Chapter 4

Segmentation-Based Classification Deep Learning Model Embedded with Explainable AI for COVID-19 Detection in Chest X-ray Scans

Abstract

COVID-19 is a highly infectious disease caused by severe acute respiratory syndrome coronavirus 2. This has caused a massive loss of life during the last two years. The standard diagnostic method has various limitations, including delayed results, high cost, and an error rate. Due to the gap in unrevealed information and research attainment, there is a need for precise, fast, cost-effective, and easily accessible diagnostic methods. The application of artificial intelligence in medical imaging is a boon for lesion detection and diagnosis. This study proposes a deep-learning-based algorithm for the automatic and rapid detection of COVID-19 in chest X-ray scans with high precision. The system was developed by a segmentation-based classification of X-ray images into five classes that cover almost every common pneumonia type. The two deep learning-based segmentation networks, namely UNet and UNet+, along with eight classification models, namely VGG16, VGG19, Xception, InceptionV3, Densenet201, NasnetMobile, Resnet50, and MobileNet, were applied to select the best-suited combination of networks. The best-performing segmentation model was UNet, which exhibited the accuracy, loss, Dice, Jaccard, and AUC of 96.35%, 0.15%, 94.88%, 90.38%, and 0.99, respectively. The best-performing classification model was Xception, which exhibited the accuracy, precision, recall, f1-score, and AUC of 97.45%, 97.46%, 97.45%, 97.43%, and 0.998, respectively. Our system outperformed existing methods by 1.13% and can be used in a clinical setting. The system can detect COVID-19 with rapid and precise results in chest X-ray scans.

4.1. Introduction

COVID-19 is a highly infectious disease caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). After the first case was identified in December 2019 in Wuhan, China [156], the virus spread rapidly worldwide, leading to the COVID-19 pandemic in March 2020 [157]. Out of coronavirus-infected people, 81% develop minor to moderate symptoms like mild pneumonia, 14% develop severe symptoms like dyspnea or hypoxia, and 5% develop acute symptoms like shock, respiratory failure, or multiorgan dysfunction such as myocardial injury or vascular injury [158].

The standard diagnostic method for COVID-19 is the detection of the virus nucleic acid in a nasopharyngeal sample by RT-PCR (real-time reverse transcription-polymerase chain reaction), RT-LAMP (reverse transcription loop-mediated isothermal amplification), or TMA (transcription-mediated amplification). However, all these procedures are laborious, rigorous, complicated, time-consuming, and costly, with a significantly high error rate [82].

Medical imaging techniques are one of the most fruitful options for detecting infections, diseases, or lesions in the internal organs or other body parts [159]. X-rays [160], CT scans [161], and ultrasounds [162] are some of the best and most common examples. Chest X-rays and CT scans are also being used to detect COVID-19 infection, the severity or stage of the infection, and the level of lung involvement or damage after the infection. Chest X-ray has an advantage over CT in having low radiation dose, being cost-effective, easily available, and having instant results [36].

The addition of artificial intelligence, especially deep learning techniques, into medical imaging is one of the miracles in the 21st century for medical diagnosis [163, 164]. Including deep learning techniques in medical imaging has significantly contributed to precise accuracy, rapid detection, and lowering of medical burden, workforce, and human error [165, 166]. The AI and CAD systems have already been approved and accepted by the medical community for several disease conditions, such as tumor detection [167, 168].

For COVID-19 detection in chest X-rays and CT, AI and deep learning methods have shown tremendous success rates and accuracy [169, 170]. Several researchers have given their valued input to achieve excellent outcomes [171]. Numerous medical communities and institutions have accepted their efforts and results. Several authors have innovated new deep learning-based classification algorithms to develop an automatic model to detect COVID-19 in CXR scans [172, 173]. However, most of them carried out classification without lung segmentation or executed only two or three-class classification [54, 56]. Some authors reported work on the segmentation-based classification model; however, multiclass classification and high accuracy were still missing [52, 59]. As a result, these systems cannot be adapted for clinical practice since they cannot meet the regulatory requirements of an error rate of <5%. Our hypothesis states that if we can design a segmentation-based classification system having an error rate < 5%, typically adopted for 510 (K) regulatory purposes, the diagnostic system can be adapted in clinical settings [174].

The segmentation-based classification models are always a better choice as they remove the unwanted areas present in scans that could be the reason for the misguidance in the network training. We use this spirit to hypothesize our vision that segmentation, when combined with classification, can get the most reliable results. Along these lines, we have developed a two-stage system, i.e., a segmentation-based classification model for COVID-19 detection and classification into multi classes, i.e., five classes.

Recently, the UNet system has shown a powerful solution for segmentation in several applications [175]. We also utilized a large number of CXR scans so that our system could be more stable and robust. Further, classification models like VGG16, VGG19, and Xception have shown their ability to classify well [126]. We have applied the eight most popular and efficient deep learning-based classification models to achieve the most precise results. Additionally, we used an explainable AI method and GRAD-CAM heatmap visualization techniques to detect and manifest the lesion present in the X-ray scans.

Figure 4.1 shows the overall methodology we opted for the experiment. The comprehensive experiment was accomplished in two phases: the first deals with segmentation and the second with classification. Four different datasets were utilized in the experiment. The first dataset with X-ray images and their corresponding masks was used to train the two segmentation models, namely, UNet and UNet+.

The other three datasets were utilized for classification purposes. The segmentation model's performance was compared to select the best model. After that, the best segmentation model was applied to the classification dataset to get the segmented lung images. Eight different deep learning networks, namely: VGG16, VGG19, Xception, InceptionV3, Densenet201, NasnetMobile, Resnet50, and MobileNet, were applied for the classification of segmented lung images into five classes: COVID-19, viral pneumonia, bacterial pneumonia, tuberculosis, and normal. The performance of all the classification models was evaluated using several parameters, including the accuracy, AUC, confusion matrix, and heatmap visualization of the images. For our COVID-19 detection system, the best combination of networks for segmentation and classification was finally chosen.

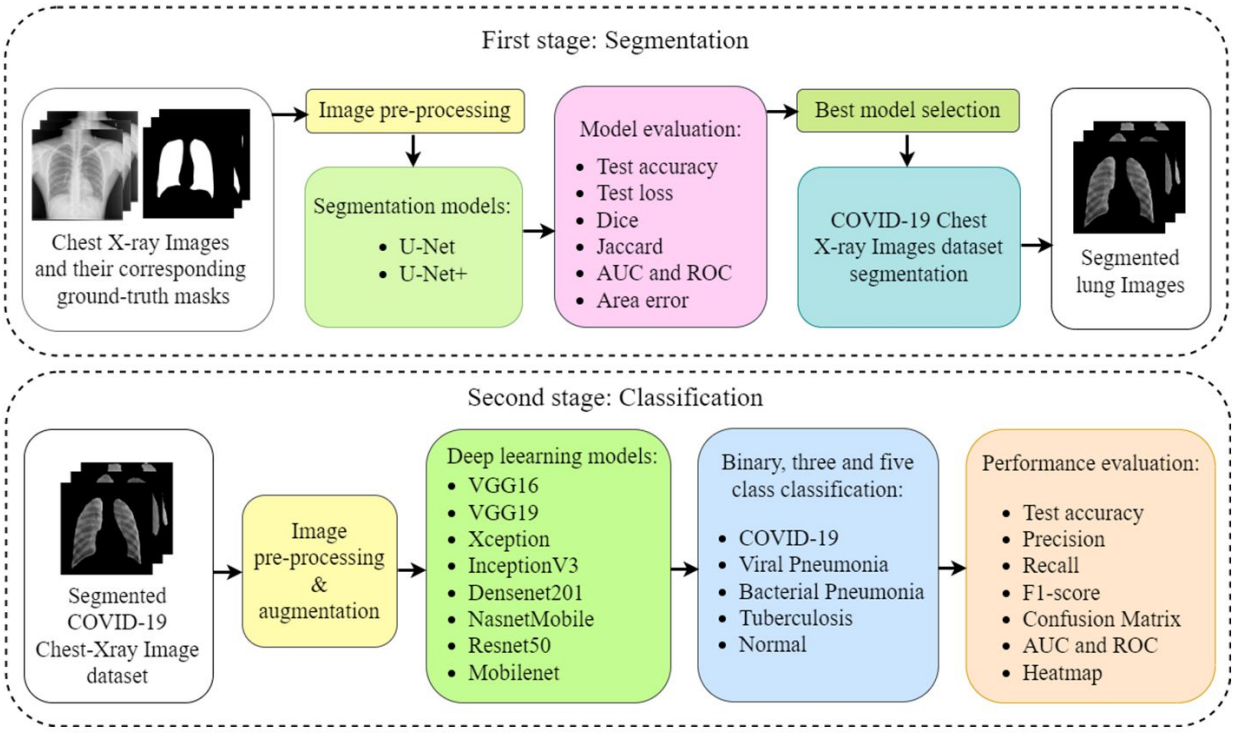


Figure 4.1: The step-wise overall schematic diagram for the proposed method.

4.2.Methodology

The entire methodology opted for developing our segmentation-based classification system has been accomplished and described into two phases or sections: (i) segmentation and (ii) classification. In each section, we have described the details of the dataset, the deep neural network architecture, the experiment protocol, the training parameters, and the evaluation metrics.

4.2.1. Data Collection and Patient Demographics

For the classification phase of the experiment, a total of 12,926 chest X-ray images were used. The images were taken from three different publicly available data sources, which are the "COVID-19 Radiography Database" [114], "Tuberculosis (TB) Chest X-ray Database" [115], and "Chest X-Ray Images (Pneumonia) " [116]. The "COVID-19 Radiography database" contains 3,616 COVID-19, 1,345 viral pneumonia, and 10,192 normal images. The dataset was created by a group of researchers and doctors from Bangladesh. Pakistan, and Malaysia. From the dataset, we have taken 3611 COVID-19, 1345 viral pneumonia, and 4490 normal images for the experiment. The "Chest X-ray images (Pneumonia)" dataset contains 5,863 images, with 2,780 bacterial pneumonia images. The chest radiographs were taken from the Guangzhou Women and Children's Medical Center, Guangzhou. From the dataset, we have taken all 2780 bacterial pneumonia radiographs for the experiment. Next, the "Tuberculosis Chest X-ray Database" contains 700 tuberculosis chest X-rays. The database was created by the collaboration of several groups of researchers and doctors. We have utilized all 700 radiographs of tuberculosis for our experiment.

Finally, 12,926 total CXR images having COVID-19, VP, BP, TB, and normal images of 3,611, 1,345, 2,780, 700, and 4,490 respectively were utilized for our classification experiment. All the CXR images were segmented before the classification. For training of the classification model, 80%, *i.e.*, 10,338 total images having COVID-19, VP, BP, TB, and normal images of 2,887, 1,075, 2,224, 560, and 3,592, respectively, were utilized. Next, for validation of the model, 10%, *i.e.*, 1,294 randomly selected images, of which COVID-19, VP, BP, TB, and normal images of 362, 135, 278, 70, and 449, respectively, were utilized. Finally, for testing the model, 10%, i.e., 1,294 randomly selected images that were not involved in training nor in validation and included COVID-19, VP, BP, TB, and normal images of 362, 135, 278, 70, and 449, respectively, were utilized.

4.2.2. The architecture of classification networks

The convolutional neural network comprises an input, hidden, and output layer. The neural network's layer works in a feed-forward manner. The intermediary layers are hidden since the activation function and final convolution hide their input and outputs. The hidden layers typically consist of convolution layers followed

by activation, pooling, and fully connected layers. The feature maps that are generated by convolution work as input for the next layer. For the classification of segmented lung images into five classes, we applied eight highly efficient deep convolutional neural networks, namely VGG16, VGG19, Xception, InceptionV3, Densenet201, NASNetMobile, Resnet50, and MobileNet. The architectures of neural networks are shown in Figures 2.3 – 2.8, Figure 4.2, and Figure 4.3. Each figure describes details about the network's hidden layers, including convolution layers, their input layer, fully connected (FC)-layers, and output layers.

Figure 2.3 represents VGG16 architecture. VGG16 is a 16-layer depth model with 13 convolution layers. It has 138 million parameters with a size of 528 MB. It performs with a speed of 4.2 ms per inference step using GPU. Figure 2.4 represents VGG19 architecture. VGG19 is a slightly larger network than VGG16 and has a depth of 19 layers with 16 convolutional layers. It is 548 MB in size with 143 million parameters. It performs with a speed of 4.4 ms per inference step. Xception is an 81-layer depth model represented in Figure 2.6. It has 22.9 million parameters with a size of 88 MB. It performs with a speed of 8.1 ms per inference step. Figure 2.7 represents InceptionV3 architecture. It is an 189 layers-depth model. InceptionV3 is a 92 MB network with 23 million parameters. Its speed is 6.9ms per inference step. Figure 2.5 represents DenseNet201 architecture. DenseNet201 is the highest in-depth, with 402 layers. However, it is smaller in size in comparison to others, having 8 million parameters with 33 MB of size. It provides a speed of 5.4 ms per inference step. Figure 2.8 represents NASNetMobile architecture. NASNetMobile is the smallest network after MobileNet in all our eight models, even after it has the highest depth after DenseNet201 with 389 layers. It has 5.3 million parameters with 23 MB in size. Its speed is 6.7 ms per inference step. Figure 4.2 represents ResNet50 architecture. ResNet50 has a depth of 107 layers. It has 25.6 million parameters with 98 MB in size. It provides a speed of 4.6 ms per inference step. MobileNet is the smallest network among all eight models represented in Figure 4.3. It has a depth of 55 layers with 4.3 million parameters and is 16 MB in size. It is the fastest among all, with a performance of 3.4 ms per inference step. Comparing all eight networks, VGG19 is the largest in size and parameters, Xception is the maximum in-depth, and MobileNet is the fastest network.

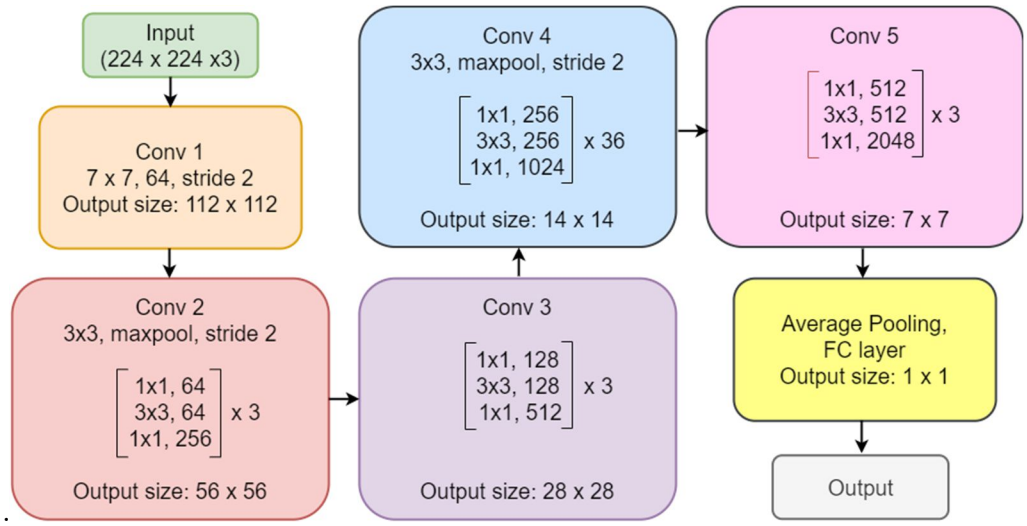


Figure 4.2: ResNet50 architecture.

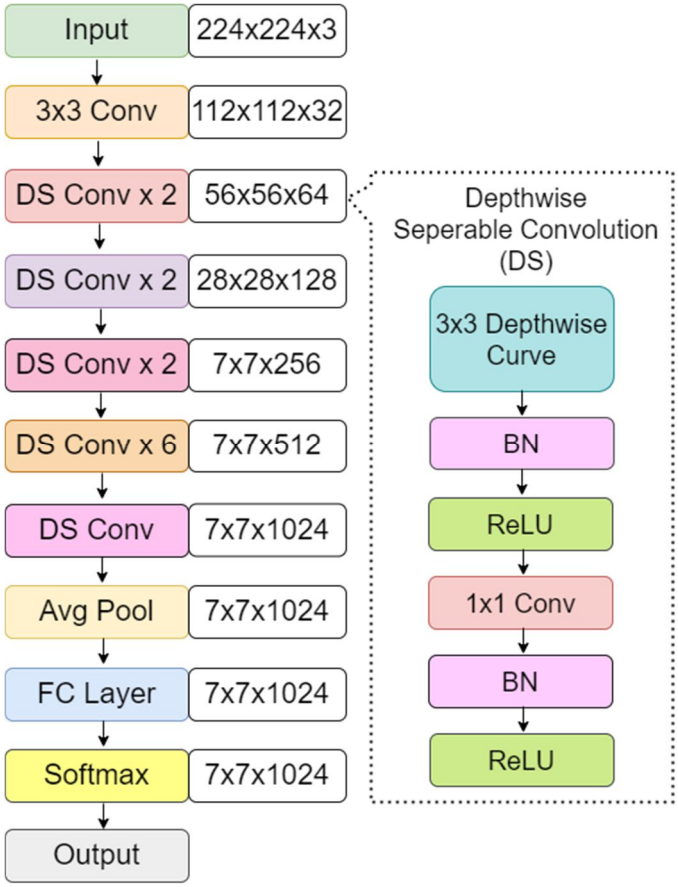


Figure 4.3: MobileNet architecture.

4.2.3. Model optimization

All the models were trained for 150 epochs with a learning rate of 0.001 and a batch size of 8 images. Model checkpoints (save best only) are applied as callbacks. Before the training, all the images were resized to a pixel value of 224 × 224. The loss function used during training was categorical cross-entropy. The categorical cross-entropy loss function can be defined as the equation below:

$$L_{CCE} = \frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} 1_{y_i \in C_c} \log a_{model}\,(y_i \in C_c) \qquad (4.1)$$

The entire experiment was carried out using Python 3.8. For training the network, we employed a workstation with 8GB NVIDIA Quadro P4000 GPU (Graphics Processing Unit). The system had an Intel Core i7 8th Generation processor and 16GB of RAM.

4.2.4. Matrices used for result evaluation

The performance of each network was evaluated on test data after the training and validation process. Five different matrices were utilized for the performance evaluation, namely accuracy, precision, recall, F1-score, and area under the curve (AUC). The mathematical equations for each matrix are given in section 2.3.6.

4.3. Results

4.3.1 Binary classification

Figure 4.4 represents the segmentation results on the classification dataset by the UNet model. As shown in the figure, the uppermost row represents the original chest X-ray images from the COVID-19 class (columns ‘a’ and ‘b') and normal class (columns 'c' and 'd'). Next, the middle row represents the corresponding segmented mask by the UNet model. Finally, the bottom row represents the corresponding segmented lung images.

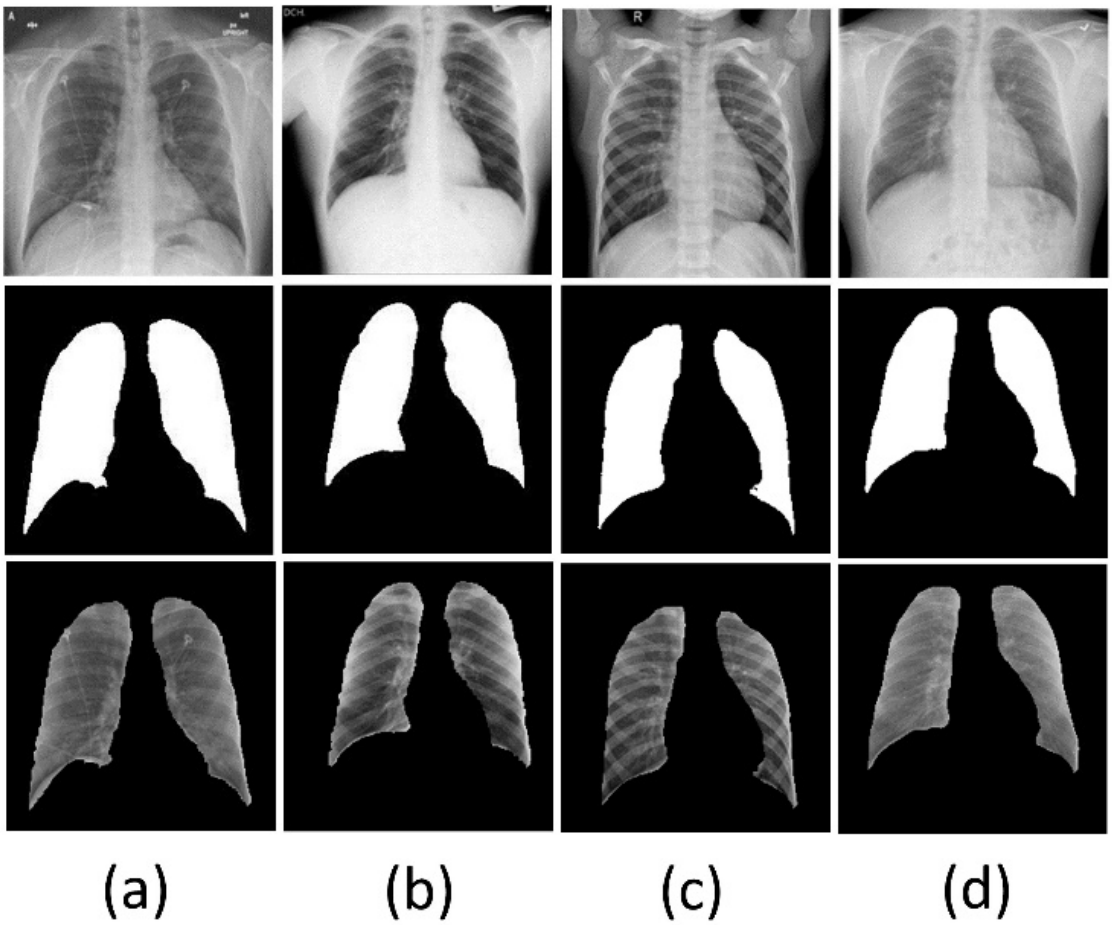


Figure 4.4: Example of Images from each class. Column (a) & (b): COVID-19, Column (c) & (d): Normal. The images in each row represent the top row: original chest X-ray images, the middle row: segmented masks, and the bottom row: segmented lung images.

Table 4.1: Performance metrics for classification by the Xception DL model.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Xception	99.01	99.01	98.98	99.00

Table 4.1 represents the performance metrics of classification by the highly efficient Xception model. The network performed with a test accuracy of 99.01% and the macro average of precision, recall, and F1-score of 99.01%, 98.98%, and 99.00%, respectively. The performance of the model was outstanding, with a high level of precision.

Learning curves are a broadly used analytical tool in machine learning. The algorithm's performance may be assessed using the curves that learn from a training dataset. The models are evaluated on the training dataset and validation dataset after each update during training and validation. The plots of the measured performance are created as learning curves. The learning-related problems may be diagnosed by assessing the learning curves of models. Figure 4.5 represents the training and validation accuracy curve by the Xception model. As the figure reveals, we reached the maximum training accuracy of 99.34% and validation accuracy of 99.08%. The graph indicates that the accuracy is improving and stabilizing along the succeeding epochs; this determines the model's precise and robust learning pattern.

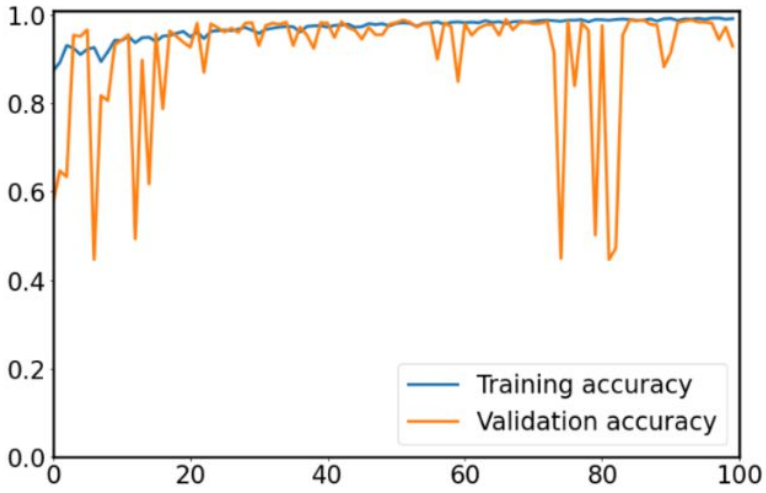


Figure 4.5: The Training and validation accuracy curve by the Xception model.

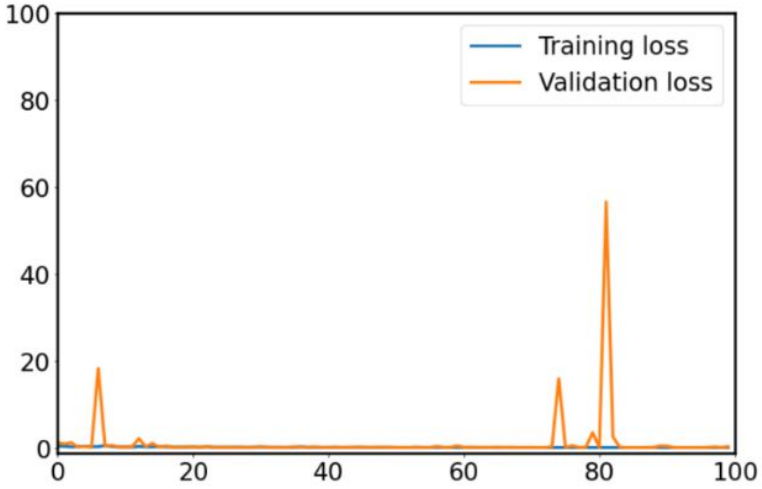


Figure 4.6: The Training and validation loss curve by the Xception model.

Limiting the loss is an essential factor for decent model learning. Low loss indicates preciseness and more stability of the model. Figure 4.6 represents the training and validation loss curve by the Xception model. The graph demonstrates that the loss is decreasing and stabilizing along the succeeding epochs.

Figure 4.7 represents the confusion matrix result of test data classification. The figure reveals that out of 362 COVID-19 images, 357 were correctly classified, and five were misclassified to normal class. Out of 449 normal images, 446 were correctly classified, and three were misclassified to the COVID-19 class.

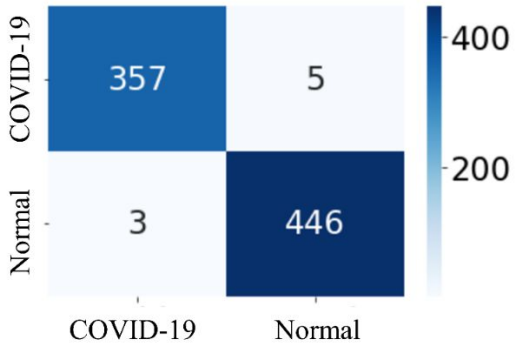


Figure 4.7: Confusion matrix for test data classification.

4.3.2. Three class classification

Figure 4.8 represents the sample images from each class after the segmentation. Colum (a) denotes the images from the COVID-19 class, (b) denotes viral pneumonia, and (c) denotes the normal class. Images from the top row denote original Chest X-ray images, the middle row signifies segmented masks by the UNet+ model, and the bottom row represents the final segmented lung images that were later used for the classification.

For classification, we applied four different tremendously efficient deep neural networks, including a hybrid model, namely InceptionResNetV2, DenseNet121, EfficientNetB1, and Xception. Table 4.2 represents the performance metrics of each network. The Xception model performed best with an accuracy of 98.52% and the weighted average of precision, recall, and F1-score of 98.53%, 98.52%, and 98.52%, respectively. The performance of the hybrid model InceptionResNetV2 was second, with an accuracy

of 96.09%. The DenseNet network performed with an accuracy of 95.24%, and the EfficientNet network performed with 94.40% accuracy.

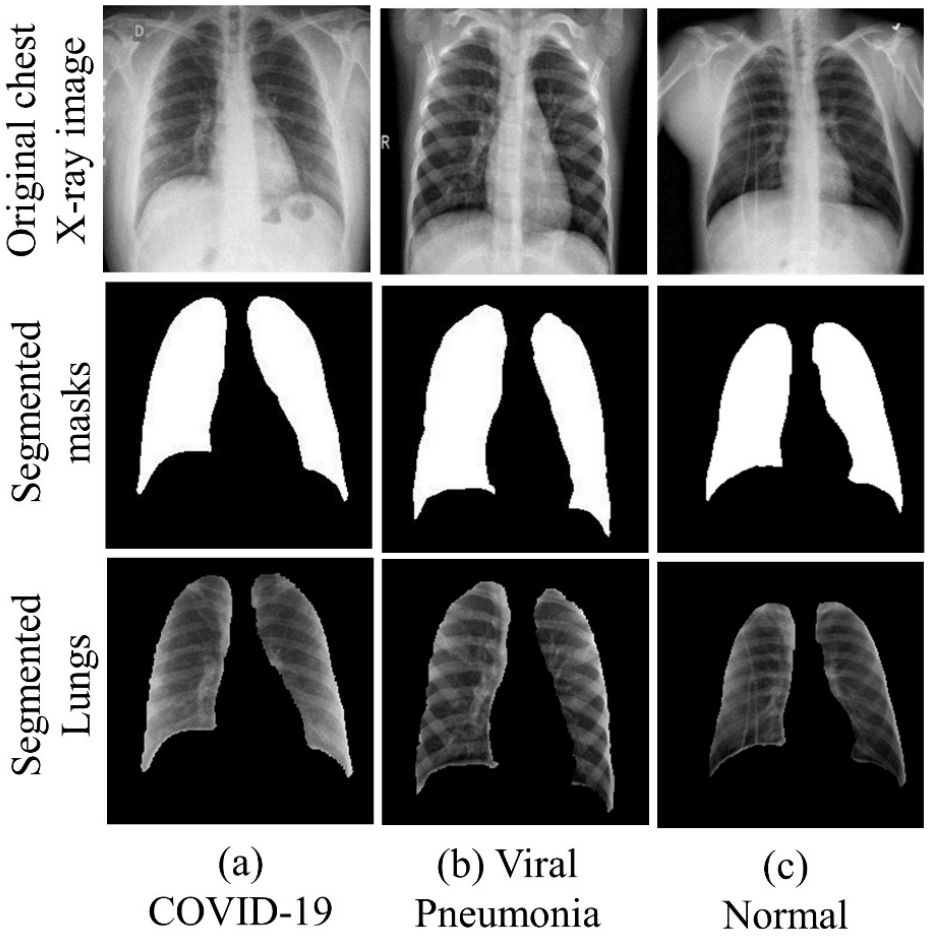


Figure 4.8: Segmented lung images from each class.

Table 4.2: Performance metrics for classification by different models.

Models	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
InceptionResNetV2	96.09	96.09	96.09	96.08
DenseNet121	95.24	95.39	95.24	95.25
EfficientNetB1	94.40	94.44	94.40	94.39
Xception	98.52	98.53	98.52	98.52

Table 4.3 represents the classification performance metrics for each class by best best-performing Xception model. The Normal images performed best in terms of precision, with 99.33%. COVID-19 images performed best in recall with 99.17%, and in the F1-score, again, normal images performed best by 98.88%.

Table 4.3: Performance metrics for classification of each class by Xception model.

Class	Precision (%)	Recall (%)	F1-score (%)
COVID-19	98.36	99.17	98.76
Viral Pneumonia	96.32	97.04	96.68
Normal	99.33	98.44	98.88

Figure 4.9 represents the training and validation accuracy curve by the best-performing Xception model. The training was done up to 150 epochs. The curve indicates that the accuracy is improving throughout the succeeding epochs.

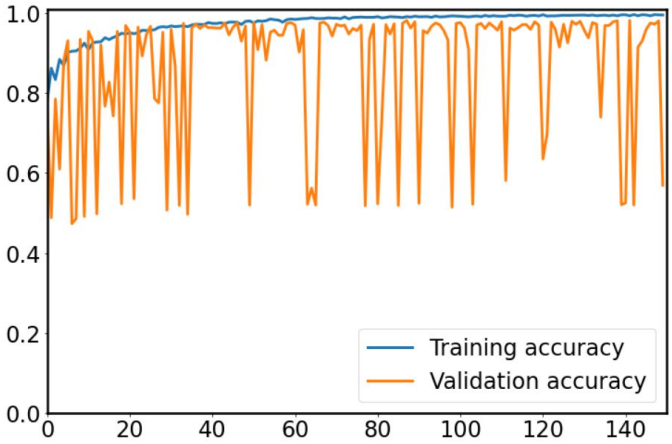


Figure 4.9: Training and Validation accuracy curve by the best performing Xception model.

Figure 4.10 represents the training and validation loss curve by the best-performing Xception model. The curve indicates that the loss decreases throughout the succeeding epochs.

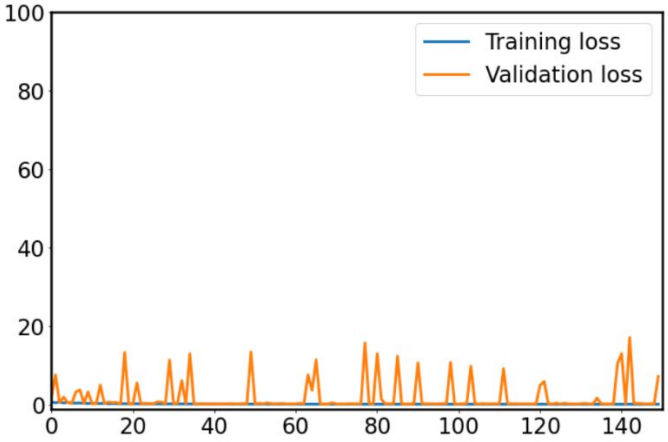


Figure 4.10 Training and Validation loss curve by the best performing Xception model.

Figure 4.11 represents the confusion matrix for the classification of the test dataset. The confusion matrix reveals that out of 362 COVID-19 images, 359 images were correctly classified as COVID-19, and three images were misclassified with two to normal and one to viral pneumonia class. Further, out of 449 normal images, 442 images were correctly classified to normal class, and seven images were misclassified with three to COVID-19 and four to viral pneumonia class. Finally, out of 135 images from viral pneumonia, class 131 were correctly classified, and four images were misclassified, with three to COVID-19 and one to normal class.

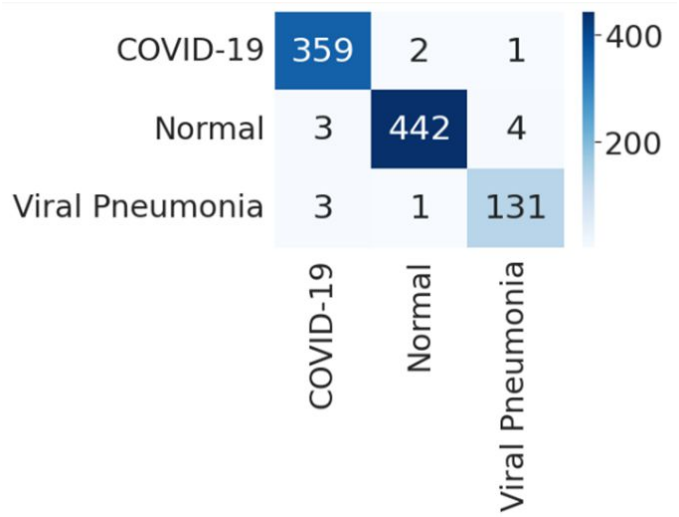
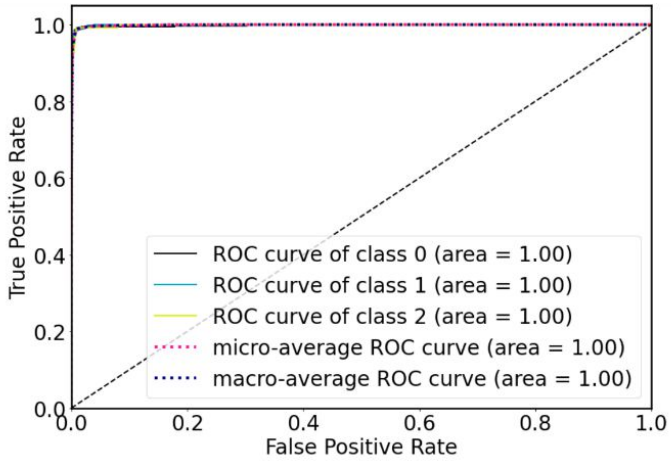


Figure 4.11: Confusion matrix for classification by Xception model.



(p<0.0001; class 0: COVID-19; class2: normal; class 3: viral pneumonia)

Figure 4.12: ROC curves and AUC values for three-class classification by the Xception model.

The ROC curves and AUC values are calculated based on inference values and actual labels for each class. The values of AUC count between 0 and 1. The higher AUC denotes a superior model. Figure 4.12 represents the ROC curves and AUC values for the classification of the test dataset by the Xception model. The AUC values for each class, i.e., COVID-19, viral pneumonia, and normal, were absorbed as 1.0 for each. The maximum AUC performance for each class denotes the best capability and implementation of the presented model.

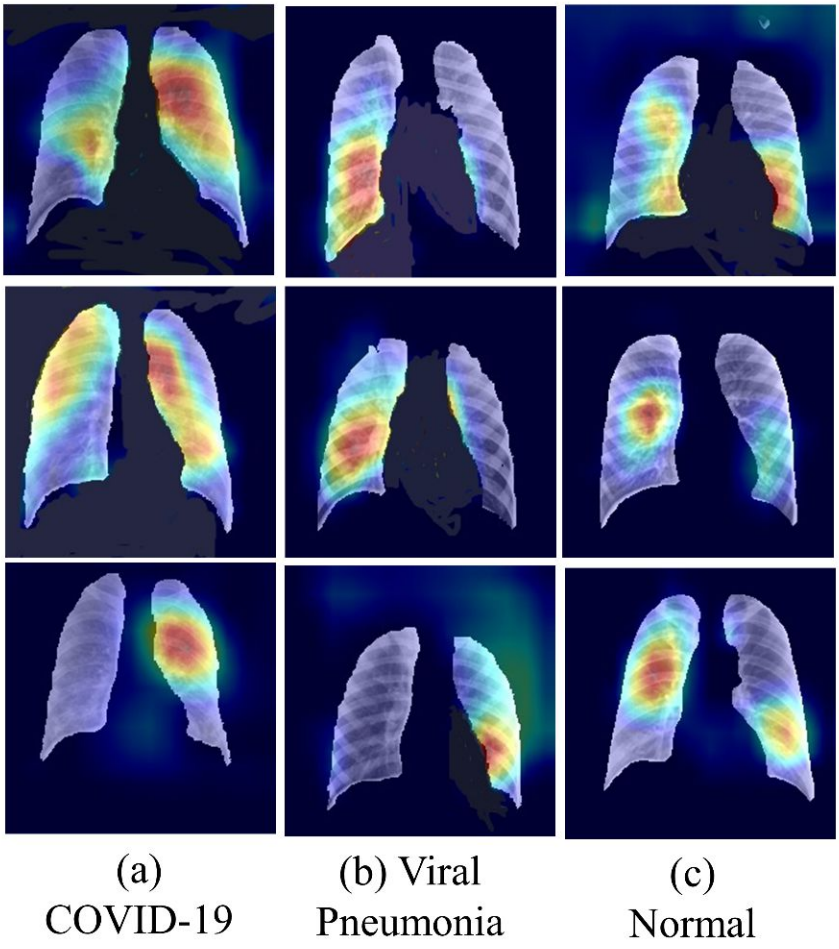


Figure 4.13: GRAD-CAM heatmap visualization for each class of images.

Figure 4.13 represents the heatmap of images from all three classes by the Xception model. Row (a) denotes COVID-19, (b) denotes viral pneumonia, and (c) denotes normal images. The heatmap images were generated using features from the final convolution layer. The Grad-CAM (Gradient-weighted Class Activation Mapping) algorithm was applied to generate the heatmap images. Grad-CAM constructs the

coarse localization map and displays the crucial locations as heatmap scans. All images from the COVID-19 class show a similar heatmap pattern. Mostly, the upper part of the lungs gets infected in the early stage of COVID-19. The heatmap images show a similar pattern and indicate the infection mostly in the upper part of the lungs in each sample image. Whereas for the viral pneumonia images, the heatmap images indicate the infection in the lower part of the images and are consistent and similar in all the viral pneumonia sample images. For the normal images, the features have been taken from all over the parts of the lungs by the model.

4.3.3 Five class Classification

After the segmentation of classification data, our next goal was to successfully classify and develop a best-suited classification model for the segmented chest X-ray images into five classes with optimal performance. To achieve the goal, we applied eight different highly efficient deep neural networks, namely VGG16, VGG19, Xception, InceptionV3, Densenet201, NasnetMobile, Resnet50, and Mobilenet, for the classification of segmented lung images into five classes: COVID-19, VP, BP, TB, and normal. Table 4.4 shows the comparison of the performance metrics of all eight CNNs. The Xception model performed best with an accuracy of 97.45% and a weighted average of Precision, Recall, and F1-score of 97.46%, 97.45%, and 97.43%, respectively. The performance of MobileNet was the second most efficient, with an accuracy of 93.66% and precision, recall, and F1-score of 93.87%, 93.66%, and 93.60%, respectively.

Table 4.4: The weighted average of performance metrics by eight different deep learning models for five-class classification of segmented chest X-ray images into COVID-19, VP, BP, TB, and normal.

DL model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
VGG16	88.25	87.88	88.25	87.82
VGG19	87.64	87.08	87.64	87.15
Xception	97.45	97.46	97.45	97.43
InceptionV3	90.88	90.89	90.88	90.47
Densenet201	82.07	82.22	82.07	80.90
NasnetMobile	92.97	93.01	92.97	92.78
Resnet50	90.03	90.00	90.03	89.70
MobileNet	93.66	93.87	93.66	93.60

Table 4.5 shows the performance metrics of each class by the best-performing Xception model. The Precision was best for the COVID-19 class, with 98.88%, whereas the Recall was best for the Bacterial Pneumonia class, with 100%, and F1-score was best for the normal class, with 98.55%.

Table 4.5: The performance metrics (precision, recall, and F-1 score) for each class for segmented chest X-ray images by the best performing Xception model.

Class	Precision (%)	Recall (%)	F1-score (%)
Bacterial Pneumonia	95.53	100.00	97.72
COVID	98.88	97.51	98.19
Normal	98.66	98.44	98.55
Tuberculosis	94.44	97.14	95.77
Viral Pneumonia	95.24	88.89	91.95

Training and Validation Curve

Figure 4.14 shows the training and validation accuracy for the best-performing Xception model. The curve indicates that training and validation accuracy improved with the successive epochs that point towards a good model.

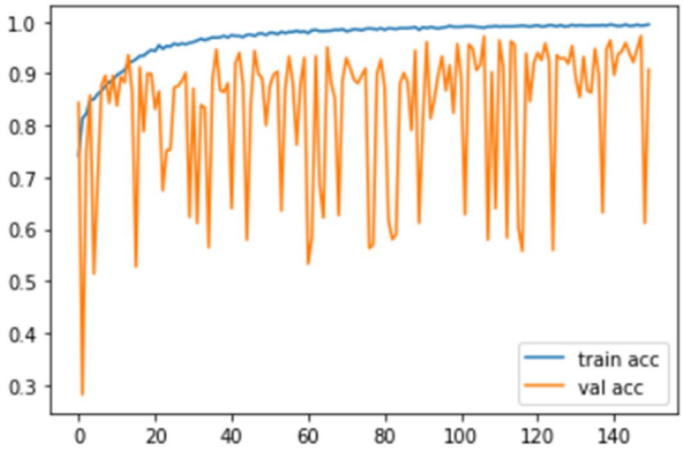


Figure 4.14: Training and validation accuracy for the best-performing Xception model. Train acc represents training accuracy, and val acc represents validation accuracy.

Figure 4.15 shows the training and validation loss curve. The curve indicates that training and validation loss is stable and reduced with successive epochs, supporting this as a good model.

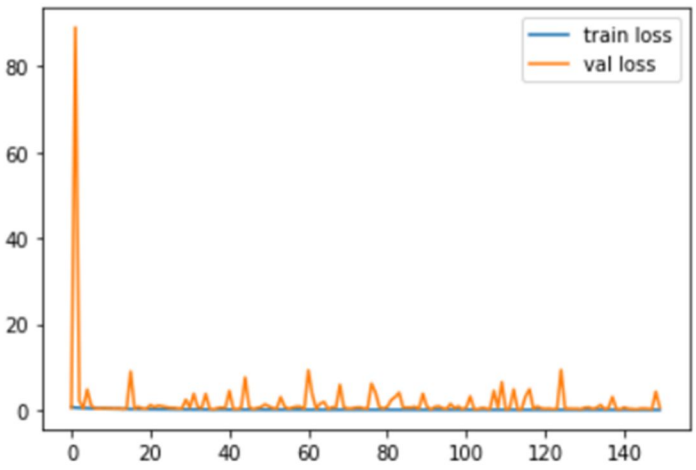


Figure 4.15: Training and validation loss for the best-performing Xception model.

Confusion Matrix

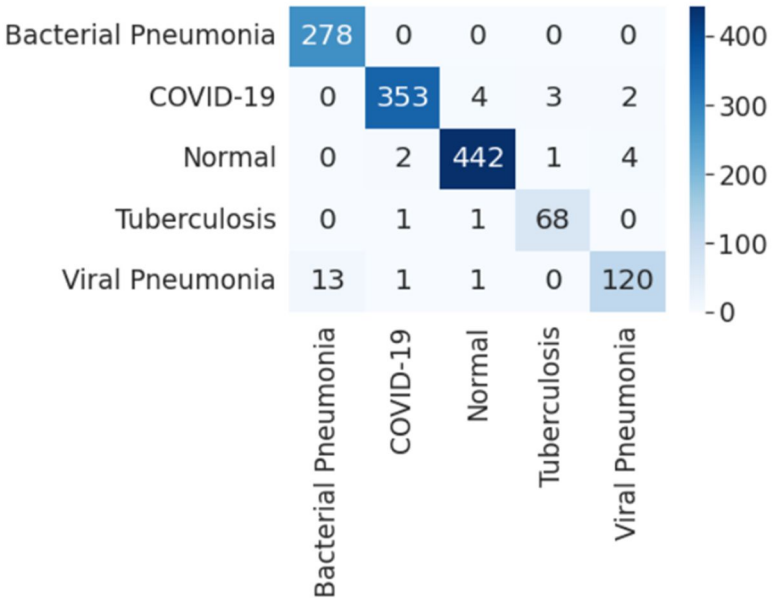


Figure 4.16: Confusion matrix for five-class classification by the Xception model.

Figure 4.16 represents the confusion matrix for the test set results by the best-performing Xception network. Results reveal that for 362 COVID-19 chest X-ray images, 353 were correctly classified, and nine were incorrectly predicted as two to viral pneumonia, three to tuberculosis, and four to normal class. Next, for the viral pneumonia class, out of a total of 135 images, 120 were correctly classified, and 15 were wrong predicted as one to COVID-19, 13 to bacterial pneumonia, and one to normal class. Further, all 278 images were correctly classified for the bacterial pneumonia class. Next, out of 70 images for the tuberculosis class,

68 were correct, and two were misclassified, with one to COVID-19 and the other to normal class. Finally, out of 449 images, 442 were correctly predicted for the normal class, and seven were misclassified with two to COVID-19, four to viral pneumonia, and one to tuberculosis class.

4.3.4. Heatmap visualization: An explainable AI model

Lesions have different characteristics such as texture, contrast, intensity variation, density changes, etc. [176]. Figure 4.17 presents the pipeline for lesion validation using heatmaps, where the input to the segmentation model is the X-ray scans that produce the segmented lungs. This segmented lung goes to the Xception-based classification model for five classes, i.e., COVID-19, viral pneumonia, bacterial pneumonia, tuberculosis, and control. The Gradient-weighted Class Activation Mapping (Grad-CAM) algorithm is applied to produce the lesion heatmap. Grad-CAM builds the coarse localization map using the gradients of the target (COVID-19 in the Xception-based classification model), thereby showing the critical regions in the form of heatmap scans. It uses the final convolution layer to produce the heatmap [177].

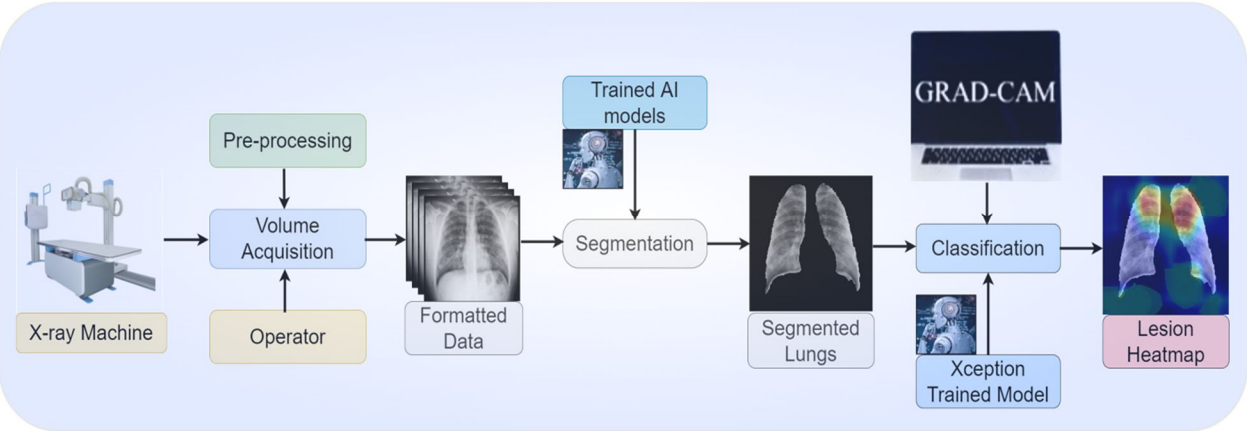


Figure 4.17: Heatmap generation using Grad-CAM and the Xception-based classifier.

Heatmaps provide information about from which part of the image the network is learning or distinguishing the images into actual classes. The coronavirus infection starts in the nose or mouth and then infects the throat, trachea, and thereafter, the lungs. That is why the upper part of the lungs is majorly

infected in most COVID-19 cases, especially during the initial infection stage. Figure 4.18 shows the sample images of the COVID-19 class that the Xception model correctly classified. The heatmap pattern of the correctly predicted COVID-19 images reveals that the network distinguishes the images and takes decisions from almost similar parts of the lungs. The model differentiates the images based mostly on the upper parts of the lungs that are majorly infected or have lesions after the coronavirus infection.

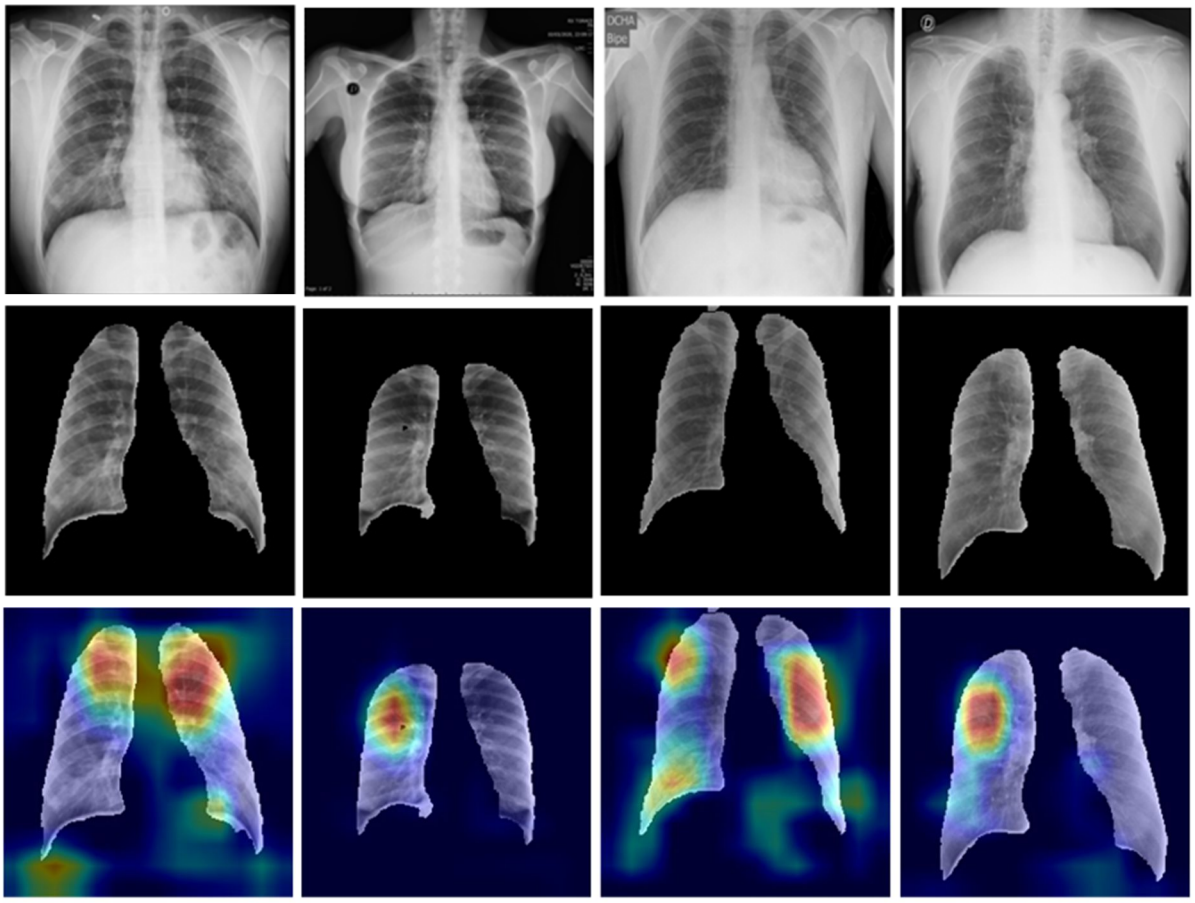


Figure 4.18: Example of miss-classified COVID-19 chest X-ray images; top row: original COVID-19 infected chest X-ray images, middle row: segmented masks, bottom row: corresponding heat map images.

Out of 362 COVID-19 images, nine (~2%) were misclassified. However, this threshold is lower than the regulatory requirement of 5% as per the 510 (K) FDA requirements. Figure 4.19 represents some erroneously predicted COVID-19 images and their heatmaps. Sometimes, the low contrast or noise present in the images may also be the reason for misclassification.

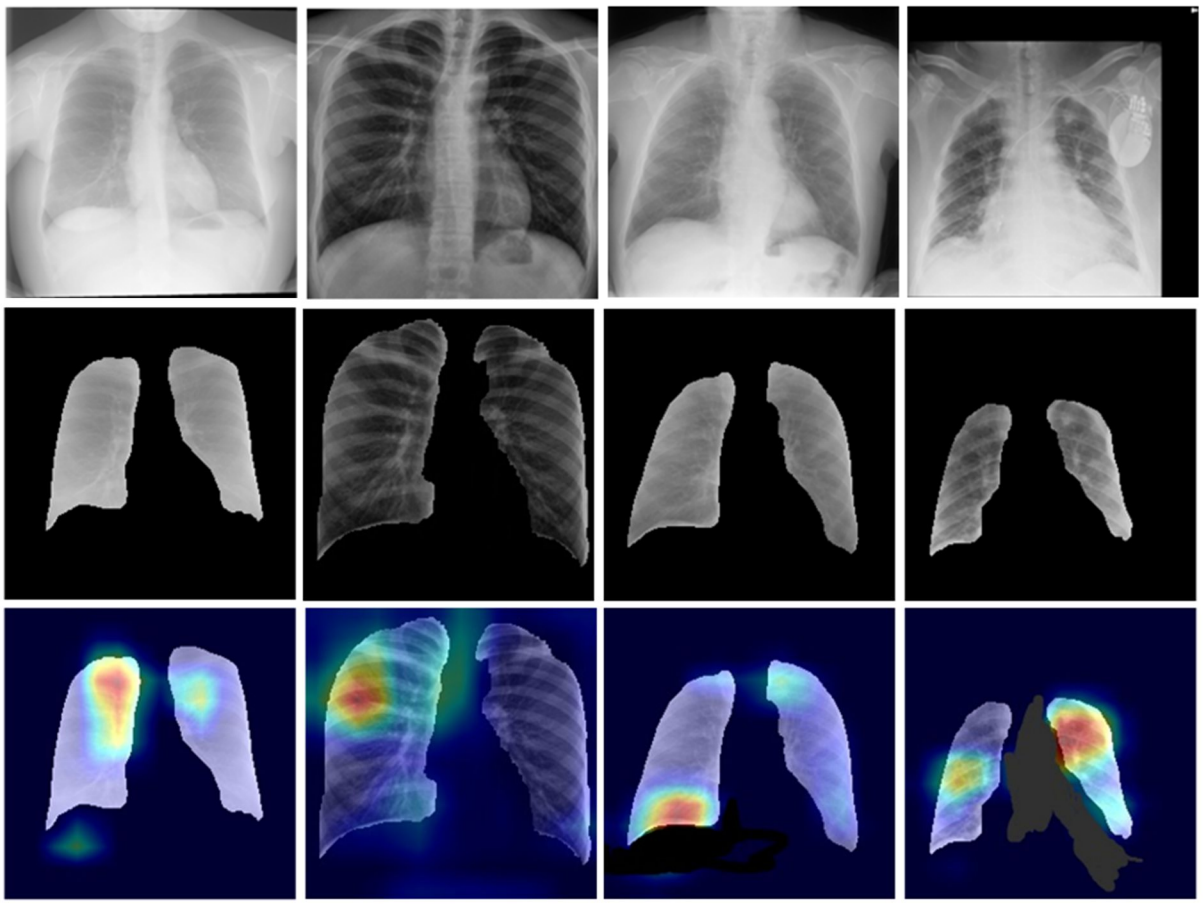
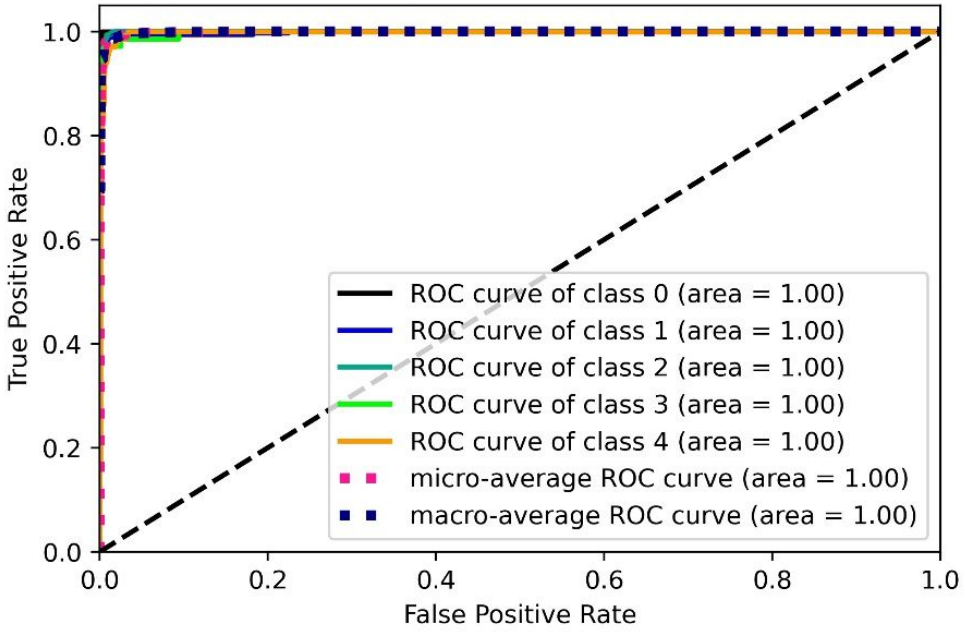


Figure 4.19: Example of miss-classified COVID-19 chest X-ray images; top row: original COVID-19 infected chest X-ray images, middle row: segmented masks, bottom row: corresponding heat map images.

4.4 Performance evaluation

We are able to design a segmentation-based classification model for COVID-19 detection. Our two-stage system has shown excellent performance with precise accuracy in detecting the lesions present in X-ray scans. However, to prove the robustness of the model against all odds, some performance evaluation is always required. Consequently, we obtained the ROC and AUC for the best-performing UNet (segmentation) with the Xception (classification) model. ROC curves are drawn using inference values and true labels for each class. The ROC and AUC for the UNet model have already been discussed earlier in Chapter 3. Figure 4.19 shows the ROC and AUC for the Xception model.



($p < 0.0001$; class 0: BP; class 1: COVID-19; class2: normal; class 3: TB; class 4: VP)

Figure 4.20: ROC and AUC for the Xception model.

4.5. Discussion

4.5.1 Principal findings

We have developed a two-stage COVID-19 detection system based on the segmentation of CXR images in the first stage and then the classification of the segmented lung in the second stage. We applied and tested two segmentation models, UNet and UNet+, in the first stage, having test accuracy, test loss, dice, Jaccard, area error, and AUC of 96.35%, 0.15%, 94.88%, 90.38%, 1.48 mm^2 and 0.99 respectively for the best performing UNet model. Next, we applied and tested eight deep neural networks for the classification: VGG16, VGG19, Xception, InceptionV3, Densenet201, NasnetMobile, Resnet50, and MobileNet of the segmented lungs. The Xception model performed best with accuracy, precision, recall, F1-score, and AUC of 97.45%, 97.46%, 97.45%, 97.43%, and 0.998, respectively. Thus, the combination of UNet and Xception is the best-suited model for our system.

4.5.2 Classification (solo)

Table 4.6 compares our classification model to the existing non-segmentation-based classification methods. Nayak et al. [54] applied the ResNet-34 network for the classification of chest X-ray images into COVID-19 and normal classes. They used 203 COVID-19 and 203 normal images taken from GitHub. They got an accuracy of 98.33% with an AUC of 0.98. Choudhury et al. [55] utilized the Kaggle dataset for the classification into three classes: COVID-19, VP, and normal by the CheXNet network. They achieved an accuracy of 97.74%. Jain et al. [56] used 490 COVID-19 and 5942 other images for classifying into three classes by the Xception model and achieved an accuracy of 97.97%. Nikolaou et al. [61] used 3616 COVID-19 images for the two and three-class classification of images. They applied the EfficientNetB0 network and achieved an accuracy of 95% for two-class and 93% for three-class classification. Yang et al. [62] applied the VGG16 network to classify into two and three classes. They utilized 3616 COVID-19 and 4845 other images and achieved an accuracy of 98% for two and 97% for three-class classification. Khan et al. [57] applied a novel Coronet model for the classification into three classes and achieved an accuracy of 95%. Hussain et al. [58] used the COVID-R dataset having 500 COVID-19 images, applied a novel CoroDet network for the classification into two, three, and four classes, and achieved an accuracy of 99.1%, 94.2%, and 91.2%, respectively, for each class-type. Timmy et al. [63] applied the ResNet-50 and Ensemble Subspace Discriminant method for the classification into five classes and achieved an accuracy of 91.6%. Khan et al. [68] applied the EfficientNetB network for the classification into four classes and achieved an accuracy of 96.13%. Our previous work [39] used 3611 COVID-19 and 13833 other images to classify them into two, three, and five classes. We applied VGG16, NasNetMobile, and DenseNet201 models and achieved an accuracy of 99.84%, 96.63%, and 92.70%, with an AUC of 1.0, 0.97, and 0.92 for two, three, and five-class classifications, respectively.

In the proposed work, we utilized 3611 COVID-19 and 9849 other class images from the Kaggle dataset. We applied the Xception model for the classification after the segmentation by the UNet model. The system performed with accuracy and an AUC of 97.45% and 0.998, respectively, for the five-class classification. We achieved the highest accuracy and AUC among all previous works for the five-class

classification. In addition, we improved the accuracy by 4.75% compared to our previous work. The proposed work has several improvements to our previous work. We have employed segmentation of chest X-ray images before the classification. Further, we have applied the explainable AI-based method and heatmap visualization of the image to detect and manifest the lesion present in the X-ray scans. Additionally, we have applied one new classifier: MobileNet, i.e., the fastest among all participating networks. Finally, as a result, we improved the accuracy by 4.75% from our previous work.

Table 4.6: Benchmarking table showing comparison of proposed and existing classification (solo) models.

Author & Year	Dataset - chest X-ray (COVID-19 images + other images)	Technique	Accuracy	AUC
Nayak et al. (2020) [54]	GitHub (203+203)	ResNet-34	2 class-98.33%	2 class-0.98
Chowdhury et al. (2020) [55]	Covid-19 Radiography Database (Kaggle) (423+3064)	CheXNet	3 class-97.74%	NA
Jain et al. (2020) [56]	Kaggle (490+5942)	Xception	3 class-97.97%	NA
Nikolaou et al. (2021) [61]	Covid-19 Radiography Database (Kaggle) (3616+11537)	EfficientNetB0	2 class-95% 3 class-93%	NA
Yang et al. (2021) [62]	Covid-19 Radiography Database (Kaggle) (3616+4845)	VGG16	2 class-98% 3 class-97%	NA
Khan et al. (2020) [57]	GitHub (284+967)	Coronet (novel CNN)	3 class-95%	NA
Hussain et al. (2020) [58]	COVID-R dataset (500+1600)	CoroDet (novel CNN)	2 class-99.1% 3 class-94.2% 4class-91.2%	NA
Timemy et al. (2021) [63]	GitHub (435+1751)	ResNet-50 + ESD(Ensemble Subspace Discriminant)	5 class- 91.6%	NA
Khan et al. (2022) [68]	Covid-19 Radiography Database (Kaggle) (3616+17449)	EfficientNetB	4 class-96.13%	N.A.
Nillmani et al. (2022) [39]	Covid-19 Radiography Database (Kaggle) (3611+13833)	VGG16, NasnetMobile, DenseNet201	2 class-99.84% 3 class-96.63 5 class-92.70	2 class-1.0 3 class- 0.97 5 class-0.92
Proposed work Nillmani et al. [155]	Covid-19 Radiography Database (Kaggle) (3611+9849)	Xception	5 class-97.45%	0.998

4.5.3 Segmentation-based classification

Table 4.7 shows the comparison of our system to the existing segmentation-based classification methods. Alom et al. [52] utilized the Kaggle dataset, having 390 COVID-19 images and 234 normal images. They applied a novel NABLA-N network for the segmentation with an accuracy, dice, and Jaccard of 94.66%, 88.46%, and 86.50%, respectively. Afterward, they applied the Inception Recurrent Residual Neural Network model for the classification of segmented lung images into two classes. They achieved a classification accuracy of 87.26% and an AUC of 0.93. Wehbe et al. [59] utilized a private dataset having 4253 COVID-19 images and 14778 normal images. They applied an ensemble network for the classification of CXR images after the segmentation. They achieved an accuracy of 83% and an AUC of 0.9 for the two-class classification. Oh et al. utilized 180 COVID-19 and 322 other images taken from Kaggle and GitHub. They applied the DenseNet103 network for the segmentation and achieved the Jaccard of 95.5%. After the segmentation, they applied the ResNet-18 model to classify the segmented lung images into four classes and achieved an accuracy of 88.9%. Teixeira et al. [50] utilized the RYDLS-20-V2 dataset, having 503 COVID-19 and 2175 images from other classes. They applied the UNet model for the segmentation with a dice coefficient of 98.2%. Following segmentation, they applied Inception V3 for classification into three classes and achieved an accuracy of 88% and an AUC of 0.9. Keidar et al. [60] applied the ensemble method for the classification of segmented lung images into two classes. Their model performed with an accuracy of 90.3% and an AUC of 0.96. Fang et al. [67] applied a novel CLseg model for segmentation and achieved a dice of 94.09%. After the segmentation, they applied a novel SC2Net model for the two-class classification of the COVIDGR 1.0 dataset and achieved an accuracy of 84.23% and an AUC of 0.94. Abdulah et al. [64] applied the Res-CR-Net model for the segmentation with dice and Jaccard of 98% each. Thereafter they classified a private dataset into two classes using an ensemble method and achieved an accuracy of 79% and an AUC of 0.85. Bhattacharyya et al. [65] used a GAN segmentation network with a VGG-19 and Random Forest classifier and achieved 96.6% accuracy for the three-class classification. Hertel et al. [69] utilized 4013 COVID-19 with 12837 other class images. They applied a ResUnet segmentation network with a dice of 95%. Following segmentation, they applied an ensemble network to classify into two and three classes. They achieved an accuracy of 91% for the two-class and 84% for the

three-class with an AUC of 0.95. Aslan et al. [70] applied an ANN-based segmentation method on the COVID-19 Radiography database (Kaggle) and a combination of DenseNet201 and SVM for the classification into three classes. They achieved an accuracy of 96.29% with an AUC of 0.99. Xu et al. [66] utilized 433 COVID-19 and 6359 other images. They applied ResUNet for the segmentation with a Jaccard of 92.50%. After that, they applied ResNet50 to classify segmented lung images into five classes. They achieved an accuracy of 96.32%.

Table 4.7: Benchmarking table showing a comparison of proposed and existing segmentation-based classification models.

Author & Year	Segmentation	Dataset - chest X-ray (COVID-19 images + other images)	Technique	Accuracy	AUC
Alom et al. 2020) [52]	NABLA-N network Accuracy - 94.66 Dice - 88.46 Jaccard - 86.50	Kaggle (390+234)	Inception Recurrent Residual Neural Network (IRRCNN) model	87.26%	0.93
Wehbe et al. (2021) [59]	N.A.	Private (4253+14778)	Ensemble CNN	2 class-83%	0.9
Oh et al. (2020) [53]	DenseNet103 Jaccard+95.5%	Kaggle+GitHub (180+322)	ResNet-18	4 class-88.9%	NA
Teixeira et al. (2021) [50]	UNet Dice+98.2%	RYDLS-20-v2 (503+2175)	Inception V3	3 class-88% (F1 score)	0.9
Keidar et al. (2021) [60]	N.A.	Private (1289+2427)	Ensemble model	2 class-90.3%	0.96
Fang et al. (2022) [67]	CLSeg Dice - 94.09	COVIDGR 1.0 dataset (426+426)	SC2Net (novel CNN)	84.23%	0.94
Abdulah et al. (2021) [64]	Res-CR-Net Dice - 98 Jaccard - 98	Private (1435+3797)	Ensemble CNN	2 class-79%	0.85
Bhattacharyya et al. (2021) [65]	GAN network Accuracy - N.A.	GitHub (342+687)	VGG-19 + Random Forest	3 class-96.6%	NA
Hertel et al. (2022) [69]	ResUnet Dice - 95	COVIDx5+MIDRC-RICORD-1C+BIMCV dataset (4013+12837)	Ensemble model	2 class-91% 3 class-84%	0.95
Aslan et al. (2022) [70]	ANN based segmentation Accuracy - N.A.	Covid-19 Radiography Database (Kaggle) (219+2905)	DensenNet201+SVM	3 class-96.29%	0.99
Xu et al. (2021) [66]	ResUNet Jaccard - 92.50	GitHub (433+6359)	ResNet50	5 class-96.32%	N.A.
Proposed work [155]	UNet Accuracy - 96.35 Dice - 94.88 Jaccard - 90.38	Covid-19 Radiography Database (Kaggle) (3611+9849)	Xception	5 class-97.45%	0.998

In our proposed work, we utilized 3611 COVID-19 and 9849 other images taken from Kaggle. We applied the UNet segmentation model and achieved an accuracy, dice, and Jaccard of 96.35%, 94.88%, and 90.35%, respectively. Thereafter, we applied the Xception model for the classification of the segmented lung into five classes. We achieved an accuracy of 97.45% and an AUC of 0.998. We achieved the highest accuracy for the segmentation-based classification among all the existing state-of-the-art methods. In addition, we achieved the highest AUC among all existing models. This makes our system the most precise. Additionally, we have used a large number of images that make our system more stable and robust.

4.5.4 A special note on segmentation-based multiclass classification system for COVID-19 detection

To date, most of the COVID-19 detection systems are based on the classification of CXR images without segmentation. However, they have shown good accuracy, but due to the unwanted region present in the chest X-ray scans, there is the likelihood of biased or inaccurate results. Segmenting the X-ray images removes the unwanted region and background noise present in the X-ray, leaving only the required lung area. Few researchers have worked on the segmentation-based classification model. However, multiclass classification has not been tried, and further, it is not robust in terms of accuracy. Next, note that in previous studies, the number of images used in the experiment for segmentation and even for classification is relatively very low. This may reduce the reliability and robustness of the system. In the proposed work, we tried to fill the gaps by developing a system-based best-suited segmentation-based classification model, keeping regulations in mind. We have used a large number of images for both the segmentation and classification experiments. Additionally, we implemented a classification method that could classify multiple types of pneumonia, including the most common lung infections that generally show similar symptoms and findings in X-rays. If screened using the naked eye by radiologists or doctors, they are very likely to misclassify the different pneumonia types. Even performing multiclass (five-class) classification, our system performed with the highest accuracy compared to any available segmentation+classification model for any class, including two classes. With segmentation, multiclass classification, involvement of a high number of images, and preciseness, our system proves its reliability, robustness, and superiority over other available approaches for medical applications in COVID-19 detection.

4.5.5 Strength, Weakness, and Extension

Our AI-powered system is capable of rapid detection of COVID-19. It takes less than one second to generate the results. Along with the fast detection, our system is more precise than any other available method. The system provides an accuracy of up to 97.45%, which is the maximum among any binary or multiclass segmentation-based classification methods. Additionally, the system design is highly cost-effective compared to current diagnostic methods. Our system requires chest X-ray images that are readily available at a meager cost. The system predicts the disease after segmenting the lung, thus highly accurately meeting regulatory requirements and our hypothesis [178]. Therefore, there is less chance of a wrong prediction as most unwanted areas and noises are removed from the X-ray images. Our system can show the infected or lesioned area in the lung by heatmap visualization that may help the radiologists or doctors and, ultimately, the patients achieve successful treatment. Since our design is AI-based, our system can learn automatically from its own mistakes or by exposure to new images. This constantly continues to enhance the performance of the system. Further, because our system can be easily updated with new sets of images at regular intervals, it can improve overall performance, especially in diverse data types. For the COVID-19 diagnosis, the setup of our system can be easily created in hospitals or other clinical centers, as it requires just a conventional computer setup and X-ray data sets. Such a system can be adopted for even long-COVID analysis [179]. Such CAD and imaging design can even be extended to the multimodality paradigm. In addition, even a low-skilled person may handle the screening setup without complex training. Our system does not require any sample handling or transportation as in screening using RT-PCR. On the contrary, in our setup, only X-ray images are required that can be transferred in seconds through the internet or other options to any place in the world.

Note that every pilot system design has some kind of challenges. We have noticed that if the resolution of the X-ray images (very low contrast) is beyond the radiologist's ability to discern pneumonia type, it can affect the AI models. However, this concern can be resolved by denoising and color normalization techniques [180, 181]. Furthermore, human error by the X-ray technician may impact the quality of the X-ray image, and ultimately, our AI model's result might be affected. Sometimes, the variation in X-ray machines and their output quality may affect the results of our system. However, this

may be overcome by training on larger data sizes and diverse types of images or by superior de-noising methods [182]. Further, retraining large databases having diverse images, our system would require a high-performing GPU [99] or supercomputer framework that may incorporate higher costs. One significant issue with the AI-based detection system, including ours, is the institutional approval for medical use. Even after many routine developments, AI-based COVID-19 detection always needs approval as the primary diagnostic method. However, the system may frequently be used as the second opinion choice.

In the extension of the work, we will train our system on more diverse and recent datasets or in a big data framework. More data sets can be collected from different machines and test the performance on the more varied datasets. Superior approaches for training, such as pruning and stochastic imaging to improve the system's performance and lower the storage [177]. Additionally, we shall use a more advanced GPU and workstations to enhance the output and lower the learning time. Newer methods, such as Tree Seed Algorithm (TSA)-optimized Artificial Neural Networks (ANN), can be tried to classify deep architectural features [183]. In another approach, the Bidirectional Long Short-Term Memories (BiLSTM) layer can be used as a hybrid pipeline that combines AlexNet with BiLSTM [184].

4.6 Summary

COVID-19 has emerged as one of the predominant challenges to saving human lives in the current circumstances. Several research groups, including medical communities, are trying to find the proper solutions to combat the disease. However, the advancement in artificial intelligence and medical imaging has made hope in lesion detection in medical images. The methods have proved their efficiency in several areas, such as tumor detection, carotid plaque detection, and much more. Numerous research groups are working on AI-based COVID-19 diagnosis systems. However, some gap was still present. In this work, we attempted to fill all the gaps and presented a better two-stage COVID-19 diagnosis system that can fulfill the regulatory requirement of <5% as per the 510 (K) FDA as a prerequisite for clinical settings. We have proposed a segmentation-based multiclass classification system to detect COVID-19 and the other three most common pneumonia, namely viral pneumonia, bacterial pneumonia, and tuberculosis, in chest X-ray scans. We applied two segmentation models, UNet and UNet+, with eight classification networks, namely

VGG16, VGG19, Xception, InceptionV3, Densenet201, NASNetMobile, Resnet50, and MobileNet. Finally, we selected the best-performing model combination, UNet for segmentation and Xception for classification. We achieved a classification accuracy of 97.45% with an AUC of 0.998 by the system. Our model outperformed all the existing state-of-the-art methods in segmentation-based classification models. Our system performed best by the mean improvement of 8.27% over all the remaining studies. Additionally, our system is completely automated and robust, yielding the highest sensitivity and specificity. The system's error rate is just ~2%, which qualifies within the regulatory bounds of less than 5%, a prerequisite for clinical settings. Further, we used heatmaps under the explainable AI paradigm for scientific validation. As our system is more precise, affordable, and accessible than the current diagnostic approaches for COVID-19 and qualifies for the regulatory requirement of the FDA, the suggested model may provide an alternative or add to the current diagnostics methods. The system may helpfully aid in rapid and accurate patient diagnosis, reducing the medical workforce and contributing to the wellness of humanity.