








Article

Multi-Scale CNN: An Explainable AI-Integrated Unique Deep Learning Framework for Lung-Affected Disease Classification

Ovi Sarkar ^{1,*} , Md. Robiul Islam ¹ , Md. Khalid Syfullah ¹ , Md. Tohidul Islam ² , Md. Faysal Ahamed ³ , Mominul Ahsan ⁴  and Julfikar Haider ^{5,*} 

- ¹ Department of Electrical & Computer Engineering, Rajshahi University of Engineering & Technology, Rajshahi 6204, Bangladesh; robiulruet00@gmail.com (M.R.I.); khalidsyfullah@gmail.com (M.K.S.)
² Department of Information & Communication Engineering, University of Rajshahi, Rajshahi 6205, Bangladesh; tanzid1971@gmail.com
³ Department of Computer Science & Engineering, International Standard University, Dhaka 1212, Bangladesh; faysalahamedjishan@gmail.com
⁴ Department of Computer Science, University of York, Deramore Lane, Heslington, York YO10 5GH, UK; mominul.ahsan2@gmail.com
⁵ Department of Engineering, Manchester Metropolitan University, Chester Street, Manchester M1 5GD, UK
* Correspondence: ovisarkareceian@gmail.com (O.S.); j.haider@mmu.ac.uk (J.H.)

Abstract: Lung-related diseases continue to be a leading cause of global mortality. Timely and precise diagnosis is crucial to save lives, but the availability of testing equipment remains a challenge, often coupled with issues of reliability. Recent research has highlighted the potential of Chest X-ray (CXR) images in identifying various lung diseases, including COVID-19, fibrosis, pneumonia, and more. In this comprehensive study, four publicly accessible datasets have been combined to create a robust dataset comprising 6650 CXR images, categorized into seven distinct disease groups. To effectively distinguish between normal and six different lung-related diseases (namely, bacterial pneumonia, COVID-19, fibrosis, lung opacity, tuberculosis, and viral pneumonia), a Deep Learning (DL) architecture called a Multi-Scale Convolutional Neural Network (MS-CNN) is introduced. The model is adapted to classify multiple numbers of lung disease classes, which is considered to be a persistent challenge in the field. While prior studies have demonstrated high accuracy in binary and limited-class scenarios, the proposed framework maintains this accuracy across a diverse range of lung conditions. The innovative model harnesses the power of combining predictions from multiple feature maps at different resolution scales, significantly enhancing disease classification accuracy. The approach aims to shorten testing duration compared to the state-of-the-art models, offering a potential solution toward expediting medical interventions for patients with lung-related diseases and integrating explainable AI (XAI) for enhancing prediction capability. The results demonstrated an impressive accuracy of 96.05%, with average values for precision, recall, F1-score, and AUC at 0.97, 0.95, 0.95, and 0.94, respectively, for the seven-class classification. The model exhibited exceptional performance across multi-class classifications, achieving accuracy rates of 100%, 99.65%, 99.21%, 98.67%, and 97.47% for two, three, four, five, and six-class scenarios, respectively. The novel approach not only surpasses many pre-existing state-of-the-art (SOTA) methodologies but also sets a new standard for the diagnosis of lung-affected diseases using multi-class CXR data. Furthermore, the integration of XAI techniques such as SHAP and Grad-CAM enhanced the transparency and interpretability of the model's predictions. The findings hold immense promise for accelerating and improving the accuracy and confidence of diagnostic decisions in the field of lung disease identification.



Citation: Sarkar, O.; Islam, M.R.; Syfullah, M.K.; Islam, M.T.; Ahamed, M.F.; Ahsan, M.; Haider, J. Multi-Scale CNN: An Explainable AI-Integrated Unique Deep Learning Framework for Lung-Affected Disease Classification. *Technologies* **2023**, *11*, 134. <https://doi.org/10.3390/technologies11050134>

Academic Editors: Yudong Zhang and Zhengchao Dong

Received: 30 August 2023

Revised: 18 September 2023

Accepted: 28 September 2023

Published: 30 September 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: COVID-19; chest X-ray (CXR) image; deep learning; multi-scale CNN; feature map; SHAP

1. Introduction

The field of medical image analysis has witnessed remarkable advancements in recent years, particularly in the context of diagnosing lung-related diseases. Among these, the severe acute respiratory syndrome caused by the coronavirus type 2 (SARS-CoV-2), commonly known as COVID-19, has posed unprecedented challenges to healthcare systems worldwide. Since its emergence in Wuhan, Hubei, China, in December 2019, COVID-19 has evolved into a global pandemic, with staggering statistics as of 30 July 2023—more than 768 million reported cases spanning 234 countries and over 6.9 million lives lost [1]. COVID-19 manifests with a spectrum of symptoms, including fever, cough, fatigue, shortness of breath, and a loss of taste and smell. Given the rapid spread of the virus, swift and accurate diagnosis is paramount in controlling its worldwide impact.

Concerning COVID-19 image classification, chest X-rays (CXRs) have emerged as a valuable tool, notably as the initial image-based strategy employed in countries like Spain [1]. When a patient is suspected of having COVID-19, a nasopharyngeal exudate sample is typically collected for reverse transcription-polymerase chain reaction (RT-PCR) analysis. Simultaneously, a chest X-ray is obtained to assess the patient's condition. The CXR plays a pivotal role in accelerating clinical evaluations, especially when PCR test results may only be available after several hours. In cases where both the clinical condition and CXR appear normal, patients may be discharged while waiting for the results of additional tests. However, if the CXR reveals abnormalities, the patient is often referred to a hospital for further evaluation.

In response to the global demand for lung-related disease testing, healthcare professionals have explored alternative diagnostic methods, particularly those relying on medical imaging techniques such as chest X-rays and computed tomography (CT) scans. These imaging modalities aid in confirming the presence of lung infection and tracking disease progression. Notably, when viral or bacterial infection affects the lungs, it manifests as distinctive radiological patterns, often referred to as ground-glass opacities (GGOs), visible in CXR images and chest CT scans.

Recent developments in deep learning (DL) have opened new avenues for predicting various lung-related diseases, including COVID-19 [2,3]. Researchers have leveraged DL-powered models to detect and classify these diseases [4,5]. Parallel to these developments, contemporary publications in system reliability research have provided rich insights, methodologies, and perspectives that can be thoughtfully integrated into the design and execution of deep learning models [6,7]. These insights contribute to embracing such models' robustness, efficiency, and reliability, when applied to the intricate domain of medical image analysis. However, existing multi-class classification models have exhibited limitations, characterized by reduced accuracy and complexity. The inherent complexity of these models has hindered their effectiveness in making precise diagnostic decisions. Existing methodologies struggle in accurate disease classification as the number of the disease class increases, impacting precision and recall rates.

To address these challenges and critical gaps in the existing research, an innovative DL architecture called a multi-scale CNN (MS-CNN) is presented. This model is specifically designed for the classification of multiple lung-related diseases, including COVID-19, bacterial pneumonia, viral pneumonia, fibrosis, lung opacity, tuberculosis, and normal cases. One of the key strengths of the proposed approach lies in its ability to maintain high accuracy, reliability, and efficiency even as the number of disease classes increase, overcoming a prevalent drawback in the existing literature. Another unique strength is that predictions from adjacent layers are carefully combined with the model's backbone, preventing the oversight of vital predictions in this innovative approach. Furthermore, it is worth noting that the proposed approach aims to significantly reduce testing time compared to the state-of-the-art (SOTA) models. This streamlined efficiency has the potential to achieve precise diagnostic results and expedite diagnostic processes particularly in real-world clinical scenarios, ensuring timely and effective medical interventions for patients with various lung-related diseases.

Moreover, XAI techniques such as SHAP and Grad-CAM have been integrated to visualize and identify the regions of CXR images that contribute most to the model's predictions. This further enhances the model's interpretability and provides valuable insights into its decision-making process. SHAP values provide insights into pixel contributions for each instance in the dataset, shedding light on the significance of different image regions in the model's decision. Grad-CAM generates heatmaps highlighting areas of interest within the images that the model relies on for classification. This additional layer of transparency enhances the reliability and trustworthiness of the deep learning model's outputs, making it a valuable tool in the clinical setting.

The major contributions of this study can be summarized as follows:

1. To create a comprehensive dataset encompassing seven distinct classes (COVID-19, normal, viral pneumonia, bacterial pneumonia, fibrosis, lung opacity, and tuberculosis), four publicly available datasets were combined.
2. An MS-CNN model is proposed to detect six lung-related disorders and healthy patients from the CXR images where predictions from different layers are combined, avoiding any instances of overlooking or omitting important predictions.
3. Predictions from several layers are concatenated to create a variety of feature maps that operate at various resolutions in order to improve the accuracy and effectiveness of multi-class predictions.
4. The performance of the proposed MS-CNN model is compared with popular TL models (VGG16 and VGG19) and other SOTA models proposed in the literature.
5. The XAI techniques were integrated to enhance the interpretability and trustworthiness of the model by providing visual insights into how the model makes predictions and highlighting the regions of importance in the chest X-ray images for different disease classifications.

The remaining parts of this paper are structured as follows. Section 2 reviews existing research works connected to this study. Section 3 explains the dataset collection and creation, preprocessing, suggested system architecture, hyperparameter settings and experimental settings, and performance metrics. Section 4 explains experimental results for dataset 1 to dataset 10 and explainable AI on multiscale-CNN interpretability using SHAP and Grad-CAM techniques. Section 5 presents discussions on comparative analysis with other published research and pre-trained models. Finally, the conclusions along with future research directions are drawn in Section 6.

2. Literature Review

Since the beginning of the COVID-19 catastrophe, investigators have developed several deep learning-based methods for accurately detecting COVID-19-positive patients using a variety of radiological imaging techniques, including CXR and CT scans. The investigation on COVID-19 diagnosis that predominantly relied on AI-based techniques, notably machine learning, and deep learning, are highlighted in this section.

To identify COVID-19 utilizing chest X-ray image classification, a deep CNN architecture was suggested by Reshi et al. [8]. The dataset used in the architecture was preprocessed using several methods throughout multiple stages, which involved balancing the dataset, having medical professionals analyze the photos, and enhancing the data. The trial outcomes demonstrated an astounding total accuracy of 99.5%, underscoring the suggested CNN model's outstanding performance in this application domain. The study by Muhammad et al. [9] presented a CNN model that had fewer model parameters but produced good accuracy. The model is made up of five primary convolution connection layers or blocks. With this model, a multi-layer fusion strategy is designed to increase the effectiveness of COVID-19 screening. Observations were made utilizing databases of lung ultrasound (LUS) images and videos that were freely available. The precision, accuracy, and retrieval rate of the suggested fusion method's data gathering were impressively high at 92.5%, 91.8%, and 93.2%, respectively. In COVID-19 screening, these efficiency metrics outperform those of current cutting-edge CNN models.

A controlled study by Mahajan et al. [10] investigated COVID-19 detection utilizing radiology-based images, specifically chest X-rays, and analyzed several detection models including VGG16, VGG19, Residual Network, and Dark-Net. For predictions, these models were compared using the Single Shot MultiBox Detector (SSD), augmented by task-specific preprocessing approaches such as CLAHE. Notably, the study indicates the efficacy of the DenseNet201 + SSD512 model, with precision and recall rates of 93.01 and 94.98, respectively.

A hybrid COVID–CheXNet model based on deep learning was developed by Al-Waisy et al. [11] to detect the COVID-19 virus in chest X-ray images. The method successfully identified COVID-19 patients with a detection accuracy rate of 99.99% demonstrating high confidence in distinguishing between healthy individuals and those infected with COVID-19 based on the X-ray images.

Srivastava et al. [12] introduced an innovative custom CNN-based CoviXNet model. This model comprises 15 carefully designed layers, emphasizing the efficiency of the architecture. Their research showcased CoviXNet’s exceptional performance in binary classification tasks related to COVID-19 detection. Notably, the model attained an accuracy rate of 99.47%, highlighting its potential as a powerful tool for diagnosing COVID-19 in medical imaging.

Nahiduzzaman et al. [13] developed a method for detecting COVID-19 cases among various lung diseases. A three-class classification approach specifically designed to identify COVID-19 cases from pneumonia and normal cases. To achieve this, the authors employed a CNN-ELM model and achieved 97.42% accuracy. CNN-ELM utilized a dataset of 12,701 samples with 512 features for model training. Additionally, 3176 data points were used to assess the model’s performance.

Yaman et al. [14] introduced the ACL model, combining attention, LSTM, and CNN for classifying healthy, COVID-19, and pneumonia cases in chest X-ray (CX-R) images. Marker-controlled watershed segmentation emphasized crucial features. The model achieved 96% accuracy with an 80:20 training:testing ratio.

A 2D-CNN model was designed to classify instances of bacterial pneumonia, COVID-19, and normal instances by Abida et al. [15]. The proposed model demonstrated high performance, achieving an impressive accuracy of 97.49%. The model was also modified for five classes (bacterial pneumonia, COVID-19, fibrosis, normal, and tuberculosis) and six classes (bacterial pneumonia, COVID-19, fibrosis, normal, tuberculosis, and viral pneumonia) and secured an accuracy of 97.81% and 96.75%, respectively. This study’s findings showcase the potential of the 2D-CNN approach for the accurate and efficient classification of different lung conditions, contributing to the field of medical imaging and disease diagnosis.

Elakkiya et al. [16] presented a novel approach for categorizing various diseases, including COVID-19, pneumonia, tuberculosis, and other specific conditions. They introduced the sharpened cosine similarity network (SCS-Net), which stands out from traditional neural networks by utilizing sharpened cosine similarity instead of dot products. In their experiments involving multi-class classification combining classes such as COVID-19, normal, pneumonia, and tuberculosis, the proposed SCS-Net demonstrated an accuracy rate of 94.05%.

Hussain et al. [17] introduced a novel CNN model named CoroDet. The primary objective of this model was to facilitate the automatic detection of COVID-19 through the utilization of raw chest X-ray and CT scan images. In their research, the authors comprehensively evaluated CoroDet’s performance, employing a four-class classification approach involving categories such as bacterial pneumonia, COVID-19, normal, and viral pneumonia, achieving an accuracy rate of 91.20%.

Al-Timemy et al. [18] presented a pipeline for classifying five classes using a combination of ResNet-50 for DF (deep features) computation and an ensemble of subspace discriminant classifiers. Through their research, this pipeline emerged as the top performer

in accurately classifying the five classes with an accuracy of 91.6% and a 95% confidence interval of 2.6%.

Some of the recent developments in CXR and CT scan dataset-based research that utilized deep learning approaches similar to the proposed work are analyzed in this section. Ghoshal and Tucker [19] developed a Bayesian convolutional neural network (BCNN) to assess uncertainties and interpretability of coronavirus identification using COVID-19 CXR images. The results demonstrate that the pre-trained VGG-16 model significantly increased detection accuracy from 85.2% to 92.9%. By developing saliency maps to understand the suggested model's outcomes better, they also established the approach's interpretability. Narin et al. [20] provided a transfer learning-based method for classifying CXR pictures into COVID-19 and normal categories, using three pre-trained models, with ResNet50 achieving the highest accuracy. Oh et al. [21] developed a patch-based approach for training and fine-tuning the ResNet18 CNN model. Jain et al. [22] used X-ray images and transfer learning-based algorithms for COVID-19 screening and found that the Xception model achieved the highest accuracy of 97.97%. Hoon et al. [23] developed a decision tree classifier based on deep learning for COVID-19 screening, achieving a 95% accuracy rate for categorizing coronavirus patients. Pereira et al. [24] proposed a deep learning-based system that used a radiography image data augmentation approach for COVID-19 identification, achieving an F1-score of 0.89. Sakib et al. [25] used a generic augmentation method and GAN to create artificial COVID-19 pictures, achieving a test data accuracy of 93.94%.

Makris et al. [26] conducted a study in which they offered numerous CNN models with transfer learning techniques to categorize three distinct categories. According to their observations, VGG16 had the maximum accuracy, with a score of 95.88%. Then, in a study by Khalid El Asnaouia et al. [27], numerous pre-trained CNN models were proposed to categorize three separate classes. According to their results, Inception ResnetV2 had the best accuracy of 92.18%. Furthermore, Saiz et al. [28] suggested a CNN VGG16 approach utilizing CLAHE. As per the study's findings, utilizing CLAHE on the database led to an accuracy level of 94% rather than an accuracy rate of 83% without such an approach. COVIDXNet, a deep learning framework proposed by the authors of [29], can facilitate radiologists in automatically diagnosing COVID-19. The suggested framework included seven distinct architectures, including a modified VGG19 and Google MobileNet's second version. Rahimzadeh et al. [30] provided a method for classifying X-ray images into three groups based on two publicly accessible datasets. They also showed how Xception and ResNet50V2 might be used to enhance classification accuracy.

While many studies have reported impressive accuracy in binary and limited-class classification scenarios, their performance consistently degrades as the number of classes increases. This phenomenon arises due to the increasing complexity of distinguishing between multiple conditions with features having minute differences. This limitation hampers the applicability of these models in real-world clinical applications where patients may exhibit diverse lung conditions. Therefore, a tailor-made and robust deep learning framework is required to perform multi-class classification of lung diseases with high accuracy and confidence for real-life scenarios.

3. Methodology

Figure 1 shows the general workflow of the proposed research work. A larger dataset with seven classes was produced by combining CXR images from publicly available sources. The dataset was split into three separate datasets for three different operations—80% of the original for training, 10% for validation, and 10% for testing. After that, the training data were appropriately preprocessed through resizing, rescaling, and augmentation. Preprocessing was performed after data splitting to ensure that information from the validation and test sets did not influence the preprocessing decisions made on the training set. This helps to maintain the integrity of the evaluation process, as the validation and test sets should represent real-world data that the model will encounter in general. Each class comprises 950 images, 760 for training, 95 for validation, and 95 for testing. To achieve the

optimum outcome, various hyperparameters were utilized. Additionally, binary, three-class, four-class, five-class, six-class, and seven-class datasets were trained with the MS-CNN model. Finally, the model's effectiveness was demonstrated by a comparative analysis using a variety of performance metrics.

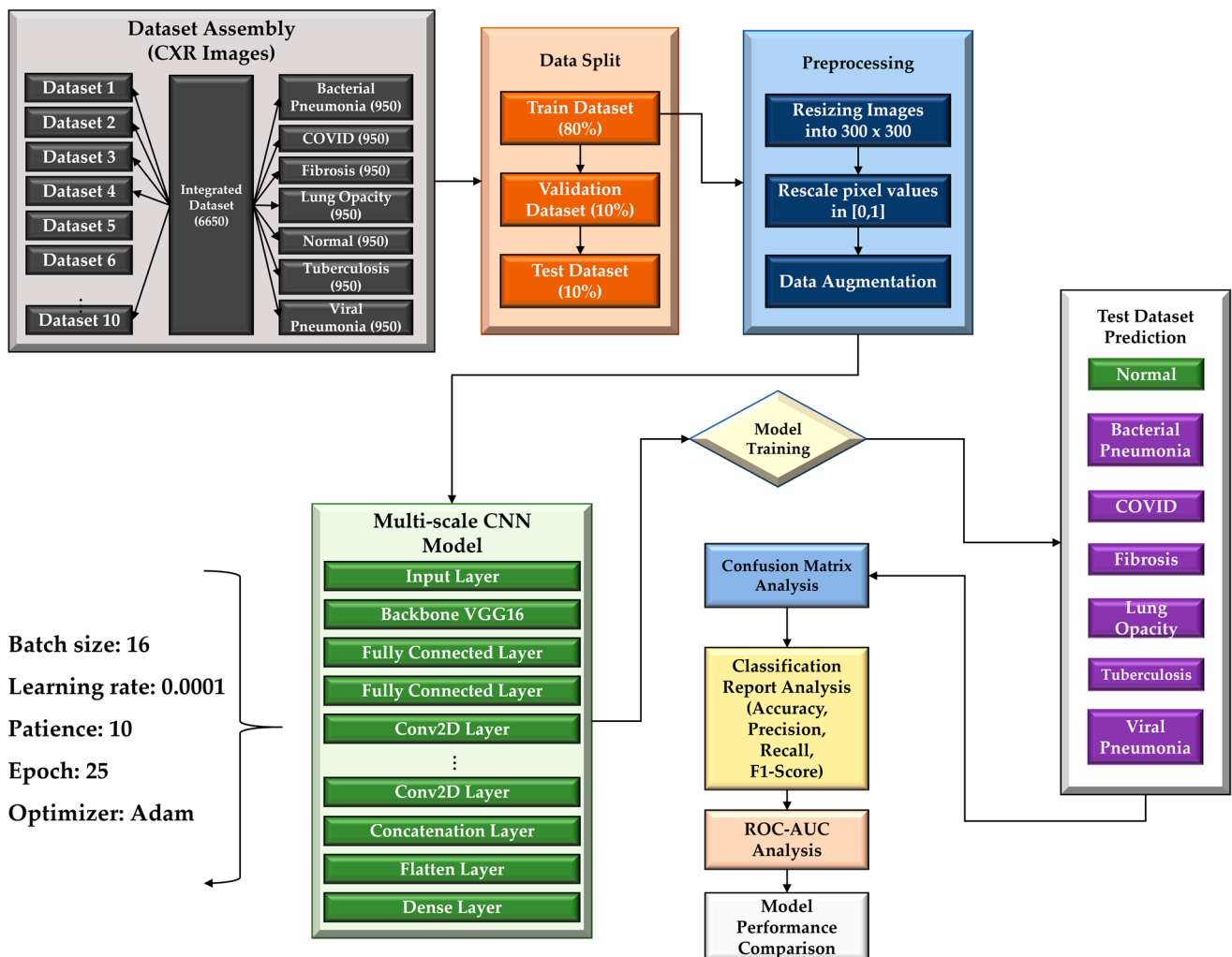


Figure 1. A schematic of the overall Multi-scale CNN system architecture.

3.1. Chest X-ray Databases

Most of the datasets used in this investigation were acquired from four distinct reputable sources. Figure 2 shows examples of Chest X-ray images of Bacterial Pneumonia, COVID-19, fibrosis, lung opacity, tuberculosis, viral pneumonia, and normal subjects utilized in the proposed work. The following public datasets of CXR images were used in this study: (1) COVID-19 Radiography Database ¹ (accessed on 16 February 2023) [31], (2) Curated Dataset for COVID-19 ² (accessed on 16 February 2023) [32], (3) NIAID TB dataset ³ (accessed on 12 May 2023) [33], and (4) NIH Chest X-ray Dataset ⁴ (accessed on 9 August 2023) [34]. The datasets are utilized in this study in the following manner

1. COVID-19 Radiography Database [31]: this dataset (available online: <https://www.kaggle.com/datasets/tawsifurrahman/covid19-radiography-database> (accessed on 16 February 2023)) provided Chest X-ray images for COVID-19-positive cases, viral pneumonia cases, lung opacity cases, and normal cases.
2. Curated Dataset for COVID-19 [32]: this dataset (available online: <https://www.kaggle.com/datasets/francismon/curated-covid19-chest-xray-dataset> (accessed on 16 February 2023)) contributed images of bacterial pneumonia.

3. NIAID TB dataset [33]: this dataset (available online: <https://tbportals.niaid.nih.gov/> (accessed on 12 May 2023)) supplied images of tuberculosis.
4. NIH Chest X-ray Dataset [34]: this dataset (available online: <https://datasets.activeloop.ai/docs/mL/datasets/nih-chest-x-ray-dataset> (accessed on 9 August 2023)) provided images of fibrosis.

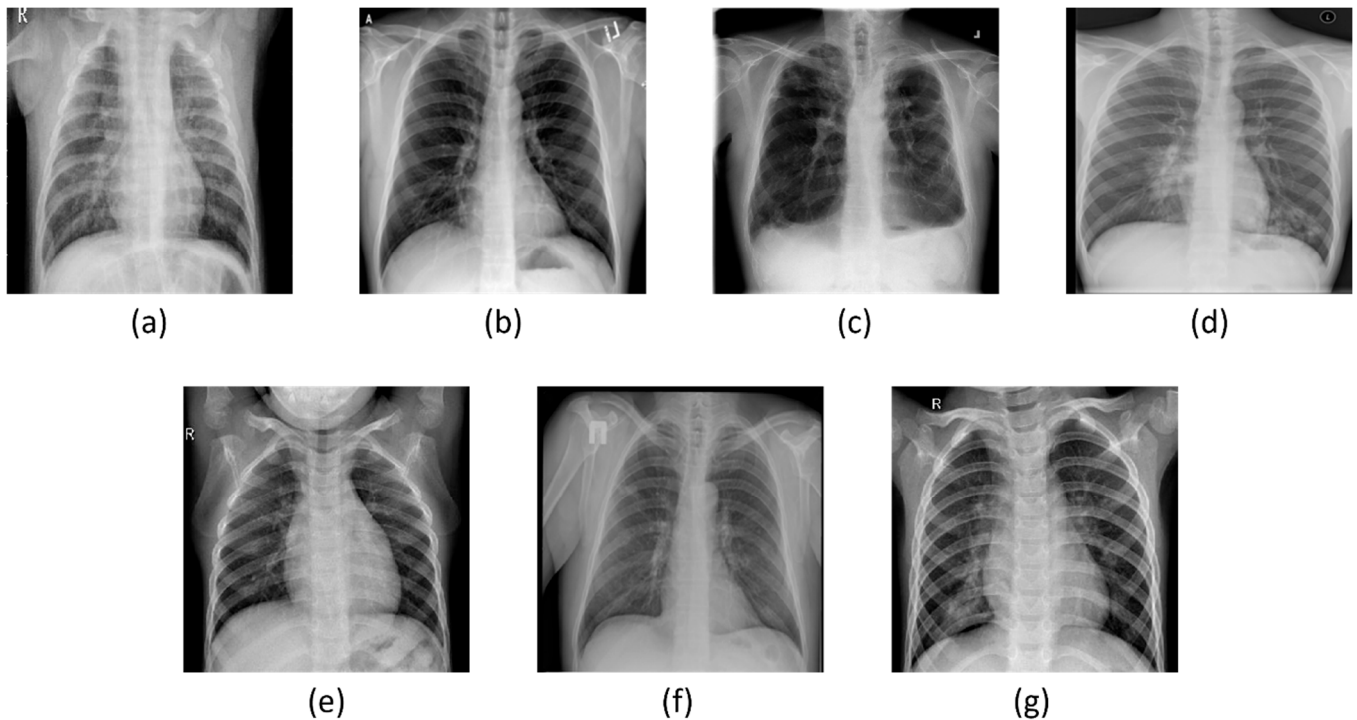


Figure 2. Sample Images of the dataset: (a) Bacterial Pneumonia, (b) COVID-19, (c) Fibrosis, (d) Lung Opacity, (e) Normal, (f) Tuberculosis, and (g) Viral Pneumonia.

3.1.1. Dataset 1

This database contains 1900 CXRs, with the images evenly divided between COVID-19 patients and healthy participants. The COVID-19 Radiography Database [31] obtains all CXRs from affected and healthy individuals. This dataset is intended to be split into two categories.

3.1.2. Dataset 2

This database contains 2850 images, 950 of which are COVID-19 images, 950 of which are Normal images, and 950 of which are Fibrosis images. The COVID-19 Radiography Database was used to obtain COVID-19 and healthy person images [31]. The 950 CXR pictures in this dataset come from the NIH Chest X-ray Dataset [34]. A three-class categorization is devised for this balanced dataset.

3.1.3. Dataset 3

This dataset contains 2850 images, 950 of which are COVID-19, 950 of which are normal, and 950 of which are tuberculosis images. The COVID-19 Radiography Database was used to obtain COVID-19 and healthy person images [31]. The 950 CXR tuberculosis images in the mix come from the NIAID TB dataset [33]. A three-class categorization is planned for this balanced dataset.

3.1.4. Dataset 4

This dataset contains 2850 images, 950 of which are COVID-19, 950 are normal, and 950 are bacterial pneumonia images. The COVID-19 Radiography Database was used to obtain COVID-19 and healthy person images [31]. The 950 CXR bacterial pneumonia images in

the mix come from the COVID-19 Curated Dataset [32]. A three-class categorization is designed for this balanced dataset.

3.1.5. Dataset 5

This dataset contains 950 COVID-19, 950 healthy individuals, 950 TB, and 950 Fibrosis images. All COVID-19 and Normal images are gathered from the COVID-19 Radiography Database [31]. The NIAID TB dataset [33] and the NIH Chest X-ray Dataset [34], respectively, served as the sources of the remaining 950 images of tuberculosis and 950 images of fibrosis. A 4-class classification is considered for this balanced dataset.

3.1.6. Dataset 6

This dataset has 950 COVID-19, 950 healthy individuals, 950 Bacterial Pneumonia, and 950 Fibrosis images. All COVID-19 and Normal images are gathered from the COVID-19 Radiography Database [31]. COVID-19 Curated Dataset [32] and the NIH Chest X-ray Dataset [34], respectively, served as the sources of the remaining 950 photos of bacterial pneumonia and 950 images of fibrosis. A 4-class classification is considered for this balanced dataset.

3.1.7. Dataset 7

In this collection, 950 COVID-19 images, 950 images of healthy individuals, 950 images of bacterial pneumonia, and 950 tuberculosis images are found. All COVID-19 and normal images are gathered from the COVID-19 Radiography Database [31]. The COVID-19 Curated Dataset [32] and the NIAID TB dataset [33], respectively, served as the source of the remaining 950 images of bacterial pneumonia and 950 images of tuberculosis. A 4-class classification is considered for this balanced dataset.

3.1.8. Dataset 8

This CXR assembly of 4750 images is spanned evenly across 950 images of COVID-19, 950 images of healthy individuals, 950 images of TB, 950 images of Bacterial Pneumonia, and 950 images of Fibrosis. The COVID-19 Radiography Database results in the images of COVID-19 and healthy persons. In addition, the COVID-19 Curated Dataset [32] is used to gather 950 images of bacterial pneumonia. While the 950 images of fibrosis are derived from the NIH Chest X-ray Dataset [34], the remaining 950 images of tuberculosis are gathered from the NIAID TB dataset [33]. The five-class categorization is considered in this regard.

3.1.9. Dataset 9

This CXR assembly of 5700 images is spanned evenly across 950 images of COVID-19, 950 images of healthy individuals, 950 images of TB, 950 images of bacterial pneumonia, 950 images of viral pneumonia, and 950 fibrosis images. The COVID-19 Radiography Database results in images of COVID-19, viral pneumonia, and healthy persons. In addition, the COVID-19 Curated Dataset [32] is used to gather 950 images of bacterial pneumonia. While the 950 images of fibrosis are derived from the NIH Chest X-ray Dataset [34], the remaining 950 images of TB are gathered from the NIAID TB dataset [33]. The six-class categorization is considered in this regard.

3.1.10. Dataset 10

This CXR assembly of 6650 images is spanned evenly across 950 images of COVID-19, 950 images of healthy individuals, 950 images of TB, 950 images of Bacterial Pneumonia, 950 images of Viral Pneumonia, 950 images of Lung Opacity, and 950 images of Fibrosis. The COVID-19 Radiography Database results in images of COVID-19, Viral Pneumonia, Lung Opacity, and healthy persons. In addition, the COVID-19 Curated Dataset [32] is used to gather 950 images of Bacterial Pneumonia. While the 950 images of Fibrosis are derived from the NIH Chest X-ray Dataset [34], the remaining 950 images of Tuberculosis

are gathered from the NIAID TB dataset [33]. The seven-class categorization is considered in this regard.

3.1.11. Dataset Splitting

As mentioned before, 80% of the introduced datasets are used for training, 10% for testing, and 10% for validation. Each class uses 760 images for training purposes for dataset 1 to dataset 10. In each class, 95 images are utilized for testing, and 95 images are used for validation. Table 1 represents details of the datasets.

Table 1. Designing of Chest X-ray Datasets.

Datasets	Number of Classes	Class Names	Samples for Training (80%)	Samples for Testing (10%)	Samples for Validation (10%)	Total Samples (100%)
Dataset 1	2	COVID Normal	760 760	95 95	95 95	1900
Dataset 2	3	Fibrosis COVID Normal	760 760 760	95 95 95	95 95 95	2850
Dataset 3	3	COVID Normal Tuberculosis	760 760 760	95 95 95	95 95 95	2850
Dataset 4	3	Bacterial Pneumonia COVID Normal	760 760 760	95 95 95	95 95 95	2850
Dataset 5	4	COVID Fibrosis Normal Tuberculosis	760 760 760 760	95 95 95 95	95 95 95 95	3800
Dataset 6	4	Bacterial Pneumonia COVID Fibrosis Normal	760 760 760 760	95 95 95 95	95 95 95 95	3800
Dataset 7	4	Bacterial Pneumonia COVID Normal Tuberculosis	760 760 760 760	95 95 95 95	95 95 95 95	3800
Dataset 8	5	Bacterial Pneumonia COVID Fibrosis Normal Tuberculosis	760 760 760 760 760	95 95 95 95 95	95 95 95 95 95	4750
Dataset 9	6	Bacterial Pneumonia COVID Fibrosis Normal Tuberculosis Viral Pneumonia	760 760 760 760 760 760	95 95 95 95 95 95	95 95 95 95 95 95	5700
Dataset 10	7	Bacterial Pneumonia COVID Fibrosis Lung Opacity Normal Tuberculosis Viral Pneumonia	760 760 760 760 760 760 760	95 95 95 95 95 95 95	95 95 95 95 95 95 95	6650

3.2. Pre-Processing and Augmentation

The images are scaled to match the input dimension for the CNN, with larger images suppressing the traits of interest most likely. To begin, all images are downsized to 300×300 pixels. Then, all pixels $[0, 1]$ are rescaled using the min–max normalization approach. Additionally, image augmentation techniques were used to address the limited number of images in the datasets and increase training efficiency while preventing model overfitting.

3.2.1. Sample-Wise Centering

This technique was applied to ensure that the mean pixel value of individual images was set to zero. It involves adjusting the brightness levels of the image while preserving the relative differences between pixels.

3.2.2. Sample-by-Sample Standard Deviation Normalization

This technique involves rescaling the pixel values based on their associated standard deviation. This normalization process helps to standardize the variability of pixel values across different images.

3.2.3. Horizontal Flipping

This technique involves creating a mirrored version of the original image by flipping it horizontally. In the context of lung images, this augmentation is relevant as lung structure and patterns can be symmetric.

3.2.4. Image Generator

The image generator uses the sample-wise center for augmentation to make the single image's mean pixel value zero. Following that, sample-by-sample standard deviation normalization is used to partition images based on their associated standard deviation value. Finally, the horizontal flip is used to flip photographs horizontally.

These augmentation techniques were specifically chosen to enhance the diversity of the dataset while ensuring that the transformations were meaningful for lung images.

3.3. Proposed Multi-Scale CNN Architecture

The proposed architecture in Figure 3 has two components: a backbone and a CNN head. The backbone is a pre-trained image classification network acting as a feature extractor. Here, the top layers of the pre-trained network are extracted, and the bottom layers are removed to provide only the low-level extracted feature maps. Using VGG-16 as the backbone, convolutional layers as the head for feature extraction in multiple scales and filter size optimizations, the model's ability to extract discriminative features from CXR images is enhanced [35]. It leverages pre-trained weights, benefits from transfer learning, captures features at multiple scales, and adapts to the specific characteristics of CXR images. Transfer learning allows the model to transfer the knowledge gained from a source task (ImageNet classification) to a target task (CXR classification) [35]. This is especially valuable when the target task has limited labeled data, as it enables the model to generalize better and achieve higher accuracy by leveraging the learned representations from a related task. These advantages contribute to improved accuracy and robustness in CXR classification tasks.

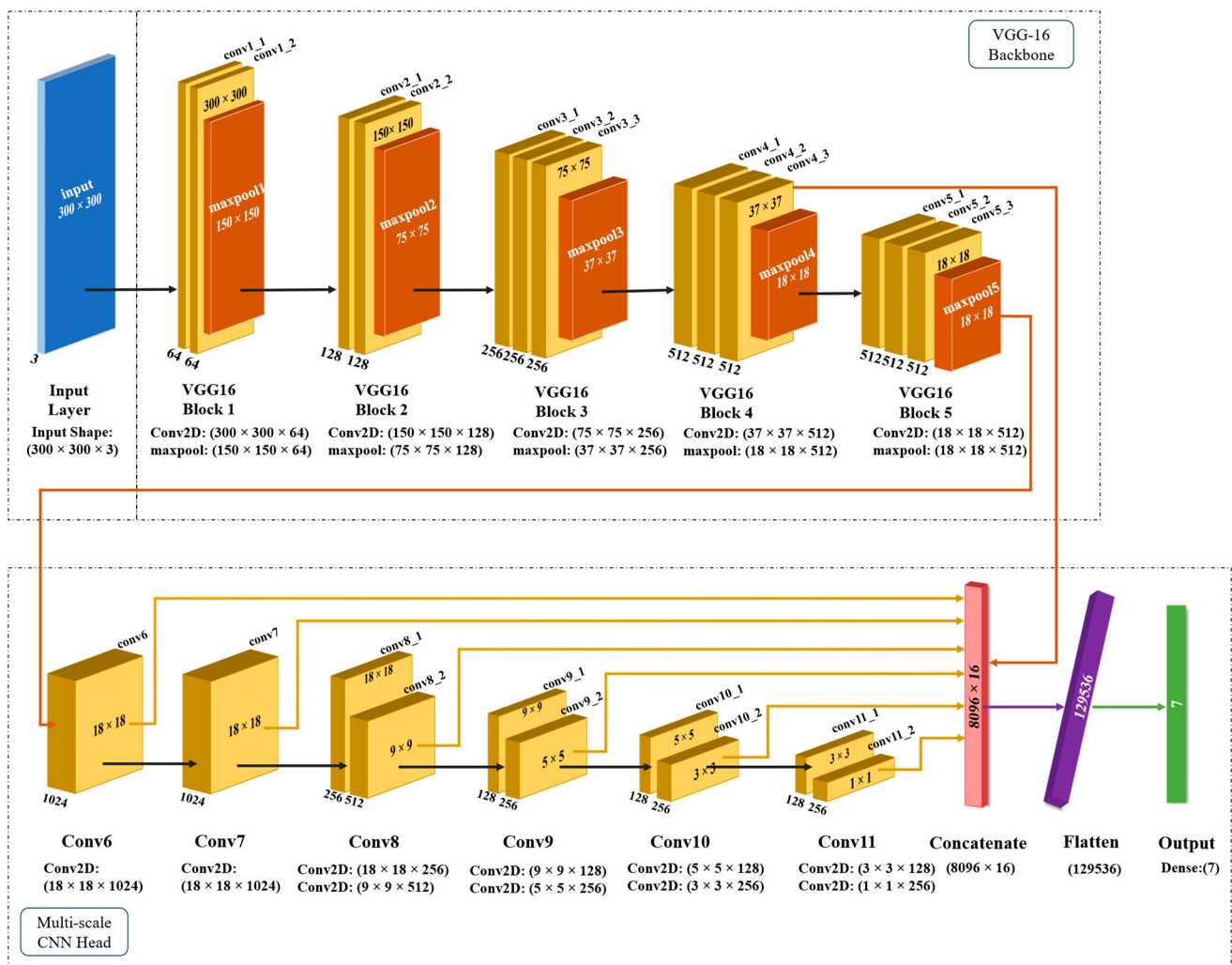


Figure 3. Block diagram of Multi-scale CNN architecture.

VGG-16 is a deep neural network architecture pre-trained on the large-scale ImageNet dataset [35]. The VGG-16 backbone consists of convolutional and max pooling layers, gradually reducing the spatial dimensions while increasing the number of channels. Using its pre-trained weights, the model can leverage the knowledge learned from millions of images to initialize its feature extraction process. This helps in capturing generic visual features useful for a wide range of image classification tasks, including CXR classification.

CNN head comprises multiple convolutional layers stacked together and added to the top of the backbone model. By incorporating these convolutional layers into the head of the model, the architecture is customized to extract features at multiple scales. The VGG16 layers closer to the input learn low-level features like edges and textures, while deeper layers in the CNN head learn more complex and abstract features. This is crucial for CXR classification, as abnormalities within the image can appear in unusual sizes. Multi-scale feature mapping enables the model to become more robust and capable of identifying abnormalities of varying sizes, improving its overall accuracy in CXR classification.

The input image size is set to 300×300 pixels, and the VGG-16 network acts as the backbone of the model consisting of five blocks. Blocks 1 to 5, the part of the VGG-16 backbone helps extract hierarchical features from the input image. In each block, the filter size is doubled on the top Conv2D layer, while the feature map size is halved in the MaxPooling2D layer. The first two blocks, named Block 1 and Block 2, include two Conv2D layers (convX_1 and convX_2) and one MaxPooling2D layer (maxpoolX). Layer output sizes of the Conv2D layers in Block 1 and Block 2 are $300 \times 300 \times 64$ and $150 \times 150 \times 128$, respec-

tively. The corresponding MaxPooling2D layers produce output sizes of $150 \times 150 \times 64$ and $75 \times 75 \times 128$. The MaxPooling2D layers downsample the feature maps, reducing their spatial dimensions. The following three blocks, named Block 3, Block 4, and Block 5, consist of three Conv2D layers and one MaxPooling2D layer (convX_1, convX_2, convX_3, and maxpoolX). The Conv2D layers progressively keep decreasing in spatial dimensions, resulting in feature maps with sizes of $75 \times 75 \times 256$ (Block 3), $37 \times 37 \times 512$ (Block 4), and $18 \times 18 \times 512$ (Block 5). The MaxPooling2D layers further perform downsampling of the feature maps. The MaxPooling2D layers produce output sizes of $37 \times 37 \times 256$, $18 \times 18 \times 512$, and $18 \times 18 \times 512$, respectively. The first effective layer responsible for Chest X-ray (CXR) classification, conv4_3, has a spatial dimension of 38×38 , representing a considerable reduction compared to the input image size. Higher-resolution feature maps play a crucial role in detecting small edges and patterns in the image.

Afterward, a CNN head consisting of Block numbers from 6 to 11 introduces additional convolutional layers to the model, increasing its complexity. This allows the model to learn more intricate and abstract features of the CXR input image. Gradually, as the Conv2D structure keeps decreasing in spatial dimensions, the resolution of the feature maps also decreases. The feature map from Block 4 (conv4_3) is connected to a Concatenate layer, and the feature map from Block 5 (maxpool5) is connected to the Conv6 block with an output size of $18 \times 18 \times 1024$. Conv6 is then connected to the Conv7 block. Following Conv7, four additional convolutional blocks (Conv8 to Conv11) are added each containing two Conv2D layers. Each block in the architecture builds upon the features extracted by the previous blocks. The Conv8 block has conv8_1 and conv8_2 layers with output sizes of $18 \times 18 \times 256$ and $9 \times 9 \times 512$, respectively. The second layer of Conv8 (conv8_2) is connected to the first layer of Conv9 (conv9_1). Similarly, Conv9 has conv9_1 and conv9_2 layers with output sizes of $9 \times 9 \times 128$ and $5 \times 5 \times 256$, respectively. This pattern reiterates with conv9_2 further connecting to conv10_1 ($5 \times 5 \times 128$), conv10_2 ($3 \times 3 \times 256$), and finally conv11_1 ($3 \times 3 \times 128$).

Lastly, a Concatenation block is used to merge the feature maps from all of the convolutional layers, namely, conv4_3, conv6, conv7, conv8_2, conv9_2, conv10_2, and conv11_2 to combine the feature maps into a single concatenated feature map. All the smaller feature maps contain different levels of information extracted from the input image at different scales and resolutions. The resulting tensor from the concatenation operation has a more significant depth, combining the channels from individual feature maps. The output size of the Concatenate layer is 8096×16 .

After the concatenation operation, the resulting tensor is passed through a flattened layer that converts the multi-dimensional tensor into a one-dimensional vector. Then, the flattened layer transforms the concatenated layer, which has a shape of 8096×16 , into a flat vector of length $8096 \times 16 = 129,536$. Following the flattened layer, the flattened vector is passed through a dense layer with SoftMax activation, providing the classification probabilities for the input image across different classes. The choice of filter sizes, number of layers, and block configurations followed the SSD300 (Single-shot Multibox Detector) feature extraction standard and the abstraction practices of VGG16. Furthermore, introducing additional convolutional blocks (Conv Blocks 6 to 11) was purposeful, aiming to allow the model to learn more intricate and abstract features from the input. The basic working principle of the MS-CNN model is presented in Algorithm 1.

Algorithm 1: Proposed Multi-scale CNN Algorithm

1. **Input:** 6650 CXR images (80% training, 10% validation, 10% testing).
2. **Output Labels:** Normal, Bacterial Pneumonia, COVID, Fibrosis, Lung Opacity, Tuberculosis, Viral Pneumonia.
3. **Begin**
4. **Preprocessing:**
 - i. Resize images: $X_{resized} = \text{resize}(X, 300 \times 300 \times 3)$
 - ii. Batch normalization: $X_{norm} = \frac{X_{resized}}{255}$
 - iii. Split: Use the normalized dataset, $X_{norm} = \{X_1, X_2, X_3, \dots, X_n\}$, to split into $\frac{[0.8]}$ datasets.
5. **Augmentation:**
 - i. Center and standard normalize: $X_{normalized} = \text{standard_normalize}(\text{center}(X_{norm}))$
 - ii. Apply horizontal flip (HF) for training.
6. **Model Construction:**
 - i. Base: Initialize model M with VGG16 base network (input size $M_{input} = 300 \times 300 \times 3$).
 - ii. Add Conv2D layers: $L_i = \text{Conv2D}(X_{train}, \text{filters} = F(128, 256, 512, 1024), \text{kernel_size} = (3, 3), \text{dilation_rate} = (1, 1), \text{strides} = (1, 1), \text{activation} = \text{ReLU})$ for $i = 1, 2, 3, 4$.
 - iii. Concatenate (Conv2D layers): $L_{concat} = \text{concatenate}(L_1, L_2, L_3, L_4 \dots L_i)$.
 - iv. Flatten: $L_{flat} = \text{flatten}(L_{concat})$.
 - v. Dense: $L_{dense} = \text{Dense}(L_{flat}, \text{activation} = \text{Softmax})$.
 - vi. Compile: $M_{compiled} = \text{compile}, \text{optimizer} = \text{Adam}, \text{learningrate} = 0.0001, \text{loss} = \text{categorical_crossentropy}$.
7. **Training Phase:**

Train: $M_{trained} = \text{train}(M_{compiled}, \text{epochs} = 25, \text{batchsize} = 16, \text{validation} = X_{val})$, monitoring validation loss.
8. **Testing Phase:** Generate labels: $y_{test} = \text{predict}(M_{trained}, X_{test})$.
9. **Performance Evaluation:**
 - i. Confusion Matrix: TP, FP samples: $CM = \text{confusion_matrix}(y_{true}, y_{test})$.
 - ii. Metrics: Accuracy, Precision, Recall, F1-score:
 $A, P, R, F1 = \text{classification_report}(y_{true}, y_{test})$.
 - iii. ROC: Area Under ROC (AUC) on X_{test} : $\text{compute}_{AUC}(X_{test}, M_{trained})$.
10. **End**

3.4. Experimental Setup and Hyperparameter Settings

The proposed model was trained using the TensorFlow API operating on a 64-bit Windows 11 Pro system. Keras and Scikit-Learn libraries facilitated seamless handling of diverse data with Multi-scale CNN model design, training, and evaluation tasks on the local machine setup mentioned in Table 2. Hyperparameters were meticulously determined through extensive experimentation on the platform. Multiple training runs, varying learning rates, patience, optimizers, epochs, and batch sizes were executed. Model performance was rigorously assessed on validation data to pinpoint the optimal combination of hyperparameters for desired model performance.

The training commenced with a conservative learning rate of 0.0001, with continuous progress monitoring. The rate of loss reduction guided adjustments to the learning rate, with a pivotal role played by a patience value of 10 in implementing effective early stopping, a technique crucial for preventing overfitting. Throughout training, the Adam optimizer dynamically adapted the learning rate for each parameter.

Experimentation revealed that training for fewer than 15 epochs resulted in underfitting, while over 50 epochs led to overfitting. Thus, a balanced choice of 25 epochs struck the right pattern-capturing balance while avoiding overfitting. Different batch sizes were explored, with smaller sizes showing promise for better generalization, albeit at a slower

training pace. Conversely, batch sizes exceeding 32 led to validation data exhibiting unstable minima. A batch size of 16 was deemed optimal, ensuring a balanced compromise between training speed, memory usage, and convergence. This meticulous hyperparameter tuning process yielded a high-performing neural model for the classification task, detailed in Tables 2 and 3 for reference.

Table 2. Experimental setup.

Name	Parameters
Programming Language	Python
Environment	Microsoft VS Code (1.74.3)
Backend	Keras with TensorFlow
Processor	Intel(R) Core (TM) i7-10700K
Installed RAM	32 GB
GPU	NVIDIA GeForce, RTX 2080 Ti 11 GB
Operating system	Windows 11 Pro
Input	Chest X-ray Images
Input Size	300 × 300

Table 3. Hyperparameters utilized in model training.

Hyperparameters	Values/Types
Epoch	25
Batch Size	16
Learning Rate	0.0001
Patience	10
Optimizer	Adam
Loss Function	Categorical Cross Entropy

3.5. Performance Metrics

A confusion matrix is a table that summarizes obtained predictions from a model with the actual ground truth labels of the dataset. A classification report is a comprehensive summary of various metrics, including precision, recall, *F1*-Score, and support (the number of occurrences of each class). The percentage of accurate predictions to the net predictions is known as accuracy (*Ac*).

$$Ac = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

Precision (*Pr*) is a metric used to evaluate the quality of the results produced by a model.

$$Pr = \frac{TP}{TP + FP} \quad (2)$$

Recall (*Rc*), also known as sensitivity or true positive rate, is a measure used to quantitatively evaluate a model's performance.

$$Rc = \frac{TP}{TP + FN} \quad (3)$$

F1 score is often considered more informative than accuracy as a performance metric when class imbalance is present.

$$F1 = 2 \times \frac{Pr \times Rc}{Pr + Rc} \quad (4)$$

where *TP*, *TN*, *FP*, and *FN*, denote true positives, true negatives, false positives, and false negatives, respectively.

4. Results

4.1. Classification of Dataset 1

The classification outputs using dataset 1 are presented in Table 4. The results served as the foundation for the proposed model including a comparison with some transfer learning models for all the applied performance assessment criteria. The proposed MS-CNN training and validation accuracy curves, ROC curves, and confusion matrix are shown in Figure 4. This graph illustrates that the model acquired a testing accuracy of 100% and a loss of 0.0131. The confusion matrix also demonstrated that the proposed model performed correctly on 100% (87 images) of the COVID-19 images and 100% (103 images) of the normal images. The model does not misclassify any normal images as COVID-19 or COVID-19 images as normal. The model also achieved an AUC value of 1.00 for identifying COVID-19 samples compared to the healthy sample. A high recall value of 1.00 indicates that the model successfully decreased false-negative rates to zero, ensuring no significant cases of COVID-19 infection cases were missed. However, the high precision value of 1.00 shows that the model has no false positive rates; hence, COVID-19-infected cases were not frequently misclassified. The performance comparison between the proposed model and other TL models is presented in Figure 5.

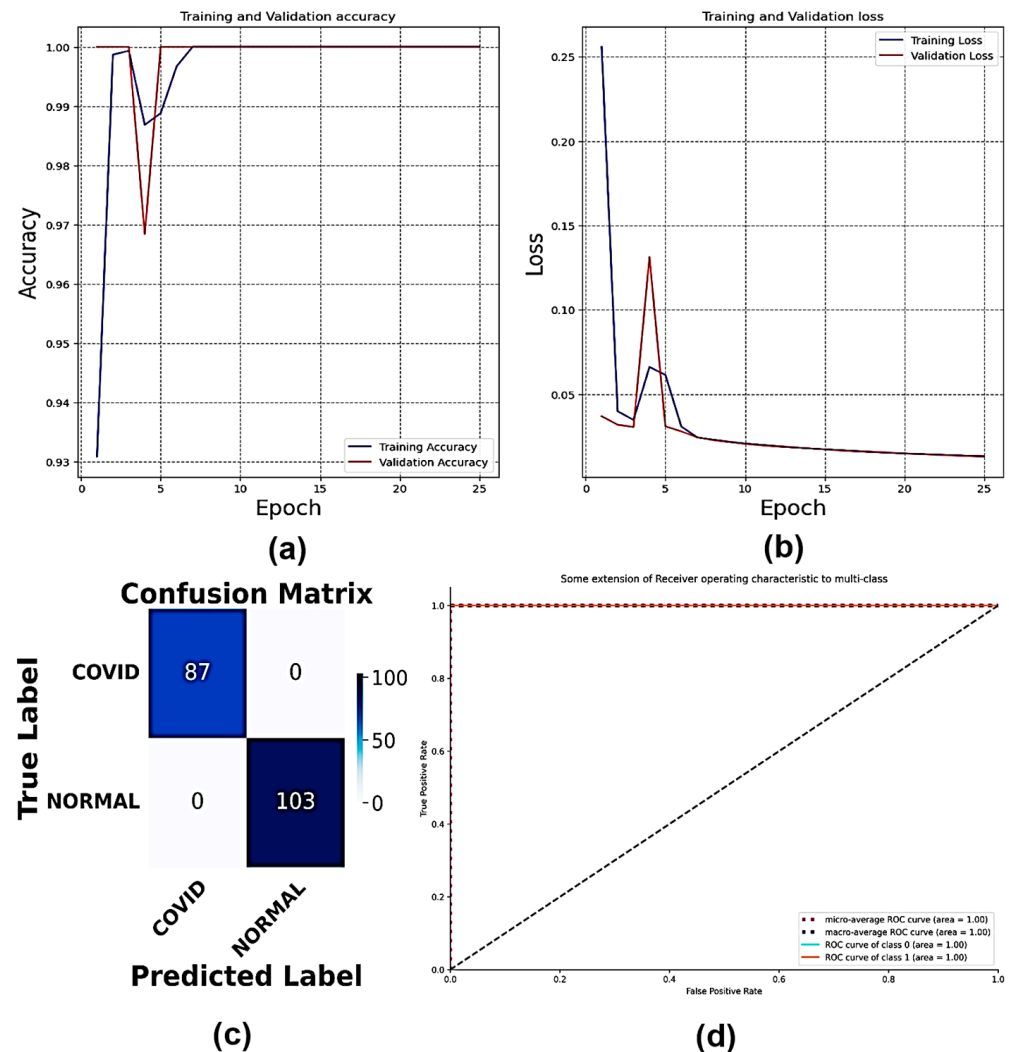
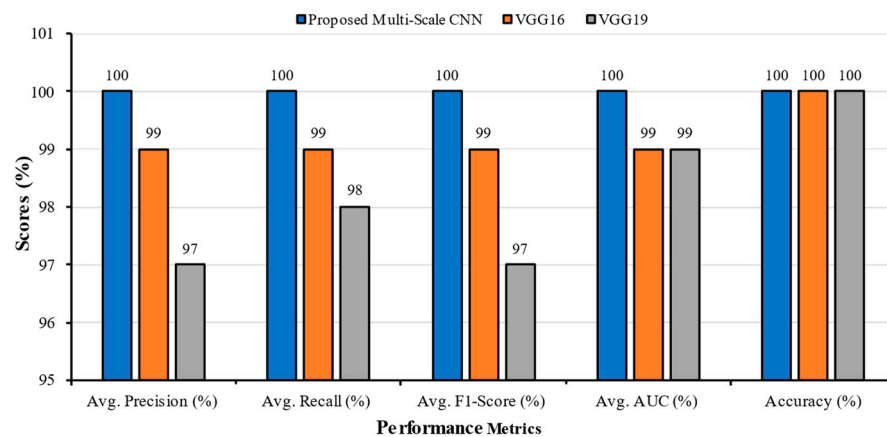


Figure 4. (a) Accuracy curves, (b) Loss curves, (c) Confusion Matrix, and (d) ROC curves of the proposed Multi-Scale CNN model with two individual classes (COVID and Normal).

Table 4. Classification performance results for Dataset 1.

Classification Models	Classes	Precision	Recall	F1-Score	AUC	Accuracy	Average Precision	Average Recall	Average F1-Score	Average AUC
Multi-Scale CNN	COVID	1.00	1.00	1.00	1.00	100	1.00	1.00	1.00	1.00
	Normal	1.00	1.00	1.00	1.00					
VGG16	COVID	1.00	0.99	0.99	1.00	100	0.99	0.99	0.99	0.99
	Normal	0.99	1.00	1.00	0.99					
VGG19	COVID	0.95	1.00	0.97	0.98	100	0.97	0.98	0.97	0.99
	Normal	1.00	0.95	0.98	1.00					

**Figure 5.** Average Precision (%), Average Recall (%), Average F1-Score (%), Average AUC (%), and Accuracy of the different models with two individual classes (COVID and Normal) for Dataset 1.

4.2. Classification of Dataset 2

The proposed MS-CNN model training and validation accuracy curves, ROC curves, and confusion matrix are shown in Figure 6. The outputs of dataset 2 are presented in Table 5 and Figure 7. The testing accuracy and loss of the proposed model were 99.65% and 0.0236, respectively. The CM indicated that the proposed model misclassified one COVID-19 image as fibrosis, two fibrosis as COVID-19, and five normal as fibrosis. The AUC values for COVID-19, Fibrosis, and Normal were 1.00, 0.99, and 0.99, respectively.

Table 5. Classification performance results for Dataset 2.

Classification Models	Classes	Precision	Recall	F1-Score	AUC	Accuracy (%)	Average Precision	Average Recall	Average F1-Score	Average AUC
Multi-Scale CNN	COVID	0.98	0.99	0.99	1.00	99.65	0.97	0.97	0.97	0.99
	Fibrosis	0.94	0.98	0.96	0.99					
	Normal	1.00	0.94	0.97	0.99					
VGG16	COVID	1.00	0.96	0.98	1.00	99.30	0.97	0.97	0.97	0.99
	Fibrosis	0.92	1.00	0.96	0.98					
	Normal	1.00	0.95	0.98	0.99					
VGG19	COVID	0.95	0.67	0.78	0.98	96.84	0.86	0.84	0.83	0.96
	Fibrosis	0.99	0.86	0.92	0.94					
	Normal	0.66	1.00	0.79	0.95					

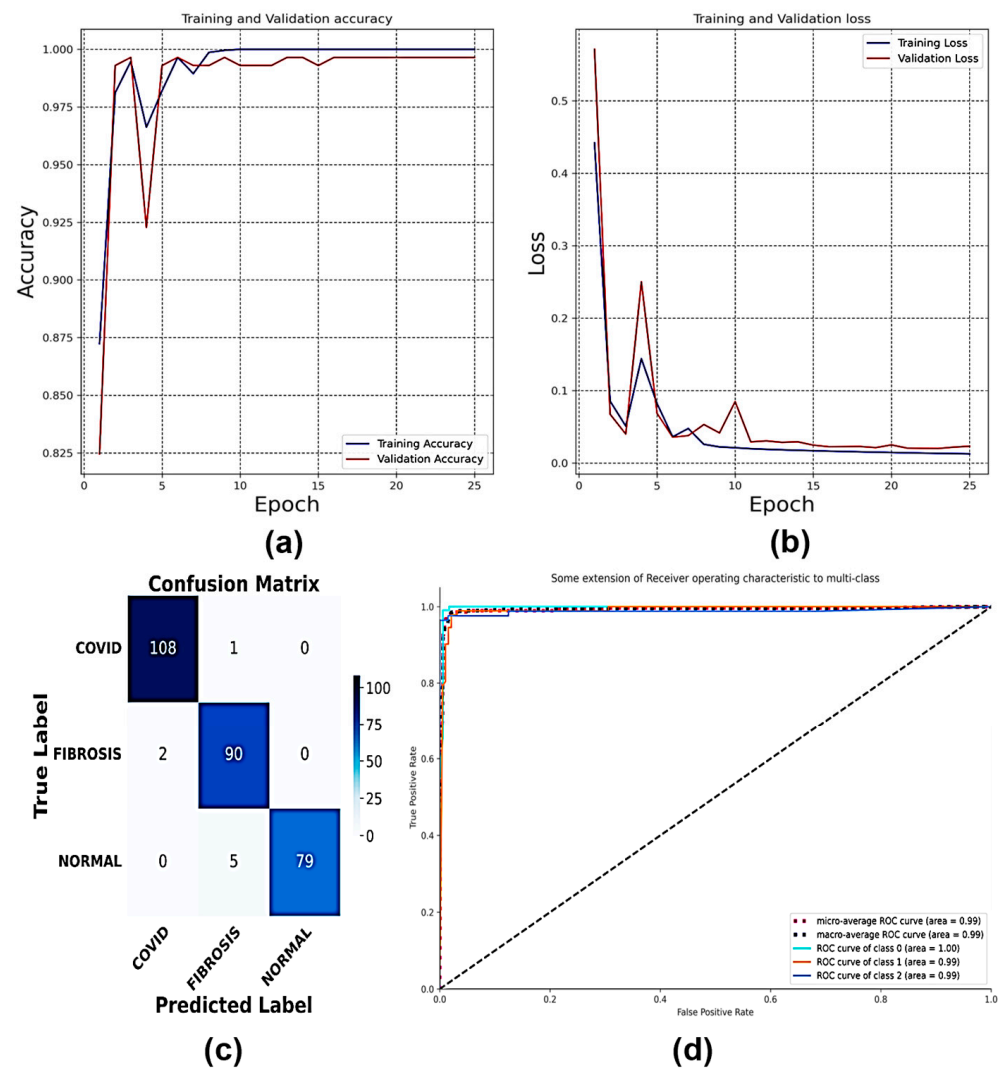


Figure 6. (a) Accuracy curves, (b) Loss curves, (c) Confusion Matrix, and (d) ROC curves of the proposed Multi-Scale CNN model with three individual classes (COVID, Normal, and Fibrosis).

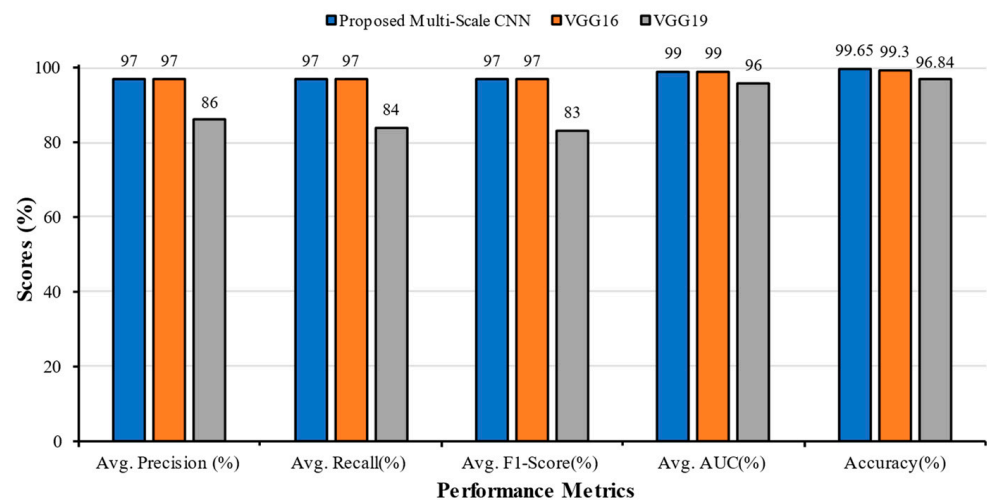


Figure 7. Average Precision (%), Average Recall (%), Average F1-Score (%), Average AUC (%), and Accuracy of the different models with three individual classes (COVID, Normal, and Fibrosis) for Dataset 2.

4.3. Classification of Dataset 3

The proposed MS-CNN model training and validation accuracy curves, ROC curves, and confusion matrix are shown in Figure 8. The outputs of the dataset 3 are presented in Table 6 and Figure 9. The testing accuracy and loss of the proposed model were 99.30% and 0.0250, respectively. The CM indicated that the proposed model misclassified two COVID-19 images as tuberculosis, one COVID-19 as Normal, and one Tuberculosis as COVID-19. The AUC values for COVID-19, tuberculosis, and Normal each were 1.00.

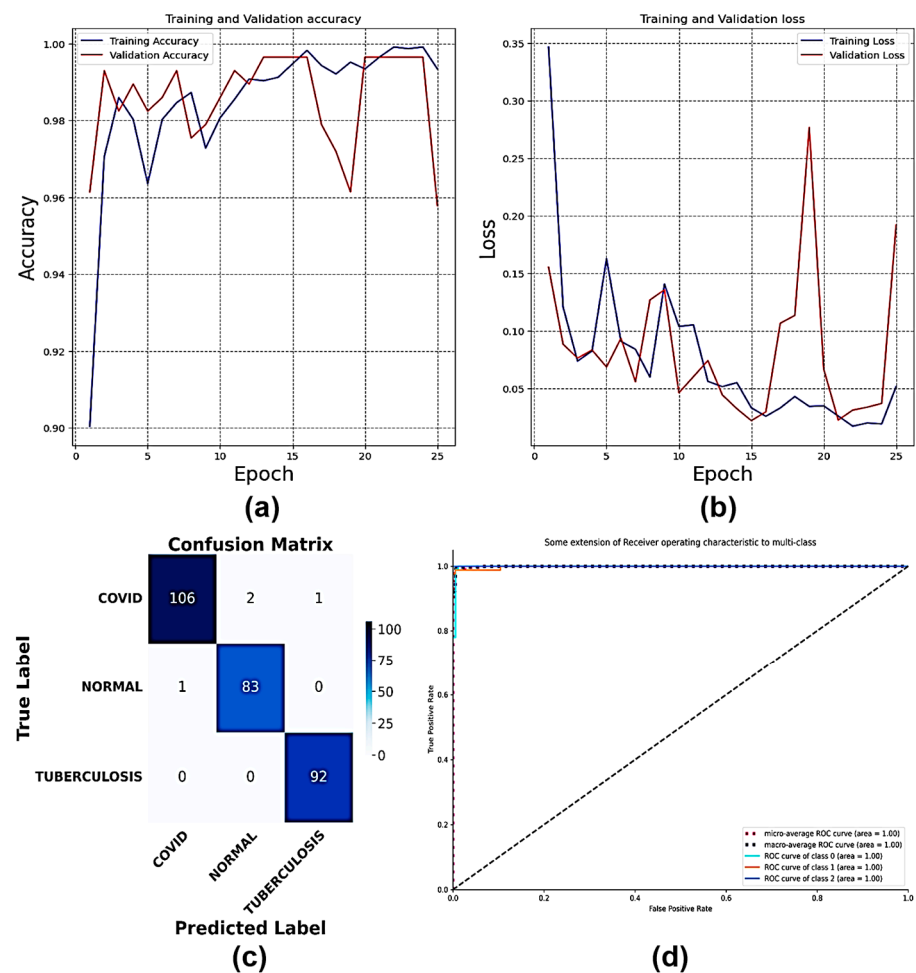


Figure 8. (a) Accuracy curves, (b) Loss curves, (c) Confusion Matrix, and (d) ROC curves of the proposed Multi-Scale CNN model with three individual classes (COVID, Normal, and Tuberculosis).

Table 6. Classification performance results for Dataset 3.

Classification Models	Classes	Precision	Recall	F1-Score	AUC	Accuracy	Average Precision	Average Recall	Average F1-Score	Average AUC
Multi-Scale CNN	COVID	0.99	0.97	0.98	1.00					
	Normal	0.98	0.99	0.98	1.00	99.30	0.99	0.99	0.99	1.00
	Tuberculosis	0.99	1.00	0.99	1.00					
VGG16	COVID	0.97	0.94	0.96	0.98					
	Normal	0.97	1.00	0.98	0.99	97.54	0.97	0.97	0.97	0.99
	Tuberculosis	0.96	0.96	0.96	0.99					
VGG19	COVID	0.96	0.92	0.94	0.95					
	Normal	0.97	0.96	0.96	0.99	95.44	0.95	0.95	0.95	0.97
	Tuberculosis	0.91	0.98	0.94	0.97					

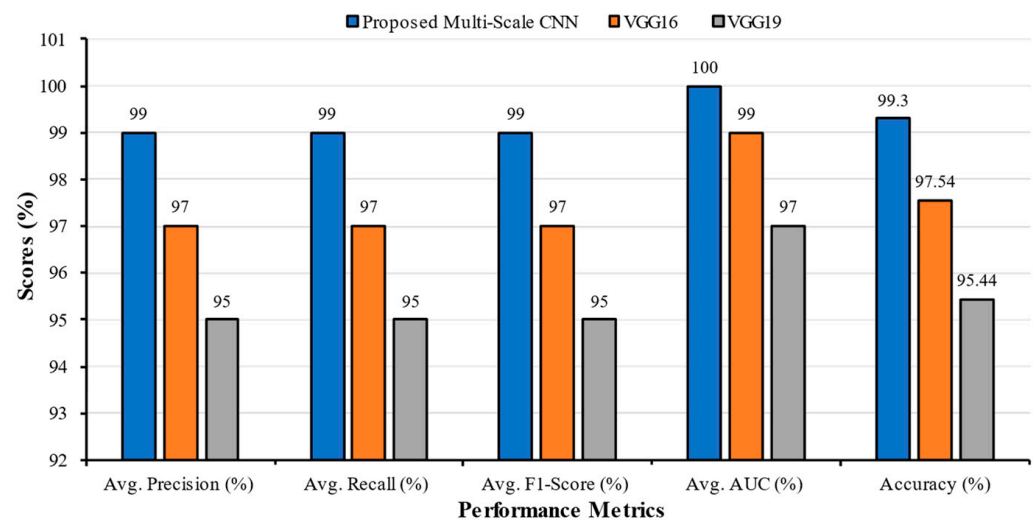


Figure 9. Average Precision (%), Average Recall (%), Average F1-Score (%), Average AUC (%), and Accuracy of the different models with three individual classes (COVID, Normal, and Tuberculosis) for Dataset 3.

4.4. Classification of Dataset 4

The proposed MS-CNN model training and validation accuracy curves, ROC curves, and confusion matrix are shown in Figure 10. The outputs of the dataset 4 are presented in Table 7 and Figure 11. The testing accuracy and loss of the proposed model were 98.60% and 0.1079, respectively. The CM indicated that the proposed model misclassified one COVID-19 image as Normal, and eight Normal as Bacterial Pneumonia. The AUC values for COVID-19, Bacterial Pneumonia, and Normal each were 1.00.

Table 7. Classification performance results for Dataset 4.

Classification Models	Classes	Precision	Recall	F1-Score	AUC	Accuracy	Average Precision	Average Recall	Average F1-Score	Average AUC
Multi-Scale CNN	Bacterial Pneumonia	0.91	1.00	0.95	1.00	98.60	0.97	0.97	0.97	1.00
	COVID	1.00	0.99	1.00	1.00					
	Normal	0.99	0.91	0.95	1.00					
VGG16	Bacterial Pneumonia	1.00	0.92	0.96	0.99	97.89	0.96	0.96	0.96	0.99
	COVID	0.96	0.99	0.98	1.00					
	Normal	0.93	0.97	0.95	0.99					
VGG19	Bacterial Pneumonia	0.82	0.89	0.85	0.98	97.19	0.91	0.90	0.90	0.97
	COVID	0.95	0.95	0.95	0.96					
	Normal	0.94	0.85	0.89	0.98					

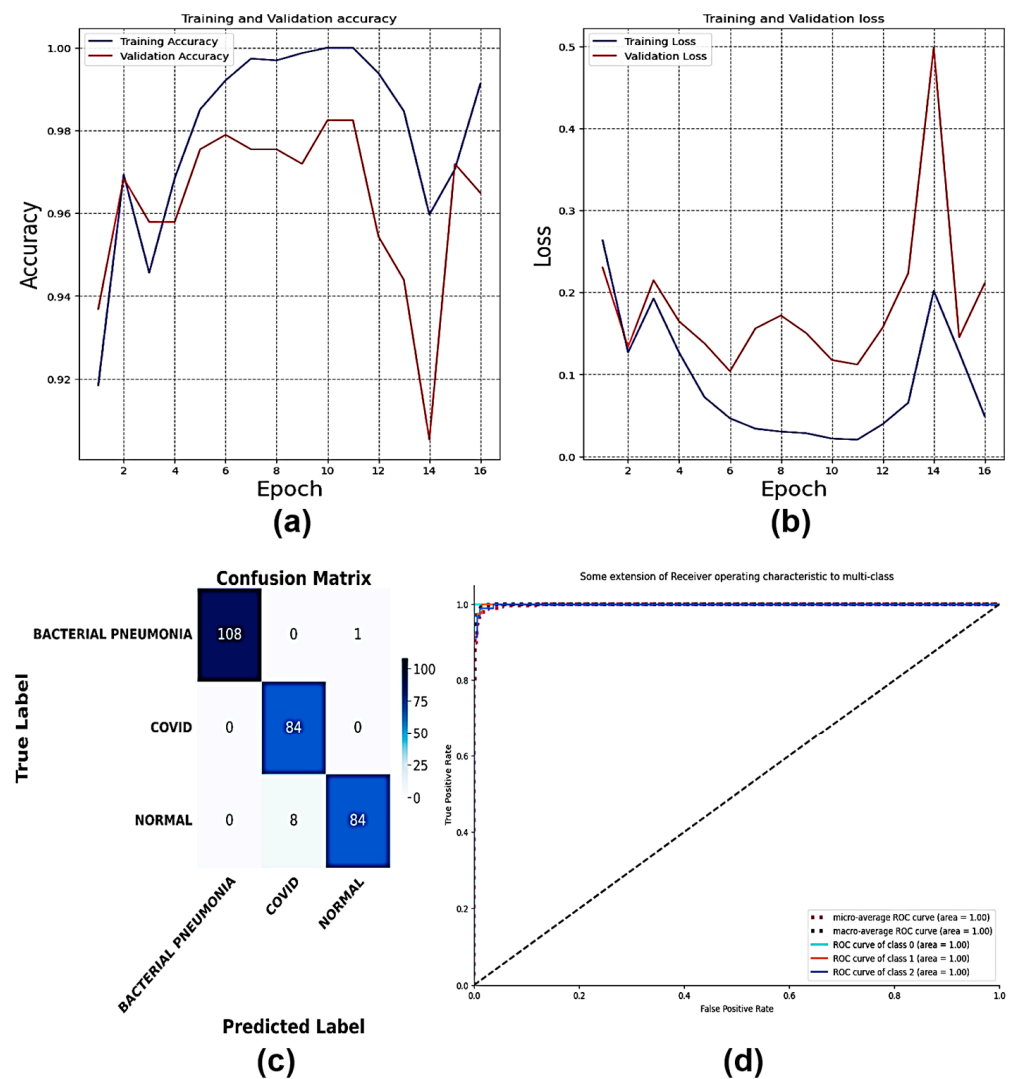


Figure 10. (a) Accuracy curves, (b) Loss curves, (c) Confusion Matrix, and (d) ROC curves of the proposed Multi-Scale CNN model with three individual classes (COVID, Bacterial Pneumonia, and Normal).

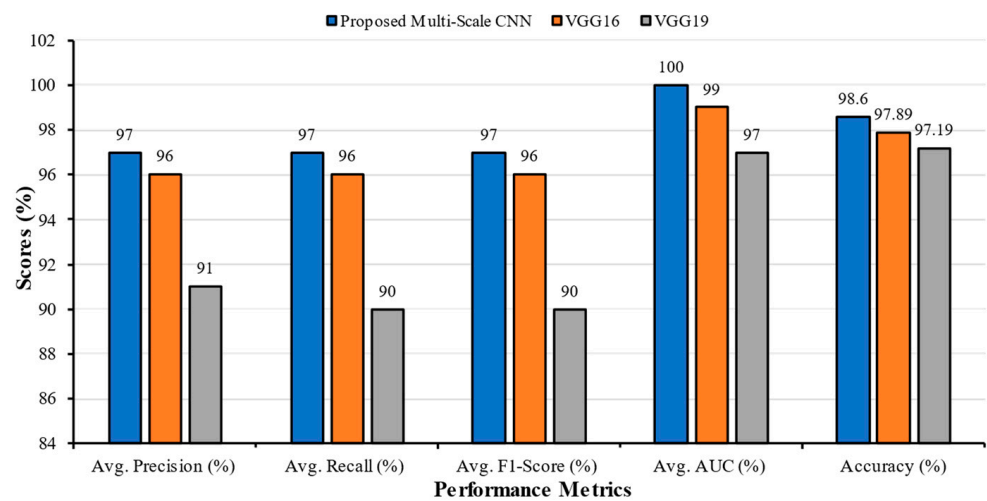


Figure 11. Average Precision (%), Average Recall (%), Average F1-Score (%), Average AUC (%), and Accuracy of the different models with three individual classes (COVID, Bacterial Pneumonia, and Normal) for Dataset 4.

4.5. Classification of Dataset 5

The proposed MS-CNN model training and validation accuracy curves, ROC curves, and confusion matrix are shown in Figure 12.

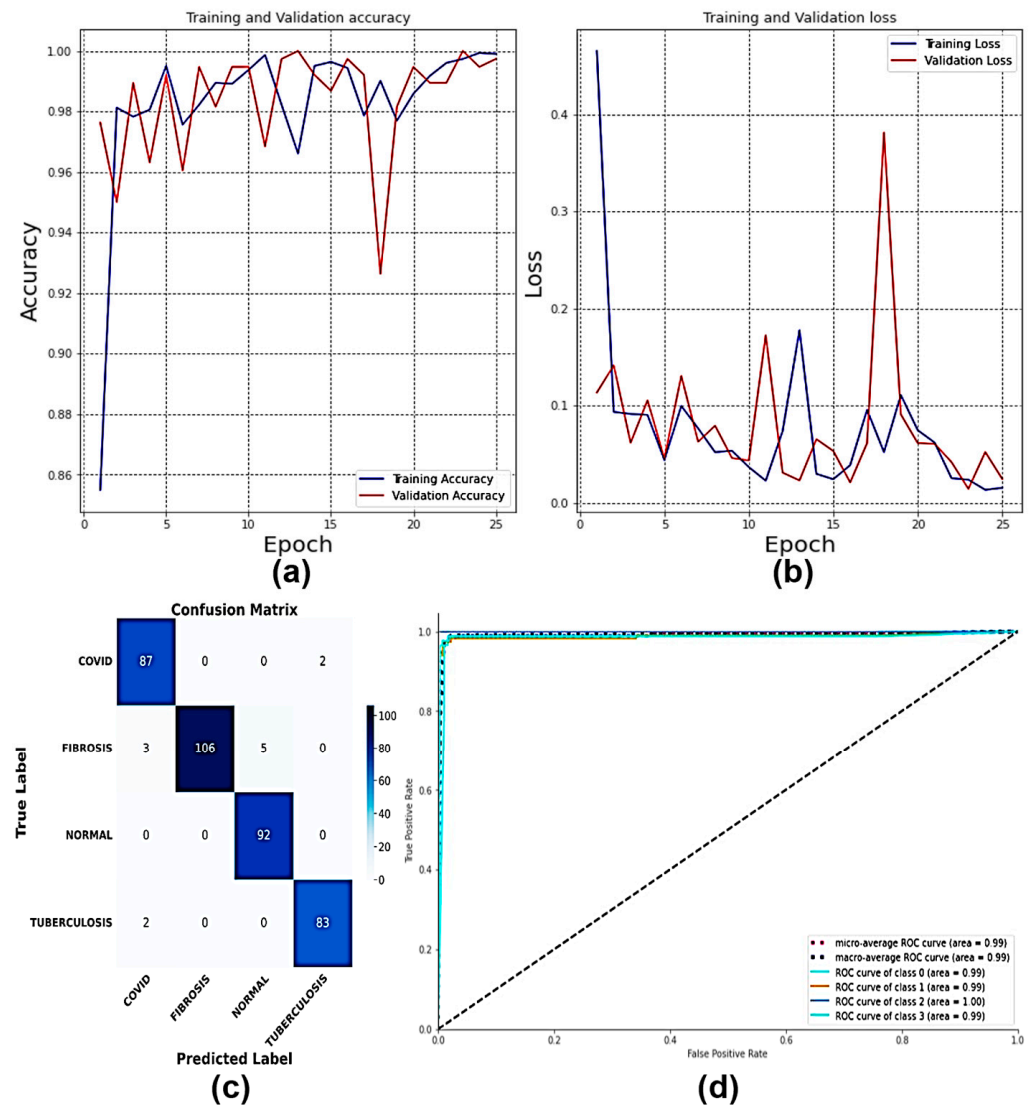
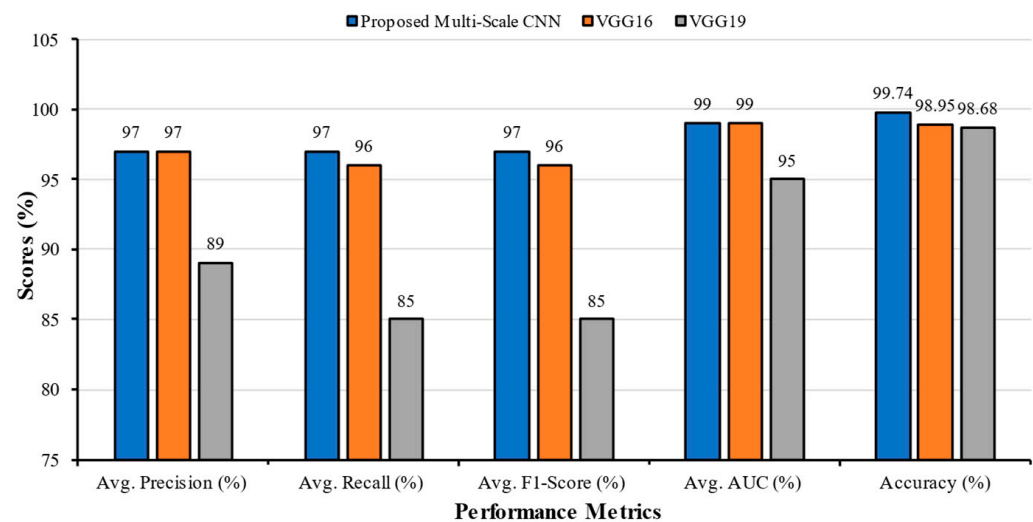


Figure 12. (a) Accuracy curves, (b) Loss curves, (c) Confusion Matrix, and (d) ROC curves of the proposed Multi-Scale CNN model with four individual classes (COVID, Fibrosis, Normal, and Tuberculosis).

The outputs of the dataset 5 are presented in Table 8 and Figure 13. The testing accuracy and loss of the proposed model were 99.74% and 0.0240, respectively. The CM indicated that the proposed model misclassified two COVID-19 images as Tuberculosis, three Fibrosis as COVID-19, five Fibrosis as Normal, and two Tuberculosis as COVID-19. The AUC values for COVID-19, Fibrosis, Normal, and Normal were 0.99, 0.99, 1.00, and 0.99, respectively.

Table 8. Classification performance results for Dataset 5.

Classification Models	Classes	Precision	Recall	F1-Score	AUC	Accuracy	Average Precision	Average Recall	Average F1-Score	Average AUC
Multi-Scale CNN	COVID	0.95	0.98	0.96	0.99	99.74	0.97	0.97	0.97	0.99
	Fibrosis	1.00	0.93	0.96	0.99					
	Normal	0.95	1.00	0.97	1.00					
	Tuberculosis	0.98	0.98	0.98	0.99					
VGG16	COVID	1.00	0.92	0.96	0.98	98.95	0.97	0.96	0.96	0.99
	Fibrosis	0.97	0.96	0.96	0.99					
	Normal	0.99	0.98	0.99	0.99					
	Tuberculosis	0.91	1.00	0.95	0.99					
VGG19	COVID	1.00	0.61	0.76	0.96	98.68	0.89	0.85	0.85	0.95
	Fibrosis	0.66	0.96	0.78	0.93					
	Normal	0.99	0.83	0.90	0.94					
	Tuberculosis	0.89	0.99	0.94	0.98					

**Figure 13.** Average Precision (%), Average Recall (%), Average F1-Score (%), Average AUC (%), and Accuracy of the different models with four individual classes (COVID, Fibrosis, Normal, and Tuberculosis) for Dataset 5.

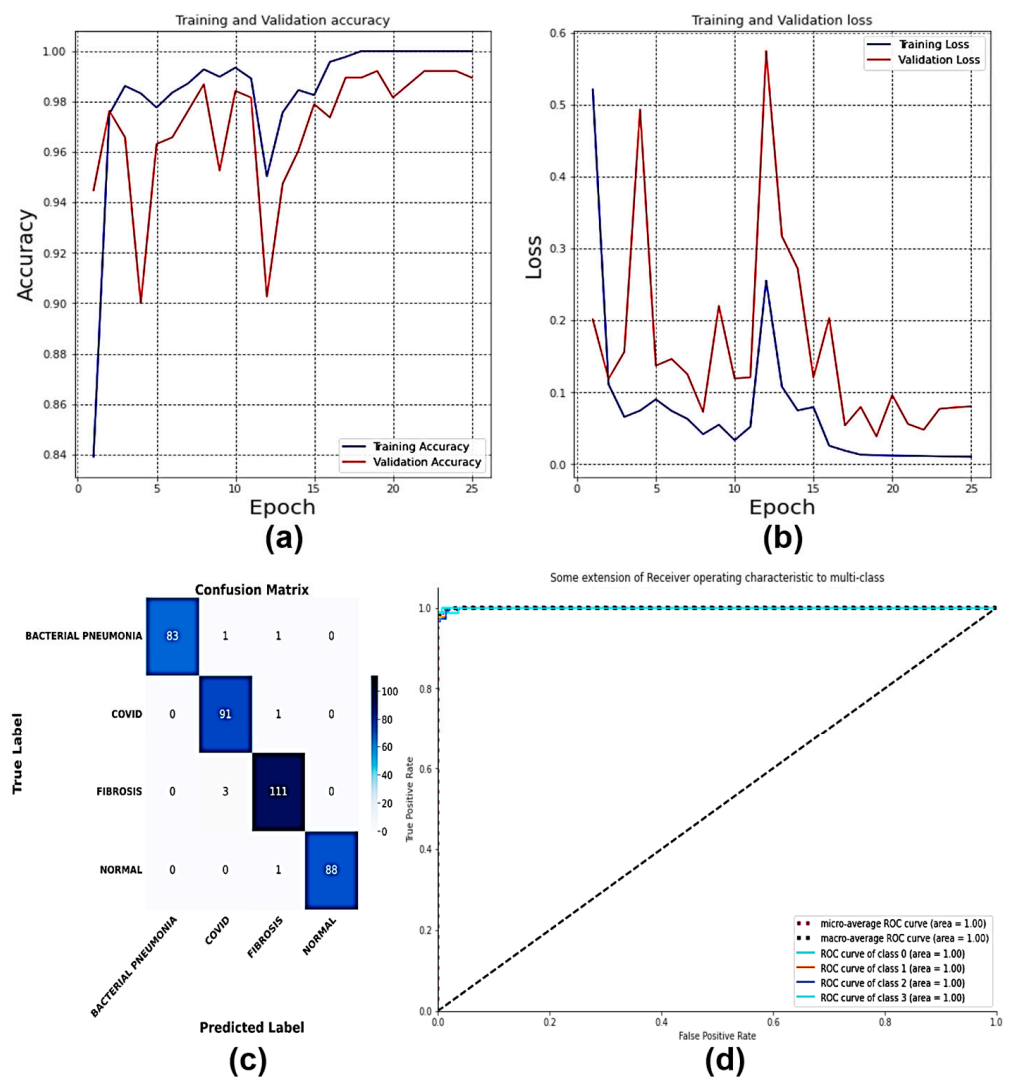
4.6. Classification of Dataset 6

The proposed Multi-Scale CNN model training and validation accuracy curves, ROC curves, and confusion matrix are shown in Figure 14. The outputs of the dataset 6 are presented in Table 9 and Figure 15.

The testing accuracy and loss of the proposed model were 99.21% and 0.0498, respectively. The CM indicated that the proposed model misclassified one Bacterial Pneumonia image as COVID-19, one Bacterial Pneumonia as Fibrosis, one COVID-19 as Fibrosis, three Fibrosis as COVID-19, and one Normal as Fibrosis. The AUC values for Bacterial Pneumonia, COVID-19, Fibrosis, and Normal each were 1.00.

Table 9. Classification performance results for Dataset 6.

Classification Models	Classes	Precision	Recall	F1-Score	AUC	Accuracy	Average Precision	Average Recall	Average F1-Score	Average AUC
Multi-Scale CNN	Bacterial Pneumonia	1.00	0.98	0.99	1.00	99.21	0.98	0.98	0.98	1.00
	COVID	0.96	0.99	0.97	1.00					
	Fibrosis	0.97	0.97	0.97	1.00					
	Normal	1.00	0.99	0.99	1.00					
VGG16	Bacterial Pneumonia	0.98	0.96	0.97	0.99	98.42	0.96	0.96	0.96	1.00
	COVID	1.00	0.88	0.94	1.00					
	Fibrosis	0.93	1.00	0.96	1.00					
	Normal	0.96	0.99	0.97	1.00					
VGG19	Bacterial Pneumonia	0.96	0.96	0.96	1.00	98.42	0.97	0.97	0.97	1.00
	COVID	1.00	0.97	0.98	1.00					
	Fibrosis	0.99	0.97	0.98	1.00					
	Normal	0.94	0.99	0.96	0.99					

**Figure 14.** (a) Accuracy curves, (b) Loss curves, (c) Confusion Matrix, and (d) ROC curves of the proposed Multi-Scale CNN model with four individual classes (Bacterial Pneumonia, COVID, Fibrosis, and Normal).

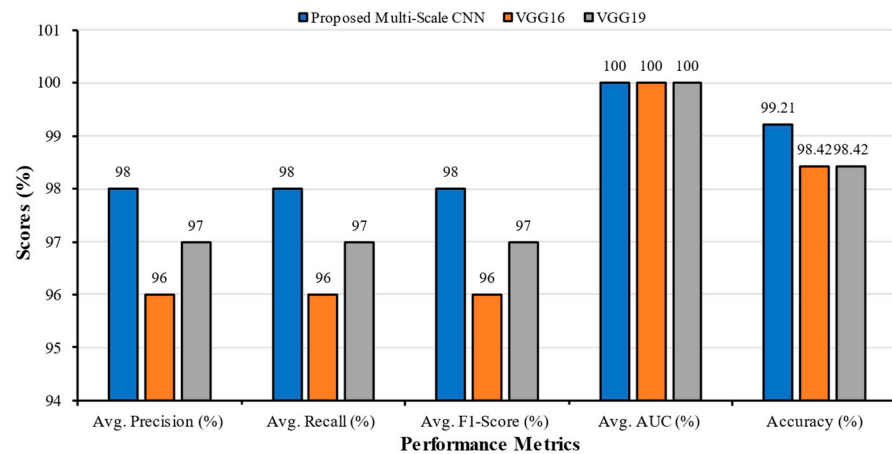


Figure 15. Average Precision (%), Average Recall (%), Average F1-Score (%), Average AUC (%), and Accuracy of the different models with four individual classes (Bacterial Pneumonia, COVID, Fibrosis, and Normal) for Dataset 6.

4.7. Classification of Dataset 7

The proposed Multi-Scale CNN model training and validation accuracy curves, ROC curves, and confusion matrix are shown in Figure 16. The outputs of the dataset 7 are presented in Table 10 and Figure 17.

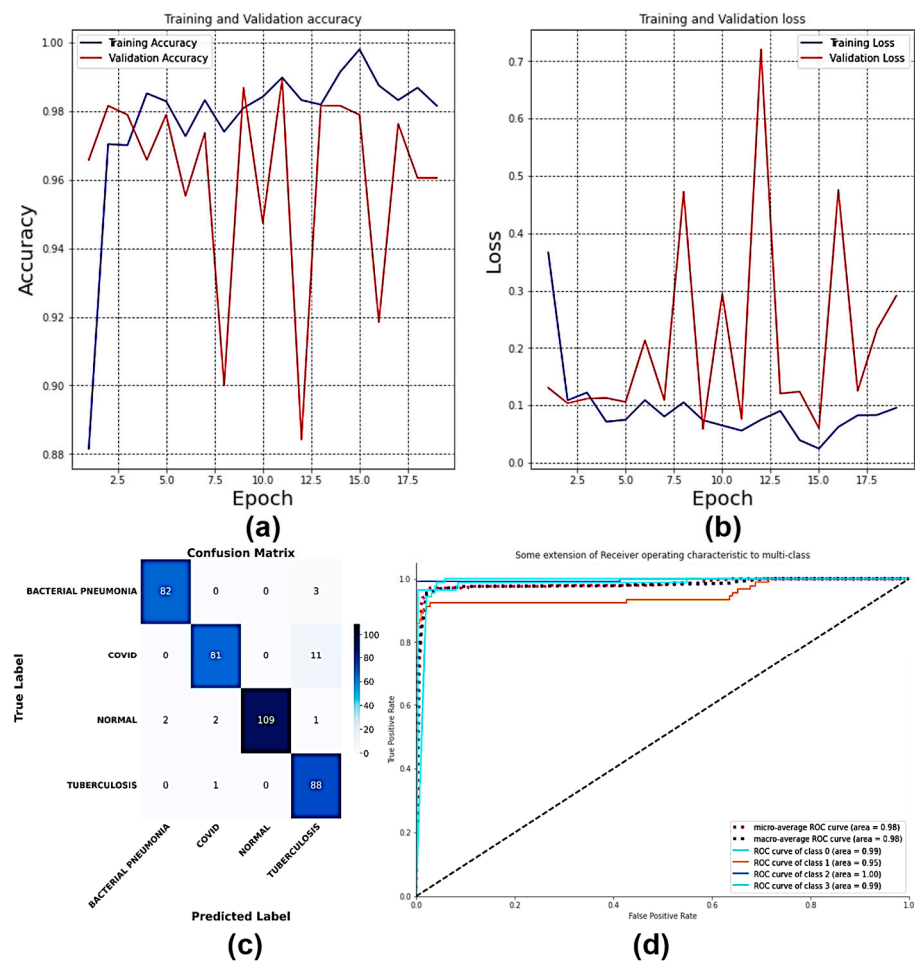
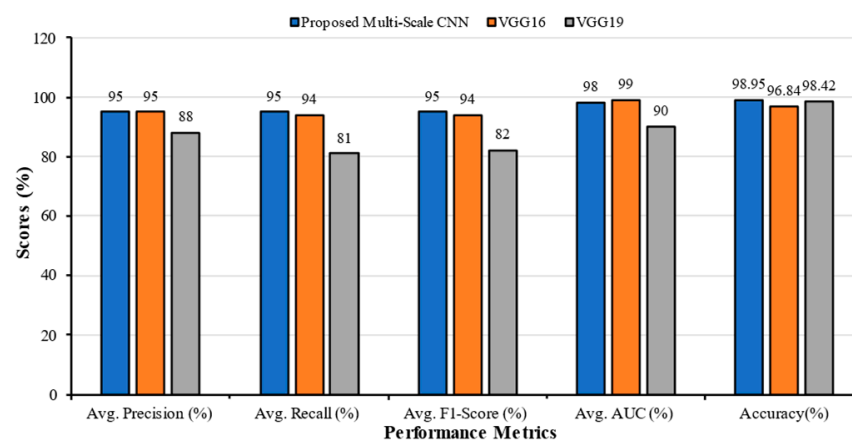


Figure 16. (a) Accuracy curves, (b) Loss curves, (c) Confusion Matrix, and (d) ROC curves of the proposed Multi-Scale CNN model with four individual classes (Bacterial Pneumonia, COVID, Normal, and Tuberculosis).

Table 10. Classification performance results for Dataset 7.

Classification Models	Classes	Precision	Recall	F1-Score	AUC	Accuracy	Average Precision	Average Recall	Average F1-Score	Average AUC
Multi-Scale CNN	Bacterial Pneumonia	0.98	0.96	0.97	0.99	98.95	0.95	0.95	0.95	0.98
	COVID	0.96	0.88	0.92	0.95					
	Normal	1.00	0.96	0.98	1.00					
	Tuberculosis	0.85	0.99	0.92	0.99					
VGG16	Bacterial Pneumonia	0.98	0.95	0.96	1.00	96.84	0.95	0.94	0.94	0.99
	COVID	0.96	0.85	0.90	0.97					
	Normal	0.98	1.00	0.99	1.00					
	Tuberculosis	0.86	0.97	0.91	0.98					
VGG19	Bacterial Pneumonia	0.98	0.68	0.81	0.86	98.42	0.88	0.81	0.82	0.90
	COVID	0.95	0.62	0.75	0.85					
	Normal	0.99	0.95	0.97	0.98					
	Tuberculosis	0.59	1.00	0.74	0.90					

**Figure 17.** Average Precision (%), Average Recall (%), Average F1-Score (%), Average AUC (%), and Accuracy of the different models with four individual classes (Bacterial Pneumonia, COVID, Normal, and Tuberculosis) for Dataset 7.

The testing accuracy and loss of the proposed model were 98.95% and 0.0589, respectively. The CM indicated that the proposed model misclassified 3 Bacterial Pneumonia images as Tuberculosis, 11 COVID-19 as Tuberculosis, 2 Normal as Bacterial Pneumonia, 2 Normal as COVID-19, 1 Normal as Tuberculosis, and 1 Tuberculosis as COVID-19. The AUC values for Bacterial Pneumonia, COVID-19, Normal, and Tuberculosis were 0.99, 0.95, 1.00, and 0.99, respectively.

4.8. Classification of Dataset 8

The proposed Multi-Scale CNN model training and validation accuracy curves, ROC curves, and confusion matrix are shown in Figure 18. The outputs of dataset 8 are presented in Table 11 and Figure 19. The testing accuracy and loss of the proposed model were 98.67% and 0.0715, respectively. The CM indicated that the proposed model misclassified one Bacterial Pneumonia image as COVID-19, two Bacterial Pneumonia as Fibrosis, one COVID-19 as Fibrosis, three Fibrosis as COVID-19, one Fibrosis as Normal, two Normal as Bacterial Pneumonia, one Tuberculosis as COVID-19, and three Tuberculosis as Fibrosis. The AUC values for Bacterial Pneumonia, COVID-19, Fibrosis, Normal, and Tuberculosis were 1.00, 1.00, 0.99, 1.00, and 0.97, respectively.

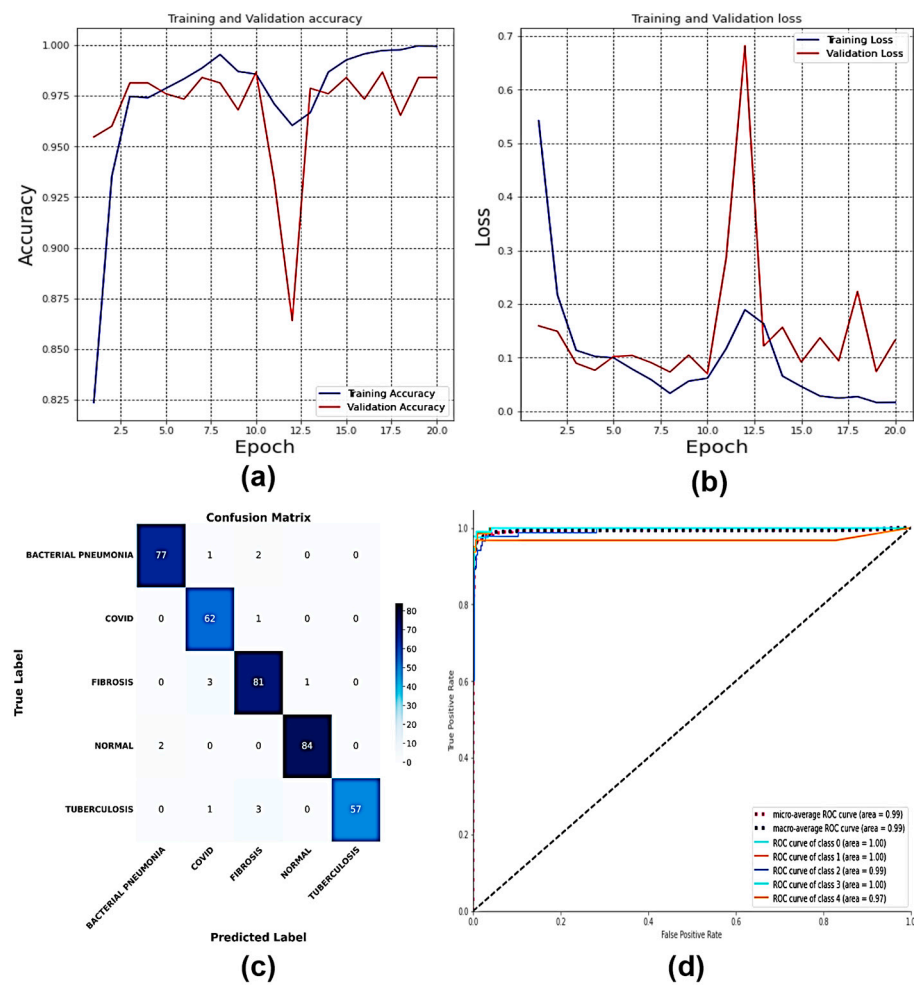


Figure 18. (a) Accuracy curves, (b) Loss curves, (c) Confusion Matrix, and (d) ROC curves of the proposed Multi-Scale CNN model with five individual classes (Bacterial Pneumonia, COVID, Fibrosis, Normal, and Tuberculosis).

Table 11. Classification performance results for Dataset 8.

Classification Models	Classes	Precision	Recall	F1-Score	AUC	Accuracy (%)	Average Precision	Average Recall	Average F1-Score	Average AUC
Multi-Scale CNN	Bacterial Pneumonia	0.97	0.96	0.97	1.00	98.67	0.96	0.96	0.96	0.99
	COVID	0.93	0.98	0.95	1.00					
	Fibrosis	0.93	0.95	0.94	0.99					
	Normal	0.99	0.98	0.98	1.00					
	Tuberculosis	1.00	0.93	0.97	0.97					
VGG16	Bacterial Pneumonia	0.98	0.96	0.97	0.99	98.32	0.96	0.96	0.96	0.99
	COVI	0.99	0.96	0.98	1.00					
	Fibrosis	0.90	1.00	0.95	0.99					
	Normal	1.00	0.90	0.95	0.98					
	Tuberculosis	0.95	0.99	0.97	0.99					
VGG19	Bacterial Pneumonia	0.90	0.98	0.94	0.99	97.68	0.95	0.95	0.95	0.99
	COVI	0.98	0.96	0.97	0.99					
	Fibrosis	0.94	0.97	0.95	0.99					
	Normal	1.00	0.83	0.91	0.99					
	Tuberculosis	0.93	0.99	0.96	0.99					

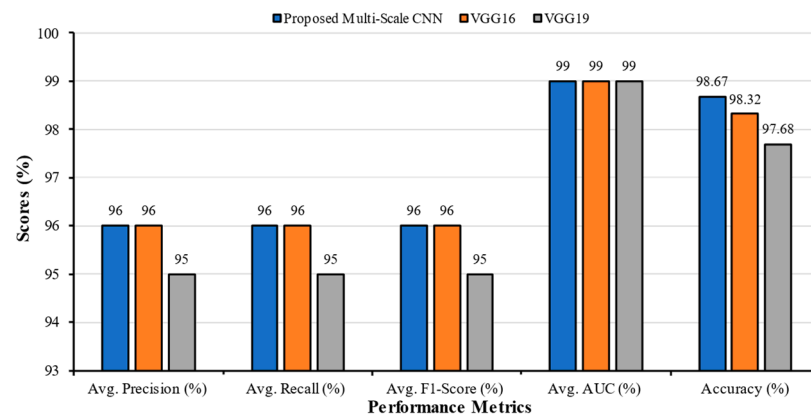


Figure 19. Average Precision (%), Average Recall (%), Average F1-Score (%), Average AUC (%), and Accuracy of the different models with five individual classes (Bacterial Pneumonia, COVID, Fibrosis, Normal, and Tuberculosis) for Dataset 8.

4.9. Classification of Dataset 9

The proposed Multi-Scale CNN model training and validation accuracy curves, ROC curves, and confusion matrix are shown in Figure 20. The outputs of dataset 9 are presented in Table 12 and Figure 21. The testing accuracy and loss of the proposed model were 97.47% and 0.0885, respectively. The CM indicated that the proposed model misclassified two Bacterial Pneumonia images as Fibrosis, three COVID-19 as Fibrosis, three Normal as Bacterial Pneumonia, three Normal as Fibrosis, two Tuberculosis as Fibrosis, seven Viral Pneumonia as Bacterial Pneumonia, one Viral Pneumonia as Fibrosis, and one Viral Pneumonia as Tuberculosis. The AUC values for Bacterial Pneumonia, COVID-19, Fibrosis, Normal, Tuberculosis, and Viral Pneumonia were 0.99, 1.00, 1.00, 1.00, 1.00, and 0.99, respectively.

Table 12. Classification performance results for Dataset 9.

Classification Models	Classes	Precision	Recall	F1-Score	AUC	Accuracy (%)	Average Precision	Average Recall	Average F1-Score	Average AUC
Multi-Scale CNN	Bacterial Pneumonia	0.90	0.98	0.94	0.99	97.47	0.96	0.95	0.95	0.99
	COVID	1.00	0.96	0.98	1.00					
	Fibrosis	0.90	1.00	0.95	1.00					
	Normal	1.00	0.91	0.95	1.00					
	Tuberculosis	0.99	0.98	0.98	1.00					
	Viral Pneumonia	1.00	0.84	0.91	0.99					
VGG16	Bacterial Pneumonia	0.88	0.70	0.78	0.93	95.79	0.90	0.90	0.89	0.97
	COVID	0.95	0.93	0.94	0.98					
	Fibrosis	0.97	0.97	0.97	1.00					
	Normal	0.99	0.92	0.95	0.99					
	Tuberculosis	0.85	0.98	0.91	0.98					
	Viral Pneumonia	0.75	0.88	0.81	0.96					
VGG19	Bacterial Pneumonia	0.71	0.72	0.72	0.91	95.61	0.78	0.72	0.72	0.91
	COVID	0.98	0.58	0.73	0.87					
	Fibrosis	0.80	0.80	0.80	0.91					
	Normal	1.00	0.54	0.70	0.95					
	Tuberculosis	0.51	1.00	0.67	0.92					
	Viral Pneumonia	0.71	0.70	0.71	0.88					

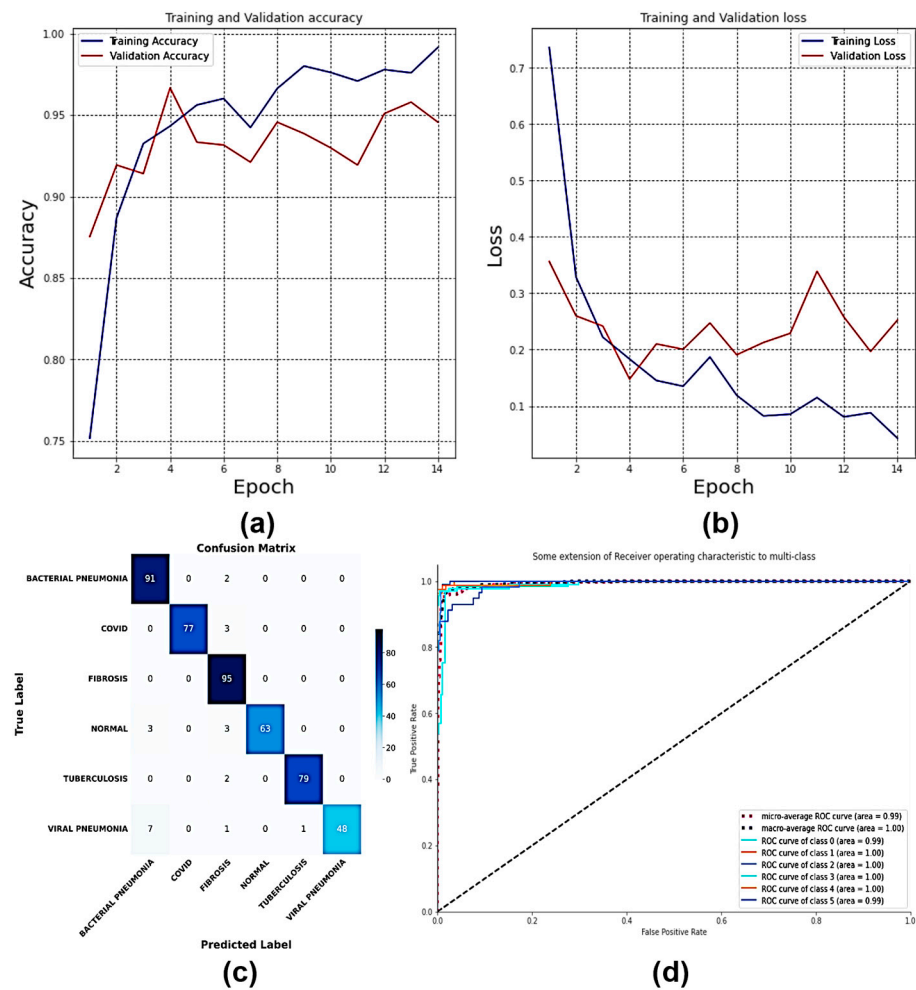


Figure 20. (a) Accuracy curves, (b) Loss curves, (c) Confusion Matrix, and (d) ROC curves of the proposed Multi-Scale CNN model with six individual classes (Bacterial Pneumonia, COVID, Fibrosis, Normal, Tuberculosis, and Viral Pneumonia).

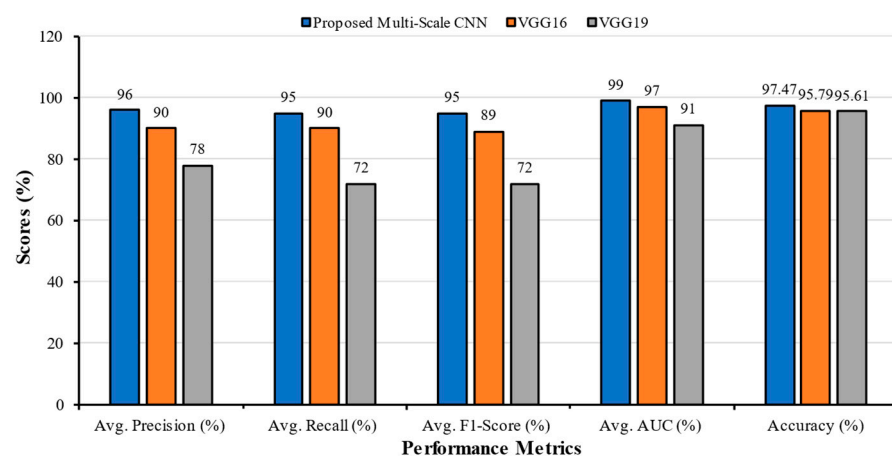


Figure 21. Average Precision (%), Average Recall (%), Average F1-Score (%), Average AUC (%), and Accuracy of the different models with six individual classes (Bacterial Pneumonia, COVID, Fibrosis, Normal, Tuberculosis, and Viral Pneumonia) for Dataset 9.

4.10. Classification of Dataset 10

The proposed MS-CNN model training and validation accuracy curves, ROC curves, and confusion matrix are shown in Figure 22. The outputs of dataset 10 are presented

in Table 13 and Figure 23. This testing accuracy and loss of the proposed model were 96.05% and 0.1386, respectively. The CM indicated that the proposed model misclassified three Bacterial Pneumonia image as Viral Pneumonia, two COVID-19 as Tuberculosis, one Fibrosis as Normal, one Fibrosis as Tuberculosis, one Lung Opacity as Bacterial Pneumonia, three Lung Opacity as COVID-19, one Lung Opacity as Tuberculosis, one Tuberculosis as Viral Pneumonia, one Viral Pneumonia as Bacterial Pneumonia, and one Viral Pneumonia as COVID-19. The AUC values for Bacterial Pneumonia, COVID-19, Fibrosis, Lung Opacity, Normal, Tuberculosis, and Viral Pneumonia were 0.99, 0.94, 0.83, 0.81, 0.99, 0.97, and 0.95, respectively. The obtained area values of the PR curve for Bacterial Pneumonia, COVID-19, Fibrosis, Lung Opacity, Normal, Tuberculosis, and Viral Pneumonia were 0.915, 0.987, 0.989, 0.980, 0.996, 0.997, and 0.915, respectively.

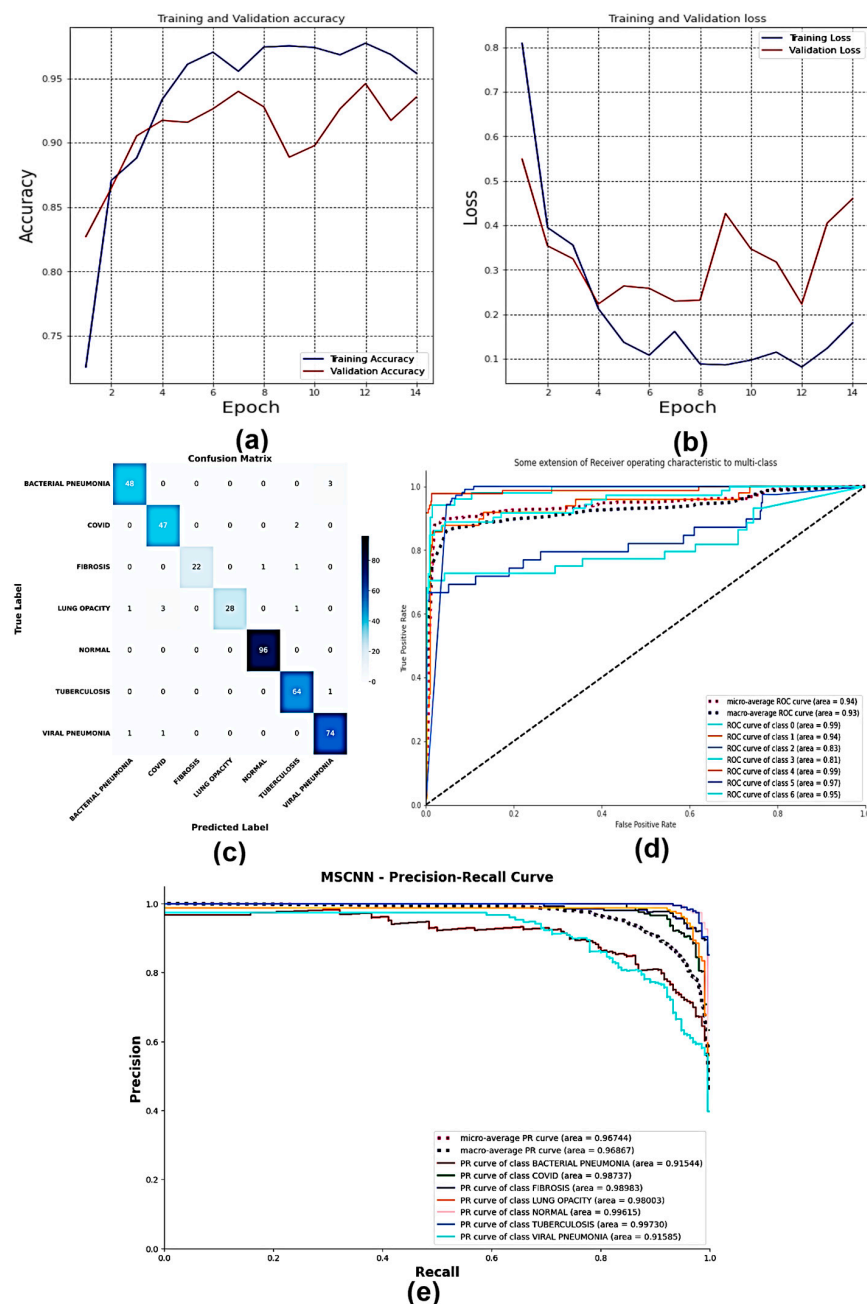
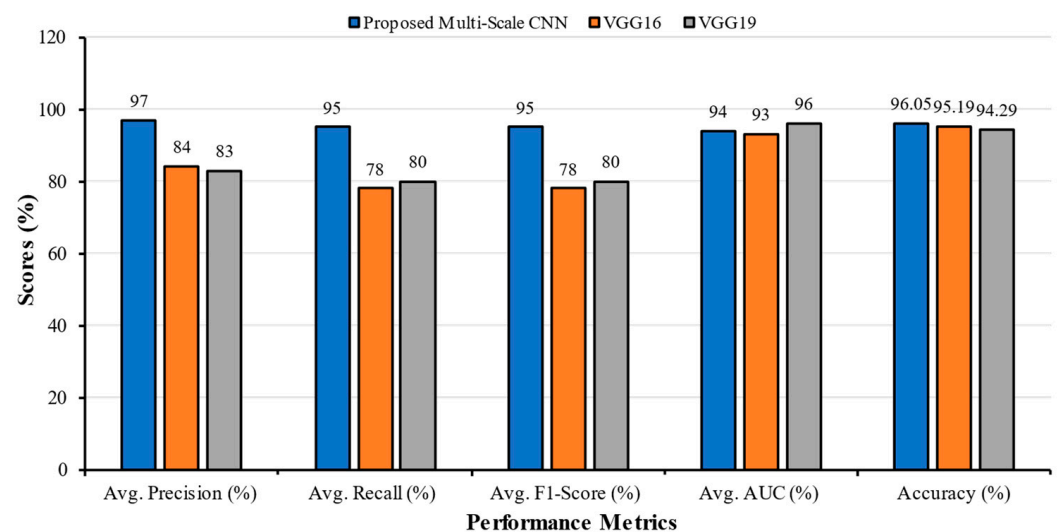


Figure 22. (a) Accuracy curves, (b) Loss curves, (c) Confusion Matrix, (d) ROC curves, and (e) Precision Recall curves of the proposed Multi-Scale CNN model with seven individual classes (Bacterial Pneumonia, COVID, Fibrosis, Lung Opacity, Normal, Tuberculosis, and Viral Pneumonia).

Table 13. Classification performance results for Dataset 10.

Classification Models	Classes	Precision	Recall	F1-Score	AUC	Accuracy (%)	Average Precision	Average Recall	Average F1-Score	Average AUC
Multi-Scale CNN	Bacterial Pneumonia	0.96	0.94	0.95	0.99	96.05	0.97	0.95	0.95	0.94
	COVID	0.92	0.96	0.94	0.94					
	Fibrosis	1.00	0.92	0.96	0.83					
	Lung Opacity	1.00	0.85	0.92	0.81					
	Normal	0.99	1.00	0.99	0.99					
	Tuberculosis	0.94	0.98	0.96	0.97					
	Viral Pneumonia	0.95	0.97	0.96	0.95					
VGG16	Bacterial Pneumonia	0.92	0.68	0.78	0.93	95.19	0.84	0.78	0.78	0.93
	COVID	0.92	0.43	0.59	0.81					
	Fibrosis	0.74	0.94	0.82	0.96					
	Lung Opacity	0.97	0.59	0.74	0.93					
	Normal	0.91	0.99	0.95	0.99					
	Tuberculosis	0.63	1.00	0.78	0.95					
	Viral Pneumonia	0.76	0.84	0.80	0.95					
VGG19	Bacterial Pneumonia	0.83	0.48	0.60	0.93	94.29	0.83	0.80	0.80	0.96
	COVID	0.96	0.60	0.74	0.96					
	Fibrosis	0.91	0.91	0.91	0.99					
	Lung Opacity	0.82	0.95	0.88	0.96					
	Normal	0.89	0.85	0.87	0.98					
	Tuberculosis	0.77	0.99	0.86	0.97					
	Viral Pneumonia	0.61	0.80	0.69	0.93					

**Figure 23.** Average Precision (%), Average Recall (%), Average F1-Score (%), Average AUC (%), and Accuracy of the different models with seven individual classes (Bacterial Pneumonia, COVID, Fibrosis, Lung Opacity, Normal, Tuberculosis, and Viral Pneumonia) for Dataset 10.

4.11. Explainable AI on MS-CNN Interpretability

For the model interpretability through different explainable AI techniques, image plots were created to visualize SHAP values generated by the explainer object [36], and Grad-CAM was used to generate a heatmap of CXR images.

The weight of the final convolution layer is typically used to create a heatmap from the original image [37]. The heatmap was then applied to the original input image to create the output image. The damaged area on the CXR is depicted in the overlay image to identify the disease category. By enabling quick viewing of the damaged region on the image, it will help medical professionals to identify problems.

To begin with, a SHAP explainer was established for the model to calculate SHAP values for a given set of instances. SHAP—partition explainer function was employed to create a specialized SHAP partition explainer explicitly designed for deep learning models. The SHAP values represent how much each pixel contributes to the model's output for every instance in the dataset. In binary classification, two sets of SHAP values correspond to the two classes. These SHAP values are organized in matrices where rows represent instances, and columns represent features. Positive values indicate features that push the prediction toward the positive class, while the negative values indicate features that push toward the negative class. Figure 24 shows an initial image plot generated using the SHAP values. The plot displays the actual image, with certain parts highlighted in shades of red and blue. Red areas signify positive contributions to the prediction of that class, while blue areas indicate negative contributions. Red regions enhance the probability of predicting a class, while blue regions diminish it.

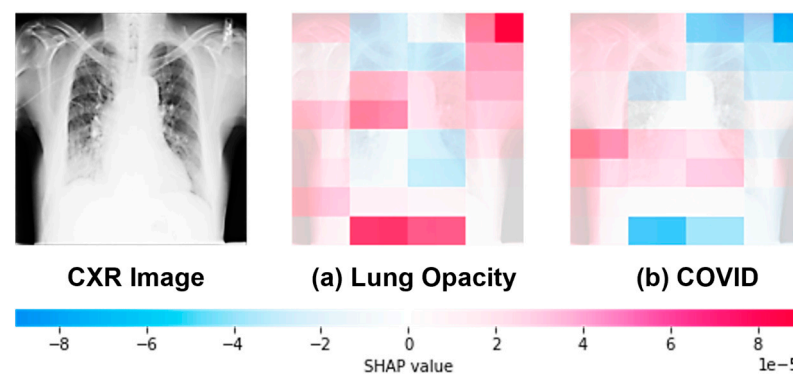


Figure 24. SHAP Partition Explainer with image plot on a lung opacity sample; top two categories that the model thinks the sample belongs to are (a) Lung opacity and (b) COVID.

Figure 24 shows a lung opacity sample prediction extracted through the MS-CNN classifier using SHAP Partition Explainer with an image plot on a lung opacity sample. The model thinks the sample belongs to the top two categories: Lung Opacity and COVID. On the x-axis of Figure 24a, the higher SHAP value to the right corresponds to a higher prediction value (“Lung Opacity” class), and the lower SHAP value to the left corresponds to a lower prediction value (not the “Lung Opacity” class). The larger the pixel value in the lung region (the redder color), the higher the SHAP value. This means that when the pixel value of the lung outermost top, left, and suitable regions in Figure 24a are more extensive, the SHAP value corresponds to a higher prediction value. Hence, the model is more likely to consider the data as a “Lung Opacity” class. On the other hand, the smaller the pixel value in lung center regions (the bluer color), the smaller the SHAP value. Hence, when the pixel value of Figure 24a is smaller inside the lung cavities, the model is less likely to consider the data as a “Lung Opacity” class. Looking at Figure 24b, the “COVID” class is the second-highest probability, where the whole left-half region of the sample corresponds to the higher prediction value of the “COVID” class, and the right-half region corresponds to the lower prediction value of the “COVID” class.

Figures 25 and 26 also show the first seven categories the model thinks the image belongs to. Figure 25 explains the same lung opacity sample for the seven class predictions used in Figure 24. The model confuses the image with not only “COVID”, as explained above, with the second-highest probability, but also with the “Fibrosis” and “Tuberculosis” classes. It can be seen in Figure 25d that the prominence of red areas (positive SHAP values) in the plot signifies a tendency toward the prediction “Lung Opacity” class, indicating the correct prediction.

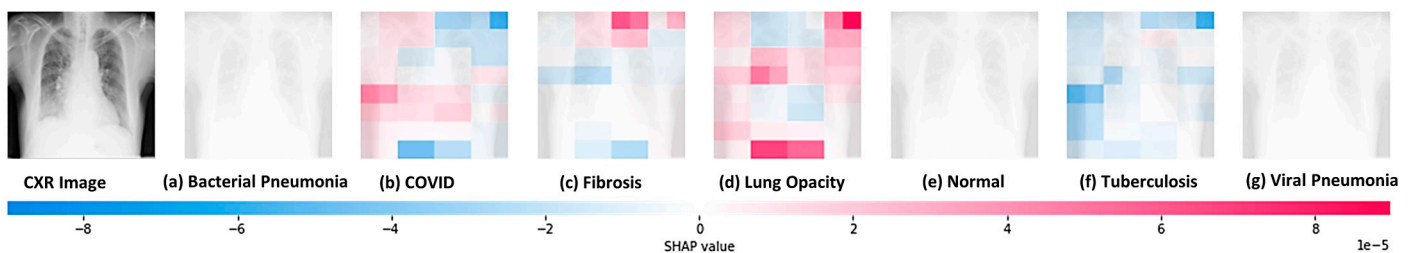


Figure 25. SHAP partition Explainer with Image Plot on a Lung Opacity sample; Predictions on all seven categories where the model thinks the sample is (a) Bacterial Pneumonia, (b) COVID, (c) Fibrosis, (d) Lung Opacity, (e) Normal, (f) Tuberculosis, and (g) Viral Pneumonia.

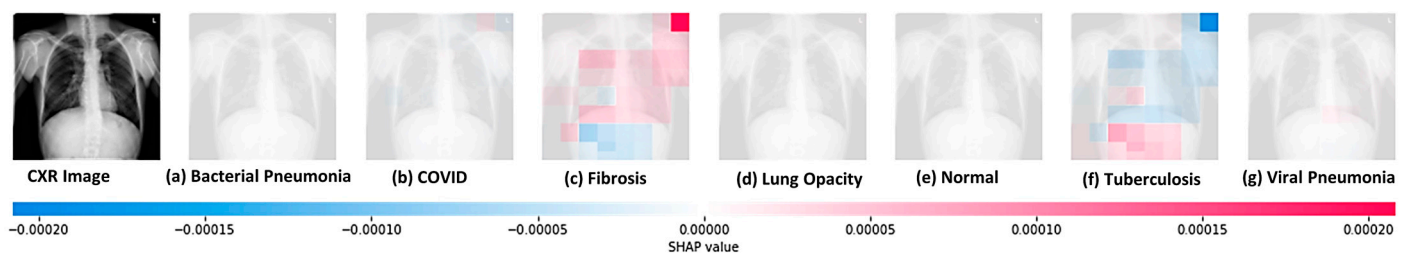


Figure 26. SHAP Partition Explainer with Image Plot on a Fibrosis sample; Predictions on all seven categories where the model thinks the sample is (a) Bacterial Pneumonia, (b) COVID, (c) Fibrosis, (d) Lung Opacity, (e) Normal, (f) Tuberculosis, and (g) Viral Pneumonia.

The sample in Figure 26 shows an essential concept about explanations for black-box models; they explain what the model is predicting but do not attempt to explain if the predictions are correct. The similarity in magnitude of red areas (positive SHAP values) in Figure 26d with the presence of blue areas (negative SHAP values) in Figure 26f creates confusion in predicting either the “Fibrosis” class or the “Tuberculosis” class. The explainer generates positive SHAP values for “Fibrosis” and negative SHAP values for the “Tuberculosis” class, where the magnitudes are similar for both SHAP values, indicating a higher probability of misclassification in model prediction for the sample.

Figure 27 shows Grad-CAM representation on example images of lung disorders where the MS-CNN model primarily detects the afflicted area as (a) Fibrosis and (b) Tuberculosis. Grad-CAM shows that a region’s more significant importance to the model is shown by its red hue, while its lesser priority is indicated by its blue color. However, caution should be taken while interpreting the heat maps. The same sample used for the SHAP explanation in Figure 26, when used in Grad-CAM, shows heatmap regions extracted from the deeper layer of the model generating heatmap for the “Fibrosis” class and “Tuberculosis” class. This indicates a higher probability of misclassification by the model.

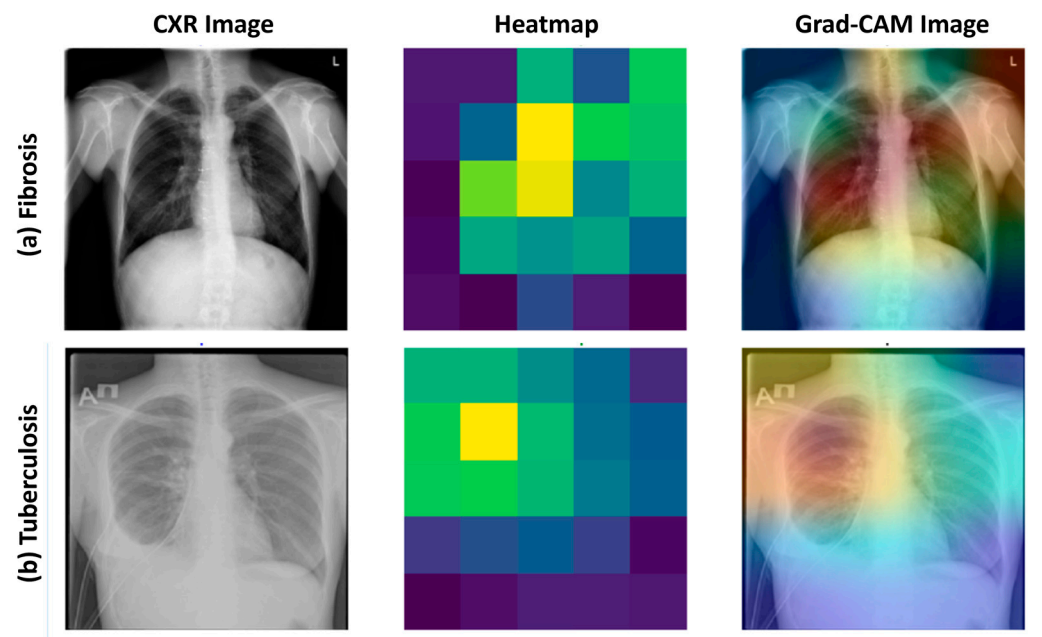


Figure 27. Original CXR, Heatmap, and Super-imposed Grad-CAM image of Multi-scale CNN Model with two individual classes for one sample: (a) Fibrosis (on top) and (b) Tuberculosis (on bottom).

5. Discussion

5.1. Comparative Analysis of Multi-Scale CNN with Different Datasets

Figure 28 compares the performance of the MS-CNN model in correctly identifying lung-related disorders for various datasets. The illustration clearly shows that the testing accuracy was lowest (96.05%) in the case of dataset 10 (seven classes), but when the number of classes was reduced, the testing accuracy improved. For example, the dataset 1 (Binary class) has an accuracy of 100.00%. However, a little discrepancy is discovered between dataset 3 (three classes), dataset 4 (three classes), and dataset 5 (four classes).

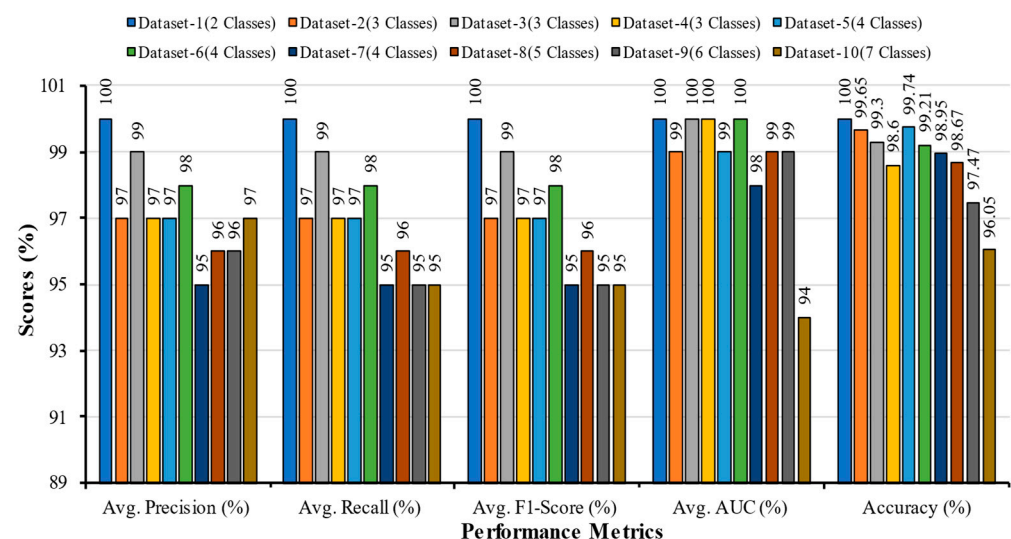


Figure 28. Comparison of performance metrics for all the ten datasets from Class-2 to Class-7 obtained by Multi-Scale CNN for identifying lung-affected diseases.

Increasing class means including additional images of the same type of lung disease in the training and testing datasets, which reduces accuracy. Higher AUC values prove the model's capacity to correctly classify lung-related disorders even from a more significant number of lung disorders.

5.2. Comparative Analysis of Multi-Scale CNN with other Research in the Literature

A comparison between the proposed MS-CNN classification technique and other research performed by deep learning algorithms based on the CXR images with two-class, three-class, four-class, five-class, six-class, and seven-class is presented in Table 14.

Table 14. Comparative analysis with different diagnostic approaches of previous works.

Research	Number of Classes	Dataset Classes	Applied Architecture	Accuracy%
Al-Waisy et al. [11]	2	COVID, Normal	COVID-CheXNet Proposed MS-CNN	99.99 100.00
Srivastava et al. [12]	2	COVID, Normal	CoviXNet Proposed MS-CNN	99.47 100.00
Abida et al. [15]	2	COVID, Normal	2D-CNN Proposed MS-CNN	98.00 100.00
Nahiduzzaman et al. [13]	3	COVID, Normal, Pneumonia	CNN-ELM Proposed MS-CNN	97.42 98.60
Yaman et al. [14]	3	COVID, Normal, Pneumonia	CNN (ACL Model) Proposed MS-CNN	96.00 98.60
Abida et al. [15]	3	Bacterial Pneumonia, COVID, Normal	2D-CNN Proposed MS-CNN	97.49 98.60
Elakkiya et al. [16]	4	COVID, Normal, Pneumonia, Tuberculosis	SCS-Net Proposed MS-CNN	94.05 98.95
Abida et al. [15]	4	Bacterial Pneumonia, COVID, Normal, Tuberculosis	2D-CNN Proposed MS-CNN	97.81 98.95
Hussain et al. [17]	4	Bacterial Pneumonia, COVID, Normal, Viral Pneumonia	CoroDet Proposed MS-CNN	91.20 98.33
Abida et al. [15]	5	Bacterial Pneumonia, COVID, Fibrosis, Normal, Tuberculosis	2D-CNN Proposed MS-CNN	96.96 98.67
Al-Timemy et al. [18]	5	Bacterial Pneumonia, COVID, Normal, Tuberculosis, Viral Pneumonia	ResNet-50 with ensemble of subspace discriminant classifier Proposed MS-CNN	91.60 97.00
Abida et al. [15]	6	Bacterial Pneumonia, COVID, Fibrosis, Normal, Tuberculosis, Viral Pneumonia	2D-CNN Proposed MS-CNN	96.75 97.47
Abida et al. [15]	7	Bacterial Pneumonia, COVID, Fibrosis, Lung Opacity, Normal, Tuberculosis, Viral Pneumonia	2D-CNN Proposed MS-CNN	93.15 96.05

Note: Bold text indicates the best values.

The COVID-CheXNet system proposed by Al-Waisy et al. [11] and the Al-Srivastava et al. [12]-proposed CoviXNet successfully diagnosed COVID-19 patients for binary classifications with an accuracy rate of 99.99% and 99.47%, respectively. In those cases, the MS-CNN model performed with 100% accuracy.

Nahiduzzaman et al. [13] employed a lightweight CNN-ELM method with only three layers in which they applied a three-class classification approach that achieved 97.42% accuracy. Yaman et al. [14] introduced the ACL model, combining attention, LSTM, and CNN for classifying healthy, COVID-19, and pneumonia cases in chest X-ray (CXR) images. The model achieved 96% accuracy on an 80:20 train/test ratio. Changing the ratio creates

an impact on the accuracy. However, in this model, every layer's outputs were merged to extract additional features to predict the exact output with a higher accuracy of 98.60%.

Abida et al. [15] designed a 2D-CNN model to classify Bacterial Pneumonia, COVID-19, Fibrosis, Lung Opacity, Normal, Tuberculosis, and Viral Pneumonia. For two-, three-, four-, five-, six-, and seven-class schemes, this model achieved 98.00%, 97.49%, 97.81%, 96.96%, 96.75%, and 93.15%, respectively. In their research, they utilized a lightweight 2D-CNN model with three Conv2D layers, which extracts features for classification, but more is needed to obtain higher accuracy in multi-class classifications. In two-, three-, and four-class schemes, they acquired good results compared to other related works, but at higher-class numbers, the accuracy is reduced. For example, in the seven-class scheme, the classification accuracy is 93.15%. However, the proposed model achieved an accuracy of 96.05%, which is nearly 3% greater than Abida et al. The proposed model used all the layer's output predictions to merge from the multiple feature maps at different resolution scales to improve class predictions.

Elakkiya et al. [16] presented a novel approach SCS-Net for categorizing COVID-19, pneumonia, tuberculosis, and normal with an accuracy of 94.05%. Hussain et al. [17] introduced CoroDet employing a four-class classification, achieving an accuracy of 91.20%. Al-Timemy et al. [18] presented a classification of five classes using a combination of ResNet-50 for DF (Deep Features) computation and an ensemble of subspace discriminant classifiers with an accuracy of 91.6%. In these cases, the proposed model achieved better scores of 98.95%, 98.33%, and 97.00%.

5.3. Comparison with Datasets of Other Literature

The MS-CNN model was further validated by training and testing the model on other datasets (balanced and imbalanced both). Model comparison is shown in Table 15. Al-Waisey et. al. [11] proposed a COVID-CheXNet framework with two deep learning methods (e.g., ResNet34 and HRNet). The authors created their own COVID-19-vs-normal dataset. The dataset contains 400 images of confirmed COVID-19 cases gathered from 4 different sources and 400 chest X-ray images of normal condition. The whole dataset is split into training, validation, and test sets with 70% for training and validation and 30% for test set evaluation. The proposed MS-CNN model was evaluated for binary-class performance on the dataset. In the literature, ResNet34 and HRNet diagnosed the COVID-19 patients with a DAR (detection accuracy rate) of 89.98% and 90%, respectively. The proposed model outperformed both ResNet34 and HRNet models with a testing accuracy of 99.38%. MS-CNN obtained an average testing accuracy, precision, recall, and f1-score of 99.38%, 99.38%, 99.38%, and 99.98%, respectively.

Table 15. Comparison with Datasets of Other Literature.

Literature No.	Model	Classes	Training Accuracy (%)	Testing Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC (%)	Training Time (min)	Test Set Evaluation Time (s)
[11]	ResNet34	2	98.9	89.88	90.83	89.14	89.98	0.154	0.216	81.00
	HRNet	2	98.9	90.00	91.31	88.98	90.13	0.121	0.136	88.00
	MS-CNN	2	100.0	99.38	99.38	99.38	99.98	0.0061	0.0779	100.00
[15]	2D-CNN	5	98.90	96.96	96.8	97.2	97.0	99.77	42.62	16.39
		6	99.04	96.75	93.43	94.43	93.86	99.39	62.74	15.20
		7	98.44	93.15	93.43	94.43	93.86	99.39	62.74	22.37
	MS-CNN	5	98.80	98.80	97.99	97.99	97.99	99.92	38.16	5.1
		6	98.75	98.10	96.55	98.75	96.49	99.91	39.81	5.3
		7	98.63	95.18	96.23	96.23	96.23	99.83	50.58	7.5

Note: Bold text indicates the best values.

A class-wise comparison has been shown against 2D-CNN architecture with transfer learning on the dataset developed by Abida et. al. [15]. The dataset contains 18,564 CXR images. The MS-CNN was tested for 5–7 class performances on the dataset of [15]. On 5, 6, and 7 classes, the proposed model outperformed 2D-CNN with an average testing accuracy of 98.80%, 98.10%, and 95.18%, respectively. Model training times were also comparably lower than the 2D-CNN with the highest seven-class training time being 50 min (almost 12 min faster than 2D-CNN).

5.4. Comparison with State-of-the-Art Models on Dataset 10

The efficiency of the proposed MS-CNN was compared with the current state-of-the-art classification architectures on dataset 10. The training dataset contains 5320 CXR images (80%), whereas the validation (10%) and testing (10%) datasets contain the rest of the 1330 images, 665 each. The proposed MS-CNN was tested for seven class performances against SOTA architectures like DenseNet, InceptionResNetV2, NasNet, ResNet, etc., using transfer learning with the top layer removed. Models such as NASNet Mobile, ResNet101V2, and ResNet152 performed best with fine-tuned weights trained with dataset 10 on the bottom layers and pre-trained ImageNet weights on the top layers. On the other hand, ResNet50 and ResNet101 performed best with fully-trained top and bottom layer weights. The rest of the pre-trained models did not need any further modifications with weight training to generate low-bias and low-variance predictions. The models were trained on Adam Optimizer for 25 epochs or less with early stop callback for patience 10. The proposed model outperformed most models with an average testing accuracy, precision, and recall of 96.05%, 97.00%, and 95.00%, respectively. The comparison of performance metrics is shown in Table 16 and computational time is presented in Figure 29. The MS-CNN model only took only 3.12 s for evaluating the whole test dataset.

Table 16. Comparison with SOTA Pretrained Models on Dataset 10 (7 classes).

Models	Training Accuracy (%)	Testing Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC (%)	Model Training Time (min)	Test Set Evaluation Time (s)
VGG16	98.50	95.19	84.00	78.00	78.00	93.00	19.74	9.72
VGG19	97.65	94.29	83.00	80.00	80.00	96.00	22.41	10.87
DenseNet121	98.38	94.81	95.29	94.36	94.82	99.64	18.45	8.94
DenseNet201	97.82	95.86	96.06	95.41	95.73	99.79	20.17	10.86
InceptionV3	93.46	83.46	85.79	81.73	83.71	98.12	15.41	6.06
Inception ResNetV2	95.81	85.34	86.94	84.06	85.47	98.53	18.01	8.64
Xception	96.56	86.99	89.14	85.79	87.43	98.80	19.88	9.25
NASNet **	96.71	84.14	84.89	82.78	83.82	97.77	14.99	7.53
ResNet50 *	77.84	75.86	78.38	73.61	75.92	97.75	15.03	6.17
ResNet50V2	86.99	73.68	82.07	69.17	75.07	95.79	14.92	5.87
ResNet101 *	83.50	82.48	84.34	80.98	82.62	98.38	26.69	6.65
ResNet101V2 **	79.62	65.79	72.12	63.01	67.26	95.80	9.61	6.18
ResNet152 **	73.87	75.94	78.03	74.51	76.22	96.43	50.07	11.44
ResNet152V2	91.24	79.55	84.24	73.53	78.52	97.42	17.46	7.88
MS-CNN	98.70	96.05	97.00	95.00	95.00	94.00	12.11	3.12

* All weights trained on Dataset 10. ** Bottom layer weights fine-tuned on Dataset 10. Bold text indicates the best values.

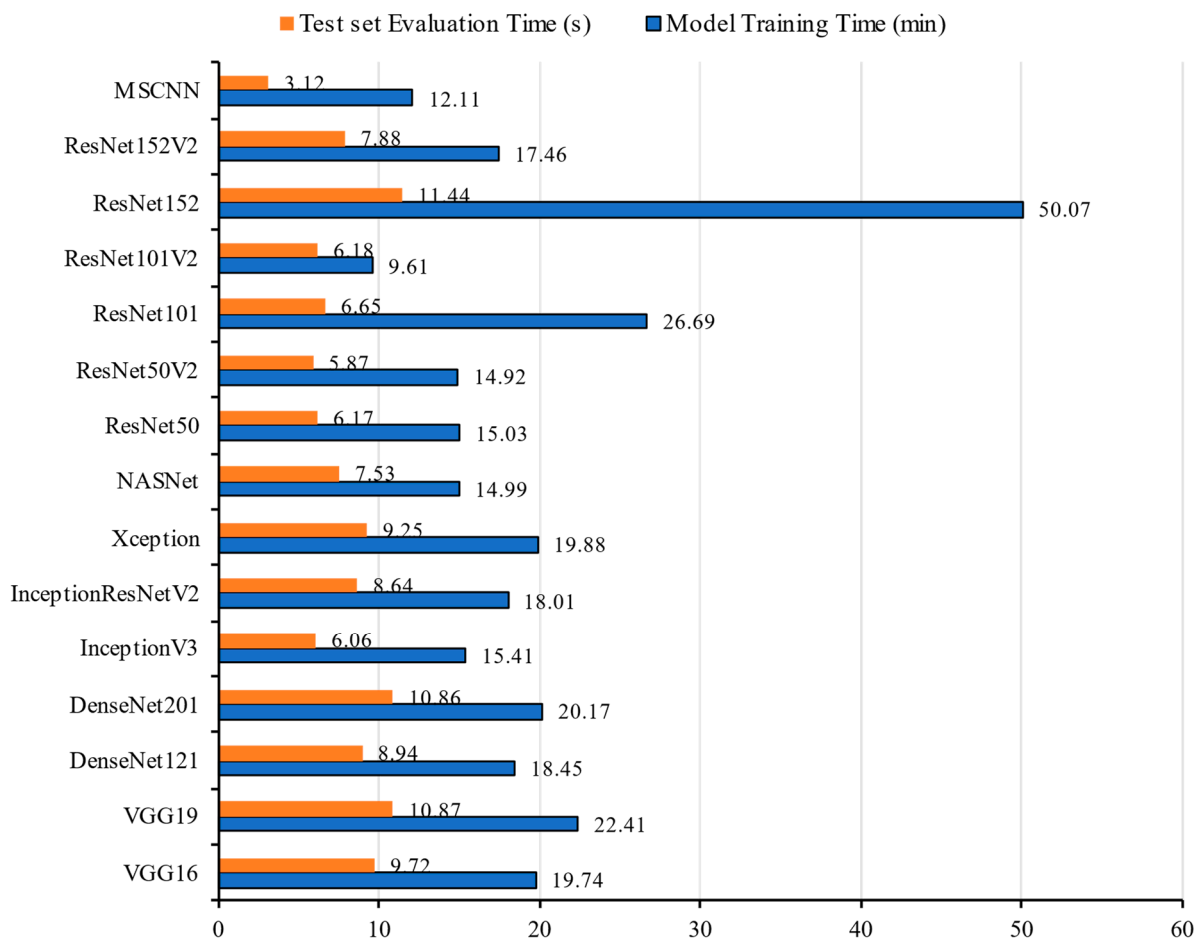


Figure 29. Comparison of computational time for all state-of-the-art (SOTA) models of Dataset 10.

5.5. Strength and Limitations

Motivated by the limitations of higher-class classification problems, the authors of this study integrated multiple databases to build a dataset of 6650 CXR images classified into a maximum of seven class classifications. In this investigation, the MS-CNN model, a pioneering deep learning framework tailored to excel in multi-class classification scenarios, was applied. Building on the strengths of existing models, this framework was designed to overcome this challenge by employing advanced architectural and optimization techniques. In the proposed model, multiple Conv2D blocks were applied and concatenated to use all the layer's output predictions to merge from the multiple feature maps at different resolution scales to improve class predictions. The resulting model achieved high accuracy across diverse lung conditions, even as the class count expanded, making it a robust tool for accurate disease classification.

The term “Multi-Scale CNN” denotes the model's ability to integrate information from different resolution scales using multiple feature maps. This allows it to capture fine and coarse-level features within images, enhancing its effectiveness in identifying lung-related diseases with varied manifestations.

In all the formed datasets from dataset 1 to dataset 10, the model was run and it was observed that in every case the scores are superior. All the details regarding this are presented in Section 5.1. In comparison with other recent research, it was observed that the proposed model performed better than others in terms of testing accuracy shown in Section 5.2. Compared with others in their respective datasets either balanced or imbalanced, the current model performed better as presented in Section 5.3. In the case of the maximum number of classes (seven classes) in dataset 10, the proposed model outperformed various

pre-trained models in discriminating lung disorders as well as healthy individuals in terms of the performance matrixes used along with the computational times shown in Section 5.4.

The lack of data on other types of lung disorders limits this study. Significant improvements can be made with greater data availability and algorithm training using radiological data from patients and nonpatients throughout the world. It should be noted that this MS-CNN model was not constructed based on a lightweight structure, but VGG-16 was employed as the backbone of the model, and some Conv2D layers were additionally used to concatenate the output, which employs more parameters than some pre-trained models. Therefore, running the model might require more hardware resources. However, this minor issue did not limit its superiority in terms of higher accuracy and shorter testing time.

6. Conclusions

In this study, a highly accurate Multi-Scale CNN architecture was designed to predict 724 distinct classes of images, encompassing COVID-19 and five other lung-affected disorders. Notably, the MS-CNN model exhibits remarkable efficiency in COVID-19 detection, resulting in significantly higher testing accuracy compared to the previous methodologies. Even as the number of classes increases, the MS-CNN consistently outperformed all previously reported models in the literature, showcasing a novel approach that addresses a persistent limitation in the existing research. Additionally, the current approach substantially shortens the testing duration in comparison with the state-of-the-art models, offering the potential for expedited medical interventions for patients with lung-related diseases. In the case of dataset 10, which comprises seven classes, the MS-CNN model achieves an impressive accuracy rate of 96.05%, complemented by precision, recall, F1-score, and AUC values averaging at 97%, 95%, 95%, and 94%, respectively. Likewise, in dataset 9, encompassing six classes, the MS-CNN demonstrates an accuracy rate of 97.47%, coupled with precision, recall, F1-score, and AUC values averaging at 96%, 95%, 95%, and 99%, respectively. Better classification scores are achieved by merging predictions from several feature maps at various resolution scales using the additional Conv2D layers with the backbone VGG16. SHAP and Grad-CAM as XAI techniques were integrated into the model, enhancing its interpretability, which ultimately brings further confidence for practical applications. As part of future development, a comprehensive plan has been devised to expand the number of disease classes in future studies.

Author Contributions: Conceptualization, O.S., M.R.I. and M.K.S.; methodology, O.S., M.K.S., M.T.I. and M.F.A.; validation, M.A. and J.H.; formal analysis, O.S., M.R.I., M.K.S., M.T.I., M.F.A., M.A. and J.H.; investigation, O.S., M.R.I., M.K.S., M.T.I. and M.F.A.; data curation, M.F.A. and M.T.I.; writing—O.S., M.K.S., M.T.I. and M.F.A.; writing—review and editing, M.R.I., J.H. and M.A.; visualization, O.S. and M.F.A.; supervision, M.R.I., M.A. and J.H.; project administration, M.A. and J.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Samples of the compounds are available from the authors.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Coronavirus Disease (COVID-19) Situation Reports. Available online: <https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports> (accessed on 16 February 2023).
2. Teixeira, L.O.; Pereira, R.M.; Bertolini, D.; Oliveira, L.S.; Nanni, L.; Cavalcanti, G.D.C.; Costa, Y.M.G. Impact of Lung Segmentation on the Diagnosis and Explanation of COVID-19 in Chest X-ray Images. *Sensors* **2021**, *21*, 7116. [CrossRef] [PubMed]
3. Kim, S.; Rim, B.; Choi, S.; Lee, A.; Min, S.; Hong, M. Deep Learning in Multi-Class Lung Diseases' Classification on Chest X-ray Images. *Diagnostics* **2022**, *12*, 915. [CrossRef]

4. Alsharif, R.; Al-Issa, Y.; Alqudah, A.M.; Qasmieh, I.A.; Mustafa, W.A.; Alquran, H. PneumoniaNet: Automated Detection and Classification of Pediatric Pneumonia Using Chest X-ray Images and CNN Approach. *Electronics* **2021**, *10*, 2949. [\[CrossRef\]](#)
5. Shamrat, F.M.J.M.; Azam, S.; Karim, A.; Islam, R.; Tasnim, Z.; Ghosh, P.; De Boer, F. LungNet22: A Fine-Tuned Model for Multiclass Classification and Prediction of Lung Disease Using X-ray Images. *J. Pers. Med.* **2022**, *12*, 680. [\[CrossRef\]](#) [\[PubMed\]](#)
6. Antosz, K.; Machado, J.; Mazurkiewicz, D.; Antonelli, D.; Soares, F. Systems Engineering: Availability and Reliability. *Appl. Sci.* **2022**, *12*, 2504. [\[CrossRef\]](#)
7. Martyshev, N.V.; Malozyomov, B.V.; Sorokova, S.N.; Efremkov, E.A.; Valuev, D.V.; Qi, M. Review Models and Methods for Determining and Predicting the Reliability of Technical Systems and Transport. *Mathematics* **2023**, *11*, 3317. [\[CrossRef\]](#)
8. Reshi, A.A.; Rustam, F.; Mehmood, A.; Alhossan, A.; Alrabiah, Z.; Ahmad, A.; Alsawailem, H.; Choi, G.S. An Efficient CNN Model for COVID-19 Disease Detection Based on X-ray Image Classification. *Complexity* **2021**, *2021*, 6621607. [\[CrossRef\]](#)
9. Muhammad, G.; Shamim Hossain, M. COVID-19 and Non-COVID-19 Classification using Multi-layers Fusion From Lung Ultrasound Images. *Inf. Fusion* **2021**, *72*, 80–88. [\[CrossRef\]](#)
10. Mahajan, S.; Raina, A.; Gao, X.Z.; Pandit, A.K. COVID-19 detection using hybrid deep learning model in chest X-rays images. *Concurr. Comput. Pract. Exp.* **2021**, *34*, e6747. [\[CrossRef\]](#)
11. Al-Waisy, A.S.; Al-Fahdawi, S.; Mohammed, M.A.; Abdulkareem, K.H.; Mostafa, S.A.; Maashi, M.S.; Arif, M.; Garcia-Zapirain, B. COVID-chexnet: Hybrid deep learning framework for identifying COVID-19 virus in chest X-rays images. *Soft Comput.* **2023**, *27*, 2657–2672. [\[CrossRef\]](#)
12. Srivastava, G.; Chauhan, A.; Jangid, M.; Chaurasia, S. Covixnet: A novel and efficient deep learning model for detection of COVID-19 using chest X-ray images. *Biomed. Signal Process Control* **2022**, *78*, 103848. [\[CrossRef\]](#) [\[PubMed\]](#)
13. Nahiduzzaman, M.; Goni, M.O.F.; Islam, M.R.; Sayeed, A.; Anower, M.S.; Ahsan, M.; Kowalski, M. Detection of various lung diseases including COVID-19 using extreme learning machine algorithm based on the features extracted from a lightweight CNN architecture. *Biocybern. Biomed. Eng.* **2023**, *43*, 528–550. [\[CrossRef\]](#)
14. Akbulut, Y. Automated Pneumonia Based Lung Diseases Classification with Robust Technique Based on a Customized Deep Learning Approach. *Diagnostics* **2023**, *13*, 260. [\[CrossRef\]](#) [\[PubMed\]](#)
15. Sultana, A.; Nahiduzzaman, M.; Bakchy, S.C.; Shahriar, S.M.; Peyal, H.I.; Chowdhury, M.E.H.; Khandakar, A.; Arselene Ayari, M.; Ahsan, M.; Haider, J. A Real Time Method for Distinguishing COVID-19 Utilizing 2D-CNN and Transfer Learning. *Sensors* **2023**, *23*, 4458. [\[CrossRef\]](#)
16. Balan, E.; Saraniya, O. 'Novel Neural Network Architecture Using Sharpened Cosine Similarity for Robust Classification of COVID-19, Pneumonia and Tuberculosis Diseases from X-rays'. *J. Intell. Fuzzy Syst.* **2023**, *44*, 6065–6078. [\[CrossRef\]](#)
17. Hussain; Hasan, M.; Rahman, M.A.; Lee, I.; Tamanna, T.; Parvez, M.Z. Corodet: A deep learning based classification for COVID-19 detection using chest X-ray images. *Chaos Solitons Fractals* **2021**, *142*, 110495. [\[CrossRef\]](#) [\[PubMed\]](#)
18. Al-Timemy, H.; Khushaba, R.N.; Mosa, Z.M.; Escudero, J. An efficient mixture of deep and machine learning models for COVID-19 and tuberculosis detection using X-ray images in resource limited settings. In *Artificial Intelligence for COVID 19*; Springer: Cham, Switzerland, 2021; pp. 77–100.
19. Ghoshal, B.; Tucker, A. Estimating Uncertainty and Interpretability in Deep Learning for Coronavirus 685 (COVID-19) Detection. *arXiv* **2020**, arXiv:2003.10769.
20. Narin, A.; Kaya, C.; Pamuk, Z. Automatic Detection of Coronavirus Disease (COVID-19) Using X-ray Images 687 and Deep Convolutional Neural Networks. *Pattern Anal. Appl.* **2020**, *24*, 1207–1220. [\[CrossRef\]](#)
21. Oh, Y.; Park, S.; Ye, J.C. Deep learning COVID-19 features on CXR using limited training data sets. *IEEE Trans. Med. Imaging* **2020**, *39*, 2688–2700. [\[CrossRef\]](#)
22. Jain, R.; Gupta, M.; Taneja, S.; Hemanth, D.J. Deep learning based detection and analysis of COVID-19 on chest X-ray images. *Appl. Intell.* **2021**, *51*, 1690–1700. [\[CrossRef\]](#)
23. Yoo, S.H.; Geng, H.; Chiu, T.L.; Yu, S.K.; Cho, D.C.; Heo, J.; Choi, M.S.; Choi, I.H.; Cung Van, C.; Nhung, N.V.; et al. Deep learning-based decision-tree classifier for COVID-19 diagnosis from chest X-ray imaging. *Front. Med.* **2020**, *7*, 427. [\[CrossRef\]](#)
24. Pereira, R.M.; Bertolini, D.; Teixeira, L.O.; Silla, C.N.; Costa, Y.M.G. COVID-19 identification in chest X-ray images on flat and hierarchical classification scenarios. *Comput. Methods Programs Biomed.* **2020**, *194*, 105532. [\[CrossRef\]](#) [\[PubMed\]](#)
25. Sakib, S.; Tazrin, T.; Fouda, M.M.; Fadlullah, Z.M.; Guizani, M. DL-CRC: Deep Learning-based chest radiograph classification for COVID-19 detection: A novel approach. *IEEE Access* **2020**, *8*, 171575–171589. [\[CrossRef\]](#) [\[PubMed\]](#)
26. Makris, A.; Kontopoulos, I.; Tserpes, K. COVID-19 detection from chest X-ray images using deep learning and Convolutional Neural Networks. In Proceedings of the 11th Hellenic Conference on Artificial Intelligence, Athens, Greece, 2–4 September 2020.
27. El Asnaoui, K.; Chawki, Y. Using X-ray images and deep learning for automated detection of coronavirus disease. *J. Biomol. Struct. Dyn.* **2020**, *39*, 3615–3626. [\[CrossRef\]](#)
28. Saiz, F.; Barandiaran, I. COVID-19 detection in chest X-ray images using a deep learning approach. *Int. J. Interact. Multimed. Artif. Intell.* **2020**, *6*, 4. [\[CrossRef\]](#)
29. Hemdan, E.E.-D.; Shouman, M.A.; Karar, M.E. COVIDX-Net: A Framework of Deep Learning Classifiers to Diagnose COVID-19 in X-ray Images. *arXiv* **2020**, arXiv:2003.11055.
30. Rahimzadeh, M.; Attar, A. A modified deep convolutional neural network for detecting COVID-19 and pneumonia from chest X-ray images based on the concatenation of xception and resnet50v2. *Inform. Med. Unlocked* **2020**, *19*, 100360. [\[CrossRef\]](#)

31. Rahman, T. COVID-19 Radiography Database. Available online: <https://www.kaggle.com/datasets/tawsifurrahman/covid19-radiography-database/versions/3> (accessed on 16 February 2023).
32. Sait, U.; Lal KV, G.; Prakash Prajapati, S.; Bhaumik, R.; Kumar, T.; Shivakumar, S.; Bhalla, K. Curated Dataset for COVID-19 Posterior-Anterior Chest Radiography Images (X-rays). *Mendeley Data* **2021**, 3.
33. Rosenthal, A.; Gabrielian, A.; Engle, E.; Hurt, D.E.; Alexandru, S.; Crudu, V.; Sergueev, E.; Kirichenko, V.; Lapitskii, V.; Snezhko, E.; et al. The TB Portals: An open-access, web-based platform for global drug-resistant-tuberculosis data sharing and analysis. *J. Clin. Microbiol.* **2017**, *55*, 3267–3282. [[CrossRef](#)]
34. NIH Chest X-ray Dataset. Available online: <https://datasets.activeloop.ai/docs/ml/datasets/nih-chest-x-ray-dataset> (accessed on 9 August 2023).
35. VGG16. Available online: https://docs.openvino.ai/latest/omz_models_model_vgg16.html (accessed on 9 August 2023).
36. Nahiduzzaman, M.; Ahamed, M.F.; Alghamdi, N.S.; Islam, S.M.R. Shap-Guided Gastrointestinal Disease Classification with Lightweight Parallel Depthwise Separable Cnn and Ridge Regression Elm. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4572243 (accessed on 31 July 2023).
37. Sarkar, O.; Islam, M.R.; Hossain, T.; Syfullah, M.K.; Islam, M.T.; Moniruzzaman, M. An empirical model of classifying lung affected diseases to detect COVID-19 using chest X-ray employing convolutional neural architecture. In Proceedings of the 2022 IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET), Kota Kinabalu, Malaysia, 13–15 September 2022.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.