Chapter 2

Multiclass Frameworks for Pneumonia Classification and Its Validation in X-ray Scans Using Seven Types of Deep Learning Models

Abstract

Background and Motivation: The novel coronavirus causing COVID-19 is exceptionally contagious and highly mutative, decimating human health and life, as well as the global economy, by consistent evolution of new pernicious variants and outbreaks. The reverse transcriptase polymerase chain reaction currently used for diagnosis has significant limitations. Furthermore, the multiclass lung classification X-ray systems with viral, bacterial, and tubercular classes, including COVID-19, are unreliable. Thus, there is a need for a robust, fast, cost-effective, and readily available diagnostic method. Method: Artificial intelligence (AI) has been shown to revolutionize all walks of life, particularly medical imaging. This study proposes a deep learning AI-based automatic multiclass detection and classification of pneumonia from chest X-ray images that are readily available and highly cost-effective. The study has designed and applied seven highly efficient pre-trained convolutional neural networks—namely, VGG16, VGG19, DenseNet201, Xception, InceptionV3, NasNetMobile, and ResNet152—for the classification of up to five classes of pneumonia. Results: The database comprised 18,603 scans with two, three, and five classes. The best results were using DenseNet201, VGG16, and VGG16, respectively having accuracies of 99.84%, 96.7%, 92.67%; sensitivity of 99.84%, 96.63%, 92.70%; specificity of 99.84, 96.63%, 92.41%; and AUC of 1.0, 0.97, 0.92 ($p < 0.0001$ for all), respectively. Our system outperformed existing methods by 1.2% for the five-class model. The online system takes <1 s while demonstrating reliability and stability. Conclusions: Deep learning AI is a powerful paradigm for multiclass pneumonia classification.

2.1 Introduction

COVID-19 is a highly contagious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) [71]. The virus was first isolated from three pneumonia patients with critical respiratory illness in December 2019 in Wuhan, China [72]. Within a short period, the virus spread globally. On March 11, 2020, the World Health Organization (WHO) declared the disease a pandemic [73]. Coronaviruses (CoVs) are a tremendously diverse family of enveloped positive-sense single-stranded RNA viruses [74]. The viruses are highly pathogenic and transmissible and spread via respiratory droplets or aerosol between individuals in close proximity[75], leading to several pathways causing damage to several organs such as the coronary and liver, causing diabetes and pulmonary embolism [76-78]. In the majority of infected cases, the person begins to exhibit symptoms like cough, fever, fatigue, and loss of smell or taste. In numerous fatal instances, the infection progresses to the lower respiratory system, including the lungs, causing illness like severe pneumonia followed by multi-organ dysfunction syndrome with several secondary infections and shock [79, 80].

Even after two years of the virus outbreak and almost ten thousand million doses of vaccination, the disease continues to destroy human health, life, and the global economy. The viruses are incredibly efficient in mutating fast and gradually converting into more deadly variants. After the severe damage caused by the Delta variant, a new variant named Omicron has evolved. WHO has already designated Omicron as a variant of concern [81]. Several notable mutations in spike proteins of the Omicron make it highly transmissible. Moreover, there is still a risk of more new mutations in Cov-2 thereafter, a more pernicious variant outbreak.

COVID-19 infection is usually detected by a reverse transcriptase polymerase chain reaction (RT-PCR) test, which is frequently followed by chest radiographs, such as X-rays and computed tomography (CT) scans [82, 83]. The reference technique for COVID-19 detection is RT-PCR; though, the procedure is laborious, complicated, rigorous, and time-consuming, with a significantly high error rate [84, 85]. The RT-PCR kit and a specific biosafety facility to host the PCR machine cost very high. Consequently, there is a substantial supply constraint. Many nations are experiencing problems with erroneous COVID-19-

positive cases caused by an inadequacy in test kit supply and a delay in the test results. These limitations of RT-PCR contribute majorly to restricting disease control and infection spread to healthy populations [86]. To counteract the spread of COVID-19, patients must undergo prompt and effective screening and get appropriate medical attention. Several medical imaging modalities, including Chest X-ray (CXR) and computed tomography (CT), can help with this [87, 88]. COVID-19 has recently been detected using CT imaging [87, 89]. However, the high patient dosage and screening expenses are the principal disadvantages of using CT imaging for diagnosis [90]. On the other hand, the CXR equipment is commonly accessible in hospitals and diagnostic centers to create a 2D projection of the thorax quickly and affordably. Radiologists already use the CXR modality to detect chest abnormalities in various lung illnesses, including pneumonia and tuberculosis. COVID-19 detection has also been done utilizing CXRs in a few patients [87, 91]. The COVID-19 patients reveal similar findings in radiographs like bilateral, peripheral, and basal predominant ground-glass opacities, septal thickening, pleural effusion, bronchiectasis, and bilateral lymphadenopathy [92-95]. As a result, CXR scans might help in the early detection of COVID-19 in the suspected person. However, one challenge is that the CXRs of various pneumonia are very similar; therefore, manually differentiating COVID-19 from other lung abnormalities is tough. Nonetheless, deep learning algorithms powered by Artificial Intelligence (AI) can efficiently extract several image-based features that radiologists may be unable to observe manually in the original CXR. Regarding image feature extraction and classification, Convolutional Neural Networks (CNNs) have proven their efficiency and are being widely implemented by the research community [96]. Nowadays, CNN-based solutions are tremendously being utilized to resolve a variety of health problems, such as brain tumor identification [97-100], lung and breast cancer detection [37, 101], Alzheimer's disease diagnosis [26], cardiovascular disease predictions [25, 102, 103], pneumonia detection [27] and much more. With promising results in several applications, deep learning techniques for chest X-rays are gaining prominence in recent days. The transfer learning technique has made the operation smoother by facilitating a highly deep CNN to be retrained quickly [104, 105].

In this work, we have designed and applied seven different deep-learning models utilizing the transfer learning method to detect multiclass COVID-19 in CXR images. We have performed the binary and multiclass classification into COVID-19 and other lung diseases, namely, viral pneumonia (VP),

bacterial pneumonia (BP), tuberculosis (TB), and normal images. Thereafter, we compared the results to get the best-suited model for their usefulness in practice. Figure 2.1 shows the overall schematic diagram of the development of the COVID-19 detection system.

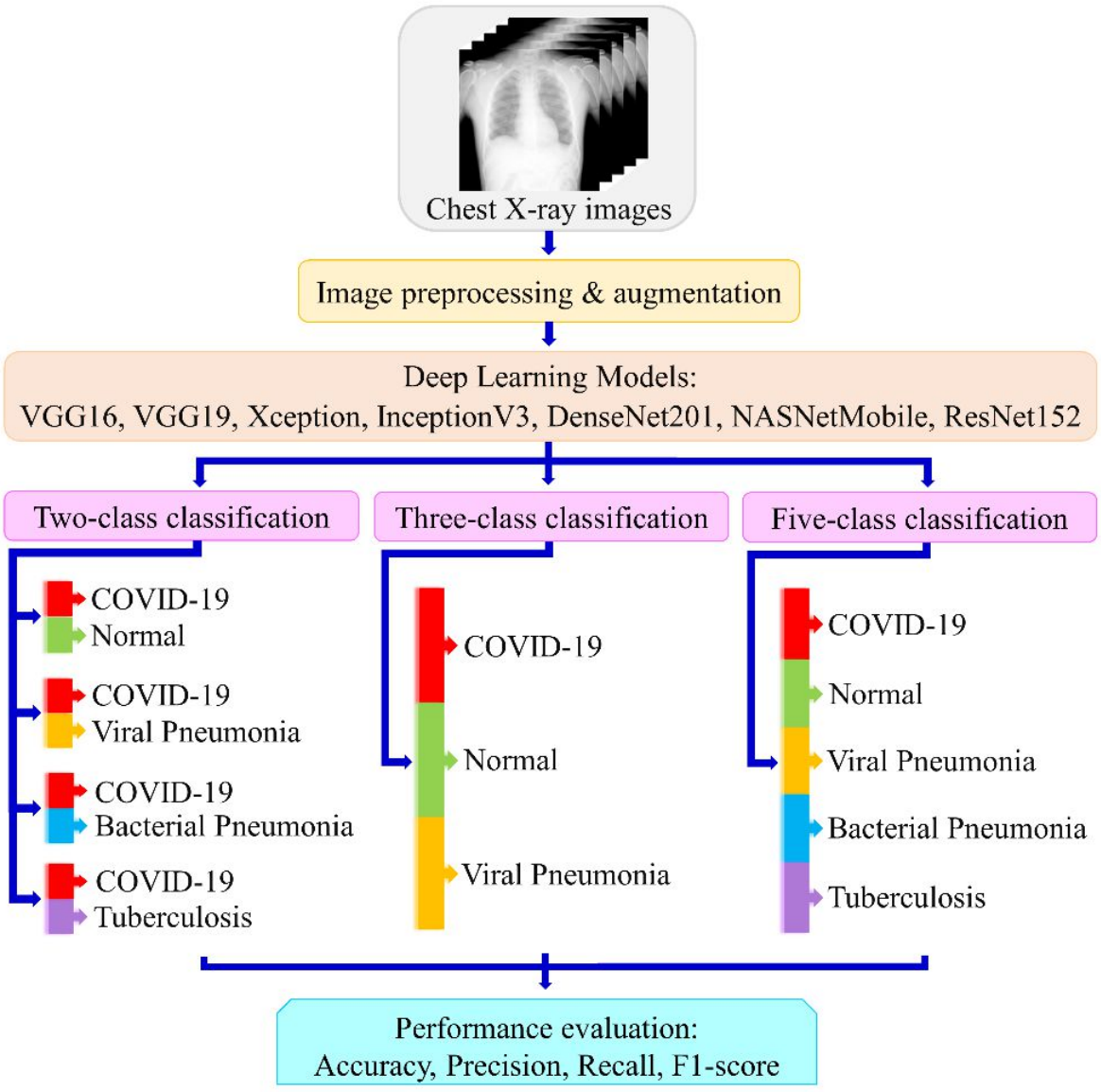


Figure 2.1: Overall schematic diagram of the proposed method for the multiclass scenario using seven deep learning models.

The whole work has been structured in a section-wise manner. In section 2.2, we have explored all the related work and contributions of different authors in this area. Section 2.3 explains the dataset, image-preprocessing, and deep-learning models. In section 2.4, the results of the experiments and their comparative performances have been provided. Section 2.5 deals with the model's performance evaluation.

Next, Section 2.6 presents the scientific validation of the proposed models on another dataset. Further, in section 2.7, we have compared the proposed models with other existing state-of-the-art methods. Finally, section 2.8 concludes the study with the future scope.

2.2 Related work

Recently, COVID-19 detection using deep learning techniques has become a very popular area. Several researchers have proposed deep learning methods for detecting disease in CXR images. However, the majority of them employed a limited dataset with a small number of COVID-19 samples. Consequently, their outputs may not be generalized, and accuracy cannot be a covenant on the larger dataset. Choudhury *et al.* [55] applied eight different deep learning pre-trained CNNs for the classification of CXR images having three classes named COVID-19, viral pneumonia, and normal, with a total of 423, 1485, and 1579 images for each class, respectively. The authors showed an accuracy of 97.74% by CheXNet for three classes with the equivalent precision, sensitivity, and F1-score of 96.61% and specificity of 98.31%. Hemdan *et al.* [106] utilized 50 CXR images with 25 confirmed COVID-19 and 25 normal for the classification using pre-trained deep CNNs and achieved the maximum accuracy of 90% using VGG16 and DenseNet201 models with a precision of 83%, recall of 100%, and F1-score of 91% for both the networks. Hussain *et al.* [58] developed a novel deep neural network (DNN) named CoroDet. The authors used CXR images having four classes named COVID-19, viral pneumonia (VP), bacterial pneumonia (BP), and normal, with an image size of 500, 400, 400, and 800 for each class, respectively. They performed the classification experiment into two-class (COVID-19 *vs*. normal), three-class (COVID-19, VP, and normal), and four-class (COVID-19, VP, BP, and normal) with the maximum accuracy of 99.1%, 94.2%, and 91.2% for each experiment respectively. Jain *et al.* [56] applied several pre-trained CNNs for the classification of CXR images into three classes: COVID-19, VP, and normal. They utilized 490 COVID-19 images and got the maximum accuracy of 97.97% using the Xception model. Mahdy *et al.* [107] recommended a deep CNN-based methodology for COVID-19 detection from chest X-ray images with an accuracy of 97.48%. Ioannis [108] *et al.* applied transfer-learning methods to classify CXR images into COVID-19, BP, and normal classes with 224, 700, and 504 images for each class, respectively. They got 96.7% accuracy,

98.66% sensitivity, and 96.46% specificity for the experiment. Sethy *et al.* [109] applied ResNet50 and SVM to classify CXRs into COVID-19, pneumonia, and normal classes. They obtained an accuracy of 95.33% for the three-class experiment. Ozturk *et al.* [110] introduced a novel network named DarkCovidNet. Using this network, the authors received an accuracy of 98.08% for two-class and 87.02% accuracy for three-class classification. Khan *et al.* [57] introduced a novel network, Coronet, inspired by Xception architecture. Using the Coronet model, the authors obtained an accuracy of 95% for three-class classification into COVID-19, VP, and normal. They also performed four-class classification into COVID-19, VP, BP, and normal with 89.6% accuracy. Wang *et al.* [111] introduced a novel DNN named COVID-Net for detecting COVID-19. The authors utilized 13,975 CXR images for the classification and achieved an accuracy of 83.5%. Afshar *et al.* [112] introduced COVID-CAPS, a capsule network to classify small-sized data of CXR images. The authors obtained an accuracy of 95.7% using COVID-CAPS. Yang *et al.* [62] applied transfer learning based on four different networks to classify CXR images into binary and three classes. The authors obtained an accuracy of 99% for binary (COVID-19 and pneumonia) and 97% accuracy for three-class (COVID-19, pneumonia, and normal) classification, both by the VGG16 network. Nayak *et al.* [54] performed binary classification into COVID-19 and normal class using 406 CXR images. Using the transfer learning method, the authors applied eight different pre-trained neural networks and obtained a maximum accuracy of 98.33% using the ResNet34 network. Regarding the fusion of machine learning and deep learning, Bhattacharya *et al.* [65] performed a three-class classification. This aimed to classify CXRs into COVID-19, pneumonia, and normal class. The authors obtained a maximum accuracy of 96.6% using a combination of VGG16 and a binary robust invariant scalable keypoints algorithm. Deb *et al.* [113] proposed a multi-model deep CNN ensemble architecture for the classification of CXRs into two classes (COVID-19 and non-COVID-19) and three classes (COVID-19, pneumonia, and normal). The authors obtained an accuracy of 98.58% for binary and 93.48% for the three-class experiment. Nikolaou *et al.* [61] developed a novel CNN by modifying pre-trained EfficientNetB0. This network was applied for the binary (COVID-19 and normal) and three-class (COVID-19, pneumonia, and normal) classification, obtaining an accuracy of 95% for binary and 93% for the three-class experiment. Oh *et al.* [53] introduced a patch-based DNN, where the network was applied for the four-class classification of CXRs into COVID-

19, BP, TB, and normal. Their database consisted of 502 images, of which 180 were COVID-19 images, and obtained a classification accuracy of 88.9%. AI-Timemy *et al.* [63] performed the five-class classification into COVID-19, VP, BP, TB, and normal class. They utilized 2,186, consisting of 435 COVID-19 images, for the experiment. The authors applied a combination of DL and ML methods and got 91.6% accuracy.

In conclusion, several recent studies have been reported for COVID-19 and other pneumonia classifications using CXR images. Most of them applied various CNN networks and achieved promising results. However, in most cases, the dataset used has a deficient number of images due to the scarcity of COVID-19 data. Hence, their results need to be verified on a larger dataset. Additionally, the classification into relevant multiclass (> three pneumonia) is rare. A rigorous experiment on classification for a larger dataset of COVID-19 and other similar lung disorders was required. In this study, we have designed and applied *seven* different deep-learning models utilizing the transfer learning method for the classification of four types of pneumonia, including COVID-19. We have used almost the largest data set of 18,603 CXR images, consisting of 3611 COVID-19, 1345 viral pneumonia, 2780 bacterial pneumonia, 700 tuberculosis, and 10,167 normal CXR images.

2.3 Methodology

We have designed and applied seven highly efficient pre-trained deep CNNs for the binary and multiclass classification of pneumonia diseases. The approaches we have opted for the experiment have been described in six subsequent sub-sections.

2.3.1 Dataset

In this experiment, 18,603 CXR images, including anterior-to-posterior (AP)/posterior-to-anterior (PA), were used. The dataset was prepared from three different publically available databases. COVID-19, viral pneumonia, and normal CXR images were taken from the Kaggle: "COVID-19 Radiography Database," [114], i.e., winner of the COVID-19 Dataset Award by Kaggle Community. The tuberculosis images were

taken from the Kaggle: “Tuberculosis (TB) Chest X-ray Database” [115]. Finally, the bacterial pneumonia images were taken from the Kaggle: “Chest X-Ray Images (Pneumonia)” [116].

COVID-19 Radiography Database

The COVID-19 radiography database includes CXR images of COVID-19 and viral pneumonia patients along with healthy persons. The dataset was created by different research groups and doctors in collaboration [55]. The first stage of release of the dataset had 219 COVID-19, 1341 normal, and 1345 viral pneumonia chest X-rays. After two updates, the current dataset has increased the number to 3,616 COVID-19, 10,192 normal, and 1,345 viral pneumonia images. The images were in Portable Network Graphics (PNG) file format with a resolution of 299x299 pixels. We have taken all the COVID-19, viral pneumonia, and normal images for our experiment.

Chest X-Ray Pneumonia Images

The Chest X-ray images (Pneumonia) dataset contains 5,863 CXR images with 2,780 bacterial pneumonia and the rest with viral pneumonia and normal images. The CXR images were taken from Guangzhou Women and Children’s Medical Center, Guangzhou, China [117]. The images were in JPEG format with variable resolutions. We have taken all the 2,780 bacterial pneumonia images for our experiment.

Tuberculosis Chest X-ray Database

The Tuberculosis Chest X-ray database contained CXR images of Tuberculosis patients along with the healthy person. The dataset was created by several research groups, along with the collaboration of medical doctors [118]. There are 700 Tuberculosis images in Portable Network Graphics (PNG) file format with a resolution of 512x512 pixels. We have taken all 700 Tuberculosis images for our experiment. Figure 2.2 shows the sample CXR images from each class. The images indicate that it is hard to manually determine the differences between them.

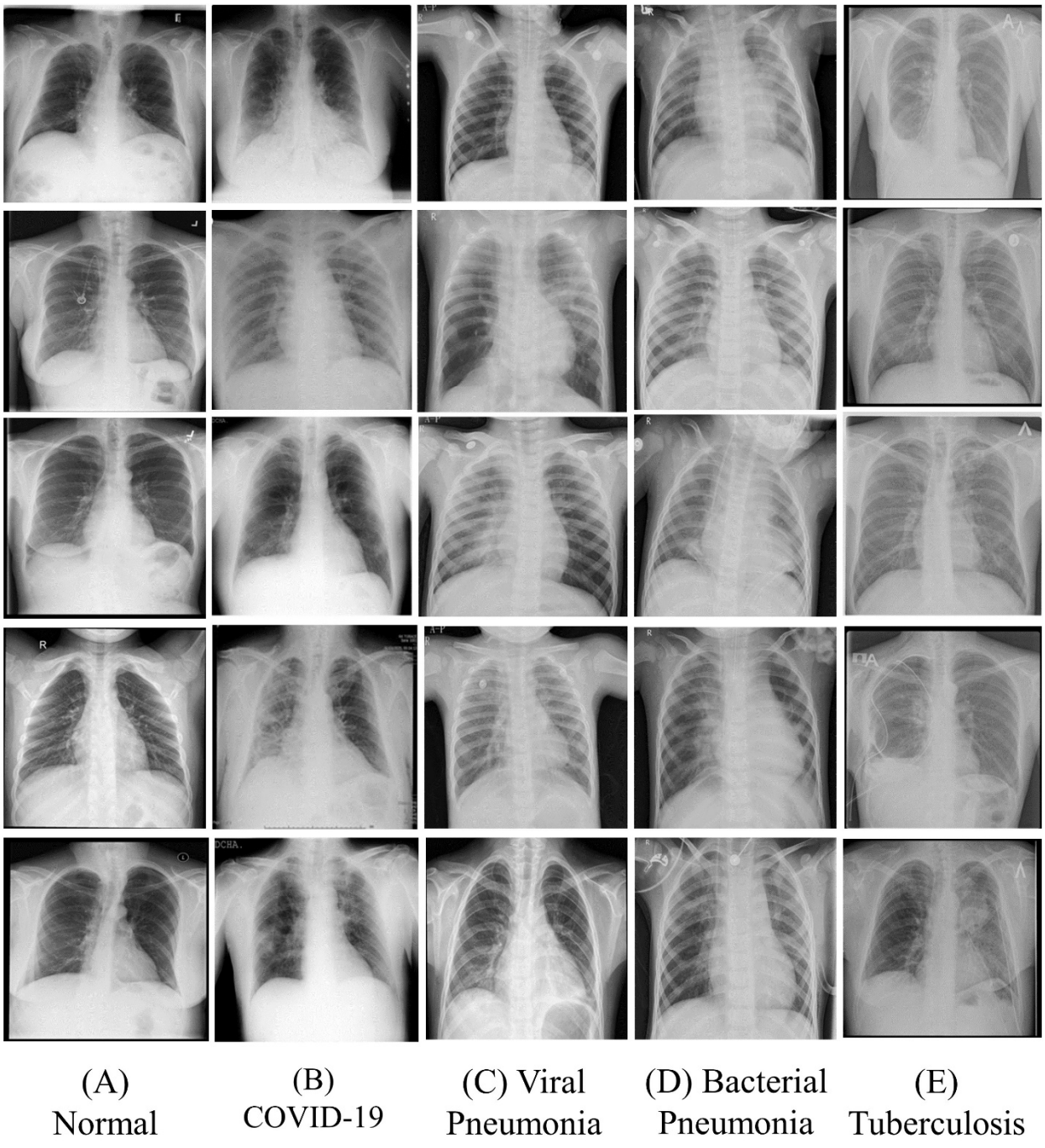


Figure 2.2: Sample chest X-ray images from each class.

2.3.2 Image Processing

All the CXR images collected from the different data sources were first converted into Portable Network Graphics (PNG) file format. Out of 18,632, 29 images, i.e., < 1%, were excluded from the experiment as outliers since they were missing details such as lung region. Some X-ray images having avoidable body

parts were cropped, displaying only the chest and lungs. Image augmentation was done for each image that participated in the training process. During image augmentation, shearing and zooming were applied to 20%, typically adapted in the imaging field [119, 120]. Images were resized to 224×224 pixels before the training process as required for the pre-trained model standards.

Finally, a total of 18,603 CXR images, including 3,611 COVID-19, 13,45 viral pneumonia, 2,780 bacterial pneumonia, 700 tuberculosis, and 10,167 normal images, were utilized for the experiments. Table 2.1 shows the experimental steps and class-wise distribution of the images. Out of the total, 80%, i.e., 14,879 images, including 2,887 COVID-19, 1,075 viral pneumonia, 2,224 bacterial pneumonia, 560 tuberculosis, and 8,133 normal images, were utilized for training the models. Next, 10%, i.e., 1,862 randomly selected images, including 362 COVID-19, 135 viral pneumonia, 278 bacterial pneumonia, 70 tuberculosis, and 1,017 normal images, were utilized for validation. Finally, 10%, i.e., 1,862 randomly selected images that were not involved in training or validation, were utilized to test the models. The test set included 362 COVID-19, 135 viral pneumonia, 278 bacterial pneumonia, 70 tuberculosis, and 1,017 normal images.

Table 2.1: Experimental steps and class-wise distribution of chest X-ray images.

Experimental Steps	Normal	COVID-19	Viral Pneumonia	Bacterial Pneumonia	Tuberculosis	Total
Training	8133	2887	1075	2224	560	14879
Validation	1017	362	135	278	70	1862
Testing	1017	362	135	278	70	1862

2.3.3 Experimental Setup

The whole experiment was organized into three phases. During the first phase of the experiment, we classified the images into two classes (i) COVID-19 and normal, (ii) COVID-19 and viral pneumonia, (iii) COVID-19 and bacterial pneumonia, and (iv) COVID-19 and tuberculosis. In the second phase of the experiment, we performed the three-class classification into viral diseases, i.e., COVID-19, viral pneumonia, and normal. In the third and final phase of the experiment, a five-class classification was done into viral and bacterial diseases, i.e., COVID-19, viral pneumonia, bacterial pneumonia, tuberculosis, and

normal. The experimental protocol consisted of 80% training, 10% validation, and 10% testing. The experiment was performed utilizing Python 3.8 on a computer with an Intel Core i7 8th Generation Processor, 16GB RAM, and 8GB NVIDIA Quadro P4000 graphics processing unit (GPU).

2.3.4 Model Architectures

Transfer learning is a machine learning approach in which a model developed for one job is used as the foundation for another task. It uses a trained model from a large dataset. Pre-trained weights are then used to train the network more quickly for an application with a smaller dataset. This eliminates the need for a large dataset and *shortens the training time* that a deep learning system requires when created from scratch. In this work, utilizing the transfer learning approach, we applied *seven* highly efficient pre-trained CNNs, namely VGG16, VGG19, Xception, InceptionV3, Densenet201, NasnetMobile, and Resnet152, for the experiment. The architecture of each network is shown in Figure 2.3 - 2.9. The Densenet offers a superior architectural design when it comes to the layering process. The feature maps of the preceding layers are utilized in all the subsequent layers. This reduces the complexity drastically, thereby improving the performance. In a conventional network, there are M connections for M layers, unlike in dense layers there are M(M+1)/2 direct connections, hence powerful and efficient. The loss function applied for two classes was binary cross-entropy, and for multiclass, it was categorical cross-entropy (CE). The activation function applied for the dense layer was sigmoid for binary and softmax for multiclass classification. The output layer was modified according to the number of classes. The models were trained for 25 epochs with a batch size of 16 images.

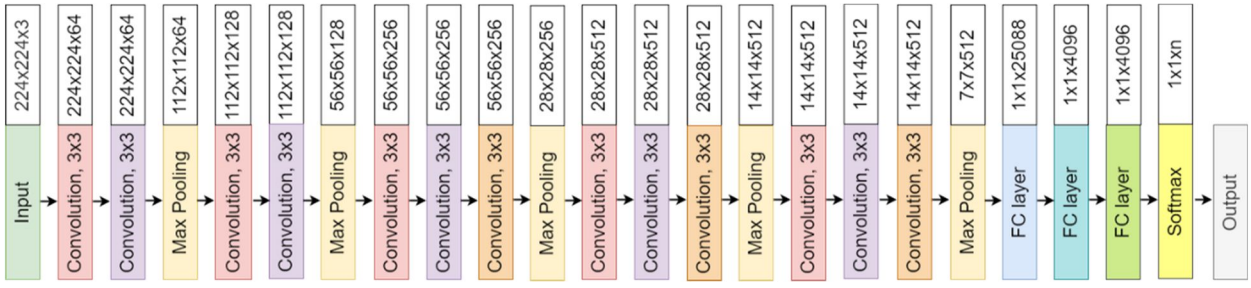


Figure 2.3: VGG16 architecture.

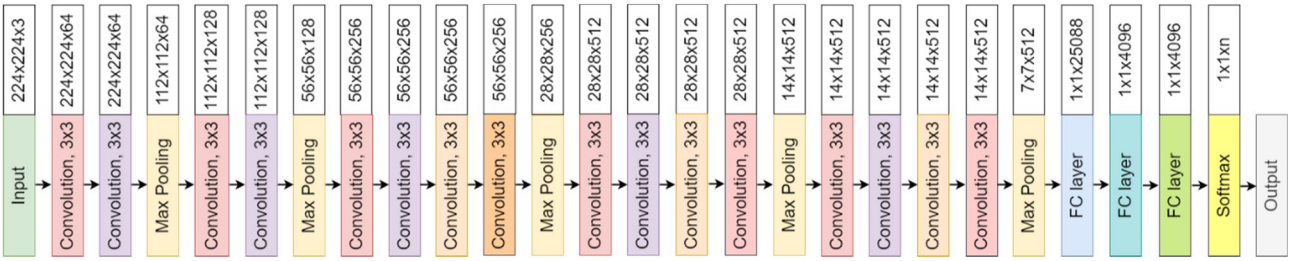


Figure 2.4: VGG19 architecture.

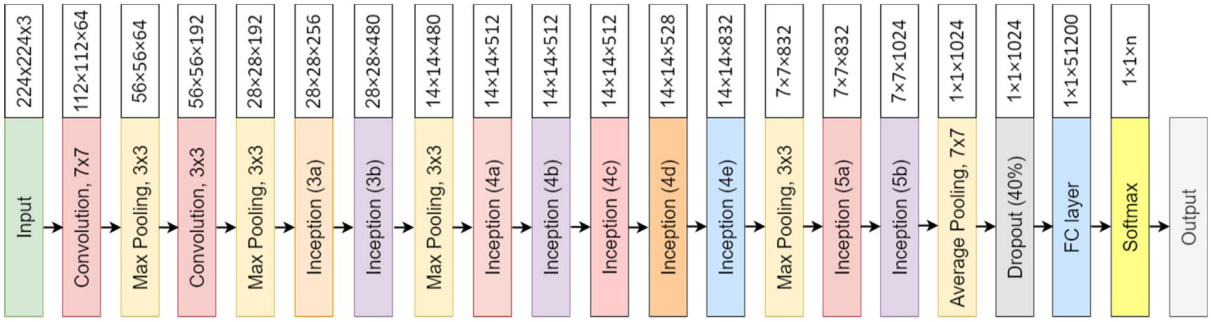


Figure 2.5: InceptionV3 architecture.

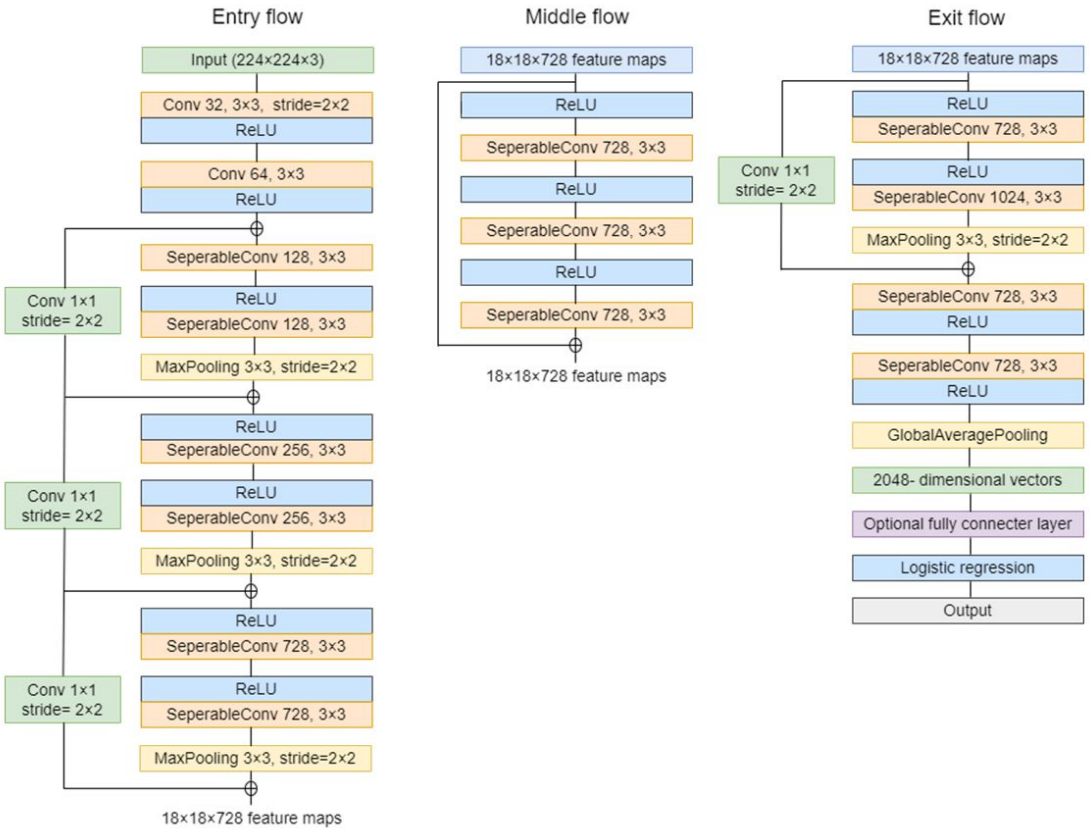


Figure 2.6: Xception architecture.

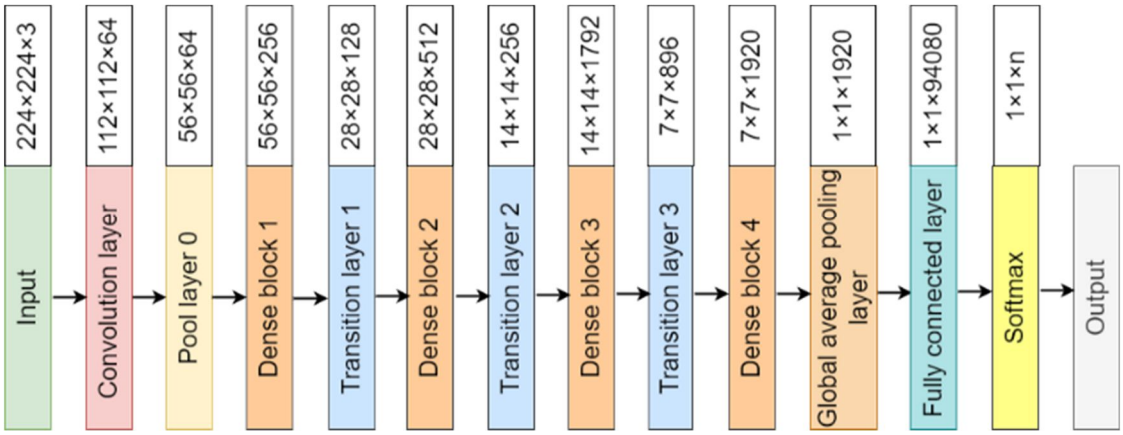


Figure 2.7: DenseNet201 architecture.

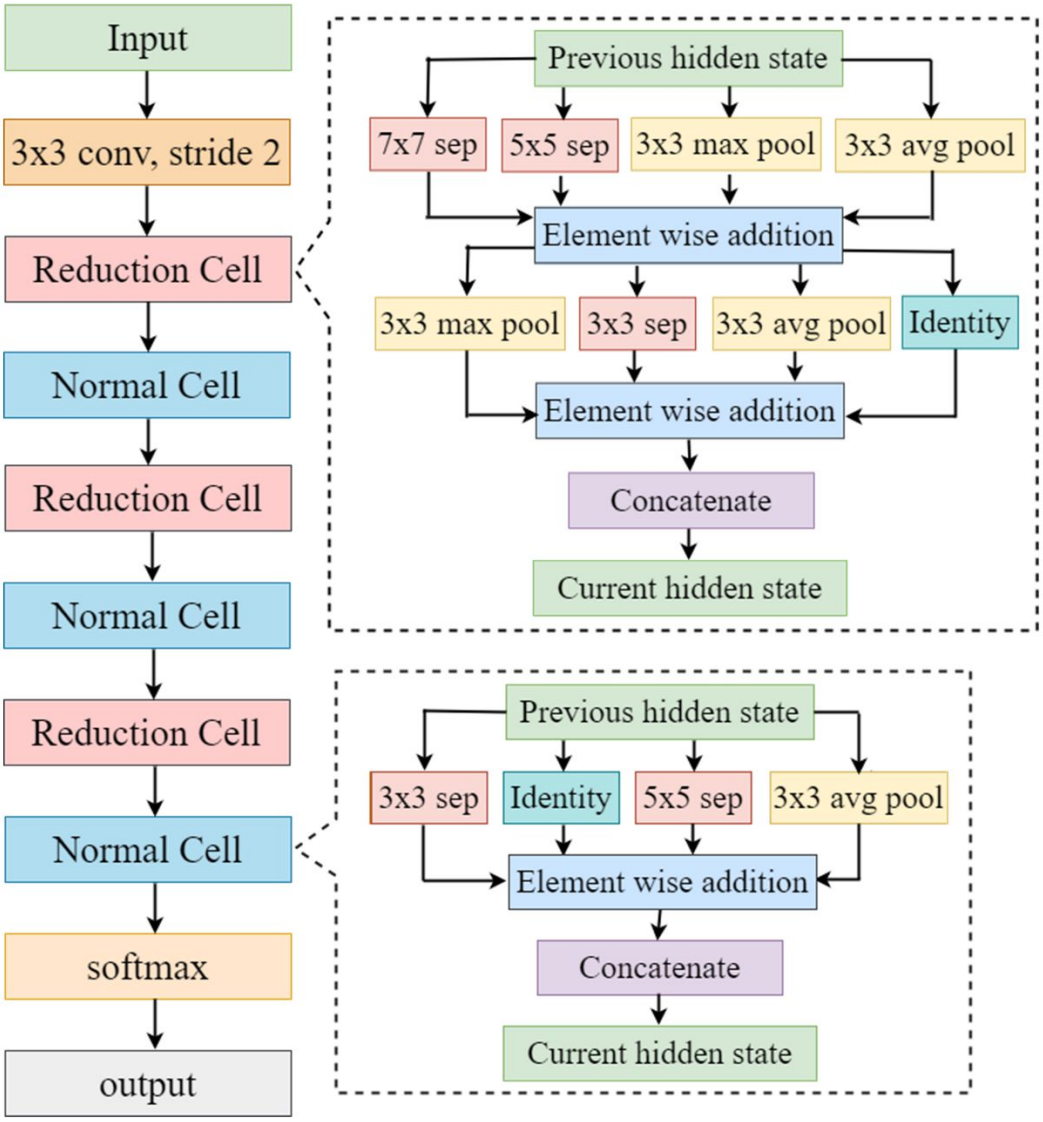


Figure 2.8 NasNetMobile architecture.

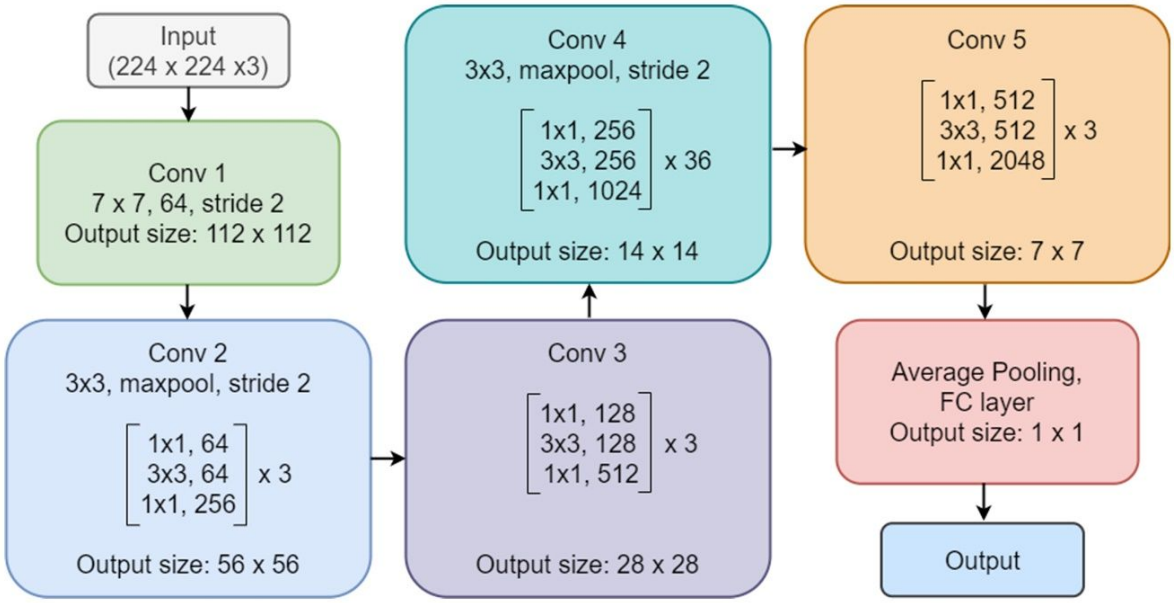


Figure 2.9: ResNet152 architecture.

2.3.5 Cross-Entropy Loss Function for models

The binary cross-entropy loss function can be defined as the following equation:

$$L_{BCE} = \frac{-1}{N}\sum_{i=1}^{N}[(y_i \times log\ a_i) + (1 - y_i) \times log(1 - a_i)] \qquad (2.1)$$

Here, y_i is the input GT label 1, (1-y_i) is GT label 0, a_i represents the Softmax classifier probability.

The categorical cross-entropy loss function can be defined as the following equation:

$$L_{CCE} = \frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} 1_{y_i \in C_c} \log a_{model}\,(y_i \in C_c) \qquad (2.2)$$

Here, N is the total number of observations (images), C is the number of categories or classes, $1_{y_i \in C_c}$ term indicates the i^{th} observation that belongs to the c^{th} category.

2.3.6 Performance Metrics used for Classification Evaluation

The following different matrices evaluated the performance of the proposed models:

(a) *Accuracy:* Accuracy is the most significant criterion for the analysis of the Convolutional Neural Network's performance. Accuracy is the sum of true positive and true negative values divided by the entire component of the confusion matrix. It is represented as given in (2.3)

$$Accuracy = \frac{True\ Positive + True\ negative}{Total\ number\ of\ cases} \quad (2.3)$$

(b) *Precision*: Precision is an important measure of the results of the CNN models. It counts how many correct positive predictions have been made. Precision is evaluated as the ratio between true positive predicted components and the sum of positive predicted components. It is represented as given in (2.4).

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (2.4)$$

(c) *Recall* (Sensitivity): Recall is another important metric for the analysis of the classifier's performance. It is defined as the ratio between the true positive predicted components and the sum of true positive and false negative predicted components. It is represented as given in (2.5).

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (2.5)$$

(d) *F1- score*: The F1-score is an important measure for assessing the test's accuracy. It is the harmonic mean between Precision and Recall. It is defined as twice the ratio between the multiplication of precision and recall and the sum of precision and recall. It is represented as given in (2.6).

$$F1 - score = \frac{2 * Precision * Recall}{Precision + Recall} \quad (2.6)$$

2.4 Results

Three different phases of the experiment were performed to compare the results of each classification possibility. In the first phase, we performed binary, then three-class, and finally, the five-class classification experiment.

2.4.1 Binary classification

The binary classification experiment deals with classifying images into COVID-19 and other classes separately. We endeavored to know how accurately the models could classify the images of different classes

from the COVID-19 class. The binary experiment was divided into four sub-phases: COVID-19 *vs*. normal, COVID-19 *vs*. viral pneumonia, COVID-19 *vs*. bacterial pneumonia, and COVID-19 *vs*. tuberculosis classification.

Binary Class Case 1: COVID-19 vs. Normal

The comparative performances of different CNNs for the binary classification into COVID-19 and normal images are shown in Table 2.2. VGG 16 network performed most efficiently with the highest accuracy, precision, recall, and f1-score among all networks. The VGG16 achieved a test accuracy of 97.24% with the weighted average of precision, recall, and f1-score of 97.26%, 97.24%, and 97.21%, respectively. The DenseNet201 performed as the second most efficient network with an accuracy of 96.01%. The performance of ResNet152 was the least efficient, with an accuracy of 78.75%. Figure 2.10 shows the training and validation accuracy, and Figure 2.11 shows the training and validation loss curve by the best-performing VGG16 model. The graphs indicate improved accuracy and reduced loss with successive epochs. Figure 2.12 shows the confusion matrix of test data classification by the VGG16 model. The confusion matrix specifies that out of 362 COVID-19 images, 331 were correctly classified, and 31 were misclassified as normal. Whereas out of 1017 normal images, 1010 were correctly predicted, and seven were misclassified as COVID-19 images.

Table 2.2: The weighted average of performance metrics by different deep learning models for COVID-19 and normal classification.

CNN models	Accuracy	Precision	Recall	F1-score
VGG16	97.24	97.26	97.24	97.21
VGG19	94.85	94.94	94.85	94.72
Xception	88.69	90.03	88.69	87.58
InceptionV3	93.33	93.32	93.33	93.32
DenseNet201	96.01	96.00	96.01	95.96
NasnetMobile	92.39	92.60	92.39	92.06
ResNet152	78.75	82.85	78.75	73.02

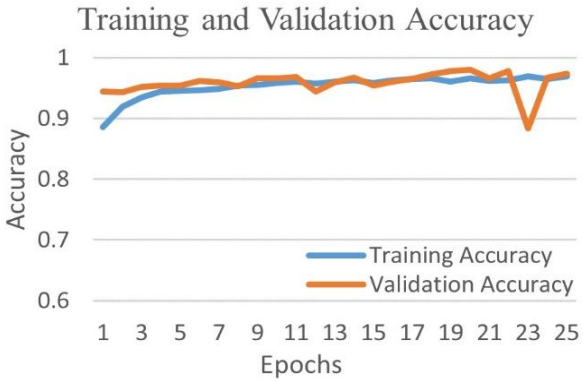


Figure 2.10: Training and validation accuracy curve by the best performing VGG16 network for COVID-19 and normal class.

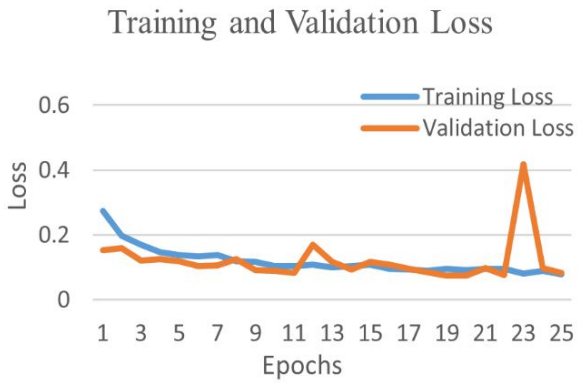


Figure 2.11: Training and validation loss curve by the best performing VGG16 network for COVID-19 and normal class.

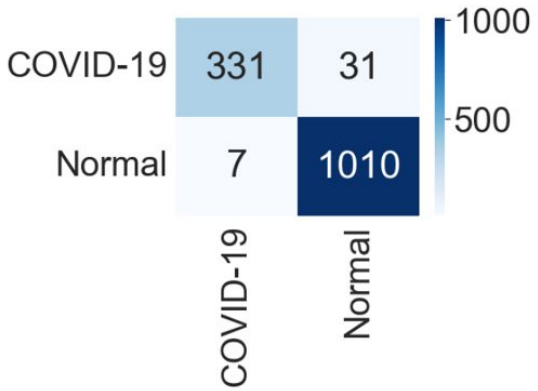


Figure 2.12: Confusion matrix for the classification into COVID-19 and normal by VGG16.

Binary Class Case 2: COVID-19 vs. Viral Pneumonia

Table 2.3 shows the comparative performances of different CNNs for binary classification into COVID-19 and viral pneumonia. The NasnetMobile network performed most efficiently with the highest accuracy, precision, recall, and f1-score among all networks. The model achieved an accuracy of 99.80% with the equivalent weighted average of precision, recall, and f1-score of 99.80% each. VGG16 model performed as the second most efficient network with an accuracy of 99.60%. The performance of the ResNet152 model was the least efficient, with an accuracy of 97.79%.

Table 2.3: The weighted average of performance metrics by different deep learning models for COVID-19 and viral pneumonia classification.

CNN models	Accuracy	Precision	Recall	F1-score
VGG16	99.60	99.60	99.60	99.60
VGG19	99.20	99.20	99.20	99.19
Xception	99.40	99.40	99.40	99.40
InceptionV3	98.99	99.01	98.99	99.00
Densenet201	99.40	99.40	99.40	99.40
NasnetMobile	99.80	99.80	99.80	99.80
Resnet152	97.79	97.80	97.79	97.77

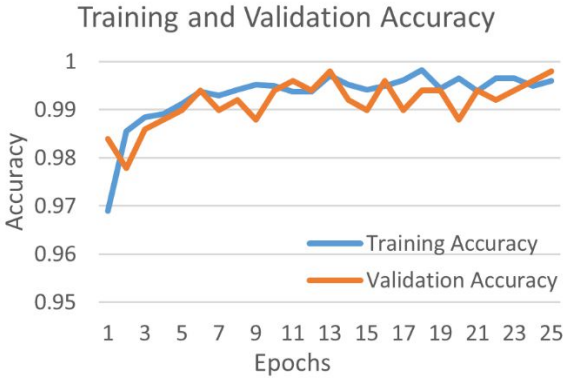


Figure 2.13: Training and validation accuracy curve by the best performing NasnetMobile model for COVID-19 and viral pneumonia class.

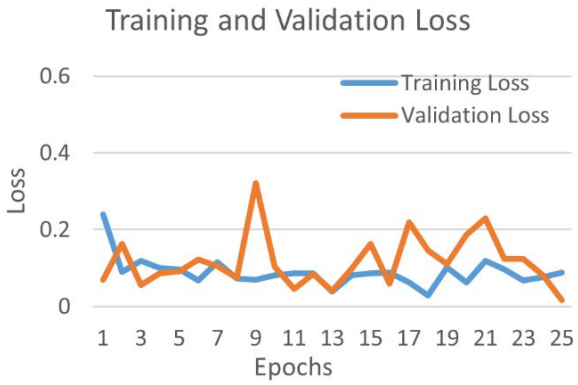


Figure 2.14: Training and validation loss curve by the best performing NasnetMobile model for COVID-19 and viral pneumonia class.

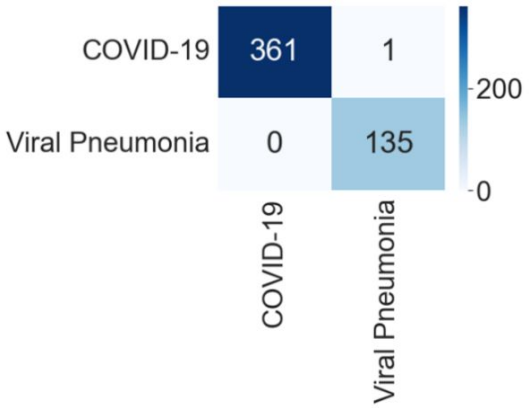


Figure 2.15: Confusion matrix for the classification into COVID-19 and viral pneumonia by NasNetMobile.

Figure 2.13 shows the training and validation accuracy, and Figure 2.14 shows the training and validation loss curve by the best-performing NasnetMobile model. The graphs specify accuracy improves and loss reduces with successive epochs. Figure 2.15 shows the confusion matrix of the test data classification by the NasNetMobile model. The confusion matrix reveals out of 362 COVID-19 images, 361 were correctly predicted, and one was misclassified to the viral pneumonia class. Further, our model correctly predicted all 135 viral pneumonia images without any false predictions.

Binary Class Case 3: COVID-19 vs. Bacterial Pneumonia

The comparative performance metrics of different CNNs for binary classification into COVID-19 and bacterial pneumonia are shown in Table 2.4. The DenseNet201 performed most efficiently with the highest

accuracy, precision, recall, and f1-score among all networks. The model achieved an accuracy of 99.84% and the equivalent weighted average of precision, recall, and f1-score of 99.84% each. The InceptionV3 and NasnetMobile performed as the second most efficient network with the equivalent accuracy of 99.53%. The ResNet152 performed the least efficiently, with an accuracy of 98.59%.

Table 2.4: The weighted average of performance metrics by different deep learning models for COVID-19 and bacterial pneumonia classification.

CNN models	Accuracy	Precision	Recall	F1-score
VGG16	99.22	99.22	99.22	99.22
VGG19	98.75	98.76	98.75	98.75
Xception	99.06	99.08	99.06	99.06
InceptionV3	99.53	99.53	99.53	99.53
Densenet201	99.84	99.84	99.84	99.84
NasnetMobile	99.53	99.53	99.53	99.53
Resnet152	98.59	98.60	98.59	98.59

Figure 2.16 shows the training and validation accuracy, and Figure 2.17 shows the training and validation loss curve by the best-performing DenseNet201 model. The graphs show accuracy improves, and loss reduces with successive epochs.

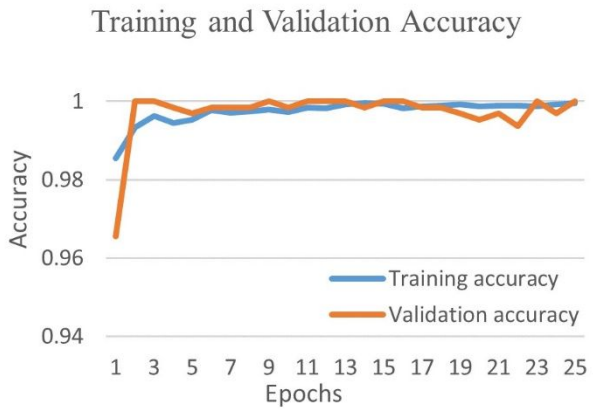


Figure 2.16: Training and validation accuracy curve by the best performing DenseNet201 model for COVID-19 and bacterial pneumonia class.

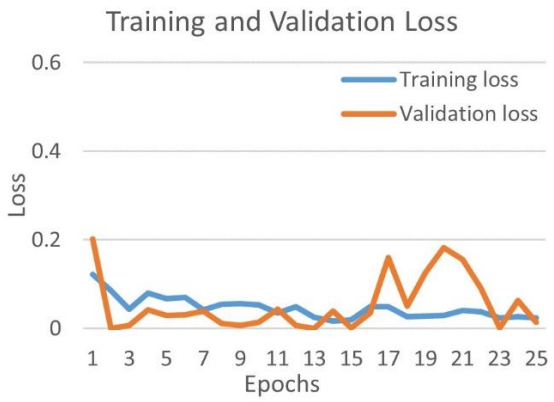


Figure 2.17: Training and validation loss of best performing Densenet201 model for COVID-19 and bacterial pneumonia class.

Figure 2.18 shows the confusion matrix of the test data classification by the DenseNet201 model. The confusion matrix specifies that out of 362 COVID-19 images, 361 were correctly predicted, and one image was misclassified to the bacterial pneumonia class. However, the model correctly predicted all 278 bacterial pneumonia images without any false predictions.

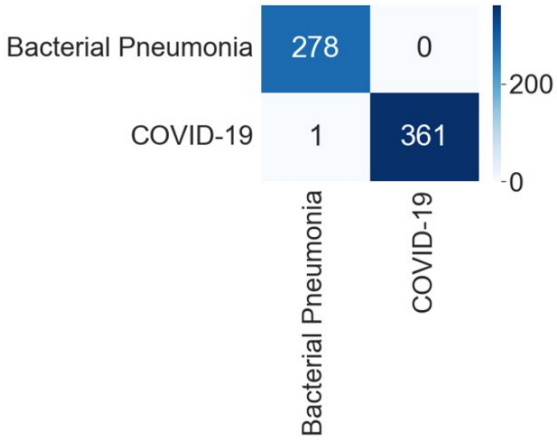


Figure 2.18: Confusion matrix for the classification into COVID-19 and bacterial pneumonia by DenseNet201.

Binary Class Case 4: COVID-19 and Tuberculosis

The comparative performance metrics of different CNNs for binary classification into COVID-19 and tuberculosis CXR images are shown in Table 2.5. VGG16 performed most efficiently with an accuracy of 99.31%, a weighted average of precision and recall of 99.31%, and an f1-score of 99.30%. VGG19 and

Xception both performed as the second most efficient models with the equivalent accuracy of 99.07%. ResNet152 performed the least efficiently, with an accuracy of 91.20%.

Table 2.5: The weighted average of performance metrics by different deep learning models for COVID-19 and tuberculosis classification.

CNN models	Accuracy	Precision	Recall	F1-score
VGG16	99.31	99.31	99.31	99.30
VGG19	99.07	99.07	99.07	99.07
Xception	99.07	99.07	99.07	99.07
InceptionV3	98.38	98.47	98.38	98.40
Densenet201	98.84	98.88	98.84	98.85
NasnetMobile	93.75	95.15	93.75	94.09
Resnet152	91.20	92.25	91.20	91.56

Figure 2.19 shows the training and validation accuracy, and Figure 2.20 shows the training and validation loss curve by the best-performing VGG16 model. The graphs indicate improved accuracy and reduced loss with successive epochs. Figure 2.21 shows the confusion matrix of the test data classification by the VGG16 model. The confusion matrix reveals the model correctly classified all 362 COVID-19 CXR images. Further, out of 70 tuberculosis CXR images, 67 were correctly predicted, and three were misclassified as COVID-19 images.

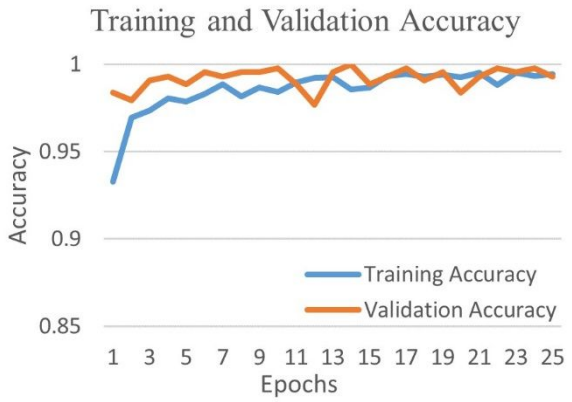


Figure 2.19: Training and validation accuracy curve by the best performing VGG16 model for COVID-19 and tuberculosis class.

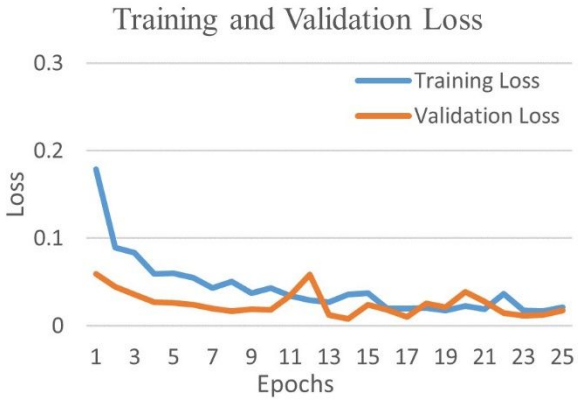


Figure 2.20: Training and validation loss curve by the best performing VGG16 model for COVID-19 and tuberculosis class.

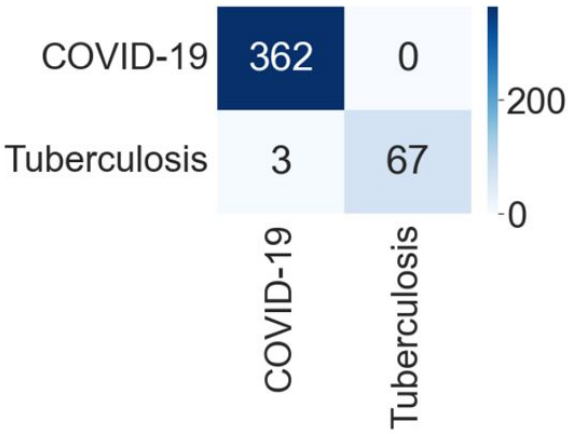


Figure 2.21: Confusion matrix for the classification into COVID-19 and tuberculosis by VGG16.

2.4.2 Three class classification into viral diseases

The comparative performance metrics of different CNNs for three-class classification into COVID-19, viral pneumonia, and normal images are shown in Table 2.6. VGG16 network performed most efficiently with an accuracy of 96.63% and an equivalent weighted average of precision, recall, and f1-score of 96.63% each. The DenseNet201 network performed as the second most efficient network with an accuracy of 95.51%. The performance of the ResNet152 model was the least efficient, with an accuracy of 77.21%.

Figure 2.22 shows the training and validation accuracy, and Figure 2.23 shows the training and validation loss curve by the best-performing VGG16 model. The graphs indicate improved accuracy and reduced loss with successive epochs.

Table 2.6: The weighted average of performance metrics by different deep learning models for three-class classification into COVID-19, viral pneumonia, and normal.

CNN models	Accuracy	Precision	Recall	F1-score
VGG16	96.63	96.63	96.63	96.63
VGG19	91.94	92.49	91.94	91.63
Xception	91.68	91.64	91.68	91.54
InceptionV3	92.54	92.47	92.54	92.43
Densenet201	95.51	95.61	95.51	95.44
NasnetMobile	92.93	93.32	92.93	92.96
Resnet152	77.21	84.70	77.21	78.57

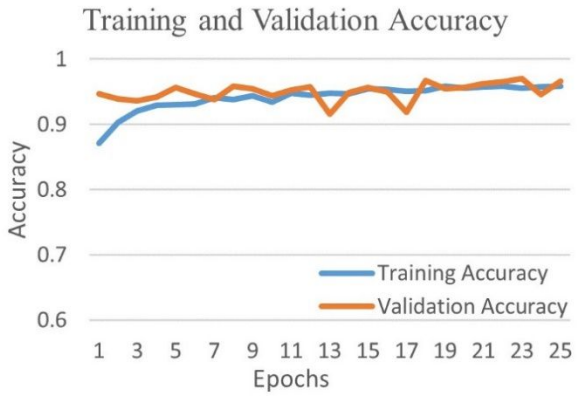


Figure 2.22: Training and validation accuracy curve by the best performing VGG16 model for three-class experiment.

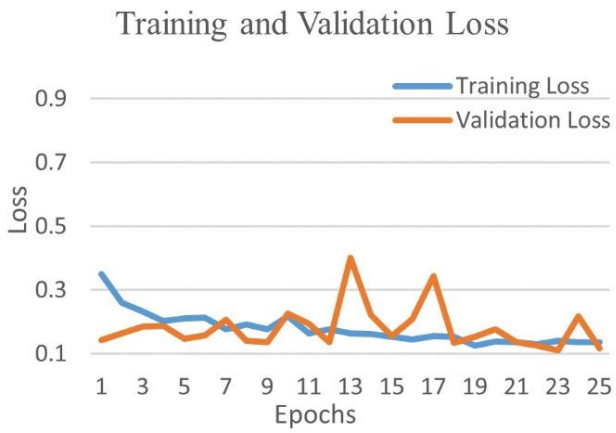


Figure 2.23: Training and validation loss curve by the best performing VGG16 model for the three-class experiment.

Figure 2.24 shows the confusion matrix of the test data classification by the VGG16 model. The confusion matrix specifies that out of 362 COVID-19 images, 339 were correctly classified, 23 were misclassified as 21 to normal, and two to the viral pneumonia class. Next, out of 1017 normal images, 994 were correctly predicted, 23 were misclassified as 18 to COVID-19, and five images to viral pneumonia class. Further, out of 135 viral pneumonia images, 130 were correctly classified, and five were misclassified as normal images.

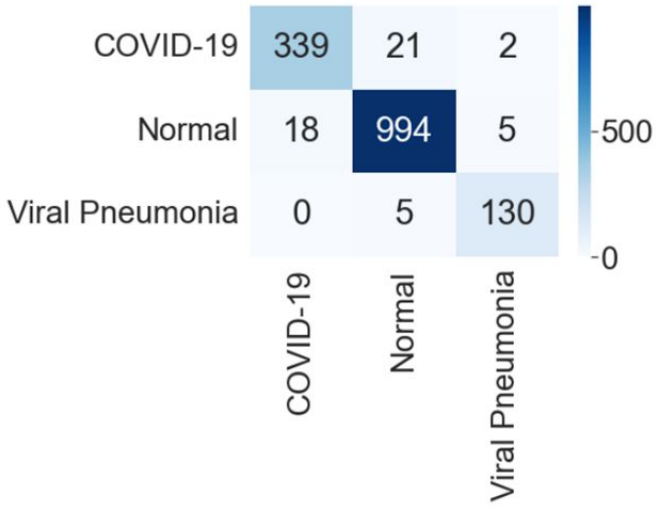


Figure 2.24: Confusion matrix for three-class classification by VGG16.

2.4.3 Five-class classification into viral and bacterial diseases

The comparative performance metrics of different networks for classification into five classes, COVID-19, viral pneumonia, bacterial pneumonia, tuberculosis, and normal images, are shown in Table 2.7. VGG16 model performed most efficiently with an accuracy of 92.70% and the weighted average of precision, recall, and f1-score of 92.41%, 92.70%, and 92.47%, respectively. The DenseNet201 performed as the second most efficient model with an accuracy of 89.10%. The performance of the ResNet152 network was the least efficient, with an accuracy of 74.70%.

Figure 2.25 shows the training and validation accuracy, and Figure 2.26 shows the training and validation loss curve by the best-performing VGG16 model. The graphs indicate accuracy improves, and loss reduces with successive epochs.

Table 2.7: The weighted average of performance metrics by different deep learning models for five-class classification into COVID-19, viral pneumonia, bacterial pneumonia, tuberculosis, and normal.

CNN models	Accuracy	Precision	Recall	F1-score
VGG16	92.70	92.41	92.70	92.47
VGG19	89.04	90.37	89.04	87.00
Xception	83.35	84.83	83.35	80.61
InceptionV3	84.00	85.54	84.00	83.44
Densenet201	89.10	89.80	89.10	88.42
NasnetMobile	87.76	88.05	87.76	86.65
Resnet152	74.70	76.80	74.70	71.60

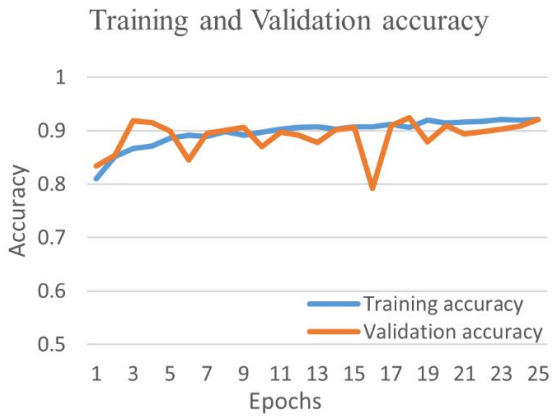


Figure 2.25: Training and validation accuracy curve by the best performing VGG16 model for five-class.

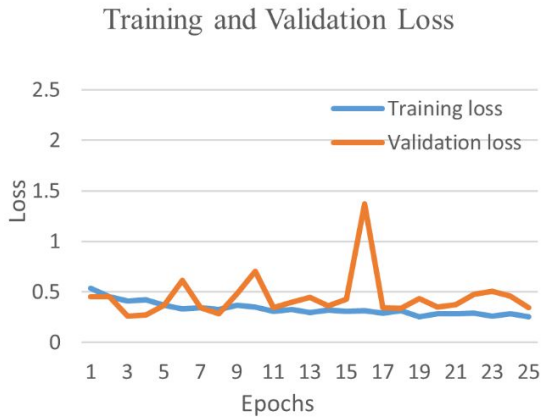


Figure 2.26: Training and validation loss curve by the best performing VGG16 model for five-class.

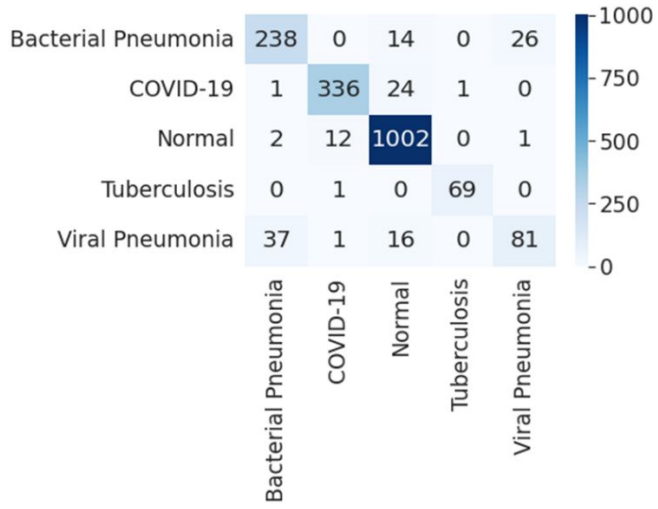
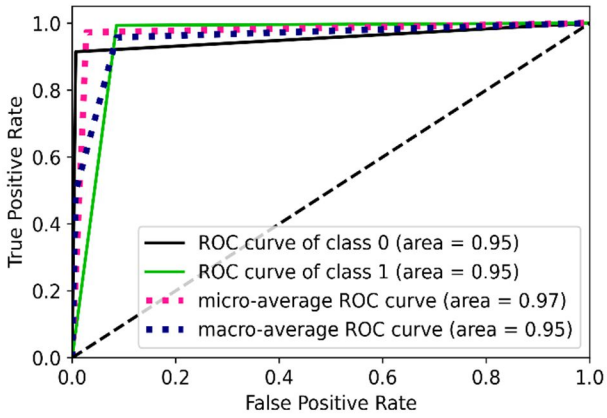


Figure 2.27: Confusion matrix for five-class classification by VGG16.

Figure 2.27 shows the confusion matrix of the test data classification by the VGG16 model. The confusion matrix reveals that out of 362 COVID-19 images, 336 were correctly predicted, and 26 were misclassified as 24 to normal, one to bacterial pneumonia, and one to tuberculosis class. Next, out of 278 bacterial pneumonia images, 238 were correctly classified, and 40 were misclassified as 14 normal and 26 viral pneumonia images. Further, out of 1017 normal, 1002 were correctly predicted, and 15 were misclassified as 12 COVID-19, two bacterial pneumonia, and one viral pneumonia image. Afterward, out of 70 tuberculosis images, 69 were correctly classified, and one image was misclassified to normal class. Finally, out of 135 viral pneumonia images, 81 were correctly predicted, 54 were misclassified as 37 bacterial pneumonia, 16 were normal, and 1 was a COVID-19 image.

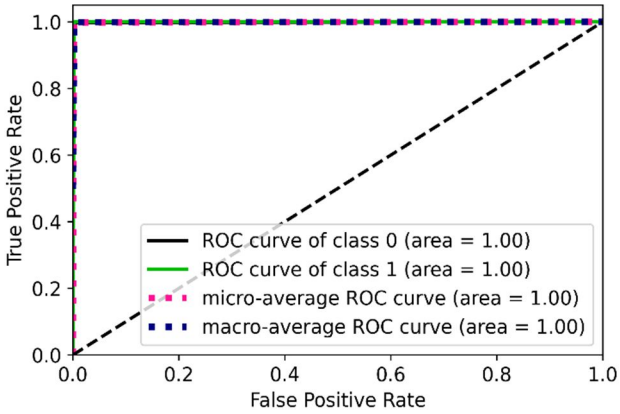
2.5 Performance Evaluation

We are able to design a multiclass system for COVID-19 classification and detection. The results of each experiment show very encouraging numbers. However, the system needs some performance evaluation to prove its robustness against all odds. Therefore, we obtained the receiver operating characteristic (ROC) curve and the Area-under-the-curve (AUC) for the best-performing model in all classification experiments.



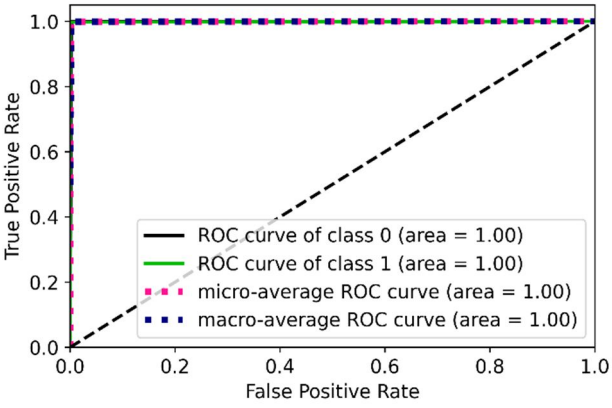
(p<0.0001; class 0: COVID-19; class 1: normal)

Figure 2.28: ROC curves and AUC values for binary classification into COVID-19 and normal by VGG16.



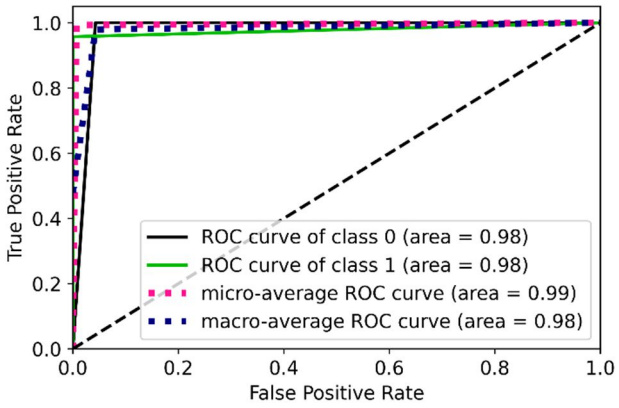
(p<0.0001; class 0: COVID-19; class 1: viral pneumonia)

Figure 2.29: ROC curves and AUC values for binary classification into COVID-19 and viral pneumonia by NasNetMobile.



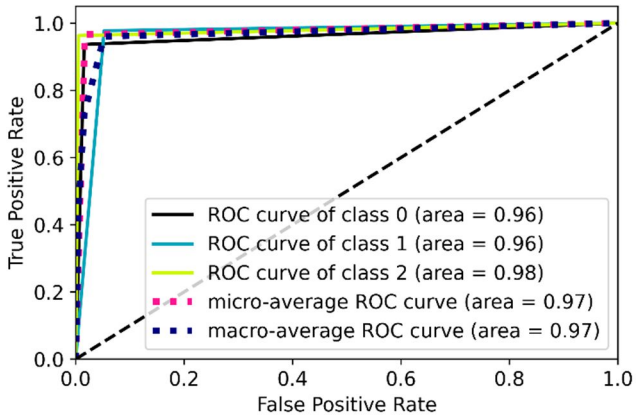
(p<0.0001; class 0: bacterial pneumonia; class 1: COVID-19)

Figure 2.30: ROC curves and AUC values for binary classification into COVID-19 and bacterial pneumonia by Densenet201.



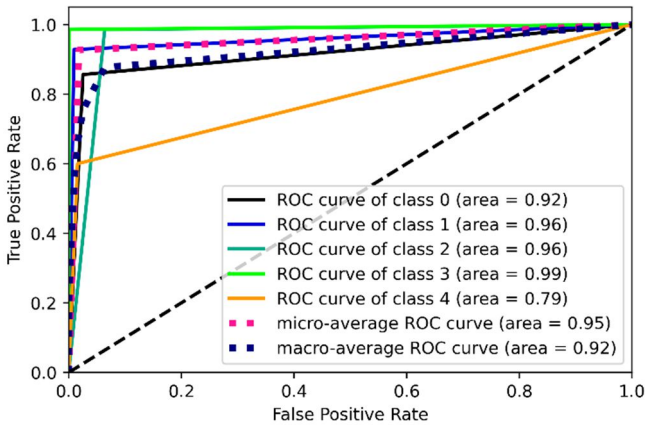
(p<0.0001; class 0: COVID-19; class 1: tuberculosis)

Figure 2.31: ROC curves and AUC values for binary classification into COVID-19 and tuberculosis by VGG16.



(p<0.0001; class 0: COVID-19; class1: normal; class 2: viral pneumonia)

Figure 2.32: ROC Curves and AUC values for three-class classification by VGG16.



(p<0.0001; class 0: bacterial pneumonia; class 1: COVID-19; class2: normal; class 3: tuberculosis; class4: viral pneumonia)

Figure 2.33: ROC Curves and AUC values for five-class classification by VGG16.

The ROC curves are drawn using each class's inference values and true labels. Figure 2.28 – 2.31 shows the four ROC curves and AUC values for best-performing models in two-class experiments. Figure 2.32 shows ROC curves and AUC values for the best-performing model (VGG16) in a three-class classification experiment. Similarly, Figure 2.33 shows ROC curves and AUC values for the best-performing model (VGG16) in five class classification experiments.

2.6 Scientific Validation

Scientific validation is a significant integrated part of the system design. The optimal model validation aims to ensure that the model is also functioning well and delivering comparable results on different dataset domains. In this work, we verified all our models on the facial biometric dataset named Faces95 from Libor Spacek's Facial Images Databases [121]. Several articles in the literature demonstrate the use of a well-known and standardized Faces95 database [122]. The database contains 72 individual images with various expressions and positions sat at a fixed distance from the camera. There are 72 classes for both men and women, with a total of 1,440 photographs. The sample images from the first eight classes are shown in Figure 2.34.

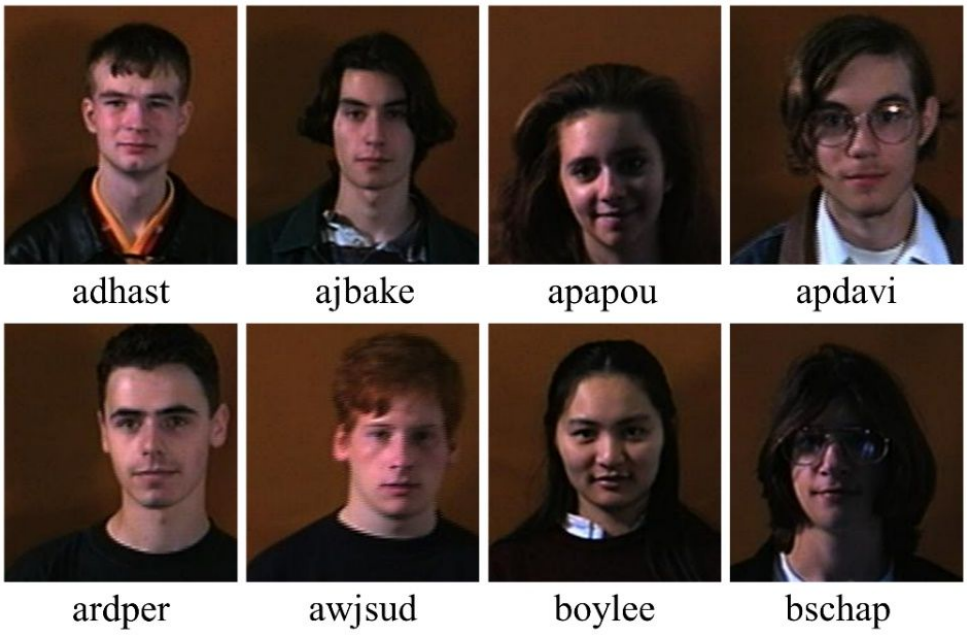


Figure 2.34: Sample images from the first eight classes of Faces95 database.

Table 2.8: The weighted average of performance metrics by different deep learning networks for facial image classification.

CNN models	Accuracy	Precision	Recall	F1-score
VGG16	98.61	99.07	98.61	98.52
VGG19	96.53	97.45	96.53	96.34
Xception	93.06	93.75	93.06	92.18
InceptionV3	95.83	97.22	95.83	95.56
DenseNet201	96.53	97.69	96.53	96.30
NasnetMobile	93.06	95.60	93.06	92.82
ResNet152	75.69	76.50	75.69	80.13

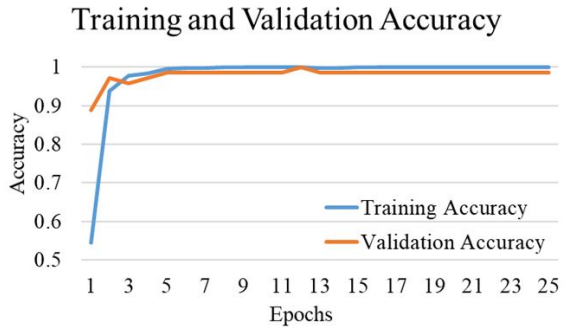


Figure 2.35: Training and validation accuracy curve by the best performing VGG16 model for Faces95 images.

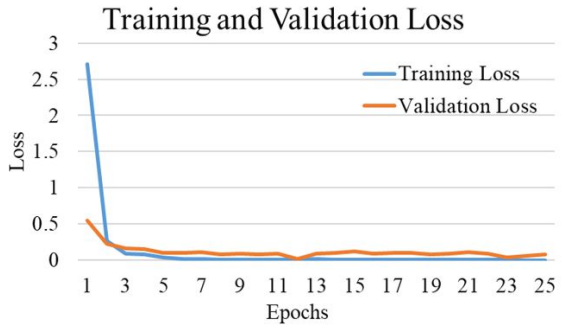


Figure 2.36: Training and validation loss curve by the best performing VGG16 model for Faces95 images.

The experiments were performed under a similar condition as CXR image classification. The loss function applied was categorical cross-entropy. The activation function used for the dense layer was softmax. The models were trained for 25 epochs with a batch size of 16 images. The training, validation, and testing were done on 80%, 10%, and 10% of the randomly selected images. The performances were also evaluated in terms of accuracy, precision, recall, and F1-score. Table 2.8 shows the comparative performance of the models. The VGG16 model performed most efficiently with an accuracy of 98.61% and precision, recall, and F1-score of 99.07%, 98.61%, and 98.52%, respectively. Figure 2.35 shows the training and validation accuracy, and Figure 2.36 shows the training and validation loss curve by the best-performing VGG16 model. The graphs indicate improved accuracy and reduced loss with successive epochs. The results support our system performing excellently on other datasets, along with the medical images, and providing outstanding results in each scenario.

2.7 Discussion

We have developed transfer learning-based deep learning models for the classification of chest X-ray images to detect COVID-19. We utilized 18,603 CXR images with 3,611 COVID-19 and the rest from viral pneumonia, bacterial pneumonia, and tuberculosis disease classes, along with normal images. We organized our experiment into three phases: (i) binary classification (COVID-19 and other classes separately), (ii) three-class classification into viral diseases (COVID-19, viral pneumonia, and normal), and (iii) five-class classification into viral and bacterial diseases (COVID-19, viral pneumonia, bacterial pneumonia, tuberculosis, and normal). We applied seven highly efficient pre-trained CNNs named VGG16, VGG19, DenseNet201, Xception, InceptionV3, NasnetMobile, and ResNet152 to achieve optimal performance in the classification of the CXR images.

2.7.1 Principal Findings

For the binary classification, we achieved the best performance by the DenseNet201 model with an accuracy of 99.84% for COVID-19 and bacterial pneumonia classification. Thereafter, the second-best performing model was NasnetMobile, which provided 99.80% accuracy for the classification of COVID-

19 and viral pneumonia. Finally, the VGG16 model performed third with 99.31% and 97.24% accuracy for the classification into COVID-19 *vs*. tuberculosis and COVID-19 vs. normal class, respectively. For three-class and five-class experiments, the VGG16 model performed best, with an accuracy of 96.63% and 92.70%, respectively. The AUC for binary classification was best for COVID-19 *vs*. viral pneumonia and COVID-19 *vs*. tuberculosis class, with a value of 1.0. Next, the AUC achieved for COVID-19 and tuberculosis was 0.98, and for COVID-19 vs. normal class was 0.95. Further, the AUC values achieved for three-class and five-class classifications were 0.97 and 0.92, respectively.

2.7.2 Benchmarking

Table 2.9 shows the benchmarking table presenting existing state-of-the-art classification methods and their comparison against the proposed method. Each row in the table shows different authors' work in this area, and the columns show the methods, number of X-ray images used, and results of the experiment. We used the highest number of images for our experiment than any other work in this area. To the best of our knowledge, our NasnetMobile model achieved the highest accuracy of 99.80% among all existing methods for binary classification of COVID-19 *vs*. viral pneumonia. Additionally, for the first time ever, we have performed the binary classification into COVID-19 *vs*. bacterial pneumonia and COVID-19 *vs*. tuberculosis disease classes with remarkable accuracy of 99.84% by Densenet201 and 99.31% by VGG16 model, respectively. Our results are very consistent with the previous studies on Densenet [123, 124]. These studies have shown superior performance of Densenet169 applied to COVID CT/X-rays. The key advantage of the Densenet is its ability to alleviate the fundamental problem of vanishing-gradient. As a result, the feature extraction process is boosted for its reuse, thereby reducing the number of parameters.

The CoroDet model by Hussain *et al.* [58] performed slightly better than our VGG16 model for binary classification between COVID-19 and normal. The authors achieved an accuracy of 99.1% in comparison to 97.24% by our model. However, our VGG16 model beat the CoroDet for three-class classification with an accuracy of 96.63% compared to 94.2%. Our model performed very close to the 97.97% accuracy by the best-performing Xception network for the three-class classification Jain et al. [56] applied. However, an advantage of our VGG16 network is that it is faster and takes less time for training.

Table 2.9: Benchmarking table showing state-of-art methods and comparing them against the proposed model.

Author & year	Method & Models	Number of images used	Classification accuracy				AUC[1]
			Two-class	Three-class[2]	Four-class[3]	Five-class[4]	
Nayak *et al.* (2020) [54]	Method: CNN with transfer learning Model: ResNet-34	C[5]: 203 Total: 406	C[5] & N[6]: 98.33%	NA[7]	NA	NA	C & N: 0.98
Chowdhury et al. (2020) [55]	Method: CNN with transfer learning Model: CheXNet	C: 423 Total: 3487	NA	97.74%	NA	NA	NA
Jain *et al.* (2020) [56]	Method: CNN with transfer learning Model: Xception	C: 490 Total: 6432	NA	97.97%	NA	NA	NA
Bhattacharyya *et al.* (2021) [65]	Method: ML[8] + DL[9] DL model: VGG-19 ML model: Random Forest	C: 342 Total: 1029	NA	96.6%	NA	NA	NA
Nikolaou *et al.* (2021) [61]	Method: CNN with transfer learning Model: EfficientNetB0	C: 3616 Total: 15153	C & N: 95%	93%	NA	NA	NA
Yang *et al.* (2021) [62]	Method: CNN with transfer learning Model: VGG16	C: 3616 Total: 8461	C & N: 98% C & VP[10]: 99%	97%	NA	NA	NA
Khan *et al.* (2020) [57]	Method: deep learning Model: Coronet (novel CNN)	C: 284 Total: 1251	NA	95%	89.6%	NA	NA
Hussain *et al.* (2020) [58]	Method: deep learning Model: CoroDet (novel CNN)	C: 500 Total: 2100	C & N: 99.1%	94.2%	91.2%	NA	NA
Oh *et al.* (2020) [53]	Method: CNN with transfer learning Model: ResNet-18	C: 180 Total: 502	NA	NA	88.9%	NA	NA
Timemy *et al.* (2021) [63]	Method: ML + DL DL model: ResNet-50 ML model: ESD[11]	C: 435 Total: 2186	NA	NA	NA	91.6%	NA
Proposed work (Nillmani *et al.*) [39]	Method: CNN with transfer learning Model: VGG16, NasnetMobile, DenseNet201	C: 3611 Total: 18603	C & N: 97.24%[12] C & VP: 99.80%[13] C & BP[14]: 99.84%[15] C & T[16]: 99.31%[12]	96.63%[12]	NA	92.70%[12]	C & N: 0.95[12] C & VP: 1.0[13] C & BP: 1.0[15] C & T: 0.98[12] Three-class[2]: 0.97[12] Five-class[4]: 0.92[12]

[1]Area under the ROC Curve; [2]COVID-19, viral pneumonia & normal; [3]COVID-19, viral pneumonia, bacterial pneumonia, & normal; [4]COVID-19, viral pneumonia, bacterial pneumonia, tuberculosis & normal; [5]COVID-19; [6]Normal; [7]Not applicable as authors have not performed such type of experiment; [8]Machine learning; [9]Deep learning; [10]viral pneumonia; [11] Ensemble Subspace Discriminant; [12]Acheived by VGG16; [13]Acheived by NasnetMobile; [14]bacterial pneumonia; [15]Acheived by DenseNet201; [16]tuberculosis.

Furthermore, for the five-class classification, our VGG16 model outperformed other existing models with an accuracy of 92.70%.

2.7.3 A special note on multiclass frameworks for pneumonia classification

Most classification experiments for COVID-19 detection have been done into binary or three classes. However, other than COVID-19, a wide range of pneumonia exists among the population, including viral, bacterial, and tuberculosis. Therefore, it is vital to distinguish COVID-19 from other diseases. A multiclass approach was apparently needed to classify COVID-19 from other pneumonia for the correct patient diagnosis. Our system is trained with the highest number of CXR images to date, includes most of the relevant pneumonia types, and is able to distinguish COVID-19 from other lung diseases with excellent accuracy.

2.7.4 Strengths, Weaknesses, and Extensions

The major strength of our system is its ability to detect COVID-19 very rapidly, and it takes just a few seconds to provide results. Further, the system is very cost-effective, as it requires only a patient's chest X-ray scan that is low-cost and readily available. Additionally, we have done six different types of classification experiments with consistently good accuracy that supports our system's robustness for practical applications.

One of the limitations of our system is its inability to detect the severity of the infection partially due to collimator noise. One can adopt denoising methods [125] as part of preprocessing. Further, predicting severity may help the physicians in treatment selection and thus in the fast and secure recovery of the patient. In addition, since we had a large database of CXR images, we did not perform k-fold cross-validation in which all images take part in training and testing at least once. In the extension of the work, we will make an effort to advance the system, as it could detect COVID-19 with the severity of the disease. In addition, we will include the heatmap images [126-128] of the disease, which will show the affected areas of the lungs. Broader advanced one-pass machine learning, like extreme learning machines [129] can be explored as more data is collected and pruning methods [130-132] to lower the storage and improve the

speed. This can also be extended for severity estimation [132] and the application of an advanced image analysis solution like stochastic imaging [133].

2.8 Summary

COVID-19 has become the prevalent challenge in the current scenario to save human lives. Several healthcare organizations are struggling to find the proper solutions. However, Artificial intelligence applications in Computer-Aided Diagnosis (CAD) have proven their efficiency and importance in resolving several medical problems. Due to various types of pneumonia, such as viral, bacterial, tuberculosis, and COVID-19, a system was apparently needed for multiclass classification as current methods offer less reliable solutions. In this work, we have designed and applied seven highly efficient pre-trained convolutional neural networks, namely, VGG16, VGG19, DenseNet201, Xception, InceptionV3, NasnetMobile, and ResNet152, for the classification of up to five classes utilizing a large database of chest X-ray scans. For the first time, we performed the binary classification into COVID-19 *vs*. bacterial pneumonia and COVID-19 *vs*. tuberculosis disease classes and achieved a powerful accuracy of 99.84% by Densenet201 and 99.31% by the VGG16 model, respectively. Our NasNetMobile and VGG16 models outperformed other existing methods for the binary (COVID-19 *vs*. viral pneumonia) and five-class (COVID-19, viral pneumonia, bacterial pneumonia, tuberculosis, and normal) classification with an accuracy of 99.80% and 92.70%, respectively. Performance with a remarkably high level of accuracy, the proposed models can provide an alternative to the current diagnostic methods for COVID-19 with a more accurate, cost-effective, and readily available system. The system may promisingly contribute to the fast diagnosis of patients, consequently lowering the medical load.