Chapter 5

Ensemble Deep Learning Models for Pneumonia Detection in Five Multiclass Frameworks of Chest X-ray Images

Abstract

Pneumonia is an inflammation of the lungs that is considered one of the most common and potentially devastating medical complications after infection, resulting in hospitalization or death. In recent years, after the COVID-19 invasion, pneumonia-caused deaths have drastically increased. The first step towards treating and managing the disease is correctly diagnosing the infection. RT-PCR is a commonly used method for detecting and quantifying the expression of viral genes in diagnostic samples. However, the technique is laborious, time-consuming, and costly, with a significantly high error rate. Additionally, the scarcity of test kits has become a major challenge during the COVID-19 pandemic. In recent years, implementing artificial intelligence and deep learning techniques for medical imaging has revolutionized the healthcare sector by automatically and precisely detecting lesions in radiological images that may be unnoted by a radiologist. In the proposed work, we have developed a deep learning-based ensemble method to classify and detect four types of pneumonia, including COVID-19, viral pneumonia, bacterial pneumonia, and tuberculosis, in chest X-ray scans. First, we applied the UNet++ network to segment the lung region from chest X-ray scans. Thereafter, we applied four classification networks, including a hybrid model: Xception, MobileNet, NasnetMobile, and InceptionResNetV2. Next, we chose the best-performing networks and ensembled them to build three combinations of ensemble models: Ensemble model-1, Ensemble model-2, and Ensemble model-3. For the first time, we employed an ensemble model to classify the chest X-ray images into five categories (four pneumonia types and a control class). The UNet++ model

performed with accuracy, Jaccard, dice, and AUC of 97.48%, 91.64%, 95.37%, and 0.99, respectively. The Ensemble model-2, comprising the best three networks, Xception, MobileNet, and InceptionResNetV2, outperformed all classification models with an accuracy of 99.61% and precision, recall, F1-score, and AUC of 99.62%, 99.61%, 99.61% and 0.9999 respectively. Our system performed best among any class's existing state-of-the-art chest X-ray image classification methods. The study also establishes a theory for the correlation between the quantity and performance of classifiers and ensemble model efficacy. Further, we investigated the visualization capabilities of the networks using the Gradient-weighted Class Activation Mapping to view the abnormalities in chest X-rays and discuss them from radiological perspectives. These findings, including the high accuracy, verify our system’s efficacy and superiority for clinician precision in COVID-19 and pneumonia diagnosis using chest X-rays. Furthermore, the system algorithm's high accuracy raises the possibility of a valuable setup, particularly in environments with a scarcity of biochemical test facilities and radiological expertise.

5.1 Introduction

Pneumonia is an inflammation of the lungs caused by bacterial, viral, or fungal infections [185]. The main symptom of pneumonia is a cough with thick phlegm and dyspnea (shortness of breath), often accompanied by fever, chills, fatigue, and chest pain [186]. It is considered one of the most common and potentially devastating medical complications after infection, resulting in hospitalization or death [187-189]. Recently, the invasion of SARS-COV2 (Severe acute respiratory syndrome coronavirus 2) that causes COVID-19 and severe pneumonia has become one of the major causes of death [79, 190]. Seven hundred sixty-two million of the world's population got infected, and 6.89 million deaths have been recorded due to SARS-COV2 [157]. The first step towards treating and managing the disease is correctly diagnosing the infection. RT-PCR is a commonly used method for detecting and quantifying the expression of viral genes in diagnostic samples [191, 192]. RT-PCR can detect viral genomes, viral RNA, DNA of infectious agents (including bacteria and viruses), tumor markers, and other molecular targets [193]. For COVID-19, the RT-PCR is the most commonly used diagnostic system [191]. However, the technique is laborious, time-consuming, and costly, with a significantly high error rate [194, 195]. Additionally, the scarcity of test kits has become a major challenge during the COVID-19 pandemic [196]. The infected person who left erroneously un-diagnosed can't be acknowledged and isolated, leading to further infecting the mass population. Thus, a fast, accurate, and cost-effective diagnosis must be required to provide appropriate medical attention to the patients.

Noninvasive medical imaging techniques such as CT, MRI, Ultrasound, and X-rays use high-frequency rays to produce images of internal body structures [197-199]. Images are derived from the interaction with radiation, which is fed through a device like a scanner or a film and can provide alternative molecular and biochemical techniques for diagnosing several diseases, including COVID-19 [200]. These techniques have already proven their efficacy in detecting abnormalities in internal body parts [201]. Traditionally, chest X-ray has been used as a powerful imaging method to detect tuberculosis and other pneumonia and their progression into the lung [202]. Recently, CT and X-rays have been principally used

to detect the lesions and progression of COVID-19 in the lungs [203, 204]. The advantage of X-ray over CT is that it is cost-effective, readily available, and has low radiation doses [204, 205].

Implementing AI-CAD (artificial intelligence-based computer-aided diagnosis) into imaging techniques has revolutionized healthcare [14]. AI-driven machine learning and deep learning techniques are capable of automatically and precisely detecting lesions in radiological images that may be unnoted by a radiologist [13]. The AI-CAD applications in radiology have proven their efficacy in diagnosing brain tumors [97, 206], breast cancer [207], prostate cancer [208], plaques [137], cardiovascular diseases [25, 102, 209], Alzheimer's disease [26], Parkinson's disease [142], tuberculosis [118], pneumonia [39, 155], and more. Implementing deep learning methods in chest X-rays can provide significant input for detecting and precisely diagnosing pneumonia infection types, including COVID-19 [139-141]. This may help diagnose patients correctly and their personalized disease management and treatment. The method requires very few resources and staffing, and the ability to deliver bulk results in a few minutes may solve the limitations of available methods cost-effectively.

In the proposed work, our objective was to develop a best-performing deep learning-based ensemble model to classify and detect four types of pneumonia, including COVID-19, in chest X-ray scans. We needed to find the best combination of classification DL networks to achieve the goal. Previous works showed that ensemble models were developed in a single combination of two, three, or four DL networks. However, an experiment showing different combinations of DL networks to get the best results and establish a correlation between the number and performance of DL networks and the efficacy of the ensemble model was lacking. In the proposed work, we have attempted to establish a correlation between the number and performance of DL networks and ensemble models. To achieve our objective first, we applied the UNet++ network to segment the lung region and remove the non-region of interest from chest X-ray images. Thereafter, we applied four classification networks, including a hybrid model: Xception, MobileNet, NasnetMobile, and inceptionResNetV2. Next, we chose the best-performing classifiers and fused them to build three different combinations of ensemble models: Ensemble deep learning model-1, the combination of the best two DL networks; Ensemble deep learning model-2, a combination of the best three

DL networks, and Ensemble deep learning model-3, the combination of all four DL networks. Finally, we compared the performances of all ensemble models and proposed a correlation between the quantity and performance of DL networks and ensemble model efficacy. Finally, we selected the best Ensemble model for further diagnostic applications of COVID-19 and other lung inflammations.

5.2 Methodology

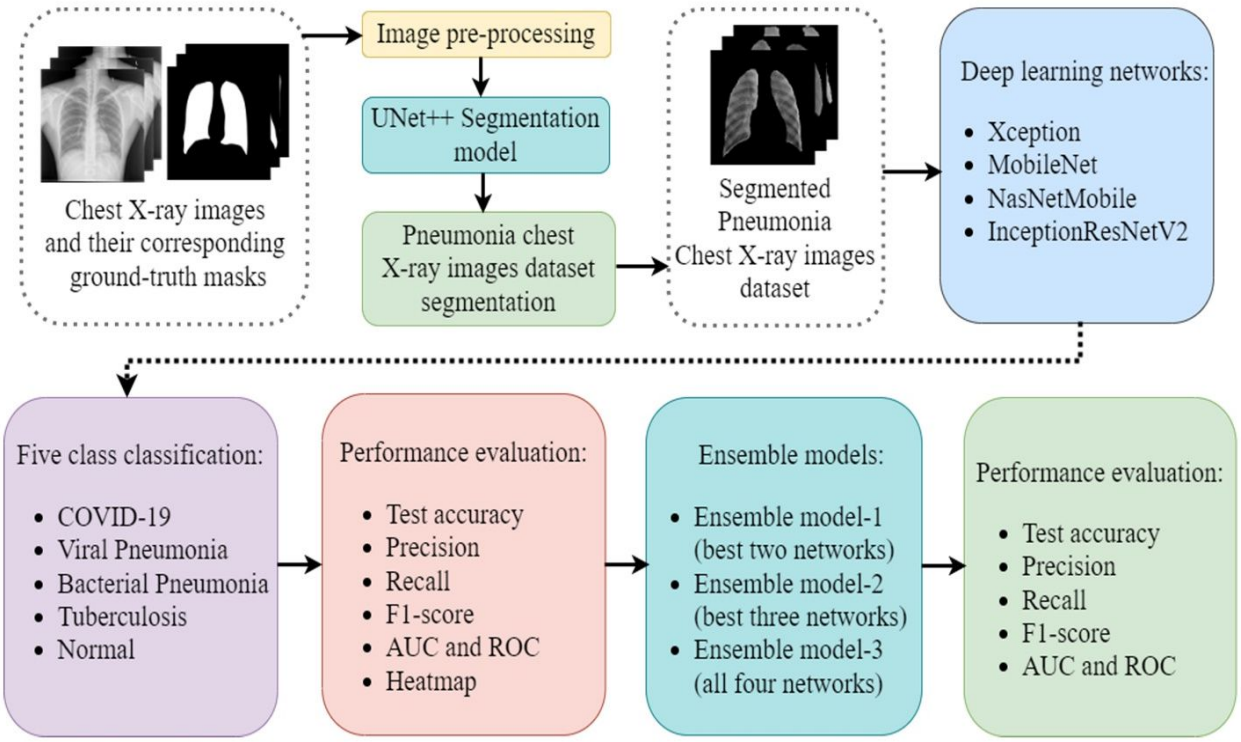


Figure 5.1: The schematic diagram representing the overall methodology.

Figure 5.1 represents the overall approach we opted for in the study. First, we designed the UNet++ segmentation model. For segmentation model training, we utilized the dataset having chest X-ray images and their corresponding masks annotated by the radiologist. The motive behind the segmentation was to automatically, accurately, and quickly segment the chest X-rays and remove the background or non-region of interest. The chest X-rays contain a large background area other than the lung region that may interrupt or misguide the AI system for precise lesion detection. Next, we tested and evaluated the segmentation model using performance metrics such as accuracy, dice, Jaccard, and AUC (area under the curve). After the segmentation model training, we applied the model to the "COVID-19 Radiography Dataset" to segment

the CXRs from four pneumonia classes: COVID-19, BP (bacterial pneumonia), VP (viral pneumonia), TB (tuberculosis), and a control class. Next, we applied four deep learning classification networks, Xception, MobileNet, InceptionResNetV2, and NasNetMobile, on the segmented lung regions to classify the segmented CXR data into five classes. After that, we evaluated the model's performance using different evaluation metrics such as accuracy, precision, recall, F1-score, and AUC. Finally, we applied the ensemble technique to enhance the performance and accuracy. We designed three ensemble models, Ensemble model-1, Ensemble model-2, and Ensemble model-3, that comprise the best-performing two, three, and all four networks, respectively. Finally, we evaluated the ensemble models' performances to finalize the best-performing Ensemble model for Chest X-ray analysis.

5.2.1 Dataset and patient demographic

We used "Chest X-ray masks and label dataset" [151] for the training segmentation model. The dataset has a total of 1408 images, with 704 chest X-ray images and 704 corresponding masks. The dataset comprised two popular chest X-ray datasets, JSRT and Montgomery, and an expert radiologist team annotated the masks. The source of chest X-rays is the National Library of Medicine, NIH, USA, and Shenzhen No. 3 People's Hospital, GMC, Shenzhen, China. The dataset contains 344 infected and 360 healthy chest X-rays.

For pneumonia classification, we utilized a total of 12,941 chest X-ray images taken from three different datasets that are The "COVID-19 Radiography dataset" [114], "Chest X-ray Images (Pneumonia)" [116] and "Tuberculosis (TB) Chest X-ray database" [115]. The "COVID-19 radiography Database" contains 3616 COVID-19, 1345 VP, and 10192 normal images from which we utilized all COVID-19 and VP images and 4500 normal images. The "Chest X-ray Images (Pneumonia) Dataset" contains a total of 5863 X-rays with 2780 BP images. The X-rays were taken from Guangzhou Women and Children's Medical Center, Guangzhou. From the dataset, we utilized all the BP images. The "Tuberculosis (TB) Chest X-ray database" contains 700 Chest X-ray images of tuberculosis patients, of which we used all the TB images. Finally, we utilized 12,941 CXR images, 3616 COVID-19, 1345 VP, 2780 BP, 700 TB, and 4500 normal images. Out of a total, 10% of randomly selected images, i.e., 360 COVID-19, 130 VP, 270 BP, 70

TB, and 450 normal images, were utilized for validation, and other 10% of randomly selected images with identical class-wise distributions were used for testing of the network.

5.2.2 Network Architecture and Optimization

UNet++ DL Model

In medical imaging, UNet is one of the most popular and efficient models for diverse radiological image segmentation. The UNet network has been proven reliable in several previous works and has been opted for by numerous researchers [19]. In the presented work, we employed a modified and more efficient version of UNet, i.e., the UNet++ model, to segment CXR images. Figure 5.2 represents the detailed architecture of the UNet++ network. As the diagram elucidates, the UNet++ network has two phases: the

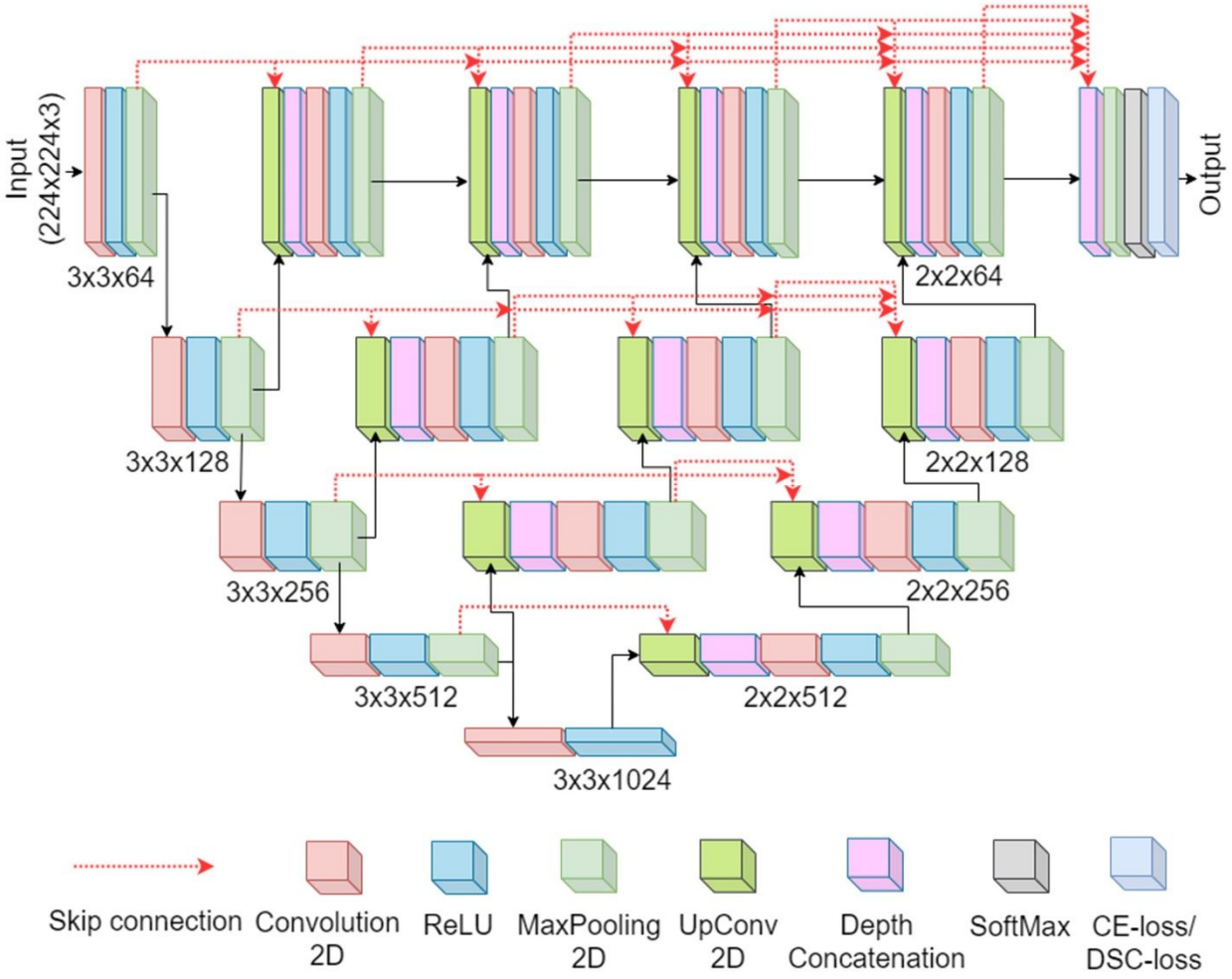


Figure 5.2: UNet++ architecture.

encoder and the decoder. The encoder phase has several convolutional layers followed by ReLU and MaxPooling. The decoder phase contains up-convolution followed by depth concatenation, softmax, and MaxPooling. The image size provided as input to the encoder is 224 × 224. The encoder extracts the features of the images in the form of numerical values between 0 and 255. Later, the features are normalized between 0 and 1. The image size gets reduced while transferring after each MaxPooling layer to the successive convolution. A bridge channel connects the encoder and decoder phases. The encoder phase transfers the image features to the decoder phase. The decoder phase decodes the numerical features into the form of an image and recreates the images, converting the pixel values <0.5 to 0 and ≥0.5 to 1. The final output images from the decoder have the binary pixel values 0 or 1, where 0 represents black, and 1 characterizes white. This results in the generation of final black and white masks. Later, these masks are used to generate the region of interest, i.e., the lung section from the CXRs. The UNet+ and UNet++ networks differ from UNet as they consist of several dense skip networks (DSN) and are connected by skip connections. The DSN comprises Upconvolution followed by concatenation, convolution, ReLU, and Maxpooling. The DSNs are employed to directly transfer image features from an encoder to a decoder; this prevents the network from losing the image information in the form of loss from neural nodes. UNet+ also has the same DSN number as UNet++ at each encoder-decoder stage. However, in UNet+, each DSN is connected to only its preceding dense network or encoder output, whereas in UNet++, each DSN is connected to all the previous dense network or encoder's output. Connection through all preceding outputs at an encoder-decoder stage enables the UNet++ to be more efficient than UNet or UNet+ for image recreation at decoder stages with reduced loss.

For the segmentation problem, we employed a five-fold cross-validation protocol. The UNet model was trained for 100 epochs in each fold. While training the model, the learning rate and dropout were applied for 0.001 and 0.25, respectively. The batch size for training and validation was maintained at four images per batch. The loss function implemented was cross-entropy, symbolized as Lce, and derived as the equation (5.1).

$$L_{CE} = [(y_i \times loga_i) + (1 - y_i) \times log(1 - a_i)] \qquad (5.1)$$

Here y_i is the GT label 1, $(1\text{-}y_i)$ is GT label 0, and a_i is the Softmax classifier probability.

Xception DL Network

Xception is an 81-layer depth model. 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 2.6 represents the detailed architecture of the Xception neural network. Xception has depth-wise separable convolutional layers. Depth-wise separable convolution has two parts: the first is a depth-wise convolution, and the second is a pointwise convolution. Depth-wise convolution performs the filtering stage, and pointwise convolution performs the combining stage. Depth-wise convolution employs convolution to a single input channel at once; however, a standard convolution applies convolution to all the channels. The Xception model is built in three phases: the entry flow, middle flow, and exit flow. The CXR images with a resolution size of 224 × 224 are given as input from the entry flow. Each flow has several neural nodes containing the convolution layers followed by ReLU and Pooling layers. The nodes are also connected with a few convolutions directly, which simultaneously transfers the features to the next neural nodes without any losses. The output from the entry flow is used as input for the middle flow, and the output from the middle flow is used as input for the exit flow. The concluding classifier layer of the exit flow is logistic regression, which classifies the images based on the features extracted by the neural nodes and predicts the class for each image as output [210].

InceptionResNetV2 DL Network

InceptionResNetV2 is constructed using both the Residual connection and the Inception structure. Several convolutional filters of various sizes are mixed with residual connections in the Inception-Resnet block. In addition to avoiding the degradation issue brought on by deep structures, the inclusion of residual connections shortens training time. The InceptionResnetV2 network architecture is depicted in Figure 5.3. The InceptionResNetV2 combines the Inception with residual connections. In the network, a filter expansion layer (1×1 convolution without activation) is employed after each Inception block to increase the dimensionality of the filter bank before the addition so that it matches the depth of the input. Batch normalization is only applied on top of the standard layers in the case of InceptionResNet and not applied on top of the summations. The training is stabilized by scaling down the residuals before adding them to the prior layer activation [211].

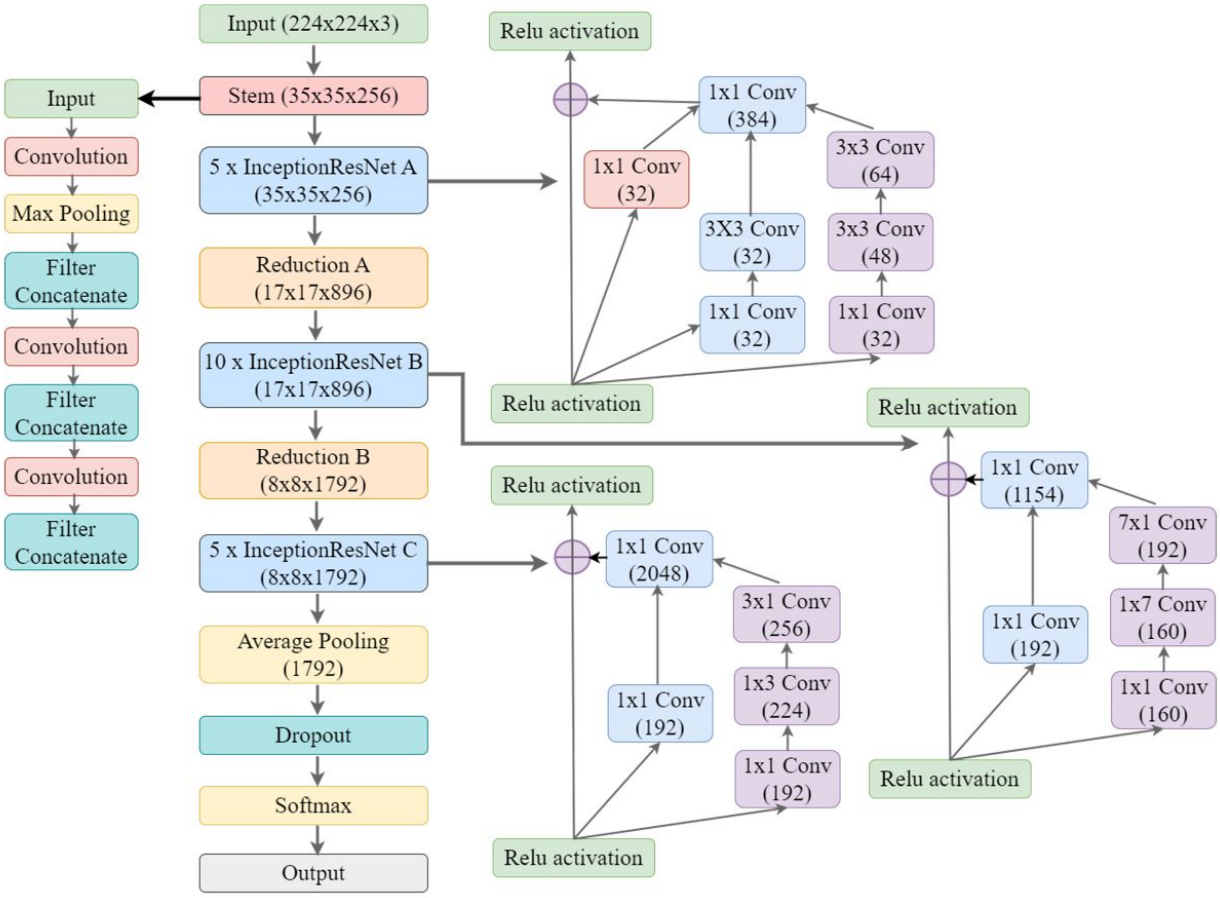


Figure 5.3: InceptionResNetV2 architecture.

MobileNet DL Network

MobileNet is a simplified architecture that builds shallow, deep convolutional neural networks using depthwise separable convolutions. It is an effective model specially designed for mobile and embedded vision applications. As shown in Figure 4.3, the foundation of MobileNet's structure is a set of depthwise separable filters. Point convolution filters and depthwise separable convolution filters make up depthwise separable convolution filters. The depthwise convolution filter executes one convolution on each input channel. The point convolution filter linearly combines the result of the depthwise convolution with 1×1 convolutions. MobileNet is the smallest network among all four networks. It is 16 MB, 55 layers deep, and has 4.3 million parameters. At a performance of 3.4 ms per inference step, it is the fastest.

NASNetMobile DL Network

NASNetMobile is the smallest network after MobileNet in all of our four networks. Figure 2.8 represents the complete architecture of the network. NASNetMobile is composed of normal cells and reduction cells. Normal cells are those convolutional layers that return a feature map of the same dimension. In contrast, the reduction cell is a convolutional layer that returns a feature map reduced by a factor of two. With these two starting hidden states, the controller RNN iteratively predicts the remaining structure of the convolutional cell. The NASNetMobile 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.

For the classification problem, 100 epochs were employed to train all the networks. The learning rate of 0.001 was maintained. The dropout rate applied was 0.25. The training and validation steps proceeded with a batch size of eight images per batch. The best-performing epoch with the highest validation accuracy was saved as the final model during the training. The loss function implemented was categorical cross-entropy symbolized as Lcce and could be derived as the equation (5.2).

$$L_{CCE} = \frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} 1_{y_i \in C_c} \log a_{model} \, (y_i \in C_c) \qquad (5.2)$$

Here, N is the total number of images, C is the number of categories or classes, $1_{y_i \in C_c}$ Denotes the ith observation that belongs to the c^{th} category.

The experiments were conducted using Python 3.9 on a workstation with an Intel Core i7 8th generation processor, 16 GB of RAM, and an 8 GB NVIDIA Quadro P4000 GPU.

Ensemble model

The ensemble method uses insights from various learning models to provide more precise and effective decision-making. The ensemble learning model aids in incorporating all of its constituent models' discriminative information. Thus, the predictions by them are superior over any of its constituent models when applied separately. Medical imaging utilizes two most common ensembling methods: averaging and voting. The averaging or weighted average is the method where the weights from multiple models are combined to predict a sample's class. However, the weighted averaging takes more time and space and is

hence more expensive. Additionally, combining the weights of different models before the prediction may impact the weights of one over another during decision-making. This may sometimes negatively impact the preciseness of results. In comparison, voting is the method where the predictions by different models for a particular sample are combined to make a final decision based on maximum votes or the highest prediction probabilities. The advantage of voting over averaging is it requires less time and space, as there is no need to combine the weights. The second advantage is that the predictions made by each model for a particular sample are completely independent, thus having no impact on the weights of each other.

$$P_{ice} = (P_{icn1} + P_{Icn2} + P_{icn3\ldots\ldots} + P_{icnx}) \div N \qquad (5.3)$$

Here, $'P'$ belongs to Prediction probability, P_{ice} represents Prediction probability by the ensemble for image *'i'* and class *'c'*. P_{icn} is the Prediction probability for image *'i'*, class *'c'* and by CNN network *'n'* where network *'n'* may be from *1* to *x* (*1* is first and *x* is last network). $'N'$ represents the total number of CNN networks used for ensembling.

In the proposed ensembling method, as represented in equation 5.3, we combined each model's prediction probabilities for a particular sample (image). Then, the highest probability for a specific class was assumed as the final result. To get the best result, we ensembled the models in all possible three ways depending upon the number and best-performed classification networks. For the first ensemble model, i.e., Ensemble model-1, we combined prediction probabilities by the best two classification networks. For the second ensemble model, i.e., Ensemble model-2, we combined prediction probabilities by the best three classification networks. Finally, we combined prediction probabilities by all four networks for the third ensemble model, i.e., Ensemble model-3. After that, we tested the performances of each ensemble model on the test dataset and selected the best-performing model for further applications.

5.2.3 Performance evaluation matrices

Performance assessment is a crucial part of analyzing the competence of the DL-CAD system. Different metrics are used for the assessment of the network's performance. For the presented approach, accuracy, precision, recall, and F1-score were used as four separate metrics to assess the network's performance. Equations (2.3) to (2.6) give the mathematical representation to calculate the metrics:

5.3 Results

Performance evaluation is the most significant part of verifying the model's performance. After training each neural network, we tested the model's performance on the unseen test dataset that was never exposed during the training and validation. The segmentation UNet++ model demonstrated the accuracy, Jaccard, dice, and AUC of 97.48%, 91.64%, 95.37%, and 0.99, respectively. Figure 5.4 represents the sample of segmented images by the UNet++ model. The topmost row denotes the original CXR scans. The middle row signifies the masks generated by the model, and the bottom row denotes the final segmented lung. Coloum a represents bacterial pneumonia, b: COVID-19, c: normal, d: tuberculosis, and e specifies sample images from the viral pneumonia class. The border of the lung region and their visualization confirms that the model performed outstandingly for segmenting the chest X-rays.

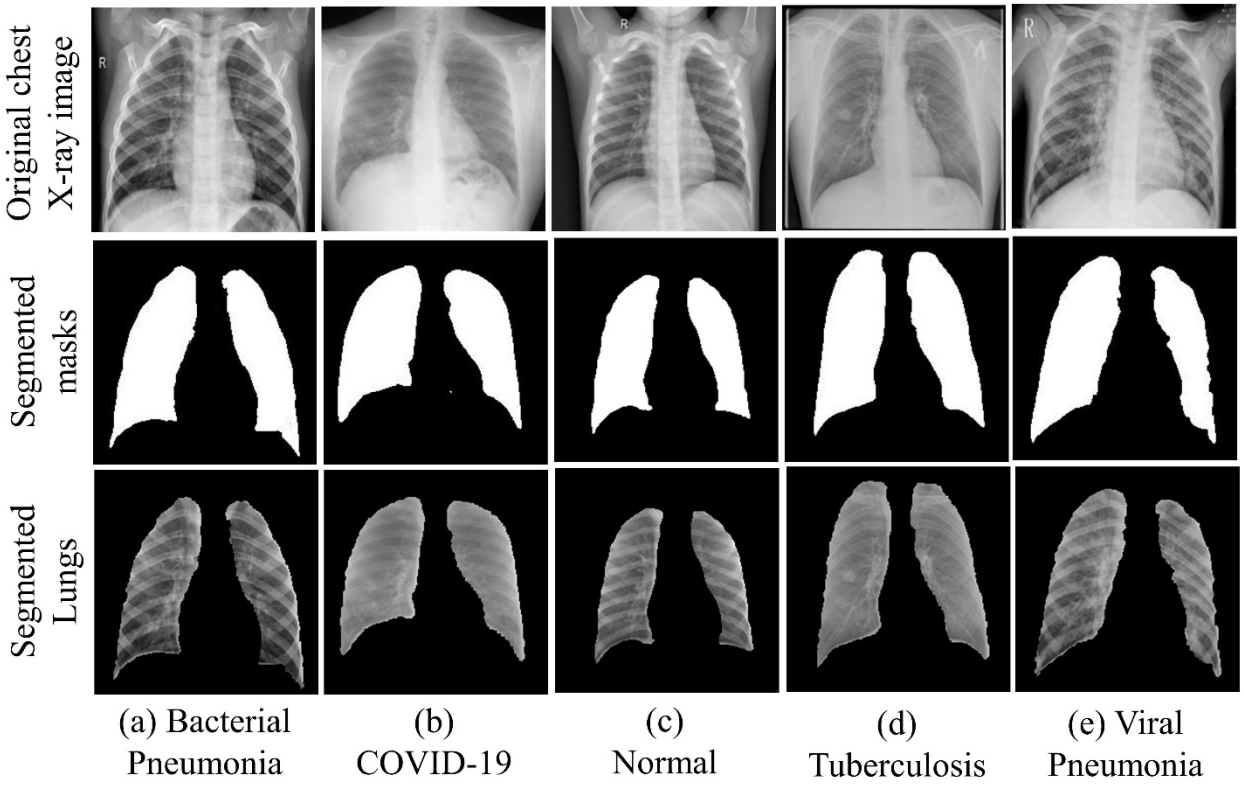


Figure 5.4: Sample of segmented chest X-ray images from each class.

We employed four classification networks to classify the segmented lung regions. The best-performing networks were selected to build the three different ensemble models: Ensemble model-1,

Ensemble model-2, and Ensemble model-3, the combination of best two, best three, and all four networks, respectively. Table 5.1 represents the weighted average of the performance metrics by each classification network and ensemble model. Xception network performed best in all four networks with an accuracy of 96.68% and precision, recall, F1-score, and AUC of 96.71%, 96.68%, 96.67%, and 0.9969, respectively.

Table 5.1: The weighted average of the performance metrics by each model.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC
Xception	96.68	96.71	96.68	96.67	0.9969
MobileNet	95.68	95.77	95.68	95.58	0.9957
InceptionResNetV2	95.06	95.17	95.06	94.92	0.9956
NasNetMobile	93.44	93.42	93.44	93.28	0.9801
Ensemble model-1 (Xception + MobileNet)	99.07	99.09	99.07	99.06	0.9998
Ensemble model-2 (Xception + MobileNet + InceptionResNetV2)	99.61	99.62	99.61	99.61	0.9999
Ensemble model-3 (Xception + MobileNet + InceptionResNetV2 + NasnetMobile)	99.46	99.47	99.46	99.46	0.9999

MobileNet network performed second with an accuracy of 95.68% and precision, recall, F1-score, and AUC of 95.77%, 95.68%, 95.58%, and 0.9957, respectively. The hybrid InceptionResNetV2 network performed third with an accuracy of 95.06% and precision, recall, F1-score, and AUC of 95.17%, 95.06%, 94.92%, and 0.9956, respectively. Finally, the NasNetMobile network performed fourth with an accuracy of 93.44% and precision, recall, F1-score, and AUC of 93.42%, 93.44%, 93.28%, and 0.9801, respectively. Next, the Ensemble model-1 that combined the best two Xception and MobileNet networks performed with an accuracy of 99.07% and precision, recall, F1-score, and AUC of 99.09%, 99.07%, 99.06%, and 0.9998 respectively. The Ensemble model-2, a combination of the best three Xception, MobileNet, and InceptionResNetV2 networks, performed with an accuracy of 99.61% and precision, recall, F1-score, and AUC of 99.62%, 99.61%, 99.61% and 0.9999 respectively. Finally, the Ensemble mod-3 that combined all the Xception, MobileNet, InceptionResNetV2, and NasNetMobile networks performed with an accuracy

of 99.46% and precision, recall, F1-score, and AUC of 99.47%, 99.46%, 99.46% and 0.9999 respectively. The results reveal that the Ensemble model-2, the combination of the best three networks, performed best in all the other ensemble and single networks. The results also state that the Ensemble model's performance depends upon the number of comprising networks and their accuracy. The high number of networks and their high accuracy enhance the ensemble model's performance. The reason behind the superior performance of Ensemble model-2 over other ensemble models was the balance between the number of the network and their accuracy. The inferior performance of Ensemble model-3 to Ensemble model-2 after comprising more classification networks also was the accuracy of the fourth network, NasNetMobile, which was slightly lower than the other three networks. It impacted the overall performance and accuracy of the Ensemble model-3.

5.3.1 Accuracy and loss curves

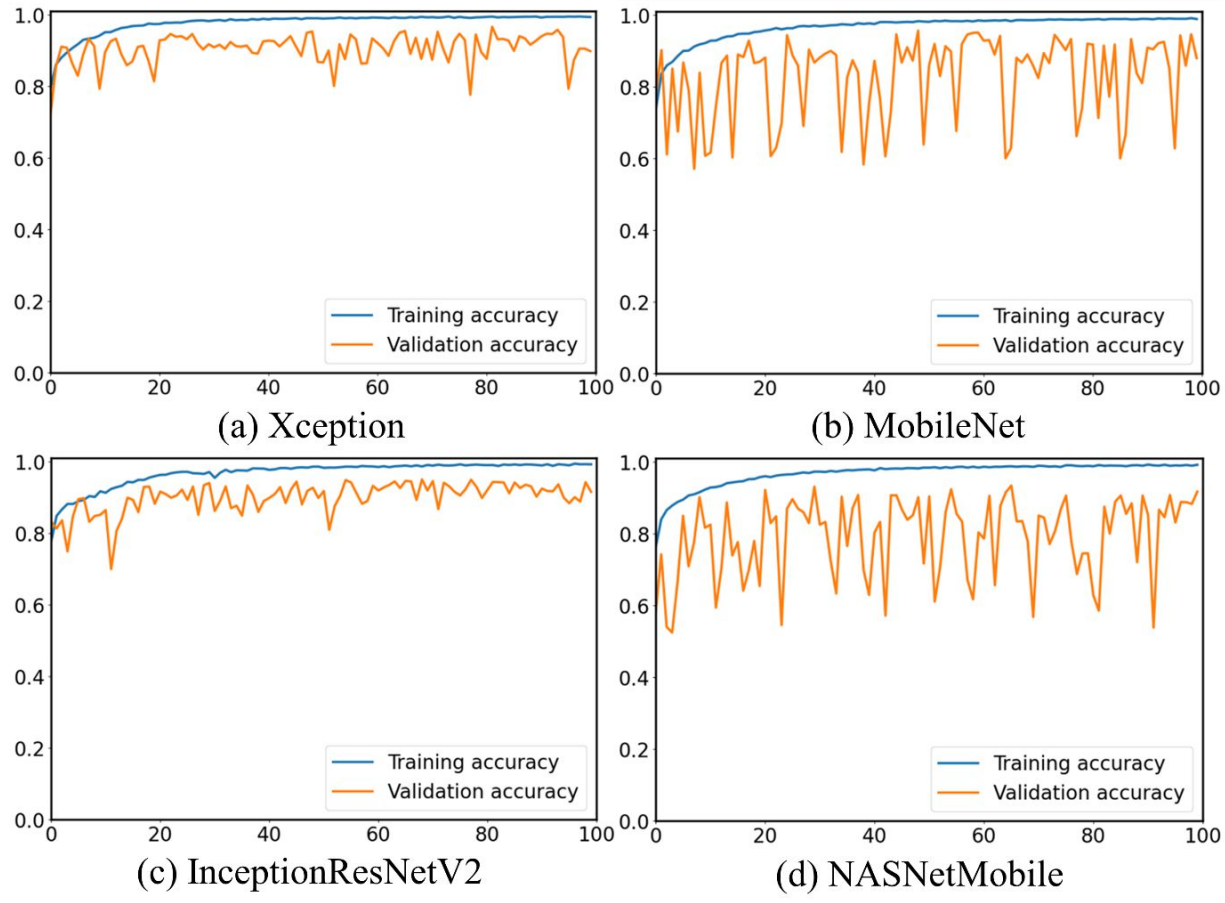


Figure 5.5: Training and validation accuracy by each network; (a) Xception, (b) MobileNet, (c) InceptionResNetV2, (d) NASNetMobile.

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 5.5 represents the training and validation accuracy of each network: (a) Xception, (b) MobileNet, (c) InceptionResNetV2, and (d) NASNetMobile. The graph of each network indicates that the accuracy is improving and stabilizing along the succeeding epochs; this determines the model's precise and robust learning pattern. The accuracy curves by Xception and InceptionResNetv2 showed the best-stabilized learning pattern.

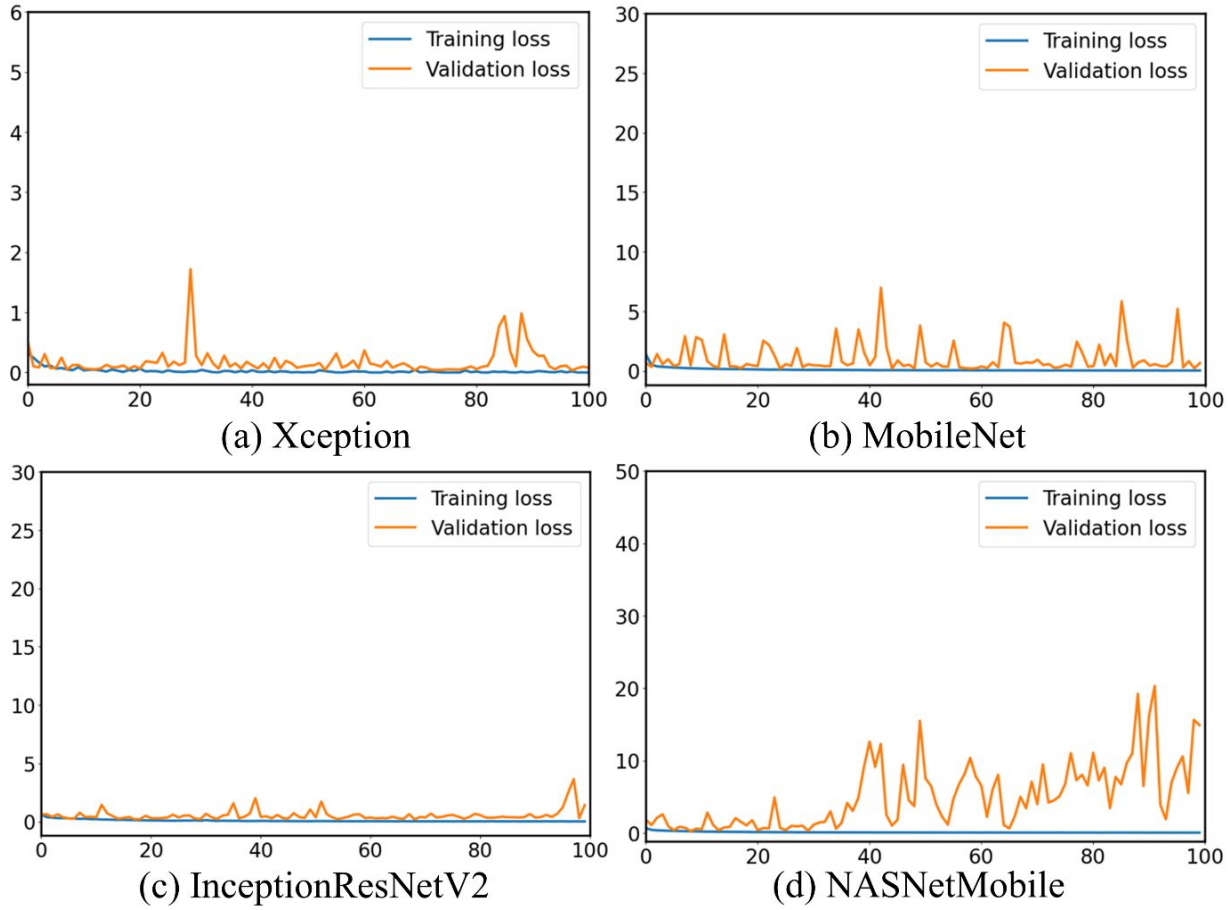


Figure 5.6: Training and validation loss by each network: (a) Xception, (b) MobileNet, (c) InceptionResNetV2, (d) NASNetMobile.

Limiting the loss is an essential factor for decent model learning. Low loss indicates preciseness and more stability of the model. Figure 5.6 represents the training and validation loss by each network: (a)

Xception, (b) MobileNet, (c) InceptionResNetV2, and (d) NASNetMobile. The graph demonstrates that the loss is decreasing and stabilizing along the succeeding epochs. The loss curves by the Xception, MobileNet, and InceptionResNetV2 showed the best-stabilized loss pattern.

5.3.2 Confusion matrix

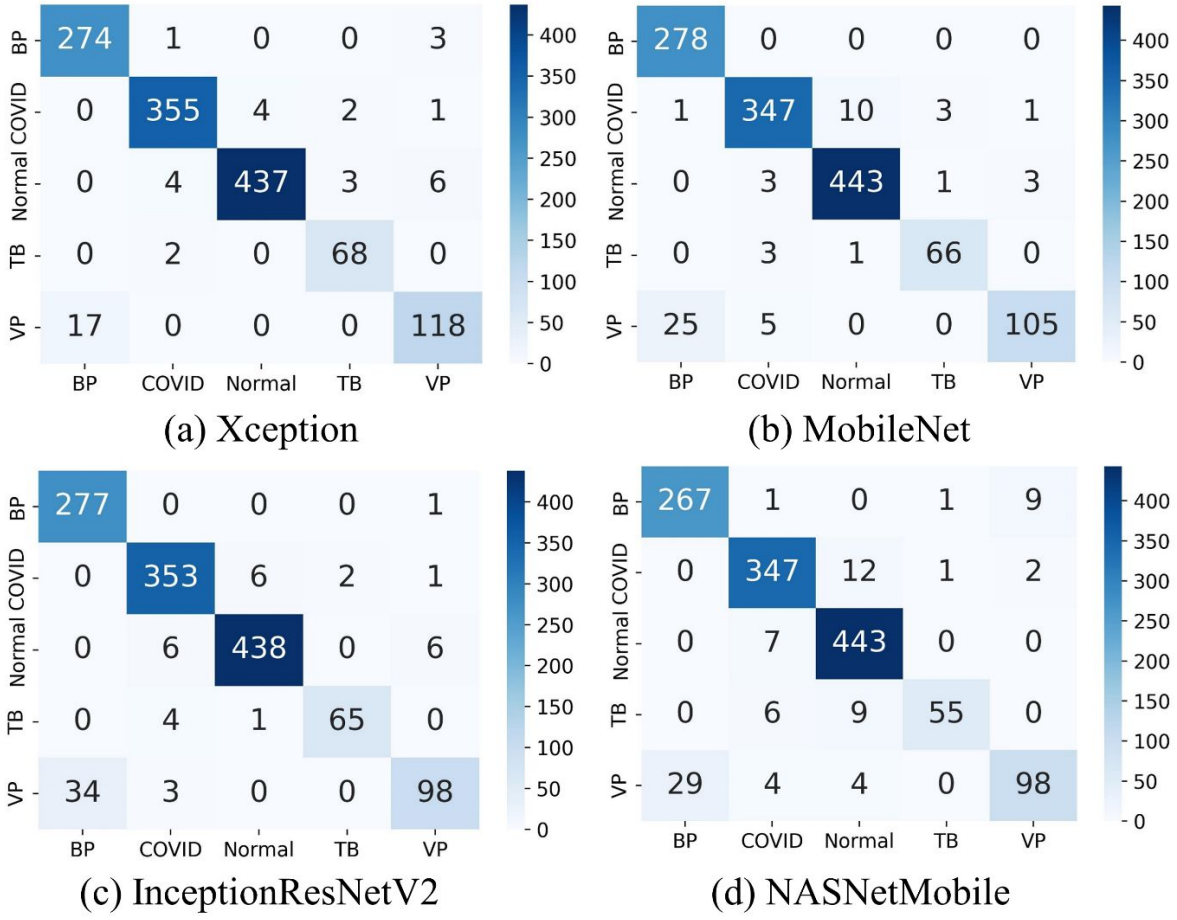


Figure 5.7: Confusion matrix of performance by each network: (a) Xception, (b) MobileNet, (c) InceptionResNetV2, (d) NASNetMobile.

A confusion matrix is a significant way of visualizing the results from a machine learning classification task. The confusion matrix provides a simple way to visualize how well each classifier performs against multiple test samples. An analysis of a model's performance on a set of test data is summarized by a confusion matrix. It is frequently used to assess how well the classification models work. These models try to predict a categorical label for each sample. The matrix shows how many true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) the model generated using the test data. Figure

5.7 represents the confusion matrix of performance by each network: (a) Xception, (b) MobileNet, (c) InceptionResNetV2, and (d) NASNetMobile. The confusion matrix demonstrates that out of 1,290 CXR images, 1,253, 1,238, 1,226, and 1,205 images were correctly classified, and 17, 52, 64, and 85 images were misclassified by the Xception, MobileNet, InceptionResNetV2, and NASNetMobile respectively. The results reveal that the Xception network performed best among all networks.

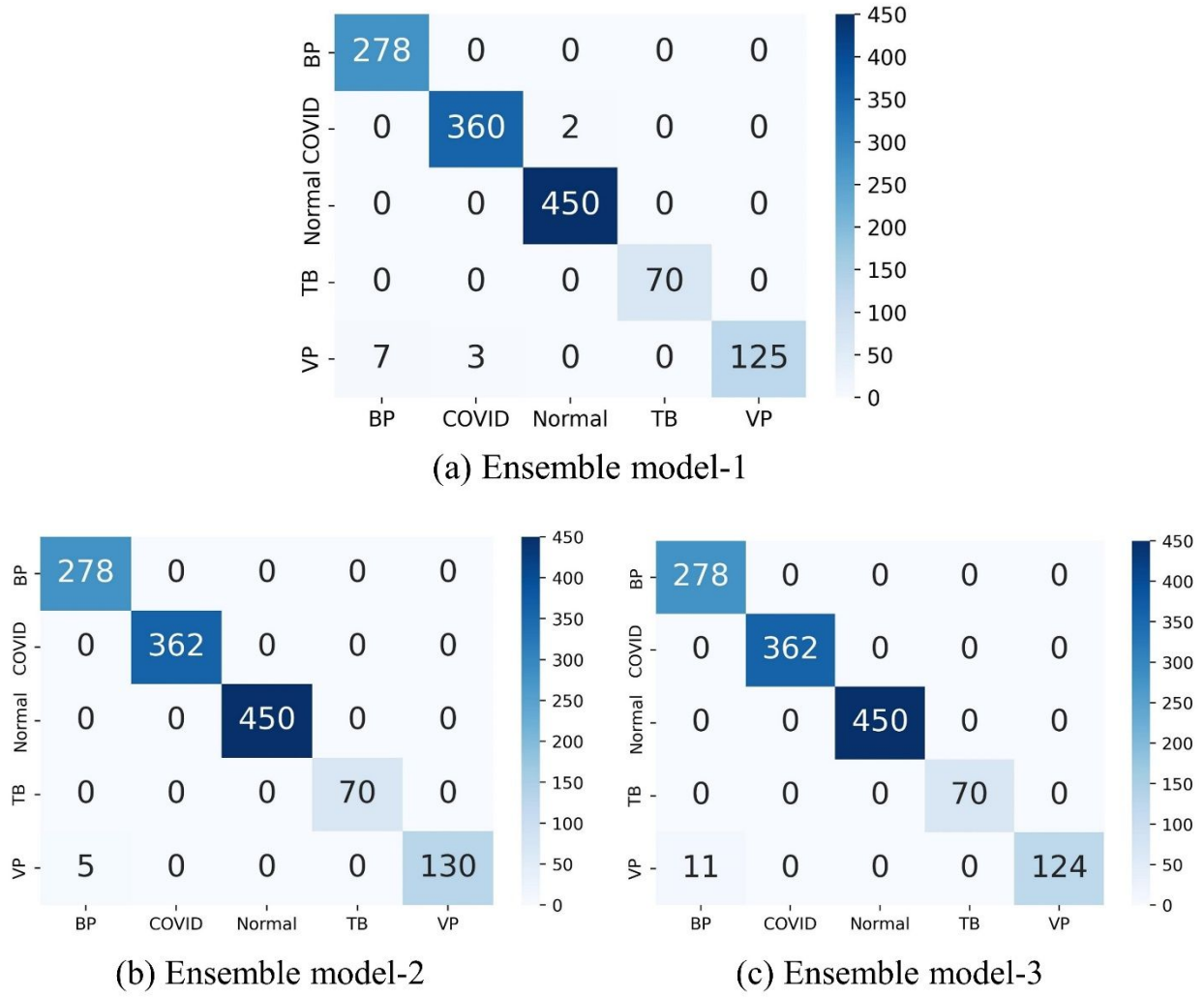


Figure 5..8: Confusion matrix of performance by each ensemble model; (a) Ensemble model-1, (b) Ensemble model-2, (c) Ensemble model-3.

Figure 5.8 represents the confusion matrix of performance by each ensemble model: (a) Ensemble model-1, (b) Ensemble model-2, and (c) Ensemble model-3. The confusion matrix reveals that out of a total of 1,290 CXR images, 1,278, 1,285, and 1,279 images were correctly classified, and 12, 5, and 11 images were misclassified by the Ensemble model-1, Ensemble model-2, and Ensemble model-3, respectively. The

results reveal that the Ensemble model-2 performed best among all ensemble models and convolutional neural networks and correctly detected all the COVID-19 images.

5.3.3 Performance evaluation

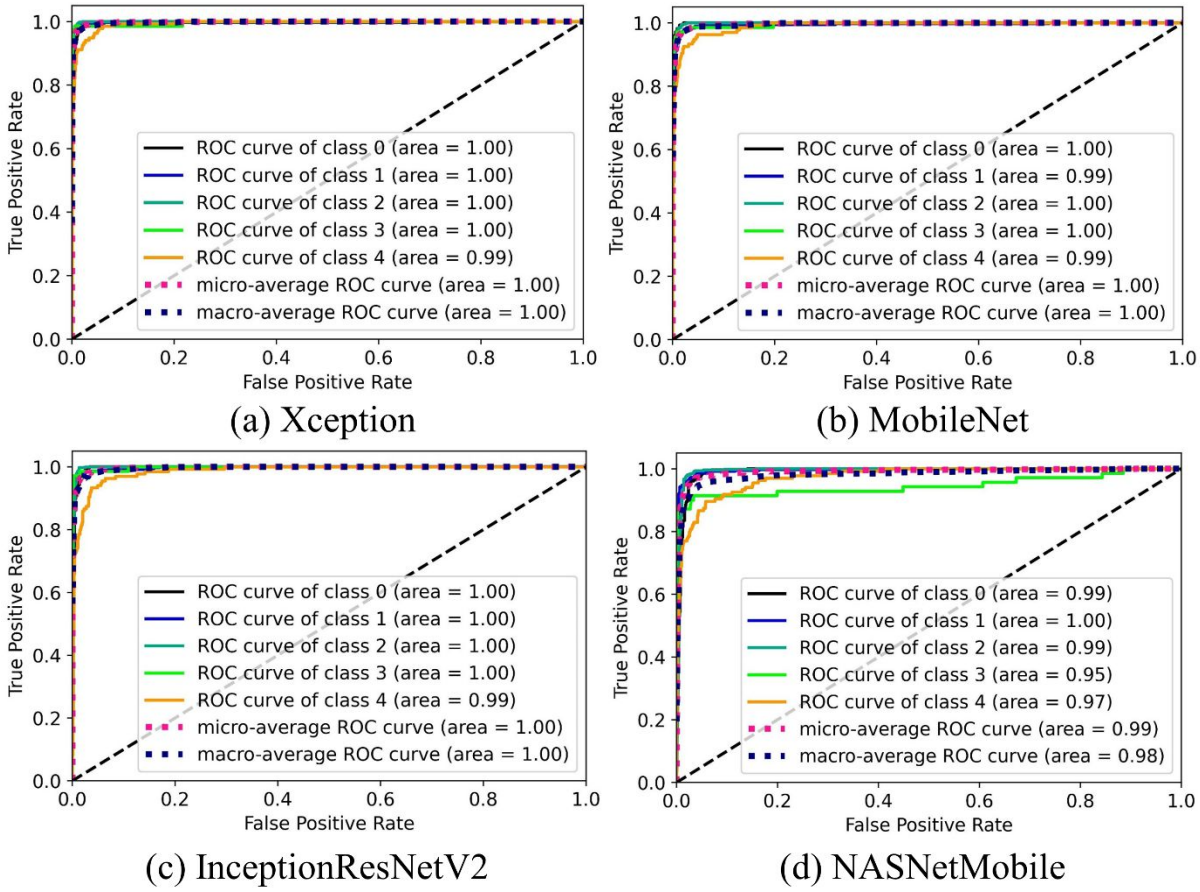
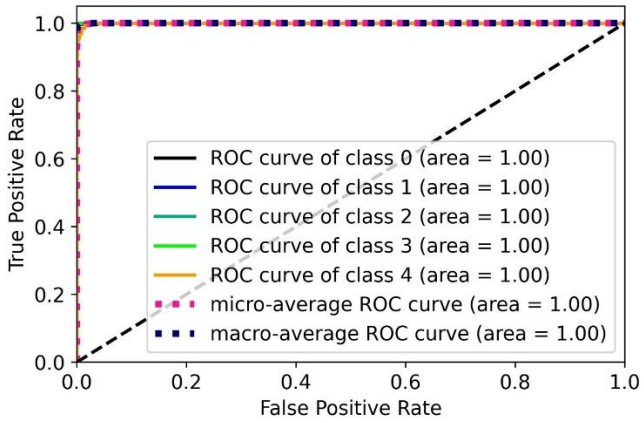


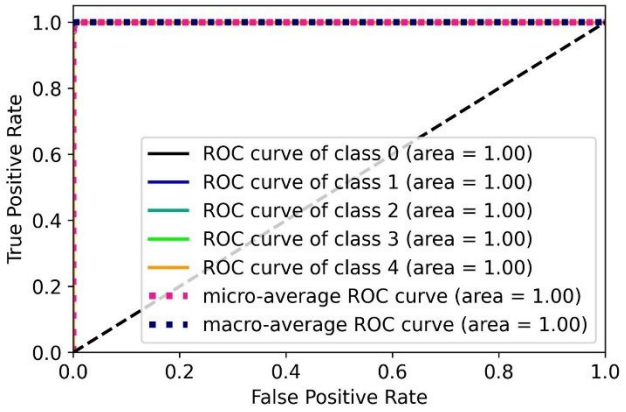
Figure 5.9: ROC and AUC for all netwirks; (a) Xception, (b) MobileNet, (c) InceptionResNetV2, (d) NASNetMobile . ($p < 0.0001$; class 0: BP; class 1: COVID-19; class2: normal; class 3: TB; class 4: VP).

Performance evaluation is an essential part of testing the robustness of the model against all odds. The ROC (receiver operating characteristics curve) and AUC (area under the ROC curve) are the most significant measures of a predictive algorithm's performance. They are the probability that an indication has a positive value greater than or equal to the null hypothesis and absolute risk (or chance) difference. ROC is the receiver's ability to identify a specific class correctly. A ROC curve is a graph showing the performance of the classification model, i.e., TPR (true positive rate) vs. FPR (false positive rate) at different classification thresholds. AUC measures the entire two-dimensional area underneath the ROC curve. AUC provides an

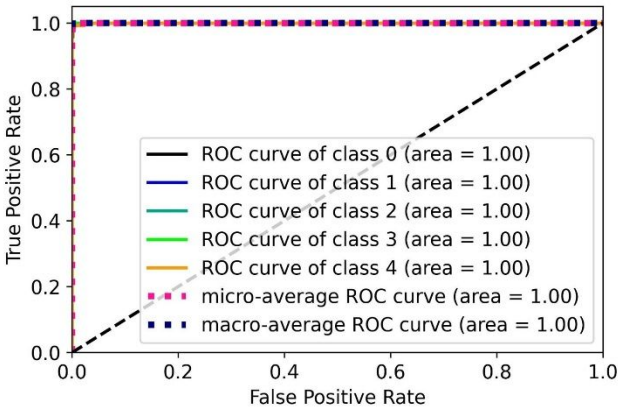
aggregate measure of performance across all possible classification thresholds. The AUC values range between 0 and 1. The AUC near 0 represents the worst performance, whereas near 1 represents the best performance of the classifier.



(a) Ensemble model-1



(b) Ensemble model-2



(c) Ensemble model-3

Figure 5.10: ROC and AUC for all ensemble models; (a) Ensemble model-1, (b) Ensemble model-2, (c) Ensemble model-3; ($p < 0.0001$; class 0: BP; class 1: COVID-19; class2: normal; class 3: TB; class 4: VP).

Figure 5.9 illustrates the ROC and AUC for all networks: (a) Xception, (b) MobileNet, (c) InceptionResNetV2, and (d) NASNetMobile. The AUC for Xception, MobileNet, InceptionResNetV2, and NASNetMobile was 0.9969, 0.9957, 0.9956, and 0.9801, respectively. The Xception performed best with the highest AUC among all networks. Figure 5.10 represents the ROC and AUC for all ensemble models: (a) Ensemble model-1, (b) Ensemble model-2, (c) Ensemble model-3. The AUC for Ensemble model-1, Ensemble model-2, and Ensemble model-3 was 0.9998, 0.9999, and 0.9999, respectively. The Ensemble model-2 performed best with the highest AUC among all ensemble models and CNNs.

5.3.4 Grad-CAM visualization

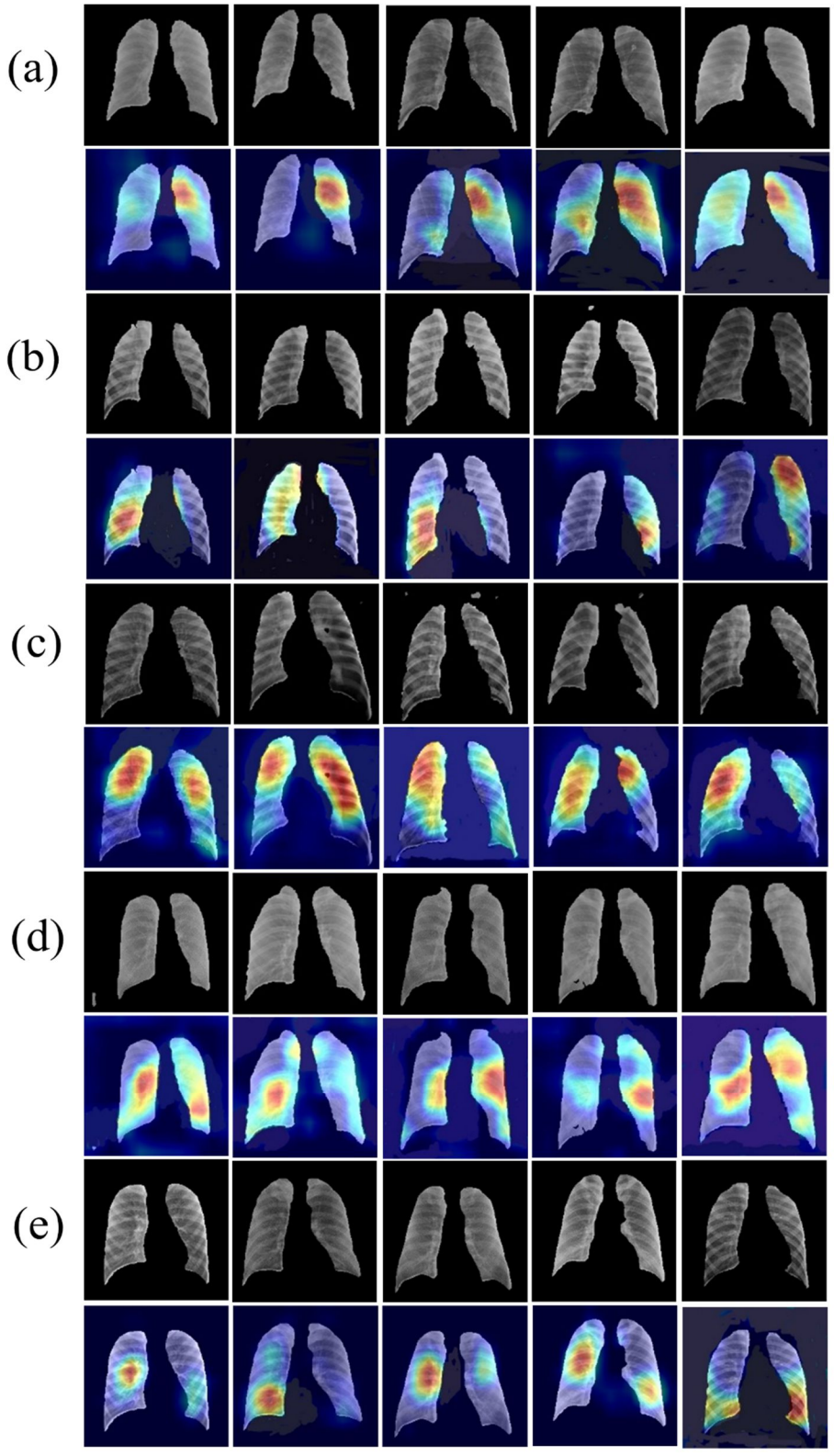


Figure 5.11: The GRAD-CAM heatmap visualization of the chest X-ray images; Row (a): original CXR images, (b): corresponding segmented lungs, (c): corresponding heatmaps.

The Gradient-weighted Class Activation Mapping (Grad-CAM) method uses gradients for the classification score and the final convolutional feature map to pinpoint the areas of an input image that influence the classification score most. The final score depends on the data the most in the locations where this gradient is large. The Grad-CAM function calculates the significance map by calculating the derivative of the reduction layer output for a particular class concerning a convolutional feature map. The Grad-CAM function automatically chooses appropriate layers to compute the significance map for classification tasks. The Grad-CAM heatmaps are a very advantageous analysis for medical images that represent the image's significant regions involved in the infection. Heatmap displays the area of lesion or infection within the images. Heatmap also signifies the region from which the network is extracting most of the significant features to determine the classes of the images. Generally, the heatmap shows a similar pattern in similar infection types.

For our system, using the gradients of the target, i.e., COVID-19 and other pneumonia in the best performing Xception-based classification model, Grad-CAM generates the coarse localization map and displays the significant locations as heatmap scans. Figure 5.11 represents the heatmap of some samples of CXR images from all five classes. Row (a) denotes COVID-19, (b) viral pneumonia, (c) bacterial pneumonia, (d) tuberculosis, and (e) represents normal class. The heatmap images from each category showed similar heatmap patterns. For COVID-19 cases, lesions were primarily present in the upper part of the lungs. In contrast, the heatmaps of other class images demonstrated completely different patterns from COVID-19 and similar patterns between images of their class.

5.4 Discussion

Previously, several authors have tried applying ensemble methods to classify pneumonia chest x-ray images. However, two major research gaps persisted: first, the number of classes was low as in most cases, authors applied classification experiments only for two or three categories of pneumonia, and the second gap was a lack of high accuracy, especially for classification in larger categories. Hashmi et al. [212] implemented a weighted ensemble classifier model built of five models: DenseNet121, Xception, InceptionV3, MobileNetV2, and ResNet18. They utilized the Guangzhou Women and Children's Medical

Table 5.2: Comparison of proposed ensemble method against previously implemented state-of-art ensemble methods.

Author & Year	Method	Dataset (total number of images)	Accuracy (%)	AUC	Accuracy improvement*(%)
Hashmi et al. [212] (2020)	Ensemble model (DenseNet121, Xception, InceptionV3, MobileNetV2, and ResNet18)	Guangzhou Women and Children's Medical Centre pneumonia dataset (5836)	2 class-98.43	0.9976	1.03
Chouhan et al. [213] (2020)	Ensemble model (DenseNet121, AlexNet, ResNet18, InceptionV3, and GoogLeNet)	Guangzhou Women and Children's Medical Centre pneumonia dataset (5836)	2 class-96.4	0.9934	3.06
Pant et al. [214] (2020)	Ensemble model (ResNet-34 based U-Net, EfficientNet-B4 based U-Net)	Guangzhou Women and Children's Medical Center pneumonia dataset (5836)	2 class-90		9.46
Das et al. [215] (2021)	Ensemble model (ResNet-50, DenseNet201, InceptionV3)	Combination of different open sources dataset (538+468)	2 class-91.62	0.91	7.84
Shorfuzzaman et al [216] (2021)	Ensemble (VGG16, ResNet-50V2, and InceptionV3)	GitHub: Cohen + Kaggle: Mooney: Chest X-Ray Images (Pneumonia) (1848)	3 class-95.49	0.95	3.97
Ahmad et al. [217] (2021)	Ensemble model, MobileNet and InceptionV3)	Kaggle: Chest X-Ray Images (Pneumonia) (4200)	4 class-96.49%	N.A.	2.97
Kundu et al. [218] (2021)	Ensemble: GoogLeNet, ResNet-18, and DenseNet-121.	Kermany: pneumonia X-ray datasets (5856)	2 class-98.81	0.9835	0.65
Ayan et al. [219] (2021)	Ensemble: ResNet50, Xception, MobileNet	Kermany: pneumonia X-ray datasets (5856)	2 class-95.83	0.9521	3.63
Wehbe et al. [59] (2021)	Ensemble (ResNet-50, DenseNet-121, Xception, InceptionV3, and EfficientNet-B2, Inception-ResNetV2)	Private (14,778)	2 class-83%	0.9	16.46
Keidar et al. [60] (2021)	Ensemble (VGG16, ReNet34, ReNet50, and ReNet152)	Private (2427)	2 class-90.3%	0.96	9.16
Abdulah et al. [64] (2021)	Ensemble (6 parallel ATTENTION + CONV-RES blocks)	Private (3797)	2 class-79%	0.85	20.46
Hertel et al. [69] (2022)	Ensemble (VGG19, ResNet152, and DenseNet201)	COVIDx5 + MIDRC-RICORD-1C + BIMCV dataset (12,837)	2 class-91% 3 class-84%	0.95	8.46
Proposed method	Ensemble (Xception + MobileNet + InceptionResNetV2 + NasnetMobile)	COVID-19 Radiography database + Chest X-ray Images (Pneumonia) + Tuberculosis (TB) Chest X-ray database (Kaggle) (12,941)	5 class-99.46%	0.9999	

*Accuracy improvement against state-of-the-art methods. #Average accuracy improvement = 7.26%

Centre pneumonia dataset containing 5,836 CXR images from pneumonia and normal classes. They employed the cross-entropy loss function with the Stochastic Gradient Descent (SGD) optimizer and

achieved a binary classification accuracy of 98.43% with an AUC of 0.9976. Chouhan et al. [213] proposed an ensemble model that differs from the weighted classifier approach implemented by Hasmi and colleagues, where the majority vote is considered the final output. The five pre-trained models, DenseNet121, AlexNet, ResNet18, InceptionV3, and GoogLeNet, were employed for ensembling, and the maximum number of predictions (normal or pneumonia) from each model was used as the output. The final result was calculated using the majority vote of each model's outputs without applying any weights. They employed the cross-entropy loss function with the Stochastic Gradient Descent (SGD) optimizer. The dataset they utilized in the study was the Guangzhou Women and Children's Medical Centre pneumonia dataset, which contains 5,836 CXR images from pneumonia and normal classes. They achieved a binary classification accuracy of 96.4% with an AUC of 0.9934.

Pant et al. [214] employed an ensemble model that included two U-Nets, one based on ResNet-34 and the other on EfficientNet-B4. They employed binary cross-entropy loss and Dice loss using the Ranger optimizer. They utilized the Guangzhou Women and Children's Medical Centre pneumonia dataset containing 5,836 CXR images from pneumonia and normal classes and achieved an accuracy of 90%. Das et al. [215] employed an ensemble model comprising three networks: DenseNet201, ResNet-50, and InceptionV3. They applied the Adam optimizer and utilized a total of 1,006 CXR images from two classes taken from different open-source datasets. They achieved an accuracy of 91.62% with an AUC of 0.91 for binary classification. Shorfuzzaman et al. [216] employed an ensemble model comprising VGG16, ResNet-50V2, and InceptionV3 CNNs. They used categorical cross entropy loss function and Adam optimizer. They utilized 1,848 CXR images taken from the Cohen dataset from GitHub and the Chest X-Ray Images (Pneumonia) dataset from Kaggle. They achieved a classification accuracy of 95.49% with an AUC of 0.95 for three class classifications into COVID-19, viral pneumonia, and normal. Ahmad et al. employed an ensemble model comprising MobileNet and InceptionV3 network. They utilized 4,200 CXR images of the Chest X-ray images (Pneumonia) dataset taken from Kaggle. They performed four class classifications and achieved an accuracy of 96.49%. Kundu et al. [218] employed an ensemble model comprising three networks: GoogLeNet, ResNet-18, and DenseNet-121. They utilized 5,856 CXR images taken from the Kermany pneumonia X-ray dataset. They achieved an accuracy of 98.81% for binary classification into

pneumonia and normal. Ayan et al. [219] also utilized 5,856 CXR images taken from the Kermany pneumonia X-ray dataset. They employed an ensemble model comprised of three networks ResNet50, Xception, and MobileNet. They achieved an accuracy of 95.83% for binary classification into pneumonia and normal. Wehbe et al. [59] employed an ensemble model that comprises six different pre-trained networks: ResNet-50, DenseNet-121, Xception, InceptionV3, and EfficientNet-B2, Inception-ResNetV2. They utilized a private dataset that contains a total of 14,778 CXR images. They achieved an accuracy of 83% with an AUC of 0.9 for the binary classification into COVID-19 and normal. Keidar et al. [60] employed an ensemble model that comprised VGG16, ReNet34, ReNet50,, and ReNet152 pre-trained networks. They implemented a nearest-neighbors approach that retrieves the images most similar to a given image using DNN-based image embeddings. They utilized a private dataset from hospitals comprising 2,427 CXR images into two classes: COVID-19 and control. They achieved a classification accuracy of 90.3% with an AUC of 0.96. Abdulah et al. [64] developed an ensemble model in which the output of 6 parallel attention and convolutional residual blocks was averaged to a single feature map of dimensions. The features map was passed to the final Softmax activation layer through a GlobalAveragePooling2D layer. They utilized a private dataset with 3,797 CXR images taken from hospitals. They achieved an accuracy of 79% with an AUC of 0.85 for binary classification into COVID-19 and normal. Hertel et al. [69] implemented an ensemble model comprising three networks VGG19, ResNet152, and DenseNet20. They utilized three different COVIDx5, MIDRC-RICORD-1C, and BIMCV chest X-ray datasets with 12,837 images. They achieved a binary classification accuracy of 91% for COVID-19 and normal and a three-class classification accuracy of 84% for COVID-19, viral pneumonia, and normal, with an AUC of 0.95.

This work developed a highly efficient system comprising a combination of segmentation and an ensembled classification model to classify CXR images into five classes having four pneumonia types, including COVID-19 and a control class. We utilized 12,941 CXR images from the "COVID-19 Radiography database," "Chest X-ray Images (Pneumonia)," and "Tuberculosis (TB) Chest X-ray database," all taken from Kaggle for the experiment. For the first time, we applied the ensemble model to classify chest X-ray images into five categories. Our ensemble model-2 achieved a classification accuracy

of 99.61% and an AUC of 0.9999, the highest among all the existing state-of-the-art methods in any class classification. Covering the diverse types of infection with highly accurate results enables our system to precisely detect most pneumonia types, including COVID-19; this will help the doctors diagnose the condition and choose the exact medications for disease management. For the first time, we applied different ensemble models in a single experiment using various combinations of classification DL networks. We applied four classifiers to build three different ensemble models: Ensemble Model-1, Ensemble model-2, and Ensemble model-3, comprising the best-performed two (Xception and MobileNet), three (Xception, MobileNet, and InceptionResNetV2), and all four (Xception, MobileNet, InceptionResNetV2, and NasNetMobile) networks, respectively. This approach facilitated us to select the best-performing ensemble model with the best network combinations. Additionally, the approach conferred a new theory that the ensemble model's performance depends upon both the number of DL networks utilized and their performance. Both factors may impact the performance of the ensemble model positively or negatively. Ensembling a good-performing DL network enhances the performance of the ensemble model; however, fusing a poor-performing network diminishes the performance of the ensemble model.

5.5 Measures taken to prevent overfitting

Overfitting is a major concern related to the DL networks training. Ensuring that our deep learning ensemble model's performance is not a result of overfitting and that it can generalize well to new, unseen data, the following strategies were applied to assess and improve the model's generalization capability:

(a) K-fold cross-validation: We used a 5-fold cross-validation technique for segmentation to assess the model's performance on multiple subsets of the training data. Our DL models consistently perform well on different folds, and it's an indicator of good generalization.

(b) Validation and test set: We used 20% of the data for segmentation and 10% of the data for classification as a validation and test set that was not used during training. This allowed us to evaluate the model's performance on unseen data. Our model's performance was consistent between the training and validation data; this denotes less likely to be overfitting.

(c) Regularization Techniques: We used regularization methods like dropout and batch normalization to reduce overfitting. We experimented with different hyperparameters to find the right balance between model complexity and generalization.

(d) Data Augmentation: Augmentation was applied to training data with rotation, translation, and flipping techniques to introduce variability and help the model generalization better.

(e) Hyperparameter Tuning: We systematically searched for the best hyperparameters, including learning rate, batch size, and model architecture; this allowed the model to be robust and generalization better.

(f) Transfer Learning: We started with pre-trained models and fine-tuned them on our specific task. Transfer learning leverages the knowledge encoded in pre-trained models, which improves generalization.

(g) Ensemble Diversity: Our ensemble members were diverse in terms of architecture, training data subsets, or hyperparameters. Diversity in the ensemble helped reduce the risk of overfitting.

(h) Peer Review and Expert Input: We involved domain experts across the world who provided an independent evaluation of the model's generalization and assessed its clinical or practical relevance.

(i) Regular Revalidation: We will periodically revalidate the model's performance as the data distribution may change over time. Retraining and updating the model helps maintain generalization.

(j) It's important to strike a balance between model complexity and generalization when building deep learning models, and the above strategies helped us achieve that balance and ensure that high performance was not merely a result of overfitting but is indicative of the model's true capability to generalize to unseen data.

5.6 Summary

In this work, we have proposed a deep learning-based segmentation and ensemble classification system to detect pneumonia, including COVID-19, in chest X-ray scans. For segmentation of chest X-rays, we implemented the UNet++, a modified version of the UNet network, and achieved the accuracy, Jaccard,

dice, and AUC of 97.48%, 91.64%, 95.37%, and 0.99, respectively. After that, we applied four classification networks, including a hybrid model: Xception, MobileNet, NasnetMobile, and InceptionResNetV2, to classify segmented lung images into five classes: COVID-19, viral pneumonia, bacterial pneumonia, tuberculosis, and control. Next, we chose the best-performing networks and ensembled them to build combinations of three Ensemble models: Ensemble model-1, Ensemble model-2, and Ensemble model-3. We employed an ensemble model to classify chest X-ray images into five categories for the first time. The Ensemble model-2, comprising the best three networks, Xception, MobileNet, and InceptionResNetV2, outperformed all classification models with an accuracy of 99.61% and precision, recall, F1-score, and AUC of 99.62%, 99.61%, 99.61% and 0.9999 respectively. Our system performed best among any class's existing state-of-the-art chest X-ray image classification methods. The system algorithm's high accuracy and efficacy raise the possibility of a valuable setup for COVID-19 and pneumonia detection, particularly in environments with a scarcity of test kits and radiological expertise.