Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Covid-MANet: Multi-task attention network for explainable diagnosis and severity assessment of COVID-19 from CXR images

Ajay Sharma, Pramod Kumar Mishra*

Department of Computer Science, Institute of Science, Banaras Hindu University, Varanasi 221005, India

ARTICLE INFO

Article history:
Received 21 September 2021
Revised 24 April 2022
Accepted 2 June 2022
Available online 6 June 2022

Keywords:
Covid-19
Lung segmentation
Infection segmentation
Chest X-ray
Deep learning
Transfer learning
Explainable AI

ABSTRACT

The devastating outbreak of Coronavirus Disease (COVID-19) cases in early 2020 led the world to face health crises. Subsequently, the exponential reproduction rate of COVID-19 disease can only be reduced by early diagnosis of COVID-19 infection cases correctly. The initial research findings reported that radiological examinations using CT and CXR modality have successfully reduced false negatives by RT-PCR test. This research study aims to develop an explainable diagnosis system for the detection and infection region quantification of COVID-19 disease. The existing research studies successfully explored deep learning approaches with higher performance measures but lacked generalization and interpretability for COVID-19 diagnosis. In this study, we address these issues by the Covid-MANet network, an automated end-to-end multi-task attention network that works for 5 classes in three stages for COVID-19 infection screening. The first stage of the Covid-MANet network localizes attention of the model to the relevant lungs region for disease recognition. The second stage of the Covid-MANet network differentiates COVID-19 cases from bacterial pneumonia, viral pneumonia, normal and tuberculosis cases, respectively. To improve the interpretation and explainability, three experiments have been conducted in exploration of the most coherent and appropriate classification approach. Moreover, the multi-scale attention model MA-DenseNet201 proposed for the classification of COVID-19 cases. The final stage of the Covid-MANet network quantifies the proportion of infection and severity of COVID-19 in the lungs. The COVID-19 cases are graded into more specific severity levels such as mild, moderate, severe, and critical as per the score assigned by the RALE scoring system. The MA-DenseNet201 classification model outperforms eight state-of-the-art CNN models, in terms of sensitivity and interpretation with lung localization network. The COVID-19 infection segmentation by UNet with DenseNet121 encoder achieves dice score of 86.15% outperforming UNet, UNet++, AttentionUNet, R2UNet, with VGG16, ResNet50 and DenseNet201 encoder. The proposed network not only classifies images based on the predicted label but also highlights the infection by segmentation/localization of model-focused regions to support explainable decisions. MA-DenseNet201 model with a segmentation-based cropping approach achieves maximum interpretation of 96% with COVID-19 sensitivity of 97.75%. Finally, based on class-varied sensitivity analysis Covid-MANet ensemble network of MA-DenseNet201, ResNet50 and MobileNet achieve 95.05% accuracy and 98.75% COVID-19 sensitivity. The proposed model is externally validated on an unseen dataset, yields 98.17% COVID-19 sensitivity.

1. Introduction

Coronavirus disease 2019 (COVID-19), has been declared a global epidemic by WHO in early March 2020. Globally, as of 11 August 2021, a corpus of 204,644,849 confirmed cases including 4323,139 deaths have been reported to WHO because of COVID-19 infection caused by SARS CoV-2 [1]. Still, these numbers are increasing and predicted to grow rapidly [12] in the upcoming months because of the higher reproduction rate of the disease. Therefore, it is essential to lessen the spread of virus which is possible only by early detection, treatment, and isolation of virus cases.

In general, three major screening methods used for early diagnosis of COVID-19 include, reverse transcriptase-polymerase chain reaction (RT-PCR), Chest X-ray (CXR) and Chest computed tomography (CT) [2]. The standard COVID-19 diagnosis RT-PCR method detects virus RNA from nasopharyngeal swab or sputum but it requires expert personnel, specific material, and laboratories for testing and is a time-consuming process. Another alternative to the RT-PCR test is the rapid antigen test, which gives a faster and

* Corresponding author.
E-mail address: mishra@bhu.ac.in (P.K. Mishra).

https://doi.org/10.1016/j.patcog.2022.108826
0031-3203/© 2022 Elsevier Ltd. All rights reserved.

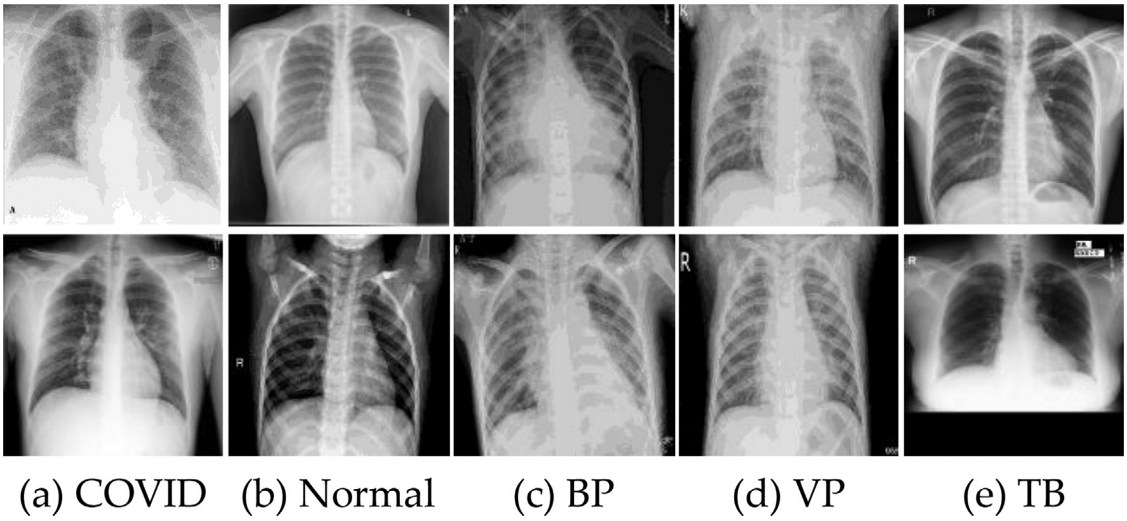


Fig. 1. Example sample images; (a) COVID-19 (b) Normal (c) BP (d) VP (e) TB classes.

less expensive diagnosis compared to PCR but has poor sensitivity to COVID-19 [3]. In particular, radiological examinations were found very useful in the diagnosis and assessment of COVID-19 disease progression. Some recent research studies were conducted by expert radiologists on sensitivity analysis of COVID-19 by RT-PCR, CT, and CXR scans. Since preliminary 3-days clinical practice performed on CT and RT-PCR test involving 51 patients describes with CT images a sensitivity of 98% outperforming RT-PCR having 71% sensitivity [4]. However, due to the rapid surge in COVID-19 cases, routine use of chest CT is not feasible because of its portability and expensive setup. Another study by wong et al. [12], involving 64 participants, reported lower sensitivity of 69% using the CXR modality compared to RT-PCR having 91% initial sensitivity. Despite having low sensitivity, more than 9% of negatively reported RT-PCR tests have shown certain abnormalities of COVID-19 in CXR radiographs. Although the evaluation by CXR is less accurate in the early studies compared to CT and RT-PCR. Still, it is considered an efficient and standard screening tool because of its low-cost, minimally-invasive, quick results, and requires simpler logistics for its implementation [16]. In addition to COVID-19, other common lung diseases like viral pneumonia, bacterial pneumonia, and tuberculosis possess almost similar symptoms as observed in COVID-19 pneumonia. Thus, it is essential to develop a system that uniquely identifies COVID-19 patients among these common lung diseases. Fig. 1 displays example samples for each disease class considered in this study.

Specifically, in this decade (2011–2021) deep learning approaches have been successfully explored for image classification and segmentation tasks. These approaches play a vital role in analyzing CXR radiographs as a standard tool for the early classification of COVID-19 disease. In 2012, the first advanced deep learning model AlexNet [5] was proposed that introduces overlap pooling, ReLU, dropout, large 11×11, and 5 × 5 kernels. In the meantime, evolution in deep learning approaches rapidly increases after AlexNet won the ImageNet [10] challenge with a 16.4% error rate. Some advanced deep learning architectures proposed include VGG [6] in 2014 uses a small 3 × 3 kernel, GoogleNet [6] in 2015 introduced the block concept, split and merge idea with different kernel sizes. Inception [7] in 2015 modified the GoogleNet model and introduced the split, transform and merge concept with asymmetric filters. ResNet [7] in 2016 introduced residual learning with skip connections, MobileNet [6] in 2017 introduced depth-wise convolution [8] followed by pointwise convolution [9], and DenseNet [11] in 2017 introduced cross-layer connections for information flow. The evolution of these deep learning models [12] improve the classification ability continuously when tested on benchmark datasets [13]. Some application areas [14] where these CNN models have been successfully employed for early diagnosis include cancer [15] classification [16], skin lesions classification [17], and the diabetic retinopathy diagnosis [18], etc.

1.1. Literature review

Recently, researchers proposed numerous AI-based methods for the detection of COVID-19 using chest x-ray and CT images. Even though the usage of deep learning approaches gave promising results having higher statistical scores. Still, these works need desirable improvements at the model development stage to achieve an explainable diagnosis for classification and infection region segmentation. For instance, the first research work proposed by Wang et al. [19] indicates that the application of deep learning models successfully identifies COVID-19 signs in CXR scans. Models successfully predict 80% correct results for the COVID-19 class. However, the reduction of false negatives is essential in building a reliable diagnosis system. In research work [20], the author projected COVID-SDNet methodology that tends to improve false negatives and classify COVID-19 into different severity labels such as mild, moderate, severe, normal-PCR+. Models composed COVIDGR-1.0 dataset available publicly and introduced segmentation-based cropping by discarding areas outside lung regions without lung segmentation in the raw image. In addition, transformation and augmentations of CXR scans based on the GAN model improved ResNet50 performance with an accuracy of 97.72%, 86.90% in severe, moderate cases. COVID-DeepNet, [21] hybrid multimodal model proposed for COVID-19 diagnosis using CXR images. Image-preprocessing with CLAHE and Butterworth bandpass filter eliminates noise and enhances contrast. Results of CNN and deep belief networks are fused for final class prediction among two classes with COVID-19 sensitivity of 99%. In the study [22], Shamsi proposed a transfer learning-based deep uncertainty aware model for CT and CXR analysis. The proposed study extracts feature by four DL approaches**,** InceptionResNetV2, ResNet50, VGG16, and DenseNet121, each passed to eight machine learning models for recognition of COVID-19. The statistical analysis reported SVM and neural net have higher accuracy and sensitivity ranging from 88% to 97% for the two-class problem. Research study by Wang et al. [23] proposed a classification model involving two stages of classification, named as Discrimination-DL stage extracting lung features and Localization-DL stage to localize features in left, right lungs regions. This network reported 98.71% accuracy of Discrimination-DL and 93.01% for localization network. In a study [24], the author developed a Graph-based approach for COVID-19 disease identification with minimal supervision required. This semi-supervised learning framework deals with vast unlabelled samples by pseudo-labeling. Results outperformed the supervised framework with better identification based on attention maps. In the reference study [25], the ConvNet model is developed based on fuzzy logic and deep learning models. Model improved learning when features extracted by a hybrid of deep and fuzzy logic approach given to multilayer perceptron for classification achieve 81% accuracy.

Further developments in the state-of-the-art involve lung segmentation, image enhancement, and the creation of localization maps by Grad-CAM to make models explainable. In a research study [26], the author explored the effect of image enhancement on COVID-19 diagnosis using CXR data. Various techniques like CLAHE, histogram equalization (HE), gamma correction, image complement, and BCET were tested on raw and segmented images. It created a dataset of 18,479 images with classes as normal, COVID-19, and pneumonia. The experimental study uses InceptionV3, ChexNet, ResNet50, ResNet101, and DenseNet201 as CNN models, where ChexNet and DenseNet201 achieve 96.29% accuracy when applied with gamma correction. Another model developed by author [27] CMTNet for recognition of COVID-19 patients and finding regions infected by a disease. The proposed model follows VGG19 backbone network for the encoder model. Two separate decoder branches for lung and disease segmentation shares a common encoder branch. The study compiled an annotated dataset of 9000 CXR for lung segmentation and 200 for disease localization. Statistical analysis gains 87.20% sensitivity and 96.80% precision. In [28], the author improved false negatives and interpretation of the COVID-Net model by tuning existing parameters. The study recognizes 87% of COVID-19 samples correctly among pneumonia and normal classes. This modified model is tested on a large corpus of 8573 COVID-19 samples. The statistical analysis shows improvement in accuracy and interpretation after lung segmentation to 91% with 87% sensitivity. In such epidemic situation, data collection is the biggest challenge to train deep learning models. In research work [29], the author proposed a transfer learning, ResNet model with modified patch-wise training and lung segmentation approach to mitigate limited dataset challenge. The model was trained to classify 5 classes i.e., COVID-19, normal, viral pneumonia, bacterial pneumonia, and TB. Lung contours are segmented with the FC-DenseNet model and ResNet-18 to classify segmented lungs among one of the pathology classes with 89% accuracy and 85% sensitivity, respectively. Ghoshal and Tucker, [30] explored the uncertainty of deep learning models based on Spearman correlation and bayesian uncertainty. Usage of bayesian CNN with drop weights and ResNet50v2 improved diagnostic interest and explainability of class activation maps. Shi et al. [31] developed EXAM, an explainable attention-based model generating results explaining the diagnostic interest of models for classification by Grad-CAM for 3-classes. EXAM developed attention by merging spatial and channel-wise features with DenseNet models. Kumar and Singh et al., [32] proposed a deep learning model for distinguishing three classes COVID-19, normal, and pneumonia. Image segmentation, and image enhancement techniques are used to develop a stacked ensemble of four CNN models using Naïve Bayes as a meta learner and Grad-CAM is used for qualitative interpretation investigation. In the study [44], Mangal et al., proposed an explainable CovidAID model diagnosing COVID-19 from viral pneumonia, bacterial pneumonia, and normal cases using the backbone model DenseNet121 outperforming COVID-Net with an accuracy of 90.5% and sensitivity of 100% with 182 PA view COVID-19 samples. Another research work done by Wang et al. [33] makes a standardized DL pipeline for the classification of COVID-19 pneumonia and lesion visualization for diagnosis. The standardization stage discards irrelevant features outside lung regions by capturing lung regions. The statistical analysis measures AUC score of the system ranges from 0.87 to 0.97 for COVID and other or viral pneumonia,0.87 for non-severe and severe COVID-19, between 0.94–0.98 for viral and other pneumonia types.

In the study [34], Tahir et al. proposed a model for infection quantification and grading of COVID-19 pneumonia. The study compiled a large dataset of 33,920 samples having 11,956 COVID-19 samples in original and augmented form. Statistical analysis compared results of segmented, non-segmented lungs and infection by state-of-the-art UNet, UNet++, and FPN (Feature pyramid learning) models. These models utilize ResNet, VGG19, DenseNet121, etc. as the backbone of encoder-decoder networks. Model localized COVID-19 infection regions with an overall Dice score of 88%. Signoroni et al. [35] proposed BS-Net model providing a deep learning-based framework for the diagnosis of COVID-19 in CXR. Model localize infection regions by assigning quantitative scoring to lung regions based on the Brixia scoring system. Weakly supervised learning is used to achieve tasks such as segmentation, score-estimation, and spatial alignment. BS-Net tested on 5000 CXR segmented images based on the UNet variant model with a 94% IoU score. Research work proposed by Gidde et al. [36] developed CovBaseAI explainable decision system by an ensemble of three deep learning models for COVID-19 diagnosis. Model validation is performed by 2 datasets having a corpus of 471 and 1401 for COVID-19/Normal CXR scans. The statistical analysis achieved an accuracy of 87% accuracy with 98% negative predictive value.

In addition to standard CXR modality, some research models have been developed using CT, ultrasound modality, or using multimodal data involving CXR and CT images both. But in real practice testing using CT images lacks portability and is expensive compared to the CXR modality. In this context, Owais et al. [37] proposed a lightweight deep learning ensemble network by combining FCNet, ShuffleNet, and MobileNet for COVID-19 diagnosis. Localization and activation map visualization enables radiologists to diagnose infection focused by model. However, the model is tested with a collection of CXR and CT datasets with a mean F1-score of 95.94% and 94.60%. The area under curve achieved is 97.99% and 97.50%, respectively. Another model [38] proposed for diagnosing community-acquired pneumonia from 3D CT volumes using a dual-sampling attention network. Dual sampling resolves class imbalance issue and attention mechanism better identifies infection in CT volumes with 3D CNN model outperforming state-of-the-art UNet model. The study processed a corpus of 2186 CT scans with labels COVID-19/NORMAL by 5-fold cross-validation. Statistical analysis attained AUC, F1-score, and accuracy of 0.944, 82%, and 87.5%, respectively. The majority of research studies involve CXR and CT modalities for the diagnosis of COVID-19 by application of transfer learning models. However, HORRY and CHAKRABORTY [39] research work developed a multimodal model for diagnosis among multiple modalities by similar transfer learning approaches. In addition to CT and CXR scans, ultrasound modality has also proven useful in the diagnosis process but has portability and cost issues for large masses. Study optimized VGG19 model, compared with other popular CNN models DenseNet, Inception, VGG16, and ResNe50 for statistical analysis. Statistical measures reveal Ultrasound modality gives superior results when compared with CT and CXR images for the 3-class classification problem. Precision values achieved are 86% for X-ray, 84% for CT, and 100% for ultrasound values having F1-score of 99%. However, the model is a black box lacking explanations and localization visualization maps using whole slide images. Another explainable research model [40] JCS proposed for the diagnosis of COVID-19 in chest CT images. JCS framework involves joint segmentation and classification to identify COVID-19 infection in CT images. The proposed work compiled a large dataset of 144,167 CT images from 350 normal CT scans and 400 COVID-19 patients. For infection quantification, 3855 CT images are annotated named COVID-CS dataset. Joint diagnosis achieves sensitivity and specificity of 95% and 93%, respectively, for classification whereas 78.5% dice score for infection segmentation.

1.2. Issues affecting results in the literature

In this section, we discuss certain shortcomings after critically analyzing state-of-the-art developments associated with existing diagnosis approaches. Despite having good results of existing methodologies still, these approaches lack explainable diagnosis since the interpretation of the model's decision is not correlated to lung regions. The major challenge with existing methodologies lies in weaker interpretation and low explainability of models for the classification of disease. Some state-of-the-art models created only backbox, not explaining whether the model focused on relevant lung regions for classification. In the beginning, models applied directly on whole slide CXR radiographs gave poor interpretation enabling classification based on out-of-lungs region. Some existing studies performed segmentation of lungs based on the UNet model, shows improvement in interpretation but there is a minor accuracy drop in these approaches. However, most of the works show a good correlation to disease but signify weaker diagnostic interest to lung regions. Indeed, the proposed segmentation-based cropping approach represents the strong correlation of disease classification to the lung regions.

Secondly, most studies revolve around the classification of two or three classes distinguishing COVID-19 from normal and pneumonia classes [19]. However, tuberculosis and virus pneumonia have similar symptoms as seen in COVID-19 but after extensive searches, only two studies were found involving five classes for classification. These studies lack interpretation and explainable nature [22] of models working with limited example datasets. However, in clinical practice, explainable diagnostic system essentially have infection region quantification and severity assessment capability in addition to detection that is missing.

Additional experimental issues in the state-of-the-art as per recommendations of work published in nature [41] have been considered to avoid pitfalls in classification tasks. The proposed study strictly considers both mandatory and non-mandatory recommendations for deep learning modeling to pass the quality screening procedure. The major issues responsible for the failure of quality screening procedure are; no explanations provided for final model selection, lack of image pre-processing steps, no clear specifications for training approach, class imbalance, absence of robustness and sensitivity analysis, no external validation, lack of performance analysis on confidence threshold intervals, interpretation and generalization. In brief, some published works present unclear implementation details such as; initial preprocessing steps [23] for conversion of the dataset into a uniform format and the range is either missing or not specified. In addition, specifications of training parameters like an optimizer, learning rate, loss function [31], and reason for selection of the best model are missing [54]. In the literature survey, all the studies involve two or more classes but few performed class-wise sensitivity analysis [35] for better behavior analysis of the model [36] for each of the classes. Despite class imbalance majority of studies uses categorical class entropy loss [36] that may prioritize the learning of the majority class [55] only. In the proposed research work, we consider all these challenges to develop a final multi-task model improving interpretation and generalization by segmentation-based cropping, explainable diagnosis by infection quantification model, localization map by Grad-CAM, and class imbalance by the multilabel classification loss function. The model robustness is maintained by a clear experimental setup, pre-processing steps, class-wise sensitivity analysis, confidence analysis, and external validation.

1.3. Research contributions

The early success of deep learning methods for COVID-19 diagnosis motivates us to further investigate CNN to minimize existing issues and improve the interpretation and explainability of diagnosis. We propose the Covid-MANet model that enables diagnosis and progression of the disease by analyzing relevant features in lung regions extracted using segmentation masks. Covid-MANet is a three-stage deep learning model that benefits from lung localization, segmentation-based classification, data-augmentation, preprocessing, post-processing, and multi-label loss to curtail any variances and imbalance among CXR radiographs collected from different repositories. Covid-MANet works for the 5-way classification problem, differentiating COVID-19 from bacterial pneumonia (BP), viral pneumonia (VP), normal, and tuberculosis (TB) cases. Precisely, the first stage in Covid-MANet takes Raw CXR radiographs as inputs, uses a segmentation model, makes a prediction of CXR lung masks and performs the post-processing of predicted masks for localization of the lung region. This forces model to attention more on segmented regions resulting in improved interpretation and generalization. In addition, segmentation models are trained with new pseudo RANZCR dataset in combination with existing datasets to enable better segmentation of VP and BP cases. In the second stage, the MA-DenseNet201 detection model improves its generalization by concentrating more on relevant lung regions extracted by the first stage. Segmentation-based cropping is conducted to further improve the interpretation of models. The final stage of the Covid-MANet model performs quantification and severity assessment of COVID-19 infection in the lungs that assist doctors to understand the progression of the disease. The severity levels assigned to the disease are mild, moderate, severe, and critical as per the RALE scoring system. In the end, Grad-CAM is used to interpret the features focused by the model for the classification of an image to a particular class in each experiment. To accomplish segmentation, UNet, UNet with residual block, and UNet with Dense block are used but UNet with dense block achieves better segmentation of lung regions compared to the other two. To perform classification in these experiments, the proposed model is compared with state-of-the-art CNN models, VGG16, VGG19, InceptionV3, MobileNet, ResNet50, NASNetMobile, DenseNet121, DenseNet201 based on performance in the existing methodologies and ImageNet database. The proposed infection segmentation system uses UNet with DenseNet121 encoder for quantification of infection region in COVID-19 classified samples. The major contributions of this study are summarized as follows:

(1) The Covid-MANet is a single generic multi-task framework for automated COVID-19 diagnosis, infection region quantification and severity assessment of COVID-19 into more specific levels as mild, moderate, severe, and critical.
(2) The proposed work provides an enhanced segmentation-based classification model with dense blocks where modified UNet architecture is investigated with two other SOTA models for automated segmentation of lung regions in CXR images. The large dataset is created by the inclusion of the RANZCR dataset that provides better segmentation of unseen VP and BP images.
(3) The Covid-MANet model improves generalization and interpretability for COVID-19 classification by introducing segmentation-based cropping and classification by the MA-DenseNet201 model with multiscale attention network outperforming state-of-the-art networks. In addition, this end-to-end framework not only classifies but also segment infection region aimed at screening the progression of the disease.

(4) The proposed work investigates the class-wise sensitivity analysis in three experiments. Based on prior awareness of various class level accuracies, we propose a weighted average ensemble approach (WAE) that outperforms state-of-the-art models for all the classes.
(5) Finally, a gradient-weighted class activation mapping (Grad-CAM) is used for explainable diagnosis to generate a localization map for each disease type investigates model interpretation in addition to COVID-19 infection map. The segmentation-based cropping approach reduces all biases and develops a generalizable model more stable as compared to whole slide and segmented images.

The structure of the whole research study is organized as; Section 2 discusses the detailed framework of the proposed model involving lung segmentation, classification and infection segmentation. Section 3 discusses the dataset resources utilized, experimental setup, and evaluation metrics for segmentation and classification. Section 4 discusses the quantitative and qualitative analysis of the proposed methodology and Section 5 ends with a conclusion.

2. Covid-MANet multi-task framework

This section presents the methodology of the Covid-MANet network developed for the diagnosis and infection quantification of COVID-19 in CXR samples. The primary focus of the multi-task network is to develop an explainable diagnosis system representing a higher correlation of disease classification to the lungs region. In the first task, the segmentation network localizes lung areas ready for input to the classification model based on segmentation-based cropping. Then, the proposed classification model classifies the image into one of pathology class and the infection segmentation network quantifies the severity of COVID-19 infection into mild, moderate, severe, or critical. The detailed discussion of the multi-task network and loss functions for segmentation and classification is explained in the following sections:

2.1. Lung segmentation network and data processing

The first task of the Covid-MANet network is to segment out lung regions since the relevant disease information lies only in the lung regions. Recall that the proposed model works for five classes, differentiates COVID-19 from other lung diseases. However, no public dataset available in all five classes in one dataset. So, a compilation of the dataset by the fusion of the COVID-19 class with other lung disease resources may possess different acquisition conditions and artifacts that affect classification results. Classification based on lung segmentation resolves all these biases, therefore resulting in higher interpretation and generalization of the model. The automated lung segmentation task achieved by UNet and its variants inspired by ResNet and DenseNet model. Fig. 2 shows the generic architecture of the segmentation network and its blocks. Fig. 2(a) represents a general encoder-decoder model for segmentation where skip connections connect encoded features to the decoder. Three segmentation models can be constructed by replacing the 3 × 3 block of a general model with structural blocks (b–d). Fig. 2(b) displays 2 convolutional layers of the UNet block [42], Fig. 2(c) represents a basic block of ResUNet model [43], Fig. 2(d) displays the building block [43] of DenseUNet model. The training of segmentation models is achieved by hybrid segmentation loss function [45] and the best model selection by IoU and Dice coefficient metrics. The references for compilation of a larger dataset compared to other models specified in Section 3.1, containing images and labeled lung annotated masks for supervision. However, the lung masks for COVID-19, TB, and normal classes are publicly available, but VP and BP classes have no annotated masks. So, another publicly available RANSCR dataset is utilized that has a similar UID compared to VP and BP, enabling better lung localization of unsupervised BP and VP classes.

2.1.1. Pre-processing and post-processing

Pre-processing is an essential step to achieve uniformity of each input class required for modeling since the dataset compiled by a fusion of different classes that may represent non-uniformity in size, range, or datatype. Pre-processing operations applied before segmentation includes resizing of input to 320 by 320, change of datatype to float32, and range consistency by normalization to [0,1]. The augmentation operations are applied to increase the size of the dataset and avoid overfitting. These operations are, image shifting range [−5.25%, −5.25%] vertically and horizontally, image rotation by [−15, 15] angle, Zoom by [0.05, 0.05] range. Augmentation operations are considered while training the model whereas model evaluation is free from augmentation. The proposed DenseUNet model outperformed ResUNet and UNet model in lung localization task with higher dice and IoU score. DenseUNet model is considered best for lung localization of unseen classification task. In addition, post-processing operations are applied to eliminate minor defects in the prediction of unseen images inside or outside ROI to improve lung localization. Three post-processing operations applied in a sequence using the Scikit-image library. These operations include first labeling the predicted mask, discarding small objects inside or outside the lungs by keeping the two largest lung regions, and finally preserving edges by dilation operation with a 5 × 5 kernel. Fig. 3 shows lung localization maps for each class, where predicted lung masks are post-processed to eliminate artifacts out-of-lung region.

2.2. Classification network and data processing

The second stage of the Covid-MANet network aims to classify the input CXR sample into one of five pathology classes based on the correct interpretation to the lung localization area. To improve interpretation and generalization three experiments have been conducted with and without lung segmentation. The classification network adopted the proposed multi-scale attention model named MA-DenseNet201 using DenseNet201 [11] as a backbone architecture that acts as a feature extractor. The reason for using a pre-trained backbone network is to avoid overfitting, better training and extraction of features even from small datasets. The proposed model is compared with state-of-the-art pre-trained deep learning networks such as VGG16, VGG19, MobileNet, ResNet50, InceptionV3, DenseNet121, DenseNet201, and NASNetMobile [6] in each of the experiment. Once lung segmentation masks of the classification dataset are ready to use after post-processing are masked with the corresponding classification image to obtain the region of lungs. These lung segmented images are given to the classification network with and without a segmentation-based cropping approach. The classification network aims to improve interpretability and generalization by training and evaluation based on three distinct experiments. The qualitative analysis measuring the interpretation of each class type based on the best model in each experiment is shown in Fig. 5.

The detailed analysis of experiments and processing steps followed to improve interpretation is discussed as follows;

2.2.1. Experiment 1: raw data

In experiment 1, classification models are trained on whole slide images without the involvement of the lung localization model. The input to the classification model is pre-processed, where images are resized and normalized to a fixed range (0,1). The visualization map generated by Grad-CAM explains result of

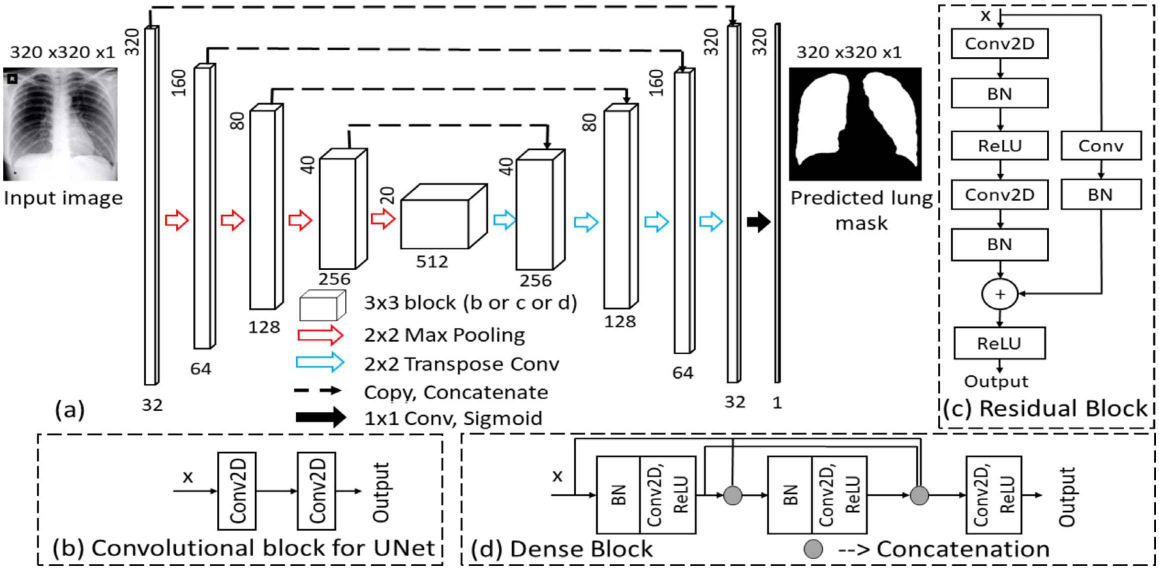


Fig. 2. Architecture of lung segmentation network, (a) basic UNet type model where 3 × 3 block is replaced by block b, c or d (b) Convolutional block of UNet (c) Residual convolutional block of ResUNet (d) Dense convolutional block of DenseUNet model.

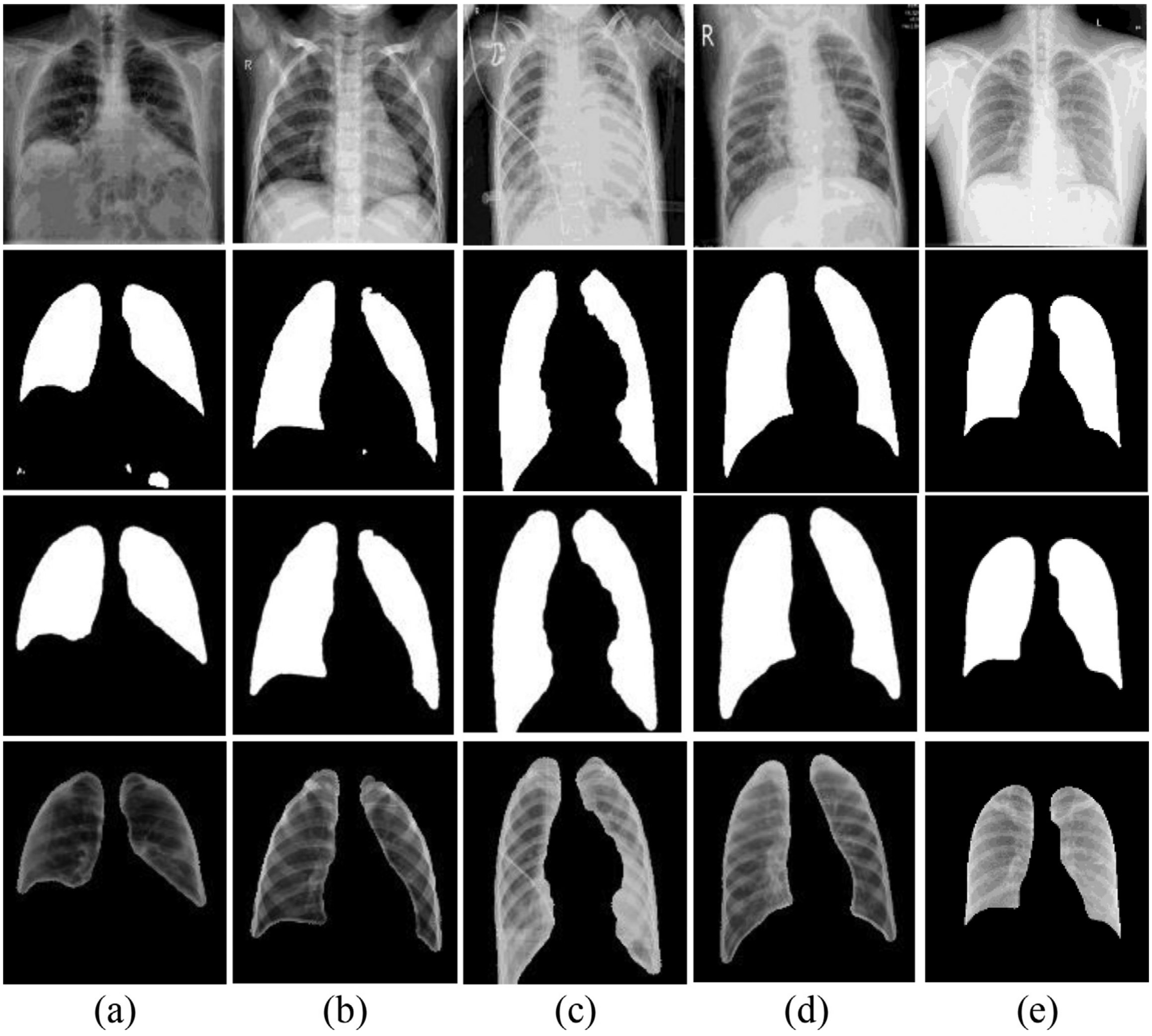


Fig. 3. Example of segmented CXR samples, (a) COVID-19, (b) Normal, (C) BP, (d) VP, (e) TB, where rows correspond to the original pre-processed image, predicted lung masks, post-processed lung masks and final segmented lung contour.

this experiment are least interpretable as the model focuses more on artifacts or information present outside lung regions.

2.2.2. Experiment 2: lung localization

The second experiment has been conducted to resolve the difficulty that arises in the first experiment by the addition of a lung localization model. The lung localization model eliminates artifact outside the lungs enabling better classification based on the interesting region of the lungs. To achieve this, classification data is first given to the stage 1 lung localization model generating post-processed lung masks. The post-processed masks are masked with corresponding raw images to zero out information outside lung regions, generating lung segmented images. These segmented images are given to classification models for training, after resizing and normalization to a consistent range. However, some models are trained with three different versions in search of the best suitable approach. Initially, lung localization images are directly given as input to the classification model. Secondly, the application of the attention module CBAM involves channel attention followed by spatial attention applied before classification layers in a residual manner. Finally, the image enhancement approach gamma correction is applied to segmented images before being given to the classification model. Experiment 2 improved the interpretation of models, but few images make incorrect interpretations considering the region outside the lungs.

2.2.3. Experiment 3: segmentation based cropping

The proposed experimental approach resolves the difficulties of experiment 2 with maximum interpretation among all three experiments. This approach uses post-processed lung localization maps generated by stage 1, masked with original images, and finds a convex hull to completely discard edges. Generation of best-fit convex hull delimits left, top, right, and bottom regions of segmented lungs. The convex hull is computed by first converting the segmented image to a grayscale image, applying thresholding operation to the grayscale image, and detecting extreme points by contour detection. Finally, extreme points are utilized to crop segmented lungs ready for classification. Cropped images are given to the classification model for disease identification just after normalization in the range (0,1). Similarly, this approach is trained on four different versions explained as; Initially, cropped images are directly given to classification models. Secondly, cropped images are enhanced by gamma correction before given to the classification model. In the third version, attention mechanism with CBAM is applied in addition to gamma correction before the classification layers. Finally, cropped images are denoised using a total variation filter followed by gamma correction to generate data for classification. After extensive experiments, we found proposed segmentation-based cropping approach gives maximum interpretation as shown by Grad-CAM activation maps. The Grad-CAM localization maps indicate better interpretation in experiment 3 corresponding to all the classes involved in the study.

2.3. Multiscale attention-based classification model

Attention and multiscale feature extraction has been used in several studies to converge model attention to more relevant regions and discard less important features. We developed an enhanced classification model considering attention maps of different scales, guiding the model in the feedforward process. Multiscale hybrid attention module with transfer learning backbone makes better decisions based on high-level feature maps. DenseNet201 model is considered as backbone because of its deep feature extraction ability for better COVID-19 diagnosis.

The proposed hybrid attention generation model aims at extracting attention maps from multiple-scale feature maps. From the DenseNet201 pre-trained model, feature maps of shapes 14×14×1024, 7 × 7*1856, and 7 × 7*1920 are used to make a corpus of 32 attention maps. The schematic representation of the MA-DenseNet201 model and multiscale hybrid attention generator is shown in Fig. 4. Multiscale hybrid attention generator module extracts feature maps f1, f2, f3 from backbone model, generates attention A1, A2, A3 by 1 × 1 convolution operation. All the feature maps are downsampled to 7 × 7 feature maps. These attention maps connect residually to generate hybrid attention map A of size 32. The feature maps generated by the DenseNet201 backbone network concatenate with a multiscale hybrid attention generator. Finally, the output map is processed by a 1 × 1 convolution layer. The classification layers added on top of this network include flatten layer, dense layer followed by a dropout layer, and final dense layer with 5 neurons and a softmax activation function.

2.4. COVID-19 infection quantification and severity assessment

In clinical diagnosis, end-to-end explainable model development not only requires blackbox classification. But it is desirable to have an automated infection segmentation module measuring the progression of COVID-19 disease [34]. However, COVID-19 infection segmentation in the lungs is much more complex compared to lung segmentation. For infection region segmentation, we adopted UNet based encoder-decoder model [42] with a dense backbone network. UNet model with DenseNet121 backbone acts as an infection segmentation model, which is compared with seven other models involving UNet, Attention-UNet, UNet++, R2UNet [45] by considering backbone architecture as VGG19 with UNet++, ResNet50, and DenseNet201 with UNet. Similar to classification, the infection segmentation model is tested on two scenarios as of classification.

Furthermore, for a better understanding of the progression of infection in the lungs, COVID-19 infection is quantified based on the predicted lungs region. The quantification of infection in the lungs is computed by the sum of predicted disease pixels divided by the total number of pixels in the lungs. In addition, infection quantification in left and right lungs is computed separately. Finally, infection score assignment and severity assessment module are introduced based on RALE [35] scoring system. According to the RALE scoring system, lungs are divided into two separate regions left and right lungs. Each lung is assigned a score between 0 and 4 based on the percentage of infection found. The score 0 is assigned for no infection involvement, score 1 for infection less than 25%, score 2 for infection between 25% to 50%, score 3 for infection between 50%–75%, and score 4 for more than 75% involvement. In addition, the severity level [20] assigned as mild if the score is 1 or 2, moderate for a score between 2 and 5, severe if the score is more than 4 and infection less than 90%, critical if the infection is more than 90%.

The detailed overview of the Covid-MANet multi-task classification and infection segmentation system for COVID-19 diagnosis is shown in Fig. 6. Firstly, a binary lung mask is generated for each class using an encoder-decoder lung segmentation model. Then the predicted lung mask is superimposed with the input image to zero out the region outside the lungs generating a lung segmented image. This segmented image is given to the classification model by discarding irrelevant regions outside the lungs based on segmentation-based cropping. The classification model predicts the image into one pathology class based on higher output probability. For infection segmentation, the original image corresponding to the predicted COVID-19 image is given to the infection segmentation model that predicts the infection mask. Then for COVID diagnosed cases, quantification of the infection region is highlighted when the predicted infection mask and lung mask are superimposed with the original image. The infection masks are predicted

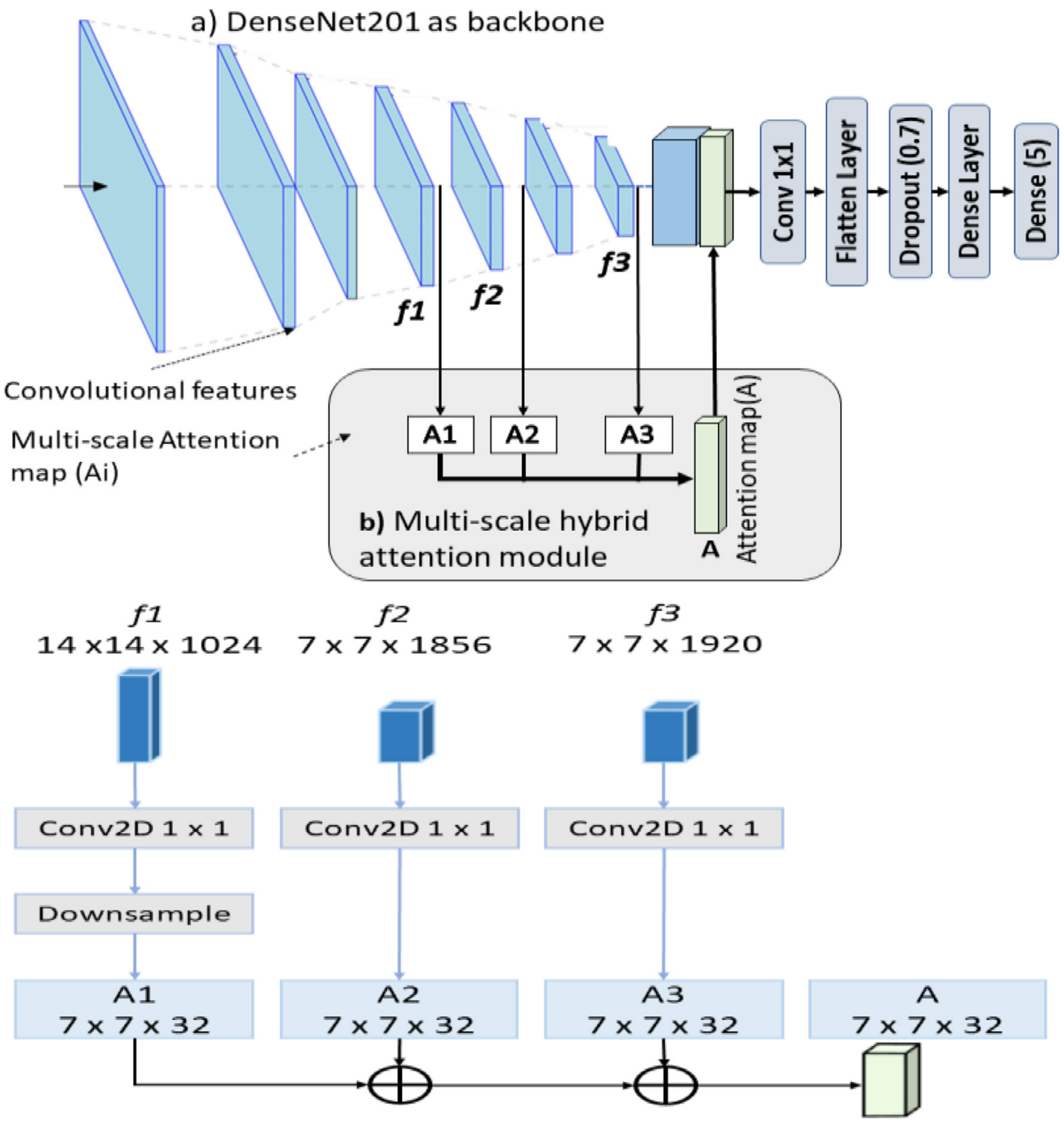


Fig. 4. The architecture of MA-DenseNet201 model adding multiscale hybrid attention module to DenseNet201 backbone.

at a threshold value of 0.5 and masked with corresponding lung masks to ensure infection regions outside the lungs are discarded. Accordingly, a COVID-19 positive sample is graded as mild, moderate, severe, and critical based on the proportion of pixels infected by the disease. Another case occurs when the model predicts the sample as COVID-19, but the infection segmentation system does not predict any disease pixel such a case is considered an *Asymptomatic* COVID-19 *infection case.* In addition, for each predicted class, an explainable diagnosis is achieved by generating the localization map focused by Grad-CAM [29] showing the region focused by model for classification.

2.5. Segmentation loss function

The loss function is one of the useful components in the training of segmentation and classification models. The primary objective of our study is to differentiate COVID-19 from other lung diseases and localize the COVID-19 infection region. The primary focus is on the lungs region, so firstly segmentation loss aims to learn the lung segmentation task and secondly infection region segmentation task. Because of higher class imbalance, we developed a hybrid loss function suitable for lung as well as infection segmentation. The segmentation loss is computed as the mean of dice loss and the intersection over union loss. Both dice coefficient and intersection over union almost measure the overlapping of predicted mask and ground truth whereas the segmentation loss function computes the error. The objective of lung and infection segmentation models is to minimize segmentation loss by learning differentiation between foreground and background class pixels. Corresponding to lung segmentation task classes belongs to {lung, background} and for COVID-19 disease segmentation these are {disease lesion, background} [34]. Moreover, segmentation loss for lung and infection segmentation models in the forward pass is computed as the mean of dice and IoU loss using Eq. (1) . The dice and intersection over union loss are computed using Eq. (2) and Eq. (3) [45], respectively. Here y_i and $p(y_i)$ represents ground truth and predicted mask output by the final layer. In addition, the loss function is updated in subsequent epochs by the backpropagation and gradient descent process.

$$Segmentation\ Loss\ (S) = \frac{1}{2} DCL\,(y, p(y)) + \frac{1}{2} IoU_{Loss}\ (y, p(y)) \quad (1)$$

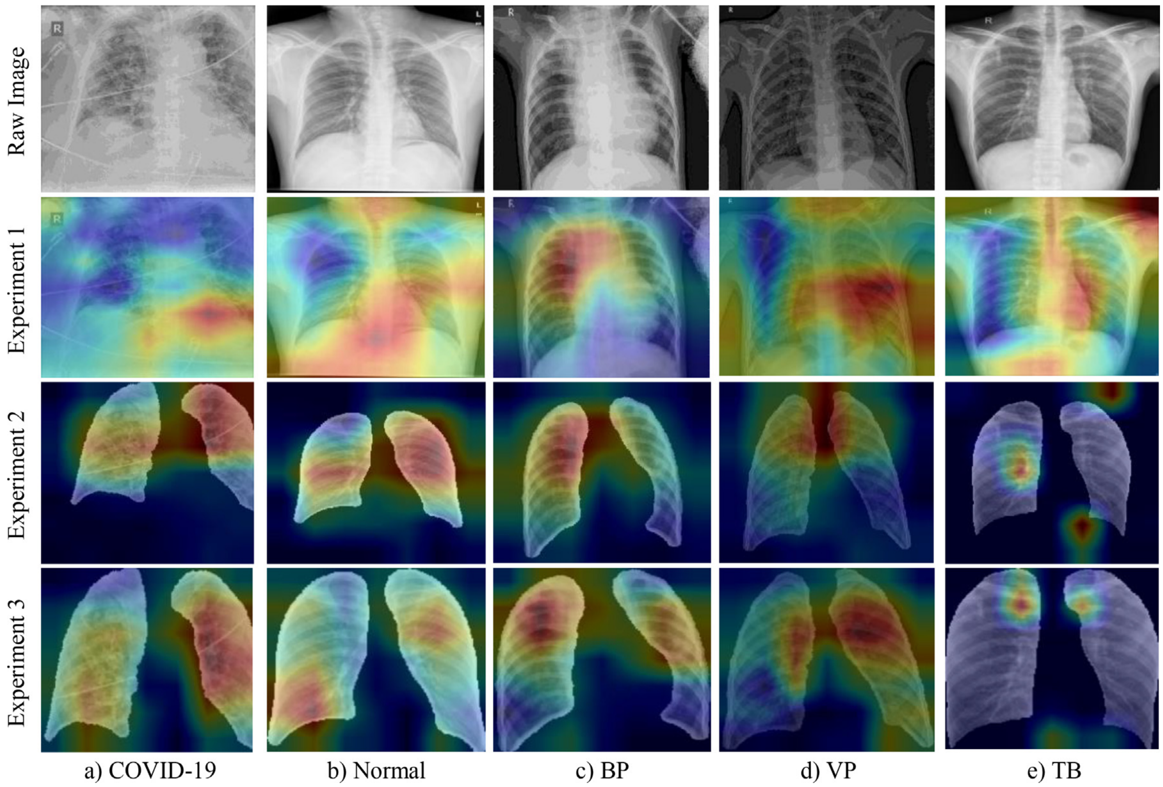


Fig. 5. Visualization of Grad-CAM maps for each class, where the first row corresponds to input images. The second, third and fourth rows show Grad-CAM activation visualization produced by the best classification model in each of the experiment indicating proposed segmentation-based cropping approach has better interpretable results.

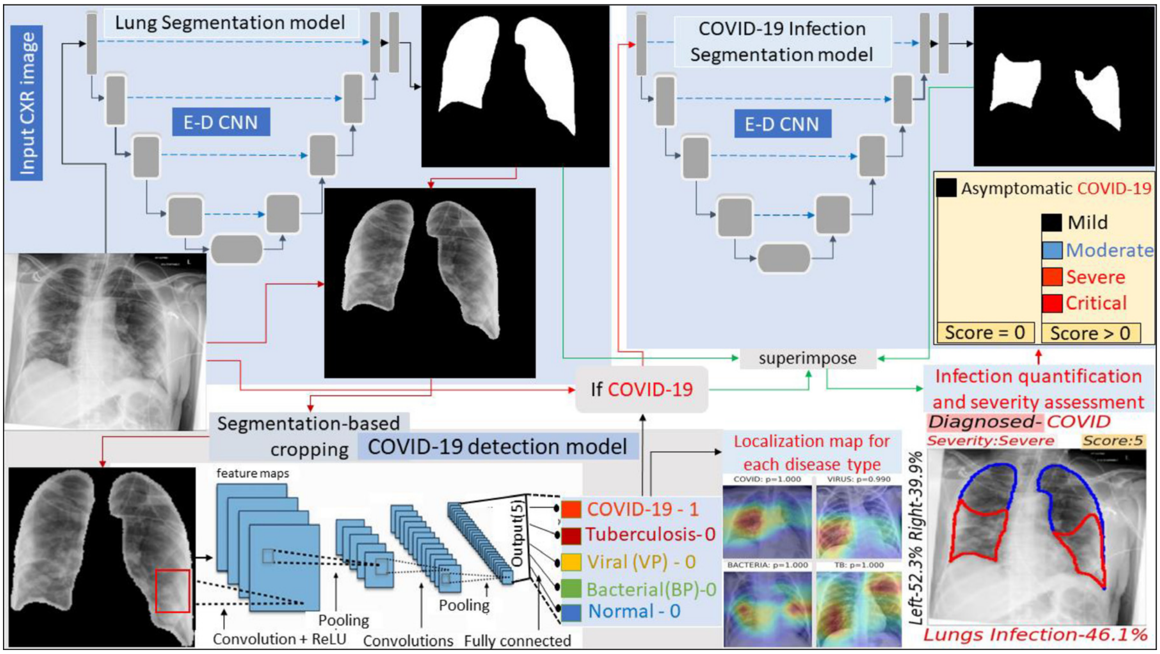


Fig. 6. The framework of Covid-MANet methodology for classification and infection segmentation of COVID-19.

$$DCL(y, p(y)) = 1 - \frac{2\sum_i^N y_i.p(y_i)}{\sum_i^N |y_i|^2 + \sum_i^N |p(y_i)|^2} \tag{2}$$

$$IoU_{Loss}\ (y, p(y)) = 1 - \frac{\sum_i^N y_i.p(y_i)}{\sum_i^N (y_i + p(y_i) - y_i * p(y_i))} \tag{3}$$

2.6. Classification loss function

The classification dataset has a large class imbalance that may affect the results of the minority class. To eliminate this problem classification models are trained using a weighted cross-entropy loss function instead of simple cross-entropy loss. The weighted loss function alters weights assigned to each training example while computation of loss. If classes are balanced, every sample

Table 1
Summary of segmentation dataset resources.

Dataset	Classes	#Images
[3] JSRT [47] / SCR [48]	Normal/Nodule	247
[2] Montgomery [49]	Normal/TB	138
[2] Shenzhen [49]	Normal/TB	566
[1] Cohen et al. [50]	COVID-19	202
[4] RANZCR clip [51]	–	6150
Total	–	*7303*
[5] QaTa-COV19 [34]	COVID-19 Infection segmentation	2951

Table 2
Summary of classification dataset resources.

Dataset used	Classes	# Of images	Total
[1] Cohen et al. [50]	COVID-19	476	476
[6] Kermany et al. [52]	BP	2780	2780
	VP	1493	1493
	Normal	1583	1986
[2] Montgomery [49]	Normal	80	
[2] Shenzhen [49]	Normal	323	
[2] Montgomery [49]	TB	58	394
[2] Shenzhen [49]	TB	336	
Total	–	–	*7129*

contributes equally to loss function but based on importance minority class samples are assigned more weights that show significant effect in training. In our case, we consider an equal contribution to all the classes. These are balanced by assigning new weights to all the classes such that positive and negative examples in each class contribute equally to the loss function computed using Eq. (4) [46]**.** Finally, these weights are fused to the classification loss that reduce false positives of the minority class. The multi-label classification loss function used for the training of the classification model is computed using Eq. (5).

$$w^k_{pos} = \frac{total\ negative\ samples}{N} \text{ and } w^k_{neg} = \frac{total\ positive\ samples}{N} \tag{4}$$

$$\begin{aligned} L = & -\frac{1}{N}\sum_{k=1}^{K}\sum_{n=1}^{N}\Big[w^k_{pos} * y^k_n * \log\left(h_\theta(x_m, k)\right) \\ & + w^k_{neg} * (1 - y^k_n) * \log\left(1 - h_\theta(x_m, k)\right)\Big] \end{aligned} \tag{5}$$

3. Experimental details

In this section, we summarize detailed specifications of datasets used for lung segmentation, infection segmentation and classification of COVID-19 from other lung diseases. In addition, implementation details concerning model development, environment setup, and evaluation measures used for segmentation and diagnosis have been discussed in detail.

3.1. Dataset description and setup

This study aims to develop an automated deep learning model that assists radiologists in the early diagnosis and infection region quantification of COVID-19. To accomplish this study, we compiled well-known publicly available datasets following proper preprocessing guidelines established and used in almost all the research studies for the segmentation and classification of COVID-19. The classification dataset is compiled by merging covid-chestxray-dataset, chest-xray-pneumonia, and Tuberculosis ChestXrayImage DataSets for distinguishing COVID-19 from normal and abnormal lungs having Tuberculosis, viral/bacterial pneumonia. For infection segmentation, QaTa-COV19 dataset is used whereas detailed characteristics and references of these publicly available datasets are summarized as follows:

3.1.1. Lung and infection segmentation dataset

Table 1 specifies detailed resources of lung and infection segmentation datasets. The supervision dataset for lung segmentation contains a corpus of 7303 image masks pairs, compiled by merging five datasets corresponding to three classes only. The compiled dataset includes a corpus of 247 PA view images from the JSRT [47] dataset and their corresponding lung masks in the SCR [48] database. Montgomery Country and Shenzhen No.3 People's Hospital [49] datasets having a corpus of 138 PA view, 566 chest radiographs with their corresponding lung masks of TB, and normal classes are used. These are created by the U.S. National Library of Medicine (USNLM), Maryland, USA in collaboration with the Department of Health and Human Services and Shenzhen No.3 People's Hospital [49] at Guangdong Medical College in Shenzhen, China. For COVID-19, Chest X-ray-dataset created by Cohen et al. [50] is utilized having a corpus of 202 lung segmented image-mask pairs available publicly. Also, we consider 6150 pseudo lung masks of the RANZCR dataset [51] with almost similar UID compared to BP and VP cases. Supervision with this dataset enables better lung segmentation of unseen VP and BP in addition to other classes.

In a real scenario, it is desirable to find the COVID-19 infection region and severity of the disease. To supervise the model for infection region segmentation, we utilized QaTa-COV19-v1[34] dataset compiled by Tampere University and Qatar University. The dataset comprises 2951 CXR images and corresponding ground truth masks for COVID-19 infection out of 4603 total COVID samples.

3.1.2. Classification dataset

Table 2 specifies detailed resources of classification datasets organized into five classes; COVID-19, normal, VP, BP, and TB. Specifically, a corpus of 476 COVID-19 CXR images of PA, AP, and AP supine view are collected from a similar dataset created by Cohen et al. [50]. For Comparison of COVID-19 with other lung diseases such as TB, VP, BP, and abnormal samples are collected from references [49,52]. Normal and pneumonia samples having a corpus of 1583 and 4273 images, taken from chest-xray-pneumonia dataset [52]. Because of the similarity between viral pneumonia and COVID-19, more specific categorization of pneumonia into 1483 viral and 2780 bacterial pneumonia are considered. In addition, more normal samples are taken from the Montgomery and Shenzhen dataset with a total sum of 1986 images. The tuberculosis samples are taken from reference with a corpus of 394 samples.

3.2. Implementation setup

The implementation details for segmentation and classification networks summarized in this section. The lung segmentation task is accomplished using UNet and its two variants. These variants use ResNet and DenseNet blocks in place of two convolution layers of UNet. Each model is trained for a minimum of 70 epochs with a batch size of 16. The loss function is minimized using adam optimizer [53] with an initial learning rate of 0.00001 and an early stopping strategy. The learning rate is reduced by a factor of 10 if the loss is not reduced for 5 consecutive epochs. Models trained with input size 320 × 320 × 1 and the dataset split for training and testing is 90% and 10%, respectively.

The classification model classifies input into one of five classes based on maximum predicted probability by a softmax activation function. The classification models trained for a maximum of 50 backpropagation epochs with a mini-batch size of 32. Again, adam optimizer with an initial learning rate of 0.00001 is used to minimize a weighted cross-entropy loss function. This supports an early stopping strategy with a patience value of 15 and the learning rate is reduced by a factor of 10 if the loss is not reduced for 7 consecutive epochs. The classification models are trained with input size 224 × 224 × 3 except for InceptionV3, which is trained with 299 × 299 × 3. In addition, the distribution statistics of the dataset into the training, testing, and validation set follow the guidelines to ensure that no patient overlap exists in the train, test, and validation set. The distribution of each class into training, testing, and validation set is specified in Table 3. However, normal, BP, and VP classes were already labeled into training and testing set but the validation set contains only 8 samples. To ensure the validity of the model, we increase the validation set by splitting training samples into a validation set.

The infection segmentation network is trained on QaTa-Cov19 dataset to localize the COVID-19 infection region. Models trained with an input size of 224 × 224 × 3 for 50 backpropagations epochs and a mini-batch size of 4. The dataset split into the proportion of 80–20% for training and testing, whereas 10% of the training dataset is used for validation. Similar segmentation loss and adam optimizer is used for training with an initial learning rate of 0.0001. All these tasks are implemented in python using NVIDIA Tesla P100 GPU. Other libraries used for implementation include Keras, TensorFlow, and scikit-image.

3.3. Statistical measures

We statistically analyzed the segmentation and classification models based on metrics computed using a confusion matrix. Lung and infection segmentation models follow pixel-level classification where negative class corresponds to the background and positive class related to the lung or infection pixel. Similarly, classification models based on sample statistics where COVID-19 classified samples are considered as positive and non-COVID as a negative class [56]. Specifically, lung and infection segmentation models are evaluated using dice similarity coefficient, Intersection over union (IoU), Sensitivity, and Precision computed [14] as per equations; (6–10). Classification models are evaluated by Accuracy, Recall, Precision, and F1-score computed as per equations; (6,9,10, and 11). Moreover, COVID-19 sensitivity and COVID-19 precision are also taken into consideration since the correct classification of COVID-19 infected samples are more desirable. These are computed in the same way as recall and precision computed for a particular class. In short, accuracy refers to the ratio of correctly classified samples/pixels to the total corpus of samples/pixels. IoU and DSC quantitatively measure the overlap between predicted lung/infection segmentation masks and ground-truth masks. However, the only difference is that the latter one (DSC) considers the double weight of true positive pixels.

Table 3
Distribution of images for classification in five infection types.

Set	COVID-19	Normal	VP	BP	TB
Training	354	1586	1220	2348	311
Testing	80	233	148	242	45
Validation	42	170	125	190	38
Total	476	1986	1493	2780	394
Training: 5816, Validation: 565, Testing: 748, Total: 7129					

Table 4
Lung segmentation networks performance.

	Model	Dice.	IOU	Recall	PPV
Without RANZCR	UNet	94.51	89.62	92.82	96.30
	ResUNet	95.20	90.89	94.81	95.62
	DenseUNet	95.49	91.40	94.26	96.78
With RANZCR	UNet	96.04	*92.40*	95.00	97.14
	ResUNet	96.29	*92.86*	94.56	98.15
	DenseUNet	*96.70*	*93.64*	*95.62*	*97.82*

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

$$Intersection\ over\ Union\ (IoU) = \frac{TP}{TP + FP + FN} \tag{7}$$

$$Dice\ Similarity\ Coef\,ficient\ (DSC) = \frac{2TP}{2TP + FP + FN} \tag{8}$$

The sensitivity refers to the rate of correctly classified positive class samples to the total number of positive class samples. The precision (PPV) refers to the proportion of correctly classified positive class samples to all the samples classified as positive. F1-score computes the harmonic mean of recall and precision. The segmentation model having a higher dice similarity coefficient and IoU score is considered as the best model. Similarly, for classification model higher accuracy, sensitivity, and COVID-19 sensitivity is desirable.

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

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

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

4. Results and analysis

In this section, we discuss results attained in all three tasks of the Covid-MANet network i.e., lung segmentation, disease recognition, and disease segmentation. The model development process performed an extensive number of experiments for improving the interpretation and recognition based on lung localization regions. The sensitivity analysis of each class has been discussed in search of a suitable model for each class separately. In addition, infection severity assessment and model robustness are cross-validated on an unseen dataset. Finally, the Grad-CAM technique is used to analyze the interpretation and explainability of models.

4.1. Results for segmentation

The Covid-MANet network accomplishes the primary aim of disease classification by analyzing segmented lungs. The performance achieved by the lung localization network with and without the involvement of RANZCR dataset on the test subset is shown in Table 4. Recall that no public dataset has annotated masks of viral and bacterial pneumonia classes. The majority of previous studies analyzed segmentation for three classes by Shenzhen CXR and Montgomery dataset masks. The addition of RANZCR dataset enables better lung localization of classification data not only for viral and bacterial pneumonia classes but upright for all the classes. Among all models, it has been seen that dense connectivity of encoder blocks outperformed other blocks because of decisions based

[50] [1] https://github.com/ieee8023/covid-chestxray-dataset
[49] [2] https://lhncbc.nlm.nih.gov/LHC-publications/pubs/TuberculosisChestXrayImageDataSets.html
[47] [3] http://db.jsrt.or.jp/eng.php
[51] [4] https://www.kaggle.com/raddar/ranzcr-clip-lung-contours
[34] [5] https://www.kaggle.com/aysendegerli/qatacov19-dataset
[52] [6] https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia

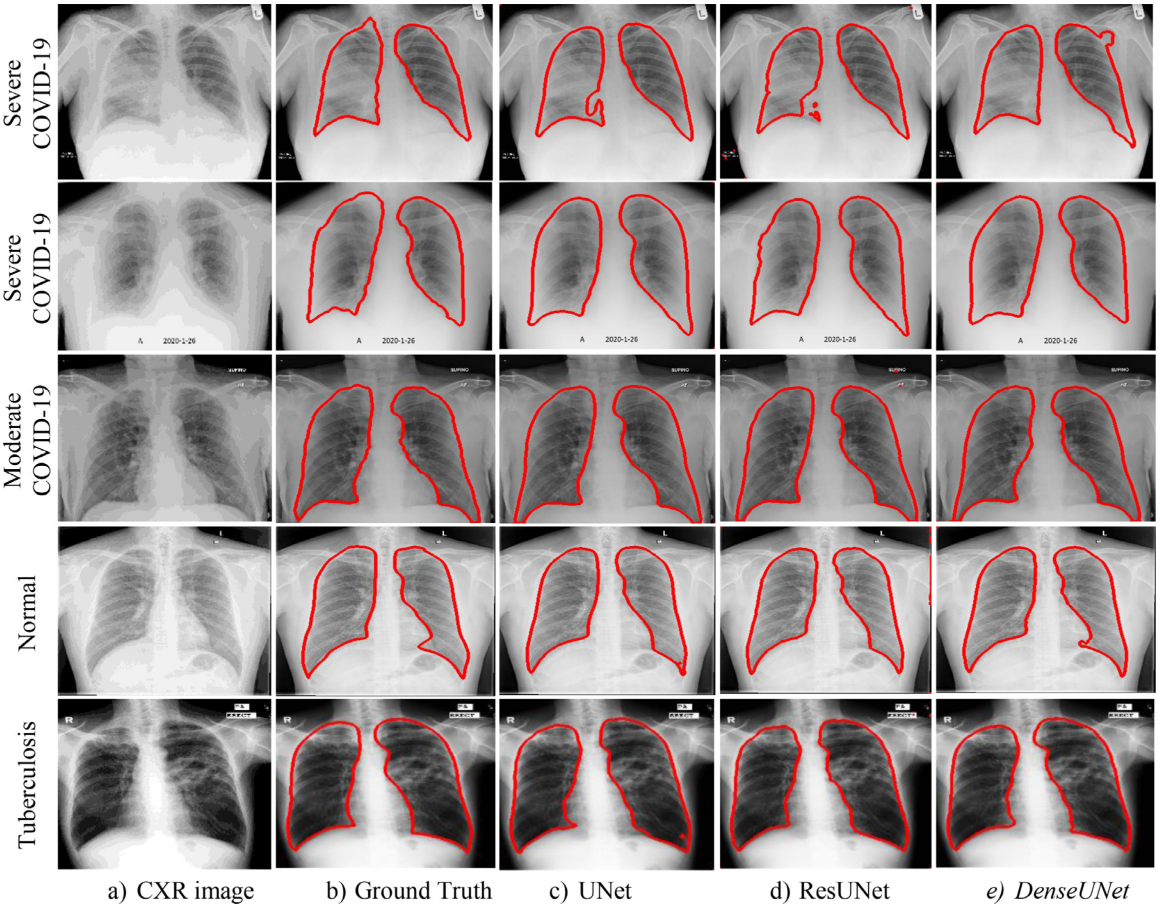


Fig. 7. Comparison of ground truth and predicted lung masks by the lung segmentation models.

on the collective knowledge from previous layers. DenseUNet holds a leading position in the lung localization task with a dice score of 96.70 and IoU of 93.64%. The quantitative results indicate improvement in the overlapping of ground truth and predicted lung mask by 2.20% with the pseudo-generated RANZCR dataset. The output of lung localization models is compared in Fig. 7 with ground truth segmentation masks. Results indicate model reliably localize lungs region of different severity levels of COVID-19 including mild, moderate, and critical in addition to non-COVID-19 cases. However, the challenge with previous approaches lacks segmentation of critical and severe cases accurately because Montgomery and Shenzhen dataset comprises TB and the normal cases of medium and high quality only. Our compiled benchmark dataset benefits to overcome the challenge of lung localization for viral, bacterial, and low-quality COVID-19 pneumonia cases. Post-processed lung segmentation mask is masked with the corresponding image to generate a lung localization map ready for disease recognition by the classification model of the Covid-MANet network.

4.2. Results of classification

The performance comparison of state-of-the-art classification models with the proposed model in each experiment is shown in Table 5. In the exploration of the most coherent approach, different model variants are evaluated on an independent test set, where the model selection process relies on accuracy, F1-score, and Covid-19 sensitivity. Recall that the proposed multi-scale hybrid classification model MA-DenseNet201 is compared with eight other deep learning models. Image enhancement and attention mechanism are evaluated in combination with comparative deep learning models. The light attention module CBAM adds 0.9 M additional trainable parameters to these networks.

The performance of comparative models in experiment 1 along with the proposed model on whole slide images indicates VGG19 has higher accuracy and F1 score. VGG19 having an accuracy of 93.85% is followed by DenseNet121 and MobileNet with 92.72% and 92.51% scores. Despite having higher accuracy values, VGG19 has less interpretation and COVID-19 sensitivity, whereas MobileNet, DenseNet121, and the proposed model have higher sensitivity for COVID-19. In addition, the layers of the VGG19 model offer less confident analysis with low accuracy for COVID-19 cases at higher threshold values compared to the proposed MobileNet model as shown in Fig. 10. The MA-DenseNet201 model is better for COVID-19 in this experiment but confusion still exists between viral and bacterial pneumonia cases.

Considering experiment 2, the performance and interpretation of the proposed model improved by classification based on lung localization regions. The performance measures of the DenseNet, MobileNet, and ResNet50 model improved after localization of lungs considering more reliable features for diagnosis whereas the VGG model decreases. Moreover, the proposed MA-DenseNet201 model outperformed all other models in terms of accuracy, F1 score, and COVID sensitivity with a score of 93.45%, 93.59%, and 97.50%, respectively. MA-DenseNet201 model is followed by the MobileNet model with accuracy and F1 score of 92.91% and 92.57%. Class-wise sensitivity and accuracy comparison at different threshold values insights that the proposed model is reliable even at higher threshold values. Compared to the previous experiment, the proposed model with a lung localization network enables reliable classification much focused on lung segmented regions and re-

Table 5
Comparison of proposed model performance with state-of-the-art variants in considered experiments.

Experiment no.	Model	Accuracy (%)	Recall (%)	Precision (%)	F1-score (%)	COVID-19 Sen. (%)	COVID-19 (PPV) (%)
Experiment 1	ResNet50	76.34	76.34	78.41	75.93	77.50	86.11
	VGG16	92.78	92.78	93.03	92.76	96.25	97.45
	VGG19	93.85	93.85	93.93	93.88	97.50	98.73
	NASNetMobile	90.91	90.91	91.05	90.76	97.50	97.50
	MobileNet	*92.51*	*92.51*	*92.74*	*92.38*	*98.75*	*98.75*
	InceptionV3	90.64	90.64	90.82	90.60	96.25	96.25
	DenseNet121	93.72	93.72	93.92	93.67	98.75	100
	DenseNet201	90.78	90.78	91.05	90.69	98.75	98.75
	MA-DenseNet201	91.84	91.84	92.16	91.69	98.75	98.75
Experiment 2	ResNet50 (S)	87.83	87.83	88.18	87.68	80.00	94.12
	ResNet50 (CBAM)	88.24	88.24	88.86	88.02	83.75	94.37
	ResNet50 (GC)	89.04	89.04	89.28	88.91	87.50	95.89
	VGG16	91.44	91.44	92.47	91.65	96.25	93.90
	VGG19 (S)	92.11	92.11	92.34	92.19	92.50	97.37
	VGG19 (CBAM)	91.84	91.84	92.36	91.99	95.00	96.20
	VGG19 (GC)	92.51	92.51	92.79	92.60	95.00	96.20
	NASNetMobile	89.57	89.57	90.49	89.84	92.50	100
	MobileNet	92.91	92.91	93.61	93.07	93.75	97.40
	MobileNet (CBAM)	92.25	92.25	93.02	92.42	95.00	97.44
	MobileNet (GC)	92.51	92.51	92.72	92.57	93.75	93.75
	InceptionV3	88.77	88.77	89.95	89.05	96.25	96.25
	InceptionV3 (CBAM)	89.71	89.71	90.79	89.96	95.00	97.44
	InceptionV3 (GC)	89.57	89.57	89.97	89.69	93.75	97.40
	DenseNet121	91.84	91.84	92.65	91.99	95	98.70
	DenseNet20	90.91	90.91	91.86	91.13	96.25	95.06
	DenseNet201 (CBAM)	91.71	91.71	92.08	91.83	95.00	97.44
	DenseNet201 (GC)	91.71	91.71	92.51	91.91	96.25	97.47
	MA-DenseNet201	*93.45*	*93.45*	*94.07*	*93.59*	*97.50*	*98.70*
Experiment 3	ResNet50	86.23	86.23	86.69	89.95	77.50	95.38
	ResNet50 (GC)	87.83	87.83	88.07	87.77	83.75	93.06
	ResNet50 (TVF + GC)	88.90	88.90	89.22	88.96	82.50	95.65
	ResNet50 (GC + CBAM)	88.37	88.37	88.58	88.04	86.25	95.83
	VGG16	89.17	89.17	90.35	89.38	96.25	97.47
	VGG19	90.11	90.11	90.79	90.28	92.50	93.67
	VGG19 (GC)	90.24	90.24	90.90	90.41	96.25	93.90
	VGG19 (TVF + GC)	90.37	90.37	91.16	90.57	92.50	96.10
	VGG19 (GC + CBAM)	89.84	89.84	90.43	90.00	93.75	96.15
	NASNetMobile	87.83	87.83	88.66	88.08	93.75	93.16
	MobileNet	90.78	90.78	91.68	90.99	95.00	96.20
	MobileNet (GC)	90.78	90.78	91.41	90.45	95.00	96.20
	MobileNet (TVF + GC)	90.64	90.64	91.10	90.78	93.75	96.15
	MobileNet (GC + CBAM)	90.78	90.78	91.18	90.90	95.00	95.00
	InceptionV3	90.78	90.78	91.50	90.95	95	97.44
	InceptionV3 (GC)	90.78	90.78	91.12	90.89	95	97.44
	InceptionV3 (TVF + GC)	90.91	90.91	91.07	90.97	93.75	97.40
	InceptionV3 (GC + CBAM)	90.51	90.51	90.80	90.61	93.75	96.15
	DenseNet121	91.84	91.84	92.24	91.97	96.25	96.25
	DenseNet201	92.11	92.11	92.31	92.18	93.75	97.40
	DenseNet201 (GC)	91.98	91.98	92.15	92.04	95.00	96.20
	DenseNet201 (TVF + GC)	91.84	91.84	92.08	91.93	95.00	97.44
	DenseNet201(GC+ CBAM)	92.38	92.38	92.68	92.47	95.00	96.20
	MA-DenseNet201	*92.78*	*92.78*	*93.67*	*92.95*	*97.50*	*97.50*

*Italics text represents best model; GC stands for Gamma correction and TVF stands for Total variation filter.

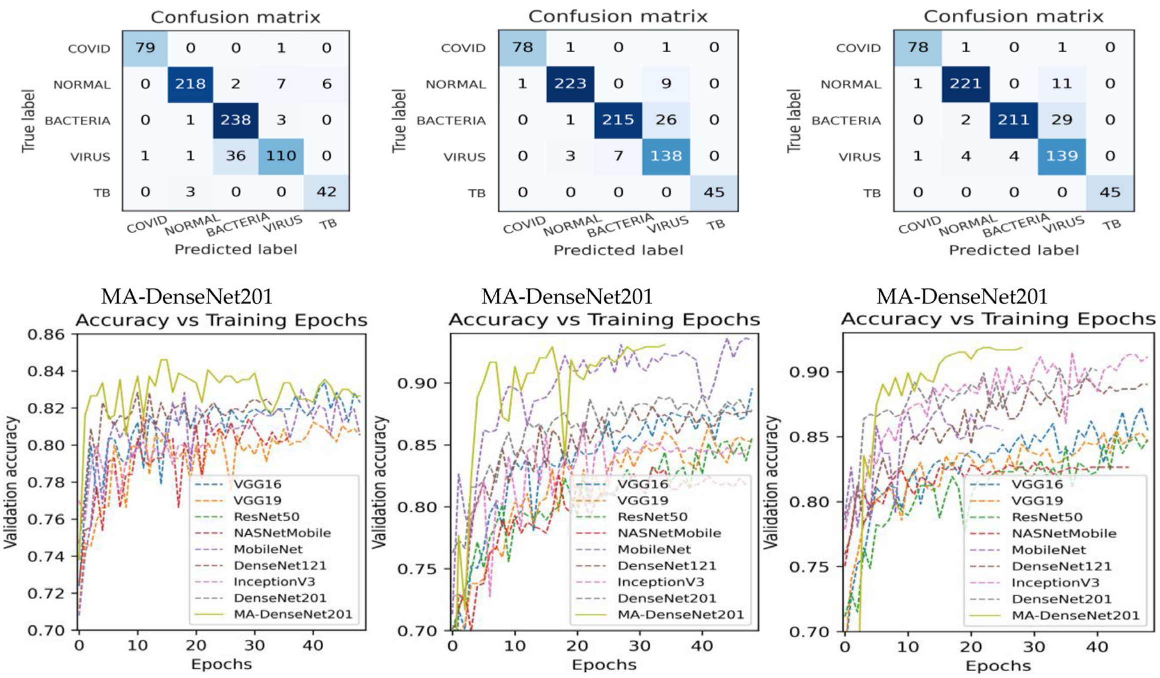
duces the confusion between viral and bacterial pneumonia cases by improving the sensitivity of viral pneumonia.

To improve interpretation, we proposed experimental analysis based on segmentation-based cropping, discarding regions outside segmented lungs. Similar to the previous experiment, the MA-DenseNet201 outperformed all other model variants with accuracy, F1 score, and COVID-19 sensitivity of 92.78%, 92.95%, and 97.50%, respectively. The proposed model is followed by the DenseNet201 and DenseNet121 models with accuracy of 92.38 and 91.84%. Moreover, the sensitivity and accuracy of the Covid-MANet approach at various threshold values are higher compared to other state-of-the-art models. In addition, interpretation of the proposed approach is maximum among all the experiments indicating the model's robustness in recognition of mild, moderate, and severe cases.

Fig. 8 shows the confusion matrix of the MA-DenseNet201 model. The training and validation growth of models corresponding to these experiments are also shown in the second row of this figure. The schematic representation of ROC curve indicates AUC of each class for the top four performing networks as shown in Fig. 9. These results reveal that the proposed model attains better AUC for all the classes after lung localization. Moreover, the validation accuracy in experiment 1 ranges between 80 and 85%, whereas it ranges from 90 to 94% in the second and third experiments.

4.2.1. Sensitivity and confidence-score analysis

In disease diagnosis, it is desirable to have higher sensitivity since it measures the proportion of correctly classified positive samples to ensure maximum cases of COVID-19 are predicted correctly. Fig. 10 shows the sensitivity of models corresponding to each class in three classification experiments. Sensitivity analysis helps to build an ensemble model based on class-varied accuracies of the best suitable model for each class. From the results of the classification scenario, we conclude that the proposed model has higher sensitivity for COVID-19, Normal, viral pneumo-



	COVID	NORMAL	BACTERIA	VIRUS	TB
COVID	79	0	0	1	0
NORMAL	0	218	2	7	6
BACTERIA	0	1	238	3	0
VIRUS	1	1	36	110	0
TB	0	3	0	0	42

	COVID	NORMAL	BACTERIA	VIRUS	TB
COVID	78	1	0	1	0
NORMAL	1	223	0	9	0
BACTERIA	0	1	215	26	0
VIRUS	0	3	7	138	0
TB	0	0	0	0	45

	COVID	NORMAL	BACTERIA	VIRUS	TB
COVID	78	1	0	1	0
NORMAL	1	221	0	11	0
BACTERIA	0	2	211	29	0
VIRUS	1	4	4	139	0
TB	0	0	0	0	45

Fig. 8. The first row shows confusion matrix of proposed model and second row representing learning curves in each best model.

Table 6
Accuracy comparison of models in each experiment at different confidence threshold values.

%age Accuracy	Threshold	ResNet50	VGG16	VGG19	NASNet Mobile	Inception V3	DenseNet121	DenseNet201	MobileNet	*Proposed*
Without Segmentation	0.5	89.17	97.19	97.54	96.47	96.44	97.45	96.25	*97.03*	96.71
	0.7	82.35	96.09	96.25	96.12	96.01	95.98	95.82	*96.73*	96.57
	0.85	80.45	92.94	91.71	94.83	93.95	94.30	95.50	*96.39*	96.25
	0.95	80	87.62	85.80	91.22	89.91	91.04	94.09	*95.50*	95.05
With Segmentation	0.5	95.72	96.55	97.03	95.88	95.85	96.73	96.68	97.16	*97.37*
	0.7	95.48	95.58	95.85	95.05	95.77	96.71	96.65	96.76	*97.35*
	0.85	93.26	93.07	93.04	93.36	94.62	95.40	95.50	96.39	*97.05*
	0.95	89.30	89.41	88.47	90.58	92.45	93.12	94.01	95.42	*96.76*
Segmentation based cropping	0.5	95.26	95.74	96.36	94.91	96.36	96.71	96.84	96.33	*97.11*
	0.7	94.49	94.86	96.01	93.82	96.01	96.17	96.76	96.49	*97.16*
	0.85	93.23	92.08	95.32	91.06	95.32	94.65	95.88	95.90	*97.08*
	0.95	89.33	87.54	92.88	87.37	92.88	92.08	94.03	94.54	*96.71*

nia, and tuberculosis ranging from 93 to 100%. But ResNet50 is better in recognition for bacterial pneumonia. The maximum sensitivity score of 96.69% is achieved for bacteria class by ResNet50, whereas the proposed model achieved 87% for BP. In addition, ResNet50 has the least sensitivity for other classes, whereas the proposed model attained 97.5% for COVID-19, 94.85% for normal, 93.92% for VP, and 100% for TB classes. The proposed model improves the sensitivity of VP cases after lung segmentation, thus reducing the risk of overlapping with COVID-19.

To further explore the robustness of the proposed model, we computed sensitivity and accuracy at different confidence threshold values of 0.5, 0.7, 0.85, and 0.95. The second row of Fig. 10 and Table 6 show the sensitivity and accuracy of experimental analysis at different confidence threshold intervals. Schematic representation insights that the sensitivity of the proposed model is stable even at higher threshold values as compared with other models. Similarly, from these observations, we conclude that the proposed model outperformed other comparative models after segmentation, but MobileNet better before segmentation.

4.3. Results of covid infection segmentation

The results of the infection segmentation models are shown in Table 7. The evaluation of the infection quantification model is tested in two configurations: serial and parallel scenarios. In the serial scheme, segmented lungs by lung localization model given to classification are directly fed into the infection segmentation model, if the sample is diagnosed as COVID-19. In the parallel scheme, an original CXR image of diagnosed COVID-19 sample is given to the infection segmentation model for quantification and severity assessment of COVID-19. UNet and UNet with DenseNet201 encoder evaluated on both the schemes. Results reveal that the parallel scheme has a better dice score of 81.86% and 85% compared to 80.84% and 83.86% of the serial scheme. Therefore, the parallel scenario is preferred to train and evaluate the remaining experiments in search of a coherent approach for infection segmentation. Similar to lung localization, the infection segmenta-

Table 7
Performance comparison of infection segmentation models.

Model	Acc.	Dice	IOU	Recall	PPV
UNet	94.89	81.86	69.59	81.35	82.75
Attention UNet	94.57	80.19	67.93	80.05	81.72
UNet++	94.34	79.28	66.65	78.72	81.44
R2UNet	95.16	82.16	70.69	82.63	82.82
UNet++ + VGG19	95.36	83.03	71.90	83.65	83.49
UNet + ResNet50	95.77	84.49	74	85.16	84.80
UNet + DenseNet121	*97.01*	*86.15*	*76.94*	*86.29*	*86.92*
UNet + DenseNet201	96.93	85	75.69	86.01	86.94

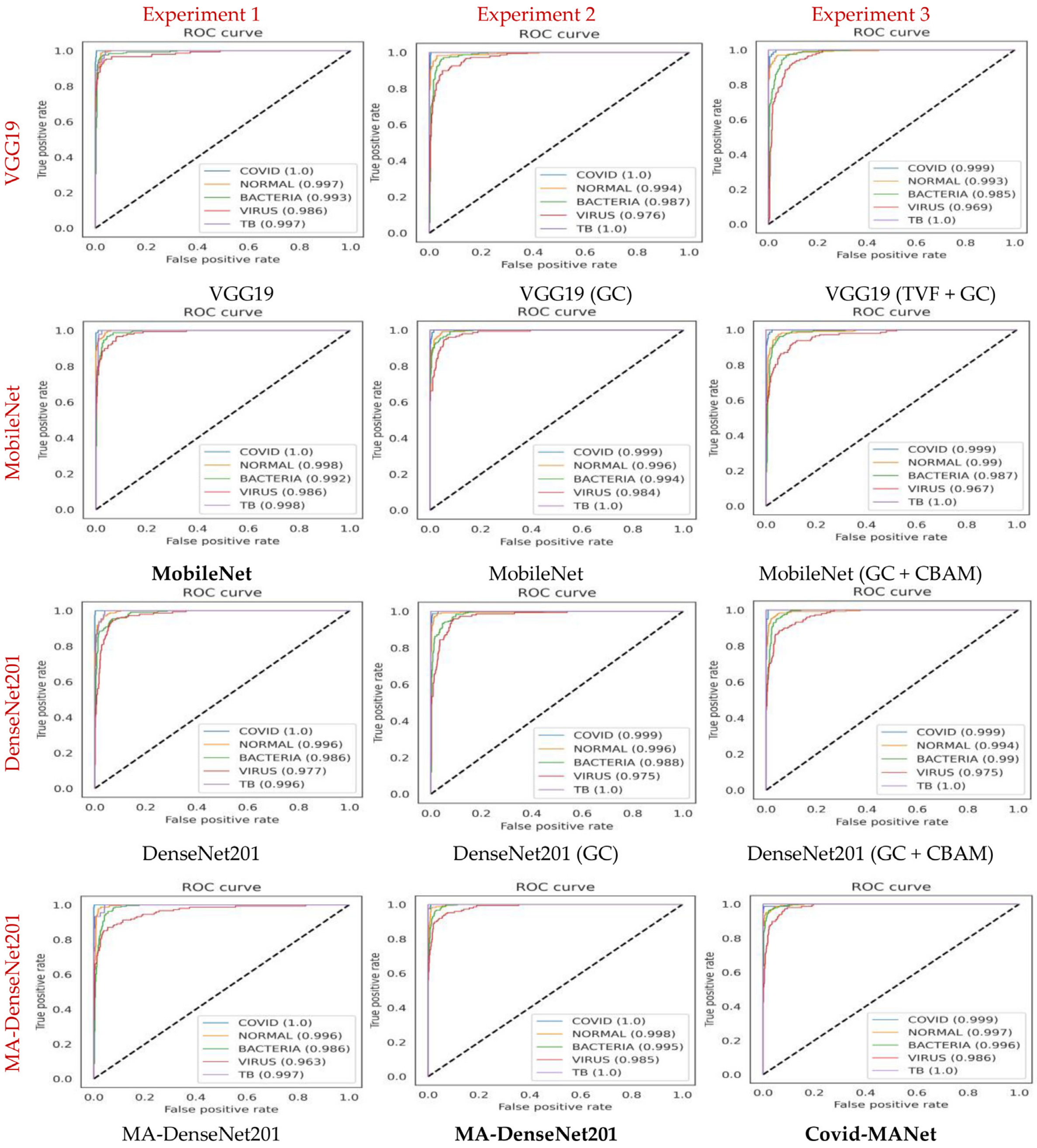


Fig. 9. Visualization of ROC map for top four performing models representing area under curve for each of class.

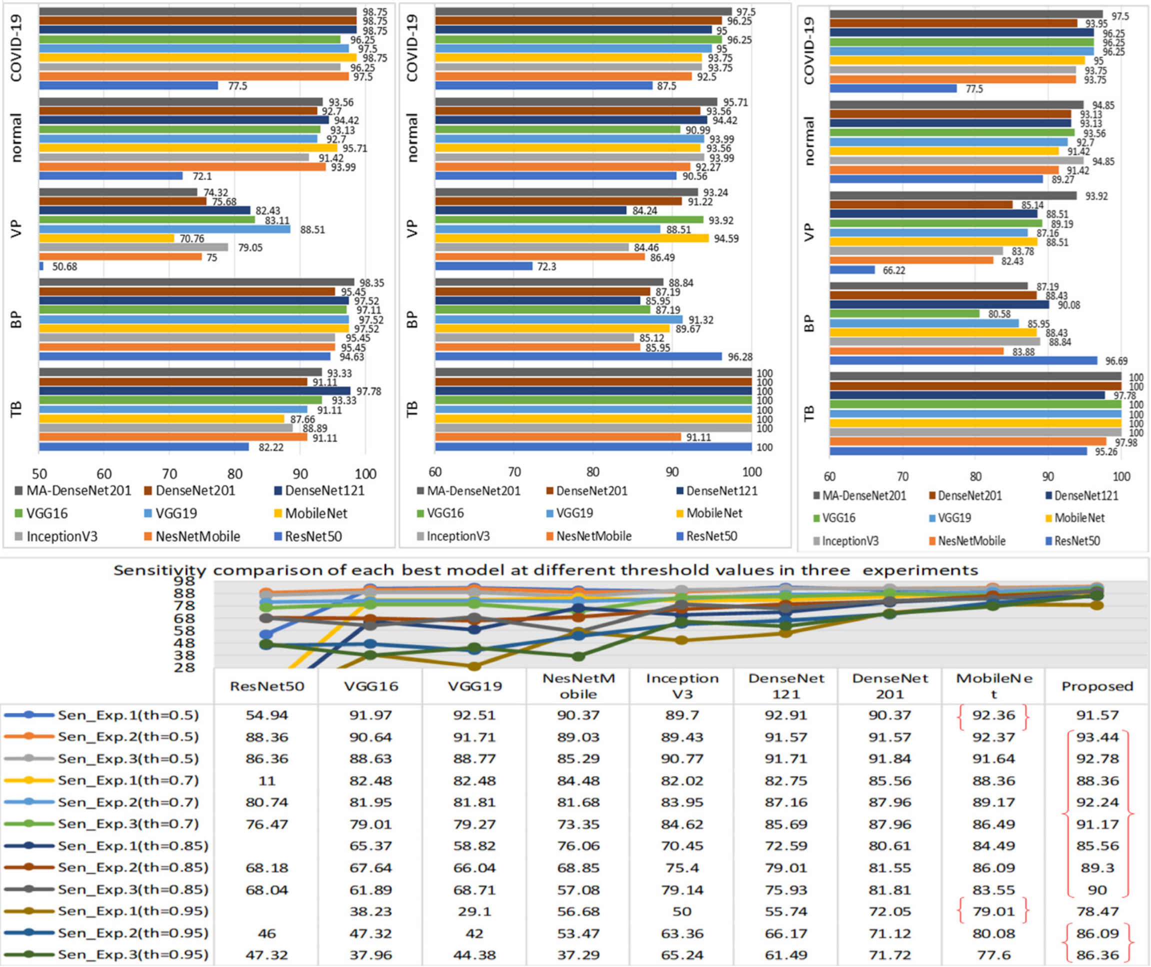
tion network achieves higher dice and IoU score in both scenarios with the DenseNet model as an encoder. The UNet model attains the highest performance when DenseNet121 is used as an encoder with a dice score of 86.15%. Besides this UNet and UNet with DenseNet201 encoder has dice scores of 81% and 85%, respectively. Fig. 11 shows the qualitative comparison of top-performing infection segmentation networks compared with ground truth infection masks. This reveals the robustness of UNet with DenseNet121 and DenseNet201 encoder networks in COVID-19 infection segmentation of mild, severe, moderate, and critical severity levels.

4.3.1. Computational complexity analysis

In this section, we discuss the complexity analysis of lung and infection segmentation networks in terms of trainable parameters and inference time. The number of trainable parameters associated with lung segmentation models, UNet, ResUNet, and DenseUNet is 7.7 M, 10.9 M, and 25 M, respectively. The number of trainable parameters for infection segmentation models AttentionUNet, R2UNet, UNet++, UNet++ with VGG19 has 8.7 M, 26.2 M, 8.6 M, 23.4 M, whereas the inference time for each batch is 14 ms, 18 ms, 34 ms, 15 ms, 30 ms, respectively. Besides, UNet with encoder ResNet50, DenseNet121, and DenseNet201 has trainable parameters of 32.5 M, 12.1 M, 26.3 M, and inference time of 24 ms, 35 ms, 55 ms. The UNet with DenseNet201 backbone has the highest inference time of 14 ms per sample and ResNet50 has the highest trainable parameters of 32.5 M. The overall worst inference time for disease recognition and segmentation is 150 ms whereas the system can process multiple batches in a second.

4.4. Hybrid classification model

In the comparison of experimental analysis, we view that despite having higher accuracy values of experiment 1 and experiment 2, has the least interpretation and generalization. These



	ResNet50	VGG16	VGG19	NesNetMobile	InceptionV3	DenseNet121	DenseNet201	MobileNet	Proposed
Sen_Exp.1(th=0.5)	54.94	91.97	92.51	90.37	89.7	92.91	90.37	92.36	91.57
Sen_Exp.2(th=0.5)	88.36	90.64	91.71	89.03	89.43	91.57	91.57	92.37	93.44
Sen_Exp.3(th=0.5)	86.36	88.63	88.77	85.29	90.77	91.71	91.84	91.64	92.78
Sen_Exp.1(th=0.7)	11	82.48	82.48	84.48	82.02	82.75	85.56	88.36	88.36
Sen_Exp.2(th=0.7)	80.74	81.95	81.81	81.68	83.95	87.16	87.96	89.17	92.24
Sen_Exp.3(th=0.7)	76.47	79.01	79.27	73.35	84.62	85.69	87.96	86.49	91.17
Sen_Exp.1(th=0.85)		65.37	58.82	76.06	70.45	72.59	80.61	84.49	85.56
Sen_Exp.2(th=0.85)	68.18	67.64	66.04	68.85	75.4	79.01	81.55	86.09	89.3
Sen_Exp.3(th=0.85)	68.04	61.89	68.71	57.08	79.14	75.93	81.81	83.55	90
Sen_Exp.1(th=0.95)		38.23	29.1	56.68	50	55.74	72.05	79.01	78.47
Sen_Exp.2(th=0.95)	46	47.32	42	53.47	63.36	66.17	71.12	80.08	86.09
Sen_Exp.3(th=0.95)	47.32	37.96	44.38	37.29	65.24	61.49	71.72	77.6	86.36

Fig. 10. The sensitivity comparison of models where first row shows result of class-wise sensitivity analysis by each of model in 1st·2nd and 3rd experiment and the second row shows sensitivity comparison at different confidence threshold values. The red marked bracket indicates the proposed model is better in all cases after lung localization approach.

models classify images by considering irrelevant features other than the lungs. But the segmentation-based cropping approach achieves comparable performance measures and higher explainability by making attention to relevant lung regions. Based on class-wise sensitivity analysis, we conclude that the Covid-MANet approach is best suitable for COVID-19, normal, TB, and virus classes compared to other models. It is seldom to have a model that is equally good for all five classes. Based on the class-level awareness, we make an ensemble model on a segmentation-based classification approach by combining the prediction ability of the proposed MA-DenseNet201 model with ResNet50 and MobileNet model. The best weights assigned to respective models are w1 = 0.1 (ResNet50), w2 = 0.2 (MobileNet) and w3 = 0.3 (MA-DenseNet201), respectively. Since, Resnet50 improves sensitivity for bacteria class, whereas MA-DenseNet201 model is better for COVID, normal, virus, and TB classes. However, the MobileNet model follows depthwise separable convolutions followed by pointwise convolutions helpful in a different set of feature extraction. The proposed Covid-MANet ensemble network attains accuracy, precision, sensitivity, F1-score, and COVID-19 sensitivity of 95.05%, 95.40%, 95.05%, 95.19%, and 98.75%, respectively. Table 8 shows the result of an ensemble approach improving the respective evaluation metrics for respective classes.

Table 8
Classification performance of Covid-MANet ensemble classification model.

Classes	Recall	Precision	F1-score	Support
COVID	98.75	97.53	98.15	80
Normal	93.99	98.64	96.02	233
VP	96.62	81.71	88.54	148
BP	92.97	99.11	95.94	242
TB	100	100	100	45
Weighted Avg	*95.05*	*95.40*	*95.19*	*748*

From a global perspective, the proposed study aims to develop an end-to-end clinical system having higher interpretation in distinguishing COVID-19 cases from healthy and other lung diseases with similar symptoms. The Covid-MANet multi-task model achieves the desired goal by three tasks; Task 1: lung segmentation, Task 2: COVID-19 detection from healthy/other lung diseases, and Task 3: Infection segmentation. DenseUNet model is best selected for lung localization, whereas for infection segmentation UNet with DenseNet121 encoder outperformed other comparative models. The proposed classification model MA-DenseNet201 achieves better results even at higher confidence intervals with lung segmentation, whereas MobileNet is better without lung segmentation followed by the MA-DenseNet201 model. The interpretation and explainability are maximum with the segmentation-based cropping approach. Also, we observed that shallow CNN models have the least interpretation and explainability compared to deep CNN models. The preliminary class-wise sensitivity analysis shows that MA-DenseNet201 has higher sensitivity for COVID-19, normal, VP, and tuberculosis. But the recognition of bacterial pneumonia is better by the ResNet50 model. Finally, to avoid mis-

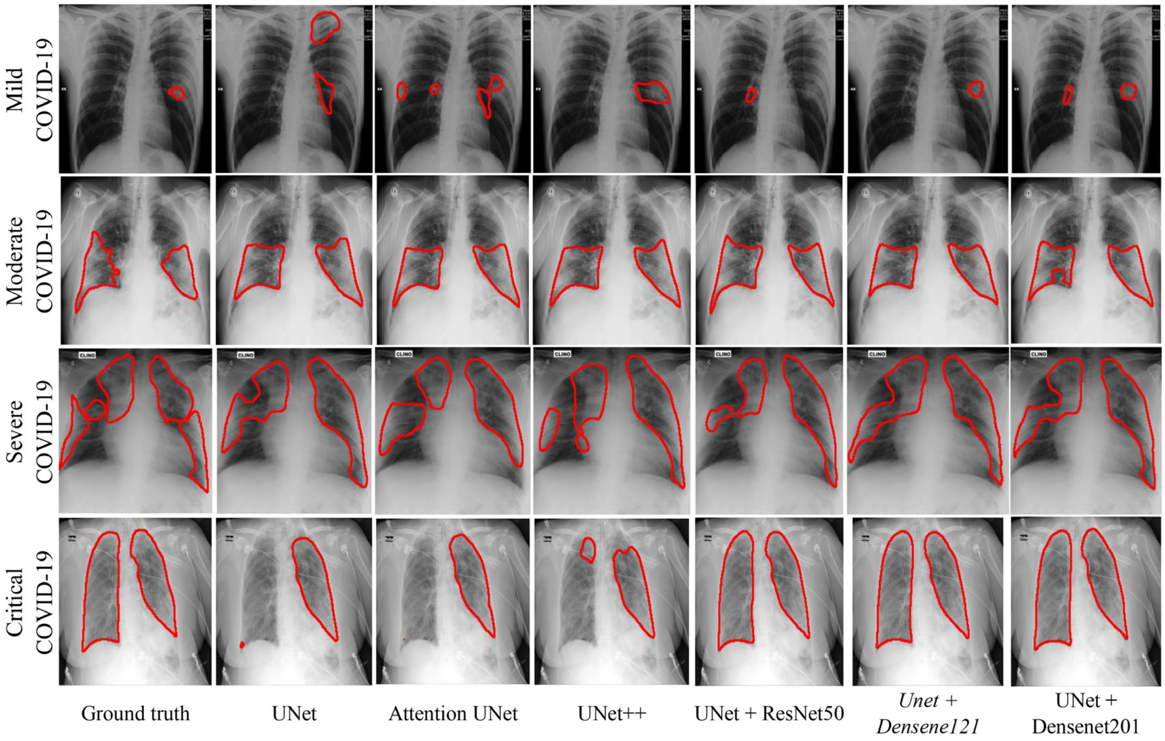


Fig. 11. The qualitative comparison of ground truth masks to the masks predicted by infection segmentation models where column 1 shows ground truth infection mask and column 2–7 shows masks predicted by infection segmentation models, respectively.

classification and overlapping, the ensemble Covid-MANet model is developed based on class varied sensitivity analysis achieving COVID-19 sensitivity of 98.75%. In addition, the infection quantification and severity assessment module assist radiologists with a better treatment plan based on the progression of the disease.

The proposed Covid-MANet model ensures correct diagnostic interest to classify images based on relevant lung regions because of the dual attention mechanism. Since relevant information lies in the lungs, firstly lung localization and secondly segmentation-based cropping impose model attention more on lungs rather than background information. We ensure the robustness of our proposed model on the basics of evaluation measures, interpretation, ROC curve, learning curves, class-wise sensitivity analysis, and attention map visualization. The proposed model outperforms other comparative models in accuracy and sensitivity with maximum interpretation by a segmentation-based cropping approach. The ROC curve reveals similar results showing the maximum AUC achieved by the proposed model in the segmentation-based cropping approach. Class-wise sensitivity analysis shows better performance of the model for COVID-19, Viral, TB, and normal cases but ResNet50 better recognizes bacterial infection in all the experiments. The proposed model is highly stable having better sensitivity and accuracy even at a higher confidence interval of 85% and 95%. The learning curve shows that the proposed model reached an early stopping condition with maximum training in fewer epochs. The reason for maximum interpretation and explainability with the proposed model is because of its multi-scale deep feature extraction ability by the DenseNet201 model given relevant lung areas. In addition, the robustness and generalization of the proposed model are cross-validated in more detail on the unseen QaTa-COV19 dataset in Section 4.5.

4.5. External validation on unseen CXR images

Precisely, to demonstrate the generalization of the proposed system in a real-world scenario, additional results on the unseen QaTa-COV19 dataset has been reported. Initially, lung localization maps are generated on QaTa-COV19 dataset to report results with and without segmentation. The dataset used for the external validation contains 2951 CXR samples of COVID-19 with ground truth infection segmentation masks. Out of 2951 COVID infected samples, COVID-19 sensitivity on unseen plain CXR samples achieved by top-performing models such as MobileNet has 96.54% (2849 correctly classified), VGG19 has 96.98% (2862 correctly classified), DenseNet201 has 95.89% (2830 correctly classified) and MA-DenseNet201 has 91.83% (2710 correctly classified). With lung segmentation MobileNet achieves 96.13% (2837 correctly classified), VGG19 has 89.96% (2655 correctly classified), DenseNet201 has 95.83% (2828 correctly classified) and MA-DenseNet201 has 98.17% (2897 correctly classified) sensitivity. Finally, with lung localization and segmentation-based cropping, MobileNet achieves 97.15% (2867 correctly classified), VGG19 has 96.23% (2840 correctly classified), DenseNet201 has 97.25% (2870 correctly classified) and MA-DenseNet201 has 97.32% (2872 correctly classified) sensitivity. However, the ensemble model gives covid-19 sensitivity of 98.20%. The results on the cross-validation dataset show similar patterns after localization of lungs as seen in preliminary experiments. These interesting observations insight that even with a limited high-quality training dataset, the proposed model gives promising results on much larger test datasets. Moreover, variability of the dataset does not affect the results of the proposed methodology since decision making of the proposed methodology focused on relevant lung regions by lung localization scheme. In contrast, whole slide images always come with some processing artifacts such as marker lines, tissue folds, and uneven sectioning that results in out-of-focus regions. The training of models on these images will produce unexpected results when encountered in test images.

The output of the infection segmentation network on the unseen classification dataset is shown in Fig. 12. The infection diagnosis system predicts the given sample into one pathology class and quantifies the percentage of infection found in the left and

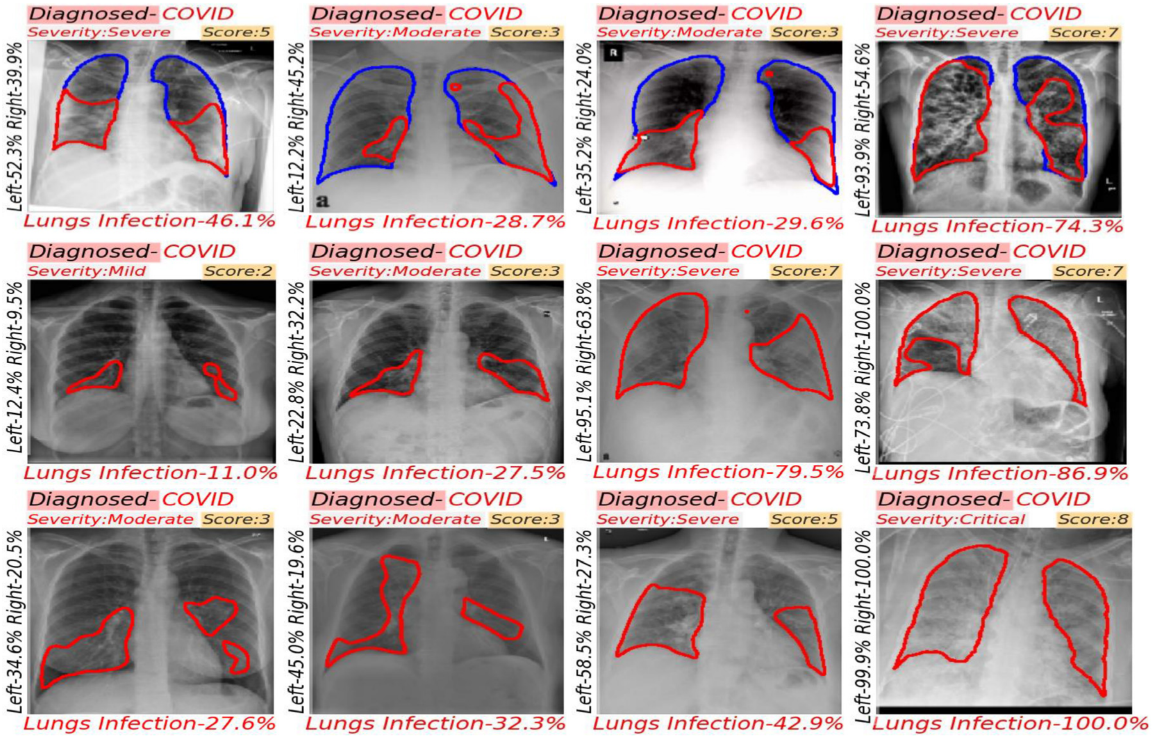


Fig. 12. The infection segmentation and severity grading results by the proposed model on COVID-19 classified samples graded as mild, moderate, severe, and critical, where the first row shows both lung and infection mask contours giving intuition for infection quantification.

right lung separately to assist the doctor in a better treatment plan. Finally, COVID-19 diagnosed sample is assigned a severity level such as; mild/moderate/severe or critical based on the score assigned according to the RALE scoring system. However, our proposed system not only classifies COVID-19 but quantifies the severity of COVID-19 infection, more useful in real-time clinical applications to understand the progression of the disease. Moreover, the proposed system can be used in real clinical practice, since the inference time in the worst case takes less than 150 ms per sample. This means number of samples can be tested in a second.

The proposed model successfully distinguishes COIVD-19 cases from healthy and other lung diseases having similar symptoms as COVID-19. The comparative results of existing state-of-the-art methods on our dataset split under a uniform approach as the proposed model with 85% and 95% confidence threshold values are shown in Table 9. The proposed Covid-MANet system constitutes of four tasks; Task 1 (T1) is semantic lung segmentation, Task 2 (T2) COVID-19 detection with and without lung segmentation and segmentation-based cropping (SBC), Task 3 (T3) is infection quantification and severity assessment, and Task 4 (T4) as the multi-label loss to curtail class imbalance. Each task plays an important role in improving the behavior of the model. MA-DenseNet201 directly applied on the whole slide image (task 2) shows less accuracy and sensitivity. The performance measures improve by lung segmentation i.e., Task 1 (T1) and Task 2 (T2)**.** The accuracy, sensitivity, and COVID-19 sensitivity of the MA-DenseNet201 model are higher than other SOTA models. The segmentation-based cropping (SBC) approach gives comparable accuracy and sensitivity values. But the major advantage of the proposed approach lies in interpretation improvement as explained in the next paragraph. In addition, parallel analysis of COVID-19 is based on the direct infection segmentation module (Task 1 and Task 3) where the sample is classified as COVID-19 if at least one pixel is classified as a disease. This task leads to more false positives and generalization errors where most VP, BP, and TB samples are diagnosed as COVID-19 with mild symptoms. This shows a certain similarity of COVID-19 infection to other lung diseases. Moreover, the proposed Covid-MANet model can better recognize the patterns of each disease type thus reducing false positives. We can say that lung segmentation and segmentation-based cropping play a pivotal role in

Table 9
Accuracy and sensitivity comparison of proposed model with existing SOTA approaches.

Model	Accuracy@Y%		Sensitivity@Y%		COVID-19 Sen.@Y%	
	Y = 95	*Y* = 85	*Y* = 95	*Y* = 85	*Y* = 95	*Y* = 85
VGG19 + FC [27]	92.88	95.32	44.38	68.71	58.8	76.2
ResNet50+FC [20]	89.33	93.23	47.32	68.04	58.8	76.2
InceptionV3+FC [26]	92.88	95.32	65.24	79.14	60	78.7
MobileNet+ FC [37]	94.54	95.90	77.6	83.56	90	93.8
DenseNet121+FC [44]	92.08	94.65	61.49	75.93	81.2	86.2
DenseNet201+FC [26]	94.03	95.88	71.72	81.81	86.2	92.5
MA-DenseNet201+T2	95.05	96.25	78.47	85.56	97.5	97.5
MA-DenseNet201+T1 + T2	*96.76*	97.05	86.09	89.3	96.5	97.5
MA-DenseNet201+T1 + T2 + SBC	96.71	*97.08*	*86.36*	*90*	*95*	*96.5*

improving the interpretation and explainability of models. This approach adds more weights to lung regions thus providing stable results not affected by the variability of the dataset as mentioned in Section 4.5. The infection segmentation model finds the percentage proportion of infection and assigns severity levels showing the progression of the disease to radiologists. Multi-label loss reduces biasness with minority class COVID-19 because of imbalance and improves the sensitivity of minority class.

The Grad-CAM analysis shows that the model focuses more on relevant lung regions for the classification of samples in each class after segmentation-based cropping. After a critical analysis, MobileNet is considered the best model in experiment 1. It achieves better performance measures at higher threshold values as compared to VGG19. Considering experiment 2 with lung localization network, MA-DenseNet201 outperforms other comparative models. In experiment 3, with segmentation-based cropping MA-DenseNet201 is considered the best model based on higher sensitivity, accuracy, and interpretation analysis. The quantitative and qualitative analysis for the COVID-19 class shows MobileNet gives 42% correct interpretation with whole slide images. But the interpretation improved by MA-DenseNet201 with lung segmentation network (Task 1 and Task 2) to 74%, which was 72%

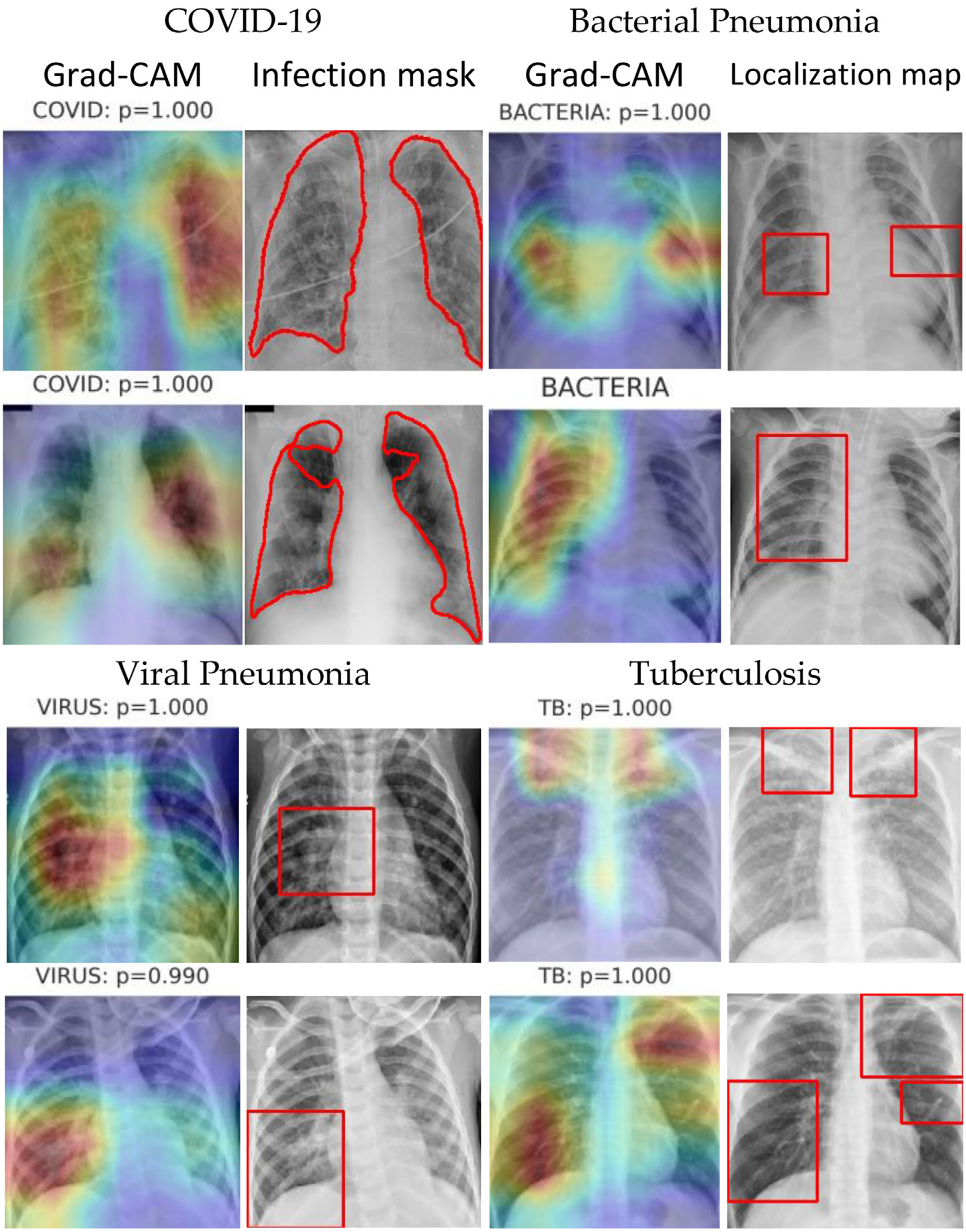


Fig. 13. Interpretation of disease map focused using Grad-CAM for each disease type by Covid-MANet improved after supervision by segmentation-based cropping.

with MobileNet. The maximum interpretation achieved by the MA-DenseNet201 model is 96% after joint analysis by lung localization and segmentation-based cropping in experiment 3. Also, the multi-label classification loss function improves covid-19 sensitivity from 93.75% to 97.15%. Fig. 13 shows the infection map and localization map generated for each disease type by the Grad-CAM visualization tool, explaining the visual decisions focused by the classification model. For COVID-19 disease, the Grad-CAM map correlates with the infection segmentation map. The localization map shows that Covid-MANet focuses more on the relevant lungs region for the classification of each disease type. The proposed model can be useful in real clinical scenarios since the annotation of lungs, disease, and understanding the progression of the disease is a challenging and time-consuming task.

5. Conclusions and future work

In the state of exponential reproduction rate of the COVID-19 pandemic, the only solution to lessen its growth rate is to perform mass screening, early diagnosis, isolation, and correct treatment plan. However, several countries face a shortage of testing kits, laboratories, and medical professionals to deal with the situation within time. At that difficult time, researchers around the world presented CXR as the standard screening tool for findings of COVID-19 in a time-efficient and cost-effective way. In this paper, we present an “explainable solution” for the recognition of COIVD-19 pneumonia from other common lung diseases using CXR radiographs. The proposed Covid-MANet model not only classifies disease but also quantifies infection that intuitions the progression of disease in the lungs. This is the most generic three-stage classification model developed so far works for five classes involving lung segmentation, COVID-19 detection, infection quantification, and severity assessment. The experiments conducted give promising results showing the robustness of the proposed model over existing studies on CXR radiographs. The Covid-MANet is a highly interpretable and generalizable model that benefits from lung segmentation and segmentation-based cropping. Moreover, internal and external validation reveals that the proposed methodology is more stable, and not affected by the variability of the dataset since model decision-making focused on relevant lung regions. In contrast, the results on the whole slide images lack generalization and interpretability because of some processing artifacts such as marker lines, uneven sectioning, and tissue folds that result in focus out-of-lungs. The advantage of the proposed methodology lies in following the real scenario of classification and infection segmentation of COVID-19 by recognition among five classes. Some studies directly classify COVID-19 based on infection detected by segmentation that works fine for healthy/COVID-19 cases but lacks when VP, BP, and TB samples having similar symptoms as COVID-19 occurs. Precisely, Covid-MANet can be applicable for clinical applications. The features of the proposed methodology are lung segmentation, proper documentation of pre-processing, training parameters, loss functions for reproducibility, explainable five-class classification model, higher interpretation, generalization, quantification, and severity grading as mild, moderate, severe, and critical based on RALE scoring system. We believe that the promising results of the proposed Covid-MANet system can guide radiologists in understanding the progression of the disease and better treatment plans in the early stages. Moreover, the Covid-MANet model can be deployed at airports, bus stands, railway stations, etc., for screening of large masses in less time.

In future work, we will extend our proposed framework to modalities other than CXR images such as ultrasound and CT images for explainable diagnosis and localization of COVID-19 infection. We believe that the proposed model is beneficial in improving the sensitivity of COVID-19 with these modalities. In addition, to extend work on CXR images, we plan to increase the corpus of the COVID-19 class by collecting samples from different institutions and exploring more public repositories. Also, we annotate VP and BP samples and increase the corpus of annotated COVID-19 samples under the guidance of radiological experts.

Ethical approval

This article does not contain any studies with human participants or animals performed by any of the authors.

Funding

There is no source of funding for this article.

Declaration of Competing Interest

The authors declare that they have no conflict of interest.

Acknowledgments

We would like to thank Dr. Edwin Hancock, Editor-in-Chief of Pattern Recognition for kind suggestions and critical comments for improving the presentation of the paper. We show our sincere gratitude to the Associate Editor and anonymous reviewer(s) whose valuable comments helped us greatly to enhance the experimental analysis. The first author is thankful to the University Grant Commission, Government of India for supporting him with a Junior Research Fellowship. The corresponding author greatly acknowledge the IoE grant of the Banaras Hindu University.

References

[1] WHO. WHO director-general’s opening remarks at the media briefing on COVID-19 - 11 march 2020. https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19—11 march-2020. 2020

[2] M.Y. Ng, E.Y. Lee, J. Yang, F. Yang, X. Li, H. Wang, M.D. Kuo, Imaging profile of the COVID-19 infection: radiologic findings and literature review, Radiol.: Cardiothorac. Imaging 2 (1) (2020) e200034.

[3] Z. Wang, Y. Xiao, Y. Li, J. Zhang, F. Lu, M. Hou, X. Liu, Automatically discriminating and localizing COVID-19 from community-acquired pneumonia on chest X-rays, Pattern Recognit. 110 (2021) 107613.

[4] Y. Fang, H. Zhang, J. Xie, M. Lin, L. Ying, P. Pang, W. Ji, Sensitivity of chest CT for COVID-19: comparison to RT-PCR, Radiology 296 (2) (2020) E115–E117.

[5] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Adv. Neural Inform. Process. Syst. 25 (2012) 1097–1105.

[6] M.A. Morid, A. Borjali, G Del Fiol, A scoping review of transfer learning research on medical image analysis using ImageNet, Comput. Biol. Med. 128 (2021) 104115.

[7] F.M. Shah, S.K.S. Joy, F. Ahmed, T. Hossain, M. Humaira, A.S. Ami, S. Ahmed, A comprehensive survey of covid-19 detection using medical images, SN Comput. Sci. 2 (6) (2021) 1–22.

[8] A. Kesarwani, K. Purohit, M. Dalui, D.R. Kisku, Measuring the degree of suitability of edge detection operators prior to an application, in: Proceedings of the 2020 IEEE Applied Signal Processing Conference (ASPCON), IEEE, 2020, pp. 128–133.

[9] K. Chadaga, C. Chakraborty, S. Prabhu, S. Umakanth, V. Bhat, N. Sampathila, Clinical and laboratory approach to diagnose COVID-19 using machine learning, Interdiscip. Sci.: Comput. Life Sci. 14 (2022) 452–470.

[10] Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., & Adam, H. (2017). Mobilenets: efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:*1704.04861.

[11] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4700–4708.

[12] K. Shwet, P.K. Mishra, A hybrid deep learning model for COVID-19 prediction and current status of clinical trials worldwide, Comput. Mater. Contin. (2021) 1896–1919.

[13] G. Kaur, P.S. Rana, V. Arora, State-of-the-art techniques using pre-operative brain MRI scans for survival prediction of glioblastoma multiforme patients and future research directions, Clin Transl. Imaging (2022) 1–35.

[14] A. Sharma, P.K. Mishra, Performance analysis of machine learning based optimized feature selection approaches for breast cancer diagnosis, International Journal of Information Technology (2021) 1–12.

[15] S. Ketu, P.K. Mishra, India perspective: CNN-LSTM hybrid deep learning model-based COVID-19 prediction and current status of medical resource availability, Soft Comput. 26 (2) (2022) 645–664.

[16] B. Gecer, S. Aksoy, E. Mercan, L.G. Shapiro, D.L. Weaver, J.G. Elmore, Detection and classification of cancer in whole slide breast histopathology images using deep convolutional networks, Pattern Recognit. 84 (2018) 345–356.

[17] L. Bi, D.D. Feng, M. Fulham, J. Kim, Multi-label classification of multi-modality skin lesion via hyper-connected convolutional neural network, Pattern Recognit. 107 (2020) 107502.

[18] A. Samanta, A. Saha, S.C. Satapathy, S.L. Fernandes, Y.D. Zhang, Automated detection of diabetic retinopathy using convolutional neural networks on a small dataset, Pattern Recognit. Lett. 135 (2020) 293–298.

[19] L. Wang, Z.Q. Lin, A. Wong, Covid-net: a tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images, Sci. Rep. 10 (1) (2020) 1–12.

[20] S. Tabik, A. Gómez-Ríos, J.L. Martín-Rodríguez, I. Sevillano-García, M. Rey-Area, D. Charte, F. Herrera, COVIDGR dataset and COVID-SDNet methodology for predicting COVID-19 based on Chest X-Ray images, IEEE J. Biomed. Health Inform. 24 (12) (2020) 3595–3605.

[21] A.S. Al-Waisy, M.A. Mohammed, S. Al-Fahdawi, M.S. Maashi, B. Garcia-Zapirain, K.H. Abdulkareem, D.N. Le, COVID-DeepNet: hybrid multimodal deep learning system for improving COVID-19 pneumonia detection in chest X-ray images, Comput. Mater. Contin. 67 (2) (2021) 2409–2429.

[22] A. Shamsi, H. Asgharnezhad, S.S. Jokandan, A. Khosravi, P.M. Kebria, D. Nahavandi, D. Srinivasan, An uncertainty-aware transfer learning-based framework for covid-19 diagnosis, IEEE Trans. Neural. Netw. Learn. Syst. 32 (4) (2021) 1408–1417.

[23] Z. Wang, Y. Xiao, Y. Li, J. Zhang, F. Lu, M. Hou, X. Liu, Automatically discriminating and localizing COVID-19 from community-acquired pneumonia on chest X-rays, Pattern Recognit. 110 (2021) 107613.

[24] A.I. Aviles-Rivero, P. Sellars, C.B. Schönlieb, N. Papadakis, GraphXCOVID: explainable deep graph diffusion pseudo-labelling for identifying COVID-19 on chest X-rays, Pattern Recognit. 122 (2022) 108274.

[25] C. Ieracitano, N. Mammone, M. Versaci, G. Varone, A.R. Ali, A. Armentano, F.C. Morabito, A fuzzy-enhanced deep learning approach for early detection of Covid-19 Pneumonia from portable chest X-ray images, Neurocomputing 481 (2022) 202–215.

[26] T. Rahman, A. Khandakar, Y. Qiblawey, A. Tahir, S. Kiranyaz, S.B.A. Kashem, M.E. Chowdhury, Exploring the effect of image enhancement techniques on COVID-19 detection using chest X-ray images, Comput. Biol. Med. 132 (2021) 104319.

[27] A. Malhotra, S. Mittal, P. Majumdar, S. Chhabra, K. Thakral, M. Vatsa, A. Agrawal, Multi-task driven explainable diagnosis of COVID-19 using chest X-ray images, Pattern Recognit. 122 (2021) 108243.

[28] J.D. Arias-Londoño, J.A. Gomez-Garcia, L. Moro-Velázquez, J.I. Godino-Llorente, Artificial Intelligence applied to chest X-Ray images for the automatic detection of COVID-19. A thoughtful evaluation approach, IEEE Access 8 (2020) 226811–226827.

[29] Y. Oh, S. Park, J.C. Ye, Deep learning covid-19 features on cxr using limited training data sets, IEEE Trans. Med. Imaging 39 (8) (2020) 2688–2700.

[30] Ghoshal, B., & Tucker, A. (2020). Estimating uncertainty and interpretability in deep learning for coronavirus (COVID-19) detection. *arXiv preprint arXiv:*2003.10769.

[31] W. Shi, L. Tong, Y. Zhuang, Y. Zhu, M.D. Wang, EXAM: an explainable attention-based model for COVID-19 automatic diagnosis, in: Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, 2020, pp. 1–6.

[32] R.K. Singh, R. Pandey, R.N. Babu, COVIDScreen: explainable deep learning framework for differential diagnosis of COVID-19 using chest X-rays, Neural Comput. Appl. 33 (14) (2021) 8871–8892.

[33] G. Wang, X. Liu, J. Shen, C. Wang, Z. Li, L. Ye, T. Lin, A deep-learning pipeline for the diagnosis and discrimination of viral, non-viral and COVID-19 pneumonia from chest X-ray images, Nature Biomed. Eng. 5 (6) (2021) 509–521.

[34] A.M. Tahir, M.E. Chowdhury, A. Khandakar, T. Rahman, Y. Qiblawey, U. Khurshid, T. Hamid, COVID-19 infection localization and severity grading from chest X-ray images, Comput. Biol. Med. 139 (2021) 105002.

[35] A. Signoroni, M. Savardi, S. Benini, N. Adami, R. Leonardi, P. Gibellini, D. Farina, BS-Net: learning COVID-19 pneumonia severity on a large chest X-ray dataset, Med. Image Anal. 71 (2021) 102046.

[36] P.S. Gidde, S.S. Prasad, A.P. Singh, N. Bhatheja, S. Prakash, P. Singh, D. Dash, Validation of expert system enhanced deep learning algorithm for automated screening for COVID-Pneumonia on chest X-rays, Sci. Rep. 11 (1) (2021) 1–12.

[37] M. Owais, H.S. Yoon, T. Mahmood, A. Haider, H. Sultan, K.R. Park, Light-weighted ensemble network with multilevel activation visualization for robust diagnosis of COVID19 pneumonia from large-scale chest radiographic database, Appl. Soft Comput. 108 (2021) 107490.

[38] X. Ouyang, J. Huo, L. Xia, F. Shan, J. Liu, Z. Mo, D. Shen, Dual-sampling attention network for diagnosis of COVID-19 from community acquired pneumonia, IEEE Trans. Med. Imaging 39 (8) (2020) 2595–2605.

[39] M.J. Horry, S. Chakraborty, M. Paul, A. Ulhaq, B. Pradhan, M. Saha, N. Shukla, COVID-19 detection through transfer learning using multimodal imaging data, IEEE Access 8 (2020) 149808–149824.

[40] Y.H. Wu, S.H. Gao, J. Mei, J. Xu, D.P. Fan, R.G. Zhang, M.M. Cheng, Jcs: an explainable covid-19 diagnosis system by joint classification and segmentation, IEEE Trans. Image Process. 30 (2021) 3113–3126.

[41] M. Roberts, D. Driggs, M. Thorpe, J. Gilbey, M. Yeung, S. Ursprung, C.B. Schönlieb, 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) (2021) 199–217.

[42] O. Ronneberger, P. Fischer, T. Brox, U-net: convolutional networks for biomedical image segmentation, in: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, Cham, 2015, pp. 234–241.

[43] M.Z. Khan, M.K. Gajendran, Y. Lee, M.A. Khan, Deep neural architectures for medical image semantic segmentation, IEEE Access 9 (2021) 83002–83024.

[44] Mangal, A., Kalia, S., Rajgopal, H., Rangarajan, K., Namboodiri, V., Banerjee, S., & Arora, C. (2020). CovidAID: COVID-19 detection using chest X-ray. *arXiv preprint arXiv:*2004.09803.

[45] N.S. Punn, S. Agarwal, Modality specific U-Net variants for biomedical image segmentation: a survey, Artif Intell Rev (2022) 1–45.

[46] A. Sharma, P.K. Mishra, Deep learning approaches for automated diagnosis of COVID-19 using imbalanced training CXR data, in: Proceedings of the International Conference on Advanced Network Technologies and Intelligent Computing, Springer, Cham, 2021, pp. 453–472.

[47] J. Shiraishi, et al., Development of a digital image database for chest radiographs with and without a lung nodule, Amer. J. Roentgenol. 174 (1) (Jan. 2000) 71–74, doi:10.2214/ajr.174.1.1740071.

[48] B. van Ginneken, M.B. Stegmann, M. Loog, Segmentation of anatomical structures in chest radiographs using supervised methods: a comparative study on a public database, Med. Image Anal. 10 (1) (2006) 19–40 Feb..

[49] S. Jaeger, S. Candemir, S. Antani, Y.X.J. Wáng, P.X. Lu, G. Thoma, Two public chest X-ray datasets for computer-aided screening of pulmonary diseases, Quant Imaging Med. Surg. 4 (2014) 475.

[50] Cohen, J.P., Morrison, P., Dao, L., Roth, K., Duong, T.Q., Ghassemi, M., 2020. COVID-19 image data collection: prospective predictions are the future. ArXiv: 2006.11988 URL: https://github.com/ieee8023/covid-chestxray-dataset.

[51] A. Pitman, D.N. Jones, D. Stuart, K. Lloydhope, K. Mallitt, P. O'rourke, The Royal Australian and New Zealand College of Radiologists (RANZCR) relative value unit workload model, its limitations and the evolution to a safety, quality and performance framework, J. Med. Imaging Radiat. Oncol. 53 (5) (2009) 450–458.

[52] Mooney, P.: Kaggle chest x-ray images (pneumonia) dataset. https://www.kaggle. com/paultimothymooney/chest-xray-pneumonia (2018)

[53] A. Chatterjee, J. Saha, J. Mukherjee, Clustering with multi-layered perceptron, Pattern Recognit. Lett. 155 (2022) 92–99.

[54] Purohit, K., Kesarwani, A., Kisku, D.R., & Dalui, M. (2020). Covid-19 detection on chest x-ray and ct scan images using multi-image augmented deep learning model. bioRxiv.

[55] M. Shorfuzzaman, M.S. Hossain, MetaCOVID: a Siamese neural network framework with contrastive loss for n-shot diagnosis of COVID-19 patients, Pattern Recognit. 113 (2021) 107700.

[56] K. Chadaga, S. Prabhu, V. Bhat, S. Umakanth, N. Sampathila, Medical diagnosis of COVID-19 using blood tests and machine learning, Journal of Physics: Conference Series 2161 (1) (2022) 012017.

Ajay Sharma is a Research Scholar under the supervision of Prof. P. K. Mishra in Department of Computer Science, Institute of Science, Banaras Hindu University, Varanasi (India). His-research interests include, Machine learning, Deep learning, Computer Vision and Biomedical Image Analysis.

P. K. Mishra is Professor at Department of Computer Science, Institute of Science, Banaras Hindu University, India. He is also a Principal Investigator of the research projects at DST Center for Interdisciplinary Mathematical Sciences, Banaras Hindu University. He is a senior member of IEEE. His-research interests include Computational Complexity, Data Mining, Computer Vision, IoT, High Performance Computing and VLSI Algorithms.