing the weighted sigmoid, and focal loss functions resolves the problem of class imbalance.

Based on the comparison of the proposed system with the systems proposed in the literature, and state-of-the-art models viz ResNet50, ResNet101, DenseNet121, InceptionV3, Xception, and EfficientNetB2, it is observed that the proposed system comprising preprocessing unit, and classification unit are more effective in the screening of chest radiographs to CV, VP, BP, and N classes. It reports an improvement of 1.30% in the value F1 score from the highest F1 score reported by the EfficientNetB2, a state-of-the-art model. Moreover, the proposed system fills the gap between the research and ground-level implementation. As per the best of our knowledge, none of the works proposed in the literature provides a validated, integrated, and ready-to-use system for mass screening of COVID-19. But, the research conducted in this manuscript offers a system that is ready for beta testing at the medical level. The system proposed in this manuscript is integrated with a web application and deployed on the cloud server 'AWS'. The system can work as an assisting tool for radiology experts. Based on the reliability and accuracy of the system a team of five radiology experts recommended its beta testing. Although the system is effective for primary screening of COVID-19 from the most related lung infections i.e. Bacterial Pneumonia, and Non-COVID Viral Pneumonia, there is a scope for further improving its classification accuracy. Presently, the system works on the AP view CXRs only, therefore in the future, the system's applicability can be extended to generate results on all other views of CXRs. Also, there is scope to develop such systems for screening of all kinds of diseases that can be captured in CXRs.

CRedit authorship contribution statement

Geeta Rani: Contributed in formal analysis, conceptualization, Investigation, finalizing the methodology, writing the original draft, review & editing the manuscript, and Project administration.

Ankit Misra: Responsible for data curation; finalizing the methodology, writing the original draft of the manuscript, software set up, and writing the code.

Vijaypal Singh Dhaka: Involved in supervision, review & editing the manuscript, and validating the results.

Deepak Buddhi: Being radiology experts, he contributed in data curation, continuous supervision, and validating the results, and analysis.

Ravindra Kumar Sharma: Being radiology experts, he contributed in data curation, continuous supervision, and validating the results, and analysis.

Ester Zumpano: Contributed in review & editing, preparing the final draft of the manuscript and funding acquisition.

Eugenio Vocaturo: Contributed in review & editing, preparing the final draft of the manuscript and supervision of the project.

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 paper.

Data availability

Data will be made available on request.

References

- Abbas, A., Abdelsamea, M. M., & Gaber, M. M. (2021). 4s-dt: self-supervised super sample decomposition for transfer learning with application to covid-19 detection. *IEEE Trans. Neural Netw. Learn. Syst.*
- Abdulrahma, S. A., & Salem, A.-B. M. (2020). Artificial intelligence for the detection of covid-19 pneumonia on chest ct using multinational datasets. *Fusion Pract. Appl.*, 2(1), 5–13.
- Afshar, P., Heidarian, S., Naderkhani, F., Oikonomou, A., Plataniotis, K. N., & Mohammedi, A. (2020). Covid-caps: a capsule network-based framework for identification of covid-19 cases from x-ray images. *Pattern Recognit. Lett.*, 138, 638–643.
- Chandra, T. B., Verma, K., Singh, B. K., Jain, D., & Netam, S. S. (2021). Coronavirus disease (covid-19) detection in chest x-ray images using majority voting based classifier ensemble. *Expert Syst. Appl.*, 165, Article 113909.
- Chattopadhyay, A., Sarkar, A., Howlader, P., & Balasubramanian, V. N. (2018). Grad-cam++: generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*. IEEE (pp. 839–847).
- Chollet, F. (2017). Xception: deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1251–1258).
- Greenspan, H., Estépar, R. S. J., Niessen, W. J., Siegel, E., & Nielsen, M. (2020). Position paper on covid-19 imaging and ai: from the clinical needs and technological challenges to initial ai solutions at the lab and national level towards a new era for ai in healthcare. *Med. Image Anal.*, 66, Article 101800.
- Gusarev, M., Kuleev, R., Khan, A., Rivera, A. R., & Khatkhat, A. M. (2017). Deep learning models for bone suppression in chest radiographs. In *2017 IEEE conference on computational intelligence in bioinformatics and computational biology (CIBCB)* (pp. 1–7). IEEE.
- Harmon, S. A., Sanford, T. H., Xu, S., Turkbey, E. B., Roth, H., Xu, Z., Yang, D., Myronenko, A., Anderson, V., Amalou, A., et al. (2020). Artificial intelligence for the detection of covid-19 pneumonia on chest ct using multinational datasets. *Nat. Commun.*, 11(1), 1–7.
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).
- Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4700–4708).
- Hussain, M., Bird, J. J., & Faria, D. R. (2018). A study on cnn transfer learning for image classification. In *UK workshop on computational intelligence* (pp. 191–202). Springer.
- Isola, P., Zhu, J.-Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1125–1134).
- Jain, R., Gupta, M., Taneja, S., & Hemanth, D. J. (2021). Deep learning based detection and analysis of covid-19 on chest x-ray images. *Appl. Intell.*, 51(3), 1690–1700.
- Kingma, D. P., & Ba, J. (2014). Adam: a method for stochastic optimization. *arXiv preprint, arXiv:1412.6980*.
- Kusakunniran, W., Karnjanapreechakorn, S., Siriapisith, T., Borwarnginn, P., Sutasananon, K., Tongdee, T., & Saiviroonporn, P. (2021). Covid-19 detection and heatmap generation in chest x-ray images. *J. Med. Imag.*, 8(S1), Article 014001.
- Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision* (pp. 2980–2988).
- Loey, M., Smarandache, F., & Khalifa, N. E. M. (2020). Within the lack of chest covid-19 x-ray dataset: a novel detection model based on gan and deep transfer learning. *Symmetry*, 12(4), 651.
- Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3431–3440).
- Lu, L., Shin, Y., Su, Y., & Karniadakis, G. E. (2019). Dying relu and initialization: theory and numerical examples. *arXiv preprint, arXiv:1903.06733*.
- Mangal, A., Kalia, S., Rajgopal, H., Rangarajan, K., Namboodiri, V., Banerjee, S., & Arora Covidaid, C. (2020). Covid-19 detection using chest x-ray. *arXiv preprint, arXiv:2004.09803*.
- Minaree, S., Kafieh, R., Sonka, M., Yazdani, S., & Soufi, G. J. (2020). Deep-covid: predicting covid-19 from chest x-ray images using deep transfer learning. *Med. Image Anal.*, 65, Article 101794.
- Nash Jr, J. F. (1950). Equilibrium points in n-person games. *Proc. Natl. Acad. Sci.*, 36(1), 48–49.
- Nguyen, N., Strnad, O., Klein, T., Luo, D., Alharbi, R., Wonka, P., Maritan, M., Mindek, P., Autin, L., Goodsell, D. S., et al. (2020). Modeling in the time of covid-19: statistical and rule-based mesoscale models. *arXiv preprint, arXiv:2005.01804*.
- Ohata, E. F., Bezerra, G. M., das Chagas, J. V. S., Neto, A. V. L., Albuquerque, A. B., de Albuquerque, V. H. C., & Reboucas Filho, P. P. (2020). Automatic detection of covid-19 infection using chest x-ray images through transfer learning. *IEEE/CAA J. Autom. Sin.*, 8(1), 239–248.
- Pradhan, N., Dhaka, V. S., Rani, G., & Chaudhary, H. (2020). Transforming view of medical images using deep learning. *Neural Comput. Appl.*, 1–12.
- Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., et al. (2017). CheXnet: radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint, arXiv:1711.05225*.
- Roberts, M., Driggs, D., Thorpe, M., Gilbey, J., Yeung, M., Ursprung, S., Aviles-Rivero, A. I., Etmann, C., McCague, C., Beer, L., et al. (2021). Common pitfalls and recommendations for using machine learning to detect and prognosticate for covid-19 using chest radiographs and ct scans. *Nat. Mach. Intell.*, 3(3), 199–217.
- Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: convolutional networks for biomedical image segmentation. In *International conference on medical image computing and computer-assisted intervention* (pp. 234–241). Springer.
- Roy, P. K., & Kumar, A. (2022). Early prediction of COVID-19 using ensemble of transfer learning. *Comput. Electr. Eng.*, Article 108018. <https://doi.org/10.1016/j.compeleceng.2022.108018>. <https://www.sciencedirect.com/science/article/pii/S004579062200283X>.
- Saddik, A., Latif, R., & Bella, A. (2021). Ecg signal monitoring based on covid-19 patients: overview. *J. Intell. Syst. Int. Things*, 2(2), 45–54.
- Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2018). Mobilenetv2: inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4510–4520).
- Sekeroglu, B., & Ozsahin, I. (2020). Detection of covid-19 from chest x-ray images using convolutional neural networks. *SLAS TECHNOL. Transl. Life Sci. Innov.*, 25(6), 553–565.
- Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint, arXiv:1409.1556*.
- Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2818–2826).
- Tan, M., & Le, Q. (2021). Efficientnetv2: smaller models and faster training. In *International conference on machine learning*. PMLR (pp. 10096–10106).
- Vieira, S. M., Kaymak, U., & Sousa, J. M. (2010). Cohen's kappa coefficient as a performance measure for feature selection. In *International conference on fuzzy systems* (pp. 1–8). IEEE.
- Yadav, G., Maheshwari, S., & Agarwal, A. (2014). Contrast limited adaptive histogram equalization based enhancement for real time video system. In *2014 international conference on advances in computing, communications and informatics (ICACCI)* (pp. 2392–2397). IEEE.
- Yu, F., & Koltun, V. (2015). Multi-scale context aggregation by dilated convolutions. *arXiv preprint, arXiv:1511.07122*.
- Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J. (2017). Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2881–2890).

- Zhong, A., Li, X., Wu, D., Ren, H., Kim, K., Kim, Y., Buch, V., Neumark, N., Bizzo, B., Tak, W. Y., et al. (2021). Deep metric learning-based image retrieval system for chest radiograph and its clinical applications in covid-19. *Med. Image Anal.*, 70, Article 101993.
- Zumpano, E., Fuduli, A., Vocaturo, E., & Avolio, M. (2021). Viral pneumonia images classification by multiple instance learning: preliminary results. In *25th international database engineering & applications symposium* (pp. 292–296).