Detection of Pulmonary diseases from Chest X-ray Images using Deep Learning Model

by

Aparajita Bose
20101209
Faria Kamal Suchi
20101476
Imam Hossain
20101417
Sajid Muntasir
20101304

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

Department of Computer Science and Engineering
BRAC University
October 2024

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Aparajita Bose
20101209

Faria Kamal Suchi
20101476

Imam Hossain
20101417

Sajid Muntasir
20101304

Approval

The thesis titled "Detection of Pulmonary Diseases from Chest X-ray Images using Deep Learning Model" submitted by

1. Aparajita Bose (20101209)
2. Faria Kamal Suchi (20101476)
3. Imam Hossain (20101417)
4. Sajid Muntasir (20101304)

Of Summer, 2024 has been accepted as satisfactory in partial fulfillment of the re-quirement for the degree of B.Sc. in Computer Science on October, 2024.

Examining Committee:

Supervisor:
(Member)

Ankan Ghosh Dastider

Lecturer(on leave)
Department of Computer Science and Engineering
Brac University

Co-Supervisor:
(Member)

Rafeed Rahman

Senior Lecturer
Department of Computer Science and Engineering
Brac University

Program Coordinator:
(Member)

Dr. Md. Golam Rabiul Alam

Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Sadia Hamid Kazi, Ph.D.

Chairperson
Department of Computer Science and Engineering
Brac University

Abstract

Deep learning models are important in efficiently identifying different pulmonary diseases from Chest X-ray Images (CXRs). Pneumonia is one of the most common lung diseases that cause death. Especially, stage 4 pneumonia can become the reason for an untimely death. Moreover, COVID-19 is still killing a lot of people all around the world. Scientists, doctors, and institutions are working on inventing the most effective way of detecting these diseases. Accurate and early detection of these diseases is essential, otherwise, they can be deadly. In this work, we will detect different pulmonary diseases like COVID-19, and Pneumonia from chest X-ray images. There are many deep learning models like CNNs, RNNs, GANs, and so on. Among them, CNN models are the best for image classification. For example, ResNet18, ResNet50, InceptionV3, VGG19, DenseNet201 and so on. However, we have not used these models. We have used models that have the highest accuracy, Recall, precision, and F1 score. The CNN models generally perform well with image data. So, we used models that are not traditional CNN models. Rather, they essentially rely on transformer architectures or a combination of transformers and CNNs. So, we have used a Swin Transformer, Vision Transformer (ViT), VoLO-D1, FocalNet, and VITamin. Transformers rely on self-attention mechanisms to determine the similarities across an image. On the other hand, CNNs use convolutional layers to extract features locally from an image. Our proposed model is a customized CNN model and it is time and cost-efficient as it provides higher accuracy faster than other models. It is deploy-friendly as the size of the model is 257 MB. Other transformer based model are bigger in size. Moreover, it has a transformer-based ecosystem and benefits. The accuracy of our customized CNN model is 98 percent and learning rate is 0.001. We have built an automated lung disease detection system to make the detection less time-consuming, cost-efficient, and error-free for developing countries.

Keywords: Pulmonary diseases; Deep learning; Chest X-ray Images; Biomedical image processing; COVID-19; Pneumonia; Swin Transformer; Vision Transformer (ViT); VOLO-D1; FocalNet; and VITamin.

Acknowledgement

Firstly, all praise to the Great Allah for whom our thesis have been completed without any major interruption.

Secondly, to our supervisor Mr. Ankan Ghosh Dastider and co-supervisor Mr. Rafeed Rahman sir for their kind support and advice in our work. They helped us whenever we needed help.

And finally to our parents, without their throughout support it may have not been possible. With their kind support and prayer we are now on the verge of our graduation.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction

Thousands of people are being infected each year with pulmonary diseases such as COVID-19, Pneumonia, Tuberculosis, and so on [21]. Specifically, the outbreak of COVID-19 as a pandemic has been taking many lives away for the last three years. Moreover, Tuberculosis(TB) is the fifth leading cause of death throughout the world having 10 million new cases and 1.5 million deaths per year [9]. In fact, most pulmonary diseases being contagious, have been extremely dreadful for human beings. The major challenge in predicting these diseases is that all of them have almost the same symptoms. Hence identifying these diseases with automated systems has become a priority for the healthcare system. Earlier, such automated systems used to be proposed and built based on Machine Learning algorithms. However, traditional machine-learning algorithms are data-dependent [26]. As a result, only data similar to the dataset can be recognized and identified accurately.

On the other hand, deep learning methods can automatically extract the needed features to identify an illness from the available dataset. Therefore, in recent times, deep learning models are considered to be more efficient than those models based on machine learning algorithms. Some of the recent works on our topic have also used different types of CNN models. For example, in paper [6], CNN models like LeNet, AlexNet, VGGNet, GooleNet, ResNet, DenseNet, and R-CNN were used to compare the error rate of each model to solve this problem.

In some papers, they have used some image enhancement algorithms to get better results. Different studies have used different classifiers. Some of them have used SVM classifiers and some of them have used decision trees. In some papers, authors have used the RNN model. Even in some papers, authors have used ANN models. Each model has a different approach to solving a problem. However, we wanted to choose our model in a way that it will extract only the necessary features and then it would process those features.

Firstly, models like the Swin Transformer integrate attention mechanisms and CNN layers. Moreover, the Swin transformer model uses selective attention to extract the feature efficiently across images. It not only balances accuracy but also reduces computational cost. Secondly, ViT (Vision Transformer) processes images by splitting them into different patches and then by applying self-attention. It effectively captures global dependencies. However, it requires large datasets to perform well. Now, VoLo-D1 incorporates local and global attention through bottleneck layers and extract features. However, this method increases the complexity. FocalNet improves attention and focuses on only the image regions that are critical. This allows this model to perform efficiently and accurately on large datasets. However, it requires a large amount of tuning and a fair amount of computational resources. VITamin incorporates CNN and transformer layers and merges local feature extraction and global attention together. As a result, this model is versatile among different types of dataset sizes and image resolutions. However, this model has a large number of architectural complexity and higher computational resources.

In our proposed method, we have used a customized CNN model that is faster than other CNN models and has an accuracy of 98 percent. Our customized model performs better than the Swin Transformer, Vision Transformer (ViT), VoLO-D1 (Vision Outlooker), FocalNet, and VITamin models. Our proposed model is small and that's why it is time and space efficient. Moreover, due to its size, our customized CNN model is deploy-friendly. We have ensured that our model is solving the problems that other researchers found while detecting pulmonary diseases. Through this work, we intend to propose a Deep Learning network that can identify as well as classify pulmonary diseases as accurately as possible by analyzing chest X-ray images. Our model accurately detects pulmonary diseases. Through our customized model, we have successfully classified pneumonia X-rays, COVID-19 X-rays, and normal X-rays. And we are confident that we can detect any pulmonary disease using our model.

1.2 Problem Statement

According to the American Lung Association [9], approximately 38 million people are suffering from different types of lung diseases all across the United States. Now, if one of the most developed countries of the world has this many cases of respiratory patients then we can not even imagine the situation of the developing countries. It has been estimated that in 2015, approximately 1.8 million people died due to Pneumonia and the majority of them were from developing countries [9].

Over the period of time, this number has increased. However, many of the deaths could have been prevented, if we could detect the disease earlier. In many cases, diseases were not even detected properly meaning patients knew that they were suffering from lung diseases but did not know the exact disease they were suffering from. Moreover, in developing countries, it is hard to have trained clinical officers all across the country.

As a result, many patients died without proper diagnosis and treatment. Furthermore, the detection of TB, Pneumonia, COVID-19, and lung cancer can be tough and time-consuming. And most of the time clinicians can not detect the disease on time and these diseases get worse. To solve these problems, we have proposed an automated lung disease detection system that will make the detection process easier and more efficient. Now, among different pulmonary diseases, there are some stages of diseases. For example, Pneumonia has four stages. They are-
Pneumonia has four stages. They are-

- Congestion
- Red Hepatization
- Grey Hepatization
- Resolution

In addition, other lung diseases like COVID and lung cancer often get confused with the different stages of TB and Pneumonia as some of them have similar symptoms. To solve this problem, we wanted to build a customized model that could detect at least pneumonia and COVID-19. Both of these diseases have similar symptoms. As a result, often doctors confuse these diseases with each other. We wanted to assemble a customized CNN model that will differentiate pneumonia and covid COVID-19 chest X-ray images as accurately as possible.

Moreover, our goal was to build a cost-effective model so that we could help developing countries and rural areas. For our proposed model, we have used a customized CNN model that provides us with the highest accuracy, sensitivity, specificity, and F1 score. For this, we have not chosen a model randomly. We have researched the previous works and the ambiguity of those papers and based on our experiment we have chosen the best model that will give us the highest accuracy and AUC from an image dataset like CXRs.

Now, we have used CXR images since it is easier to detect abnormalities and lesions in the lung. Also, X-rays are inexpensive and that's why patients mostly opt for X-rays rather than CT-Scan and MRIs. Moreover, clinicians also prefer X-rays as it is easier to perform X-rays than CT-Scan and MRI. As a result, we found a variety of datasets related to X-rays.

- Diagnosing lung diseases early is difficult, especially for developing countries.
- Developing countries do not have enough trained doctors and medical officers.
- Since different lung diseases have similar symptoms, this often leads to incorrect diagnoses of lung diseases. As a result, it sometimes becomes deadly for the patients.
- Many automated systems for pulmonary disease detection exist, but most work with CT scans and MRIs.

- CT scans and MRIs are not accessible to many people who live in rural areas.
- As per our knowledge, there are limited automated tools to diagnose lung diseases quickly and accurately. Especially in developing countries, there are no automated systems.
- Detection of any pulmonary disease becomes not only expensive but also time-consuming for developing countries and rural areas.
- Available datasets are mostly smaller in size, making it difficult to build a reliable system.

1.3 Research Objectives

Achieving high accuracy in detecting various lung diseases from chest X-ray image datasets was the prime approach to take for this research work. We have trained a dataset with a variety of chest X-ray images so that our model can successfully perform the classification of different lung diseases from the dataset. Also, we wanted to pursue as much accuracy as possible in detecting these diseases. Our main objective is to make the disease detection process easier for the clinician as detecting different lung diseases can be time-consuming and tough. Sometimes many clinicians make mistakes while detecting a particular disease as multiple diseases have the same symptoms and conditions.

Moreover, in many developing countries, properly trained clinicians are not available. An automated detection system can reduce mistakes and time and it can increase the accuracy rate to detect diseases. Our main objectives are-

- We wanted to learn knowledge about various Machine Learning models, Deep Learning models, and Image processing architecture and models.
- We tried to find the accuracy of each model and the loopholes of each model. In addition, we have tried to find a suitable model for our automated detection system and tried to keep the accuracy level as high as possible.
- In this paper, we have only worked with two lung diseases: pneumonia and COVID-19. In the future, we will try to incorporate more lung diseases such as TB, Lung cancer, etc.
- We wanted to build a cost-effective system to detect different lung diseases using chest X-ray images.
- We wanted to create a model that can accurately classify pneumonia, COVID-19, and Normal CXRs.
- We planned to build a model that is smaller in size.
- We increased the accuracy, F1 score, recall, and parameters for pulmonary disease detection.

- We wanted to construct a model that would be able to handle large datasets.
- We planned to build an automated detection system for resource-limited areas.
- Lastly, we will find weaknesses in our model and improve it to get the highest accuracy rate.

Chapter 2

Related Work

2.1 Detailed Literature Review

The following are some of the previous works we have reviewed to assess and enrich our ideas for this research work.
Santosh et al. (2022) reviewed the past 5 years of machine learning and deep learning models to refer to the lackings of ML models and highlight the reason for the rising DL models over ML models [26]. The paper also evaluates DL models by analyzing various datasets. Some of them are as follows:

Serial no.	Datasets
1.	Montgomery County Dataset (MC, USA) [1]
2.	National Institute of Tuberculosis and Respiratory Diseases Dataset (NITRD, India)[2]
3.	Japanese Society of Radiological Technology Dataset (JSRT, Japan)[3]
4.	Belarus Tuberculosis Dataset (Belarus)[4]
5.	Shenzhen Hospital Dataset (SH, China)[5]
6.	Radiological Society of North America Dataset (RSNA, USA)[6]
7.	Chest X-rayS - NIH (MD, USA)[7]
8.	Mendeley Dataset (UK)[8]

Table 2.1: Available Datasets used for comparison [26]

These datasets were analyzed and visualized by a variety of methods individually to ensure the accuracy of the built models/networks over 80%. Some of the methods used are- CNN, GoogLenet, AlexNet, ResNet, CheXNet, CAD4TB, CAD4TBV3.07, VGG (having different numbers of layers), DenseNet, etc. This review paper also shows by comparison that using different methods for the same data collection results in different accuracy levels. Hence assessed the existing DL models in terms of efficiency. Moreover, through the review of 54 research articles published during 2016-2020, this paper also presented the evolution of DL models used in research activities from binary DL-based algorithms to models that improve the individual processes of a network like visualization, segmentation, augmentation, etc.

Similarly, in the article published in 2021, Murali Krishna Puttagunta and S. Ravi have shown the increment in the use of CAD(computer-aided diagnostic) systems for early-stage TB detection [17]. This system helps to improve detection by DL-based CNN models during the screening process. With the help of various available datasets, this article also compares the error rate of proposed variations of CNN models such as LeNet, AlexNet, VGGNet, GoogleNet, ResNet, DenseNet, and R-CNN.

Likewise, Rajaraman et al. worked with the goal of improving the segmentation process for datasets that might have images with weak localization [18]. To ensure the work can achieve the needed accuracy, the authors have followed a specific sequence of tasks to process the input data through the proposed network. In this paper, we have seen works on the knowledge transfer topic, localization and statistical analysis, and so on.

Bhandari et al. have proposed an XAI-based single CNN model to detect and classify some of the lung diseases that are possible to detect from chest X-Ray images [21]. Compared to other research and datasets, this work is evaluated in terms of efficiency and accuracy. This work has achieved an accuracy of more than 90%. The goal of this research is to introduce categorizing diseases within image processing with the help of trained datasets. The following Table 2.2 presents some cases from the paper [21] which shows the possibility of classifying the images for different cases.

The article by Liu et al.(2017) designed a CNN model to detect and classify TB manifestations in X-ray images [5]. By revising the AlexNet and GoogLeNet architectures for image Classification, the proposed model in this work is designed to improve the accuracy of the outcome. Moreover, this work is open to unbalanced datasets by applying the shuffle sampling technique in the augmentation of data.

The research work by Rahman et al.(2020) classifies TB and normal chest X-Ray images automatically by different CNN models and compares their efficiency [13]. This research shows that CNN models with segmentation techniques have better accuracy levels than those which do not apply segmentation to the input data while processing. Moreover, to utilize the outcome of segmentation Score-CAM visualization technique is used for visualizing the output in this particular work.

The architecture used in the article by Pasa et al. (2019) to detect Tuberculosis from chest X-Ray, basically consists of 5 convolutional blocks. To process the input into more precise output, the network of the architecture contains one global average pooling layer and another fully-connected softmax layer [9]. To randomize and generate the most possible accurate result, the network takes preprocessed data as input from a dataset trained with batch normalization. Moreover, as an aid to the visualization factor, two techniques called saliency maps and gradient class activation maps (grad-CAM) were used on the output of the system.

Cases	**Class**	**Methods**	**Accuracy (%)**

TB = 4248
Normal = 453
	2	Transfer learning with AlexNet and GoogLeNet	85.68

Normal = 8851
COVID-19 = 180
Pneumonia = 6054
	3	Ensemble of Xception and ResNet50	91.40

Normal = 310
PneumoniaB = 330
PneumoniaV = 327
COVID-19 = 284
	4	CNN-based CoroNet	89.60

Normal = 1583
COVID-19 = 576
Pneumonia = 4273
TB = 155
	4	Custom CNN	94.53

Normal = 310
PneumoniaB = 330
PneumoniaV = 327
COVID-19 = 284
	4	Attention-based VGG	85.43

Normal = 1341
COVID-19 = 864
Pneumonia = 1345
	3	Inception V3 with Transfer learning	93.00

Normal = 439
COVID-19 = 435
PneumoniaB = 439
PneumoniaV = 439
TB = 434
	5	Transfer learning with Resnet18	91.60

Normal = 1583
COVID-19 = 576
Pneumonia = 4273
TB = 700
	4	Custom CNN and GoogLeNet	95.94

Table 2.2: Comparison of various methods and their accuracy for lung disease classification.

In the research work by the U.K. Lopes and J.F. Valiati (2017), three different methods had been shown for the detection of TB using pre-trained Convolutional Neural Networks [6]. For using pre-trained CNNs for TB detection three approaches are described. The CNNs were trained in ImageNet and performed decently while the detection of this disease. In several cases, pre-trained CNNs were performing better than the finely adjusted CNNs. In this paper, for feature extraction, they have used three separate architectures of pre-trained CNNs such as GoogLenet, ResNet, and VggNet. Moreover, they used an SVM classifier to identify whether or not the images contained TB. Furthermore, in proposal 1, they have done simple CNN feature extraction. Here, proposal 1 has a loophole as the lung images were being resized to fit in the CNN input layer, it was losing a lot of information and that was also reducing the chances of detecting TB correctly. And to solve this problem, they introduced the second proposal and this proposal suggests the three CNN architectures that were mentioned above, were used as feature extractors. Still, this time they did not resize the CRs. In addition, they divided the CRs into subdomains and the sizes of the CRs were equal to the network layer. In the third proposal, they suggested an ensemble classifier for better results with higher accuracy. In this proposal, they created ensemble classifiers by merging the SVMs which were trained with the features that were extracted from the three architectures of CNNs that they mentioned earlier. For proposals 1 & 3, the accuracy and the AUC of the Shenzhen dataset were higher than the Montgomery dataset. For proposal 2, the AUC for Montgomery dataset was 0.908 and the AUC for Shenzhen dataset was 0.926. One of the problems of this paper was ensembles were created using a simple voting scheme. However, there is no information about the changes that will occur after using different voting methods and CNN architecture.

In the article by Rajaraman et al., it is determined that the effectiveness of knowledge transfer had been gained by combining several modality-specific deep learning models together to improve the process of Tuberculosis detection [14]. A customized CNN model and a few pre-trained CNNs were trained to learn the modality-specific features from various large-scale CXR datasets. The predictions or the results of the best-performing models were combined using various ensemble methods to determine the improved performance of the models in classifying the TB-infected CXRs and the non-TB-infected CXRs. The models were assessed through 5 cross-validations to reduce overfitting and improve robustness, and generalization. In this paper, they have shown that the accuracy and AUC of the ensemble of the top 3 pre-trained models both are 95%. However, one limitation of this paper is that these ensembles were evaluated with a small dataset.

From the work by Urooj et al. (2022), we can summarize that sometimes TB gets misclassified with other diseases due to similar symptoms and similar radiographic patterns of CXR images and this leads to false treatment [28]. The current Computer-Aided Detection had some limitations as those were only evaluated by non-deep learning models. In this paper, the authors proposed a method to develop a reliable TB detection system depending on stochastic learning with an artificial neural network (ANN) model with some random variations using the CXR dataset. In this proposed method, the model learns features from CXR images and collects the parameters of an ANN model, and here they randomly mix the training dataset before every iteration. As this method focuses on randomness, it achieves higher accuracy. The reason for this proposed method was to detect changes in CXR and identify the different levels of TB just by extracting deep geometric contexts like shape, size, etc from the CXR. The proposed method of this paper performed better compared to other methods. The accuracy of the proposed method is 98.45%.

In the article by Munadi et al.(2020), the authors suggested a different approach to detecting Tuberculosis [11]. Deep learning needs a huge number of high-quality images to diagnose TB efficiently. However, many CXRs have low contrast issues. Moreover, most of the time, the images are not high quality at all. As a result, it affects the diagnosis. This paper suggested that if they increase the quality of the images, they would be able to upgrade the performance of deep learning models. Additionally, the paper by Afzali et al.(2019), also mentioned three image enhancement algorithms, and those algorithms are Unsharp Masking, High-Frequency Emphasis Filtering, and Contrast Limited Adaptive Histogram Equalization [8]. The enhanced CXR images were given to the pre-trained models like ResNet and EfficientNet models for transfer learning. The accuracy and AUC both were 89.92% and 94.8% respectively.

In the paper Verma [15], the authors tried to find the optimal feature vectors for TB detection from Chest X-ray images or CXR images. This paper mainly focuses on the contour-based shape descriptors and Two Dimensional Principal Component Analysis meaning this paper had taken a different approach in selecting features from CXRs for TB detection. Other prior studies have worked with texture-based features for TB detection instead of contour-based features. They achieved 92.86% accuracy and 91.67% AUC.

From the research of the American Lung Association, the authors suggested a framework to classify pulmonary Tuberculosis, Bacterial pneumonia, and Viral Pneumonia from CXR images (n.d.) [33]. This analysis was performed by using a neural network classifier. Several data augmentation methods were used to pre-process the data and these methods improved the classification accuracy of the suggested model. In this paper, it is visible that the proposed framework was able to efficiently classify different pulmonary infections and the accuracy of this proposed model is 99.01%.

The paper by Haq et al. (2022) presented an approach for diagnosing TB from CT scan images [23]. Large volumes of pathology and radiology data can be processed using machine learning algorithms, which allows for quicker decision-making. This approach provides more accuracy and efficiency for detecting and identifying diseases and significantly reduces both cost and time. For the research 100 abnormal (TB infected) and 100 normal CT scan images of lungs were acquired from the Department of Radiology, Bahawal Victoria Hospital, Bahawalpur, Punjab, Pakistan. Classification between normal and infected TB images was done using multiple supervised learning classifiers. Their accuracy is more than 95%.

The study by Li et al. (2020) served the purpose of establishing and validating a deep learning system that produces quantitative CT scan reports for the recognition of pulmonary tuberculosis [16]. The dataset used in this study included 501

CT imaging files from 223 patients with active PTB as well as 501 datasets used as negative samples that were drawn from a healthy population. For the inspection of the images four 3D convolutions neural network (CNN) models were trained and evaluated. The 3D CNN model was used to identify the lesion region. After that, the infection probability was calculated using the Noisy-Or Bayesian function. The study concluded that this new method might serve as an effective reference for decision-making by clinical doctors. However, a specific limitation of the study model might be less sensitivity to trivial PTB lesions. The doctors still needed to review the full CT scan to confirm the result.

The paper by Venkataramana et al. (2022) shows how COVID-19, TB and pneumonia can be classified using deep learning [29]. Though lung diseases can be detected from CXRs. The X-rays will be then classified into pneumonia, tuberculosis, or COVID-19 groups. The paper claims that their model is cost-effective and can perform faster and more accurately, and can be used efficiently for mass screening to detect COVID-19 in people.

The article by Kaila et al.(2007) has acknowledged that MRI can detect asymptomatic lesions as symptoms of spine TB even if it may not be the most cost-effective measure and might be unable to identify multiple-level noncontiguous TB. This test confirms the higher presence of multiple-level noncontiguous TB breaking away from previous beliefs [3].

The article by Zacharia et al.(2003) determines the role of MRI in detecting ankle tuberculosis, which is extremely rare when occurring by itself without involvement in any other body parts. On a 0.5 T scanner, the findings showed various irregularities in the case of ankle TB, symptoms not usually found in other cases that were successfully detected by MRI, unlike CT scans [2].

The article by Desai(1994) tests the results of the MRI of 24 routinely tested patients with possible spinal Tuberculosis that were treated accordingly. Results showed that all patients but one responded to the treatment provided, showing the high effectiveness of MRI and confirming a nearly 100-year-old quote by Massart and Ducroquet made in 1926, where they stated that Pott's disease abscesses are not hidden to X-ray and similar imaging [1].

The article by Rizzi et al.(2011) sets out to compare MR imaging to the apparent gold standard for assessing morphological changes in the lungs. The study found that the CT scan and the MRI had little differences in their assessments, with the MRI being able to detect finer details of the irregularities due to the higher resolution, minus the radiation inflicted on the patient by the CT scan [4].

The article by Jianbing et al. (2019) explores the uses of MRI with advanced motion correction to detect lung tissue changes and TB-induced lesions. Using the Multi-Vane technique on 63 TB subject samples resulted in a 100% detection of TB with satisfactory quality, albeit it was less effective at identifying calcified lesions [10].

The article by Yusuf et al.(2022) has shown a high rate of success, removing the need for unnecessary and complicated processes such as biopsy. It also makes the diagnosis of TB spondylitis much less complicated, as the previous absence of any symptoms in the lungs would make it challenging to detect spondylitis [31].

The study of the article by Khokhar et al.(2022) is aimed to determine the diagnostic accuracy of MRI for the detection of spinal tuberculosis (TB) [24]. This cross-sectional study was conducted from January 2020 to August 2020 on 150 patients with suspected spinal tuberculosis. Patients underwent MRI scans of the entire spine. For the diagnosis of spinal tuberculosis, MRI was performed using 1.5 Tesla MR. According to MRI accuracy, there were 83 true positives (55.3%), 10 false positives (6.67%), 8 false negatives (5.33%), and 49 true negatives (32.67%). The sensitivity and specificity were 91.2% and 83.1% respectively.

In paper [19], Tripathi et al.(2021) have proposed a model that had convolutional layers, ReLU activations, pooling layers, and a fully connected layer. The layer meaning the fully connected layer has 15 output units. According to the author, every unit will predict the prospect of getting any of these 15 diseases. Their datasets consist of fifteen different classes named Atelectasis, Cardiomegaly, Effusion, Infiltration, Mass, Nodule, Pneumonia, Pneumothorax, Consolidation, Edema, Emphysema, Fibrosis, Pleural Thickening, Hernia, and No Finding images used to train this model. This model has a decent average accuracy for multiclassification which is approximately 89.77%.

In paper [32], Alshmrani et al.(2023) have suggested a deep learning model that will perform a multi-class classification of diseases like pneumonia, lung cancer, TB, and so on. They have used 5 classes. They have used CXR images as their dataset. Firstly, the authors have resized all the images then they have normalized them. Lastly, they randomly split the images to get fitted into Deep learning requirements. For the classification, the authors have used VGG 19 which is a pre-trained model, and for feature extraction, they have used 3 blocks of CNN and a fully connected layer. So, the accuracy of the VGG19 + CNN model is 96.48%. The F1 score is also 95.62% and the AUC is 99.82%.

Ref	Dataset and Source Code	Performance	Category
[31]	**Datasets:** Chest X-ray images: 1) 146 images of COVID-19 from [32] 2) 420 images of pneumonia from [33] **Source code:** Not available.	Accuracy (%): 1) VGG16: 96.88 2) VGG19: 95.31 3) Inception ResnetV2: 89.06 4) Xception: 95.31 5) InceptionV3: 92.66 6) MobileNet: 89.06 7) DenseNet121: 92 8) Ensemble: 98	Transfer Learning - Fine-tuning
[34]	**Datasets:** Chest X-ray images: 1) A publicly available dataset of pneumonia cases [33] 2) Kaggle dataset 3) A publicly available dataset of pneumonia and COVID-19 cases **Source code:** Not available.	Accuracy: 97%	Transfer Learning - Feature extraction
[36]	**Datasets:** Chest X-ray images: 1) A publicly available dataset of COVID-19 cases [33] 2) A dataset by Qarar University 3) Pneumonia cases dataset **Source code:** Not available.	Accuracy: 99.2%	Training from scratch - Single Model
[12]	**Datasets:** Chest X-ray images from two publicly available datasets: 1) 127 COVID-19 positive cases from [33] 2) 1000 images selected randomly from [7] for two classes: no findings and pneumonia.	Results for two-class classification: Accuracy: 98.08% F1-score: 96.51% Results for three-class classification: Accuracy: 87.02% F1-score: 87.37%	Training from scratch - Multiple Models

Table 2.3: Summary of reviewed Datasets.[20]

Chapter 3

Prediction Modeling using Decision Tree

3.1 Description of the Data

As we have discussed before, we have explored various datasets. Among them, we have chosen a dataset named "COVID19+PNEUMONIA+NORMAL Chest X-Ray Image Dataset"[27][25]. This dataset has three classes. The classes are Pneumonia, COVID-19 and Normal. Here we have used Chest X-Ray or CXR images as our dataset as they are quite available compared to MRI or CT-scans.

In this dataset, there were no distinctions for age and sex or the severity of any diseases. This dataset only has CXRs of patients who are suffering from lung diseases like Pneumonia and COVID-19. Moreover, this dataset also contains CXR images of normal lungs which means unafftected/healthy lungs. The dimension of each image in the dataset is (256*256) and the format of each image is PNG.

Class	**Number of Images (per class)**
Covid	1626
Pneumonia	1800
Normal	1802

Table 3.1: Number of images per class in the dataset

In our existing dataset, there are a total of 1626 CXR images for COVID19 1300 of which is used to train the model, 163 of them are used for testing purpose and 163 of them are used to validate the model. Similarly, the dataset has total 1802 images of normal CXRs with healthy lungs and 1800 images of Pneumonia affected CXRs. It is necessary to mention that from the total normal CXR images 1442 are for training, 180 are for testing and rest of the 180 images are for validation purpose. Moreover, all the Pneumonia CXR images are divided into train, test and validation by 1440, 180 and 180 respectively. To summarize, the dataset we are using for our research work does not have noteworthy imbalance which helps in efficient and prompt detection of the disease.

3.2 Data Partitioning

Category	Train Data	Test Data	Validation Data	Total
Covid	1300	163	163	1626
Pneumonia	1440	180	180	1800
Normal	1442	180	180	1802

Table 3.2: Distribution of Train, Test, and Validation Data per Category

It is necessary to mention that from the total normal CXR images 1442 are for training, 180 are for testing and the rest of the 180 images are for validation purposes. Moreover, all the Pneumonia CXR images are divided into train, test, and validation by 1440, 180, and 180 respectively. To summarize, the dataset we are using for our research work does not have a noteworthy imbalance which helps in the efficient and prompt detection of the disease.

3.3 Preliminary Analysis

Preliminary analysis on the raw image dataset is a necessary step for prediction through computer vision. Pre-processing of image data helps in increasing efficiency, time and computational complexity of the classification model. Moreover, it can help in transforming the input data for the classification model into a standard format and can be used to handle corrupted and imbalanced data. To ensure the efficiency of our dataset and standardize the data, we have utilized the following pre-processing methods:-

1. Resize and Rescale: Resizing and rescaling is required for a dataset which has images of different shapes or sizes. Generally, the computer vision based classification models support input data of consistent shape. Therefore, the model cannot function with a dataset having images of various shapes. Moreover, resizing all the images of a dataset to exclude redundant information reduces time and space complexity for the dataset which results in increased model efficiency. For our dataset, we have used an annotation tool called ”Roboflow” in order to adjust the dimension of all the images of our dataset to 512*512 for getting better performance during the training phase of the model. Before loading the 512*512 images, we have again resized the images to 224*224. The pixel values of all the images of the dataset are also normalized to ensure faster convergence of the image dataset into having a uniform and standard dimension as well as to avoid biased training.

2. Data Augmentation: Data augmentation is a pre-processing technique which brings more variation to an image dataset upon implementation. It helps the classification model to perform better while making a prediction for random unseen images. Data augmentation can be achieved by rotating, flipping or cropping the already existing images of a dataset. Here, we have implemented cropping the images and horizontal flipping on our dataset for augmentation purpose.

Thus the image dataset we have used, transforms into a consistent, structured and standard format. As a result, the classification model does not have to interpret the raw, imbalanced and poor quality data and can perform training and image processing effectively.

3.4 Description of the Model

In this study, we have proposed a model that will detect pulmonary diseases with higher accuracy, time and cost efficiency. In our proposed method, we have used a custom CNN model for classifying the biomedical images of our dataset. We have implemented other pre-trained models such as- Vision Transformer, Swin Transformer, VOLO-D1, FocalNet and Vitamin. These models will be transfer learning based models. Now, we know that in transfer learning, a model is already trained on large datasets like ImageNet. After implementing a pre-trained model on our dataset, the model achieves a certain level of accuracy in prediction. However, we have achieved better accuracy for the pre-trained models by utilizing fine tuning process.

3.4.1 Proposed Model(Custom CNN)

Convolutional Neural Networks (CNNs) are widely used in image analysis for tasks like classification. They automatically learn features such as edges, textures, and complex patterns through convolutional layers. This study used a custom CNN for chest disease classification using X-ray images, achieving 99.61% training accuracy and 97.97% testing accuracy in just 20 epochs. The custom CNN's architecture balances computational efficiency and accuracy, making it effective in medical image classification.

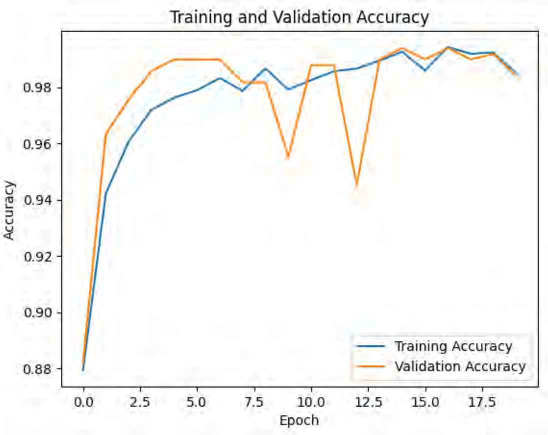


Figure 3.1: Training and Validation Accuracy

Custom CNN Architecture

The custom CNN uses a traditional architecture with enhancements for performance and overfitting prevention.

- **Convolutional Layers:**

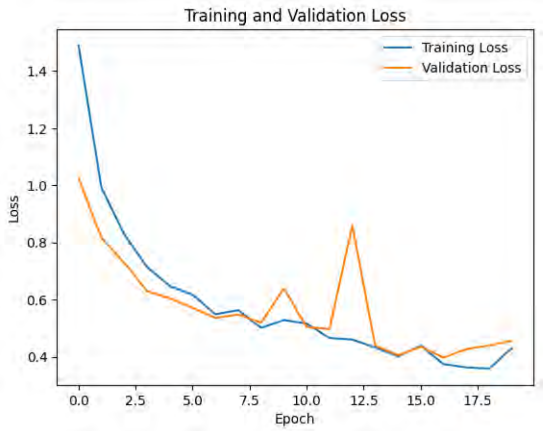


Figure 3.2: Training and Validation Loss

  - **First Block:** Uses 64 filters (3x3) for basic feature detection like edges in X-rays. Batch normalization and ReLU activation improve learning. Max pooling downsamples feature maps.
  - **Second and Third Blocks:** With 128 and 256 filters, respectively, these layers identify complex patterns like lung abnormalities. Max pooling ensures that critical features are retained.

- **Fully Connected Layers:**
  - **First Layer:** 256 neurons with 20% dropout for overfitting prevention.
  - **Second Layer:** 128 neurons with 10% dropout for further regularization.

- **Output Layer:** A softmax layer with three neurons (healthy, pneumonia, tuberculosis).

- **Activation and Regularization:** ReLU for non-linearity, batch normalization for training stability, and dropout to prevent overfitting.

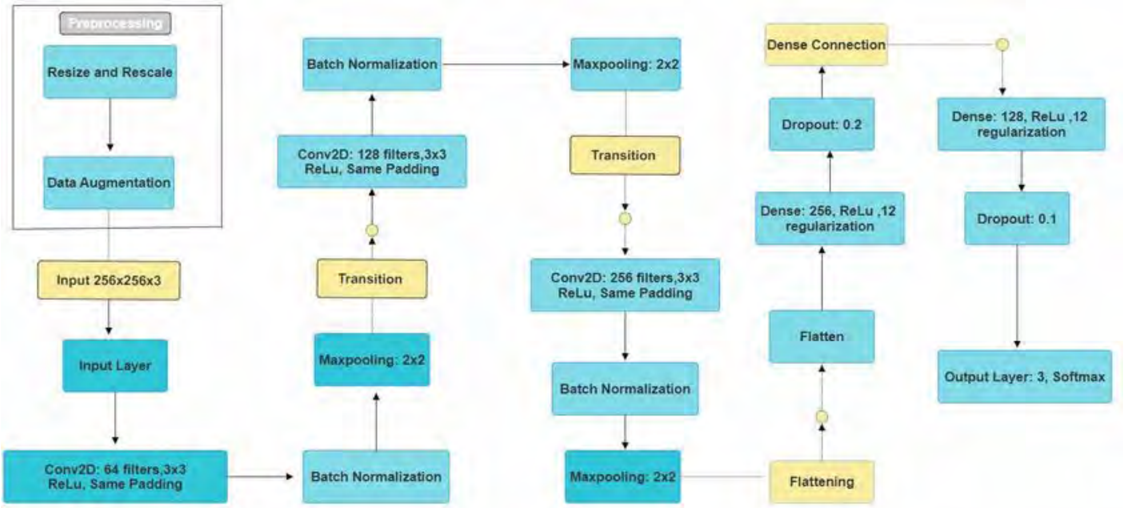


Figure 3.3: Custom CNN model workflow

Why Custom CNN Outperformed Other Models

- **Fewer Epochs, Higher Accuracy:**
  Achieving 99.61% training and 97.97% testing accuracy in just 20 epochs, faster than transformer-based models requiring over 50 epochs.

- **Localized Feature Extraction:**
  CNNs excel at capturing local features like small lesions in X-rays, while transformers often miss these details.

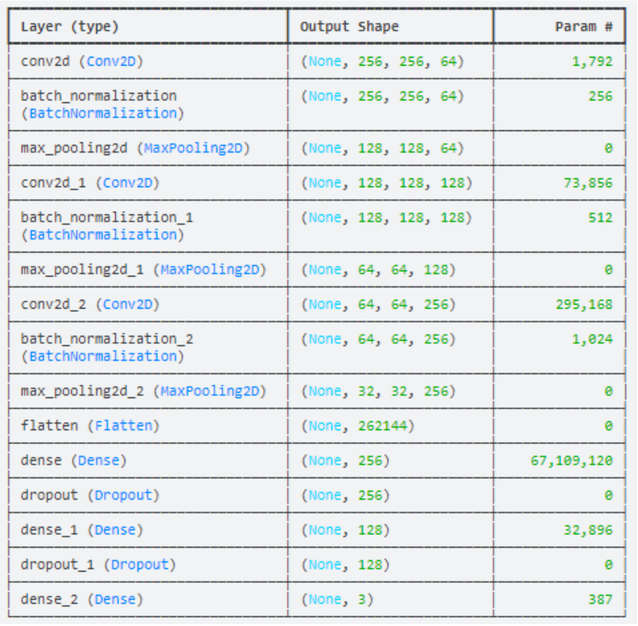
- **Efficient Computation:**
  CNNs are computationally efficient, handling high-resolution images faster with fewer resources.

- **Robust to Overfitting:**
  Dropout layers and batch normalization help prevent overfitting, with the model generalizing well to unseen data.

- **Balanced Architecture:**
  The custom CNN's three convolutional blocks efficiently capture essential details without adding complexity.

- **Locality Bias:**
  CNNs prioritize nearby pixels, beneficial for chest disease classification where abnormalities are localized.

- **Data Efficiency:**
  CNNs perform well with smaller datasets, unlike transformers that need more data.



Layer (type)	Output Shape	Param #
conv2d (Conv2D)	(None, 256, 256, 64)	1,792
batch_normalization (BatchNormalization)	(None, 256, 256, 64)	256
max_pooling2d (MaxPooling2D)	(None, 128, 128, 64)	0
conv2d_1 (Conv2D)	(None, 128, 128, 128)	73,856
batch_normalization_1 (BatchNormalization)	(None, 128, 128, 128)	512
max_pooling2d_1 (MaxPooling2D)	(None, 64, 64, 128)	0
conv2d_2 (Conv2D)	(None, 64, 64, 256)	295,168
batch_normalization_2 (BatchNormalization)	(None, 64, 64, 256)	1,024
max_pooling2d_2 (MaxPooling2D)	(None, 32, 32, 256)	0
flatten (Flatten)	(None, 262144)	0
dense (Dense)	(None, 256)	67,109,120
dropout (Dropout)	(None, 256)	0
dense_1 (Dense)	(None, 128)	32,896
dropout_1 (Dropout)	(None, 128)	0
dense_2 (Dense)	(None, 3)	387

Figure 3.4: Custom CNN model visual

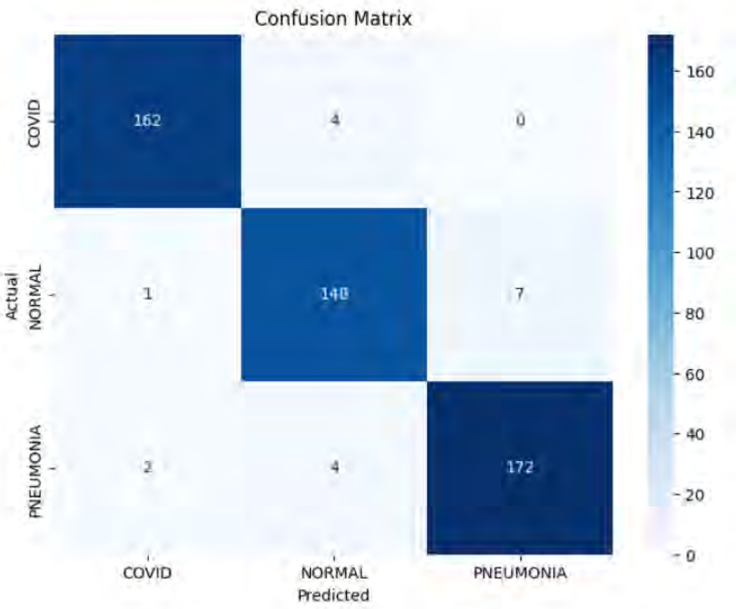


Figure 3.5: Confusion Matrix of custom CNN model

Training Process of Custom CNN

- **Dataset Preparation:** X-rays were resized to 256x256 pixels, with data augmentation like flipping and brightness adjustments to improve generalization.
- **Training Parameters:** Adam optimizer (learning rate 1e-4) with a batch size of 32. A learning rate scheduler reduced the rate when validation loss plateaued.
- **Training Epochs and Early Stopping:** The custom CNN achieved peak performance in 20 epochs, with early stopping to avoid overfitting.

Why our model is better?

We have used Data augmentation and rescaling as it helped the model to generalize data better. Moreover, data augmentation and rescaling controls overfitting as it automatically increases the variety of data. Then we applied multiple convolution layers. Multiple convolutional layers increase the chance of the model to extract complex features from the input image. Furthermore, we have used batch normalization which assists in training faster and helps to stabilize the learning process. It normalizes the input of each layer. So, it normalizes the inputs of each convolutional layer. After that max-pooling helped us to reduce the spatial dimensions. As a result, it reduces the computational complexity without reducing any key feature from the input.

Cost Efficient

Our solution highlights our custom CNN model as a cheaper alternative, with a clear cost advantage over other models like Vision Transformer (ViT), Swin Transformer, VoLo-D1, FocalNet, and VITamin. The CNN model is notably smaller (just 267 MB) compared to others (2.5 GB to 3.8 GB), allowing it to be trained in only 20 epochs, while the others require 50 or more. Additionally, the custom CNN model has lower computational complexity and memory usage, making it computationally efficient and well-suited for deployment in resource-limited environments. Despite being smaller and faster, it maintains competitive accuracy, balancing efficiency and accuracy for real-time medical applications. This overhead efficiency makes your CNN model ideal for embedded systems with limited hardware and poor area connectivity.

Model	Size	Epochs	Training Time	Complexity	Memory Usage
Custom CNN	267 MB	20	Faster	Low	Low
Vision Transformer (ViT)	3.6 GB	50	Slower	High	High
Swin Transformer	3.8 GB	50	Slower	Medium	Medium
VoLo-D1	2.5 GB	50	Moderate	Medium	High
FocalNet	3.2 GB	50	Moderate	High	High
Vitamin	2.7 GB	50	Slower	Medium	Medium

Table 3.3: Comparison of Model Sizes, Training Times, and Complexity

Then the Relu activation function helped to accelerate the learning process. It also enabled the model to learn different composite patterns. L2 regularization helped the model reduce the overfitting in dense layers as it penalized large weights. In addition, the dropout layers reduce the risk of overfitting as they drop the neurons while training and add regularization. Then the flattened layer converts the two-dimensional feature map into a one-dimensional feature vector. Lastly, the softmax activation was done as it effectively performs multi-class classification. Our model's architecture is flexible and we can edit and expand it further. By adding additional layers and filters, we can make our model enable to classify different image classifying tasks. Our future goal is to make this model efficient in classifying different stages of Pneumonia and TB.

Our customized CNN model achieved the same level of accuracy in just 20 epochs while other models took at least 50 epochs to reach similar accuracy. This proves that our model is faster and provides better results within less time. As our model takes less time to perform, we may use this model with a large dataset or with limited resources. Moreover, as fewer epochs are needed, it means our model converges faster than other models. Also, it takes less time to train. It means we can work with larger datasets and we may get higher accuracy with larger datasets as well. We can use this model in any real-time situation as it takes a shorter time to train and needs minimum resources. Our model has elements like Batch Normalization, L2 Regularization, and Dropout and this helped the model to converge faster without disturbing the accuracy and performance.

Criterion	**Custom CNN**	**Vision Transformer (ViT)**	**Swin Transformer**	**VoLo**
Convergence Speed	20 epochs for peak accuracy	50+ epochs for convergence	50+ epochs for convergence	50+ epochs for convergence
Training Time	Faster due to fewer epochs	Slower due to complex attention mechanisms	Slower due to attention windows	Slower due to local and global attention
Computational Complexity	Low, convolutional layers focus on local features	High, due to self-attention mechanisms	Medium, hybrid of convolution and self-attention	Medium, combining local and global attention
Model Size	Small and deploy-friendly	Large, requires significant resources	Medium, hierarchical structure reduces size	Larger than CNN, combines CNN and transformer features
Memory Usage	Low memory usage	High memory usage	Medium, reduced memory compared to ViT	High, due to combination of CNN and transformer
Suitability for Small Datasets	Excellent	Poor, needs large datasets	Medium, performs better with small datasets than ViT	Medium, better with larger datasets

Table 3.4: Comparison of Custom CNN with ViT, Swin Transformer, and VoLo (Part 1)

Criterion	**Custom CNN**	**Vision Transformer (ViT)**	**Swin Transformer**	**FocalNet**
Local Feature Extraction	Excellent, localized feature extraction	Medium, focuses on global features	Good, windowed attention captures local features	Excellent, focuses on critical image regions
Global Feature Extraction	Limited	Excellent, self-attention captures global dependencies	Good, hierarchical structure captures global features	Excellent, focuses on critical regions globally
Overfitting Prevention	Efficient with dropout and batch normalization	Prone to overfitting, requires regularization	Balanced, windowed attention reduces irrelevant focus	Moderate, needs tuning to prevent overfitting
Training Resource Efficiency	High, fewer resources needed	Low, requires significant resources	Medium, more efficient than ViT	Medium, requires significant computational resources
Deployment Feasibility	Excellent, small and easy to deploy	Poor, large and complex	Medium, smaller than ViT	Medium, large resources required for effective performance

Table 3.5: Comparison of Custom CNN with ViT, Swin Transformer, and FocalNet (Part 2)

Reason for certain Misclassifications

COVID-19 Misclassified as Pneumonia

- Ground-glass opacities are found in both diseases.
- Diffused opacity patterns are common in COVID-19 and pneumonia.
- Lung lesions have similar shapes and locations.

Pneumonia Misclassified as COVID-19

- Overlapping lung infiltrates are typical in both conditions.
- Peripheral lung involvement is frequent in both.
- Patterns of lung consolidations and opacities appear similar.

Normal Misclassified as Pneumonia

- Minor anomalies or noise in normal X-rays are interpreted as disease.
- Image artifacts or postural variations are misinterpreted as infections.
- Oversensitivity to small changes, likely due to model bias or noise.

Pneumonia Misclassified as Normal

- Early-stage pneumonia presents with subtle or faint features.
- Small, localized lesions are missed by the model.
- Weak contrast in X-ray images leads to missed abnormalities.

COVID-19 Misclassified as Normal

- Mild or atypical presentations with minimal lung changes.
- Poor-quality images or preprocessing lose important information.
- Subtle opacities or abnormal patterns aren't emphasized enough by the model.

3.5 Vision Transformer

The Vision Transformer (ViT) represents a significant shift in vision computing, utilizing transformers, originally designed for natural language processing (NLP), to tackle image understanding tasks. Unlike traditional convolutional neural networks (CNNs), which focus on specific regions of an image, ViT decomposes the input image into uniform-sized patches and treats them as tokens—similar to words in a sentence—using self-attention mechanisms. Researchers have demonstrated excellent performance in various image classification tasks using ViT, especially with large dataset training. For this study, ViT introduces a novel approach to chest disease classification via X-ray images, allowing for better global feature extraction combined with contextual relations.

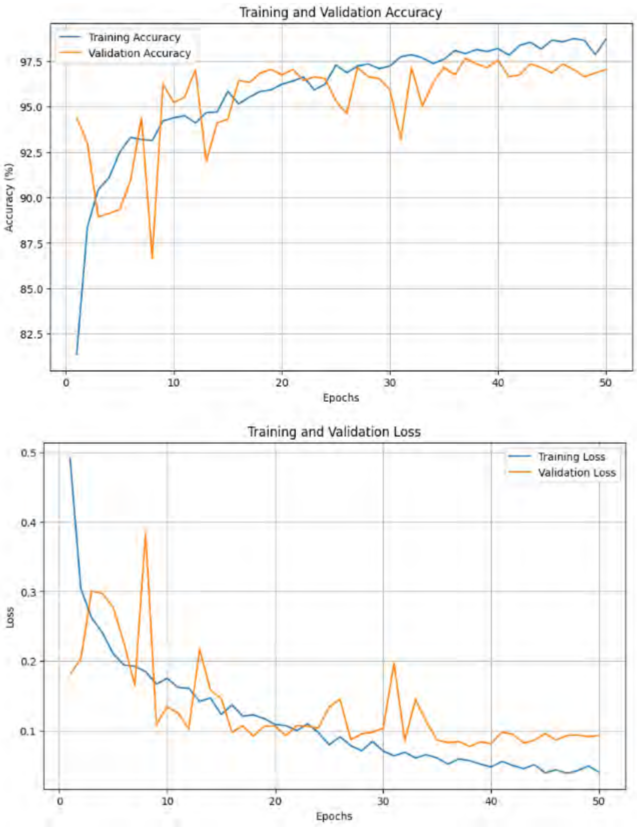


Figure 3.6: Accuracy and Loss graph for ViT

3.5.1 ViT Architecture

The main novelty of ViT lies in how it processes images. Instead of relying on convolutions, transformer layers are used to capture relationships between different image components.

Image Patch Embedding

- **Patch Splitting**: The input image is split into patches. For example, a 256x256 X-ray image can be divided into 16x16 sub-regions.
- **Linear Projection**: The patches are linearly mapped to a fixed-size embedding (e.g., 768 for base ViT). This process converts the 2D image patches into 1D patch embeddings.

Positional Encoding

Each patch embedding is further enriched with positional encoding to maintain spatial information. The transformer encoder then processes the sequence of patches through multi-head self-attention (MHSA) and feedforward neural networks. MHSA captures both local and global features, helping the model understand interconnections between patches in X-ray images.

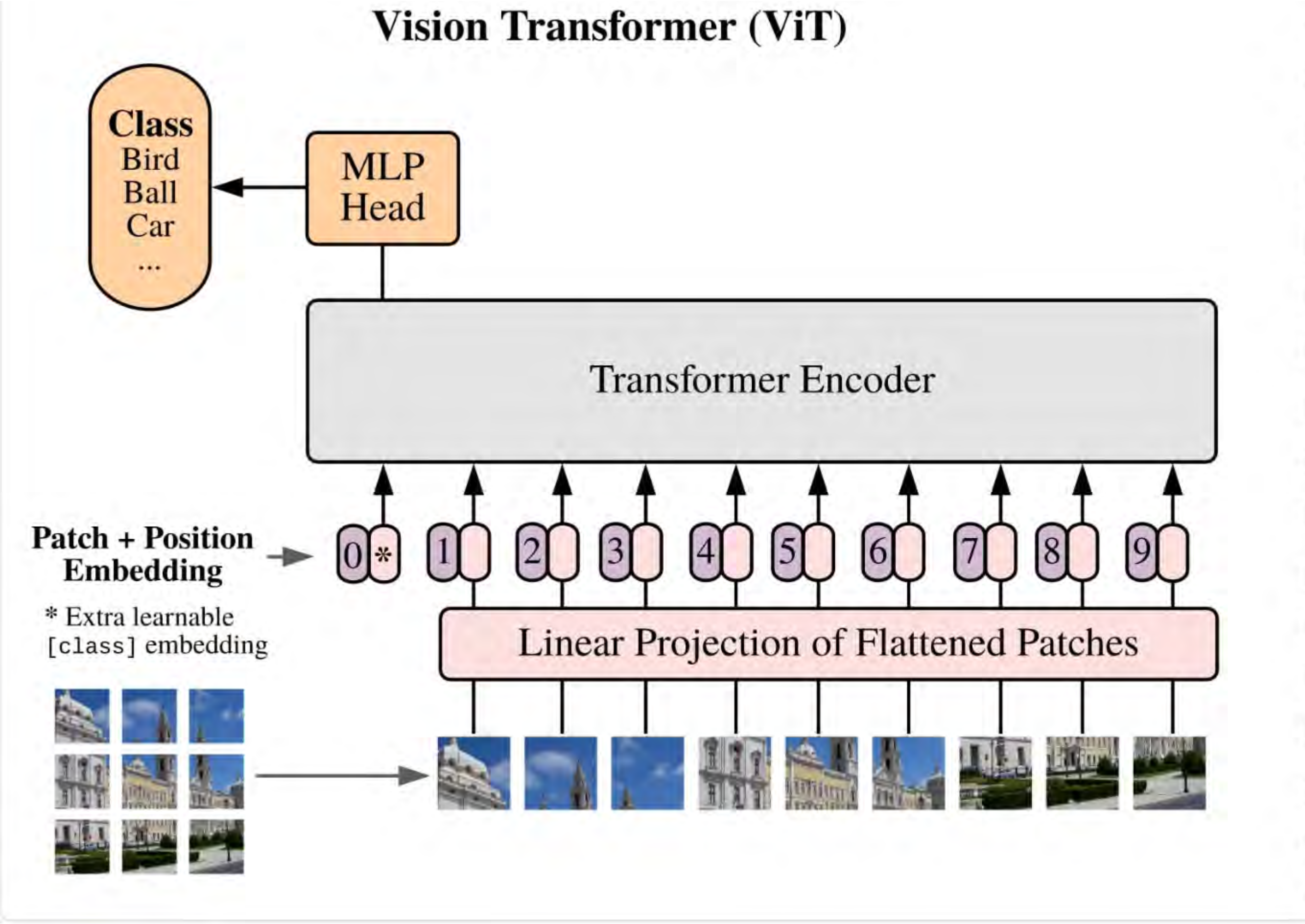

Vision Transformer (ViT)
Class
Bird
Ball
Car
...
MLP Head
Transformer Encoder
Patch + Position Embedding
* Extra learnable [class] embedding
0 *
1
2
3
4
5
6
7
8
9
Linear Projection of Flattened Patches

Figure 3.7: ViT Architecture

Transformer Encoder Layers

The image patches, combined with positional encodings, are processed through several transformer encoder layers. Each encoder layer consists of:

- **Multi-head Self-Attention (MHSA)**: Allows the model to focus on different regions of the image simultaneously, identifying both local and global patterns.
- **Feedforward Neural Networks**: After applying self-attention, fully connected feedforward networks capture more complex image representations.
- **Layer Normalization and Residual Connections**: These mechanisms stabilize training and ensure a consistent flow of gradients.

Classification Head

The classification head in ViT aggregates global information from all patch embeddings using a CLS token, which is then processed by fully connected layers to predict the class (e.g., healthy, pneumonia, or tuberculosis).

Pretraining and Fine-tuning

ViT benefits significantly from pretraining on large-scale datasets such as ImageNet. In this study, the ViT model was pretrained and then fine-tuned on chest X-ray datasets to adapt to the specific patterns of medical images, enabling the model to detect small deviations indicative of disease.

3.5.2 Training Process of ViT on Chest X-ray Dataset

The X-ray images were resized to 256x256 pixels and divided into 16x16 patches. These patches were normalized into 768-dimensional vectors, with positional encoding added to retain spatial information. The model was trained using the AdamW optimizer, with a warm-up rate to prevent overfitting. A batch size of 32 was selected to optimize GPU utilization without consuming excessive memory.

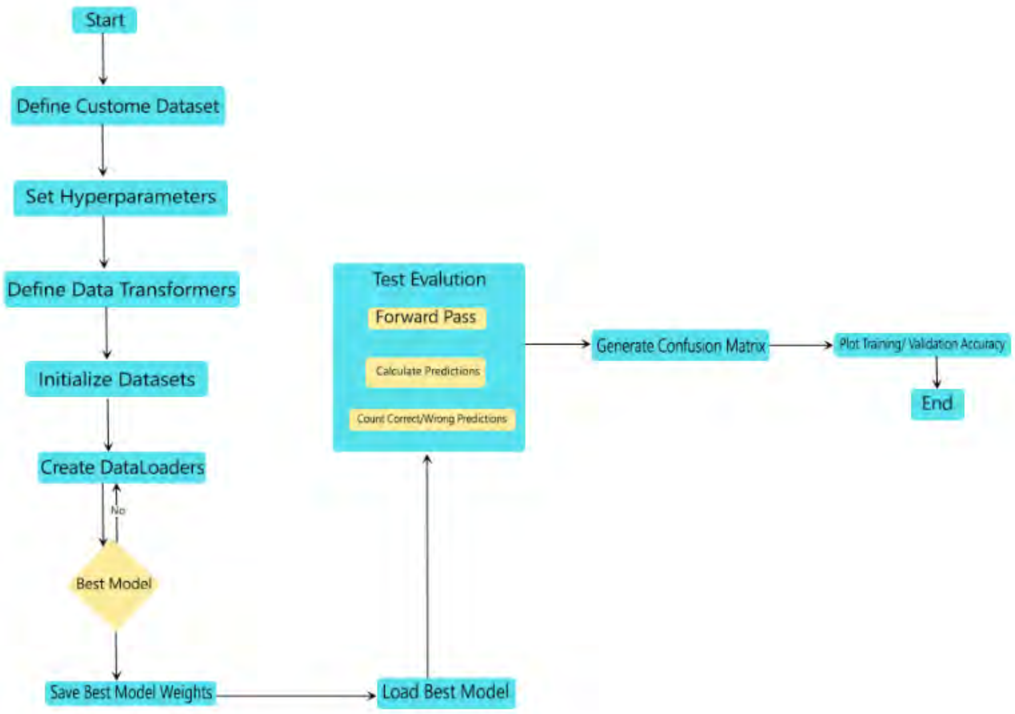


Figure 3.8: Workflow of ViT [35]

3.5.3 Performance of ViT on Chest Disease Classification

ViT achieved a training accuracy of 98.72%, demonstrating its ability to learn both global and regional features from X-ray images, including fine details such as small tissues and abnormalities. The test accuracy was 96.4%, reflecting the model's capacity to generalize well to unseen data, even with challenging image quality.

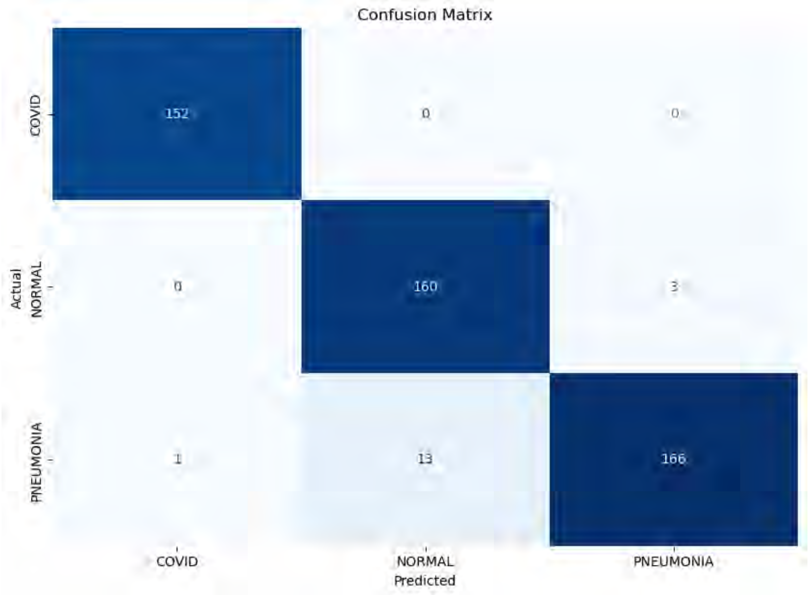


Figure 3.9: Confusion Matrix of ViT

3.5.4 Generalization and Overfitting

ViT maintained a high training accuracy throughout, with minimal overfitting. Techniques such as data augmentation and dropout were applied during training to prevent overfitting. The small gap between training and testing accuracy (2%) indicates the model's robustness, though further fine-tuning could potentially improve performance.

3.5.5 Strengths of Vision Transformer (ViT)

ViT's attention mechanism significantly improves the detection of long-range spatial information in images, which is particularly beneficial for medical imaging tasks like chest X-ray analysis, where diseases may not be immediately visible. Unlike CNNs, which focus on specific regions, ViT captures both local and global features, resulting in a comprehensive understanding of the image. Pretraining enhances the model's performance by leveraging knowledge from large image datasets, which can be especially useful when medical data is limited. Additionally, ViT's self-attention mechanism provides interpretability by generating attention maps, which can help clinicians understand the model's decision-making process.

3.6 Swin Transformer

The Swin Transformer (Shifted Window Transformer) is a hierarchical vision transformer specifically designed for large-scale image analysis. Previous work on the Vision Transformer (ViT) demonstrated impressive performance by using attention mechanisms for image classification, but its quadratic complexity limited its use on large-scale images. The Swin Transformer addresses this limitation by incorporating a local attention mechanism through the window shift operation, which allows for better feature extraction both locally and globally, while being more computationally efficient. Swin Transformer is well-suited for classifying chest diseases from X-ray images, with the ability to capture multi-scale features and process high-resolution images efficiently, similar to convolutional neural networks (CNNs). In this study, the Swin Transformer achieved a training accuracy of 98.66% and a testing accuracy of 97.2%.

Although Swin Transformer is a new model design, it builds on top of the transformer network by incorporating inductive biases that enable efficient training and near state-of-the-art performance on high-resolution medical image data, such as X-rays. Unlike ViT, which treats images as flat structures, Swin Transformer processes images hierarchically, making it more aligned with CNNs in terms of capturing multi-scale features. Patch merging layers downsample the feature map at each level of the hierarchy, increasing the channel dimension recursively along a deep convolutional stage. This allows the model to effectively capture both local and global features in the image.

3.6.1 Swin Transformer Architecture

The Swin Transformer improves upon the traditional transformer architecture by making it more scalable and flexible, especially for high-resolution tasks like X-ray image analysis. Unlike ViT's flat structure, Swin Transformer processes images at multiple resolutions to capture multi-scale features. At each stage of the hierarchy, patch merging layers downsample the feature maps, reducing spatial resolution but increasing the channel dimension to capture more global information. A key innovation is its window-based multi-head self-attention (W-MSA), where self-attention is applied within non-overlapping local windows, reducing the computational complexity from quadratic to linear. To enable information flow between windows, shifted windows are used between layers.

Swin Transformer divides input images into patches (e.g., a 256x256 X-ray image is divided into 4x4 windows of 16x16 patches) and applies self-attention within these regions. Patch merging performs pooling-like operations at each stage, reducing spatial dimensions and increasing the number of channels, similar to CNNs. Each set of windows is processed with standard multi-head self-attention (MHSA), followed by feedforward layers and normalization, enabling efficient processing of high-resolution images. The feature maps are then passed to a supervised head, consisting of fully connected layers and a softmax function, to predict image classes (e.g., healthy, pneumonia, tuberculosis). Pretraining on large datasets like ImageNet, followed

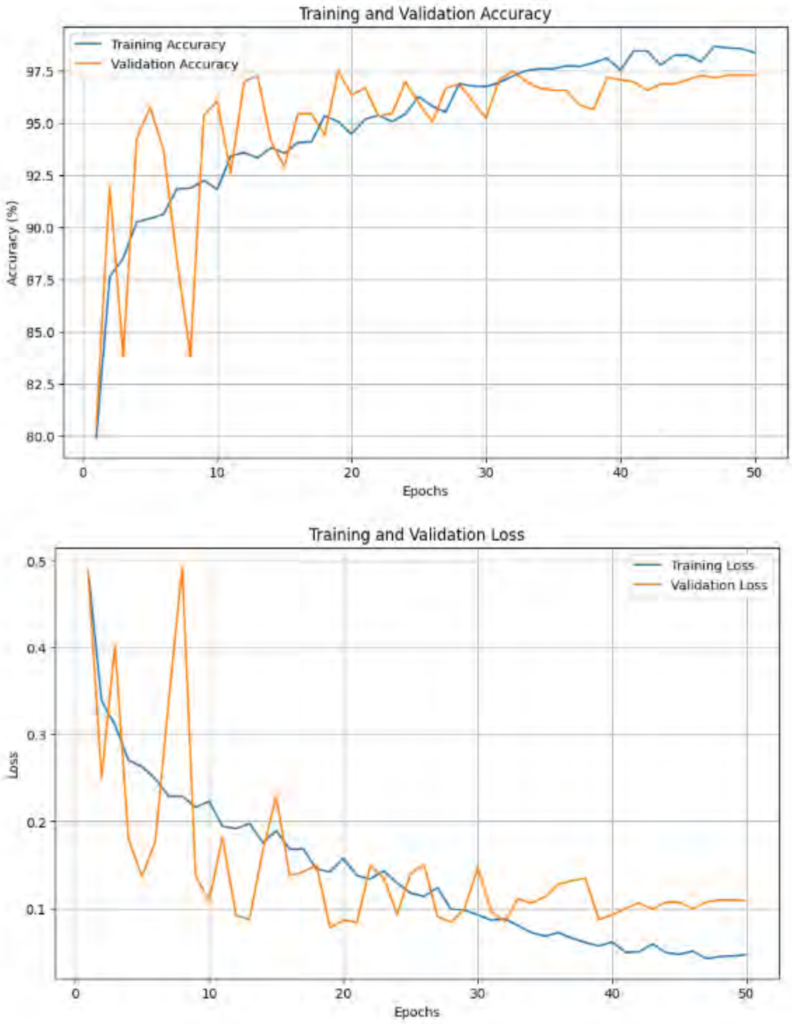


Figure 3.10: Accuracy and Loss graph for Swin Transformer

by fine-tuning on the X-ray dataset, allows the model to generalize and adapt to medical features, such as lung opacity or masses.

In this study, chest X-ray images were resized to 256x256 pixels, and Swin Transformer processed them by decomposing them into local windows of patches. Each patch was processed locally within its window, while window shifts between layers enabled cross-window dependencies. Data augmentations, such as horizontal flipping, random cropping, and brightness correction, were applied during training to prevent overfitting. The AdamW optimizer was used to update the model's parameters, starting with a learning rate of 1e-4 and using a cosine annealing schedule following a warm-up phase. A batch size of 32 was chosen to balance memory optimization and training speed. The model was trained for 50 epochs, achieving convergence with a training accuracy of 98.66%. Checkpoints were saved after each epoch, with early stopping checks performed every five epochs. Despite its reduced complexity in the final layers, Swin Transformer's multi-scale nature consumed significant computational resources.

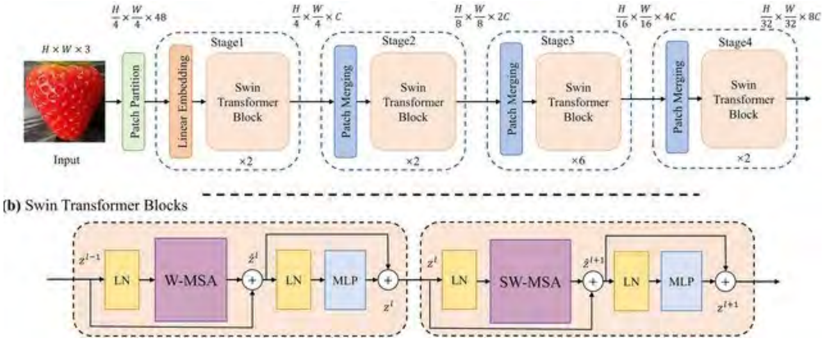


Figure 3.11: Swin Transformer Architecture

3.6.2 Training Process of Swin Transformer on Chest X-ray Dataset

In this study, the chest X-ray images were resized to 256 × 256 pixels, and Swin Transformer decomposed them into local windows of patches, processing each patch within its window locally while allowing cross-window dependencies through window shifts between layers. During training, data augmentations such as horizontal flipping, random cropping, and brightness correction were applied to address overfitting. The model's parameters were updated using the AdamW optimizer, starting with a 1e-4 learning rate using a cosine annealing schedule following a warm-up phase. A batch size of 32 was chosen to balance memory optimization and training speed. Swin Transformer was trained for 50 epochs, achieving convergence with a training accuracy of 98.66%, with checkpoints saved after each epoch and early stopping checks every five epochs. The complete 50 epochs were necessary to reach optimal performance, leveraging its hierarchical design and window-based attention. Despite its reduced complexity in final layers, its multi-scale nature consumed significant computational resources.

3.6.3 Performance of Swin Transformer on Chest Disease Classification

Swin Transformer achieved a training accuracy of 98.66%, demonstrating its effectiveness in learning both local and global features from chest X-ray images, making it highly suitable for medical classification tasks. The testing accuracy of 97.2% reflects the model's strong generalization ability. The small gap between training and testing accuracies indicates that overfitting was avoided. The window-based attention mechanism allowed the model to focus on critical regions of the X-ray, such as the lungs and heart, while filtering out irrelevant background noise, contributing to its robust performance on unseen data.

3.6.4 Strengths of Swin Transformer

Swin Transformer's window-based attention mechanism significantly reduces computational complexity compared to ViT, making it particularly efficient for handling high-resolution medical images like chest X-rays, where processing detailed information is crucial. Its ability to extract multi-scale features allows the model to capture both local features, such as small lesions, and global features, such as the overall

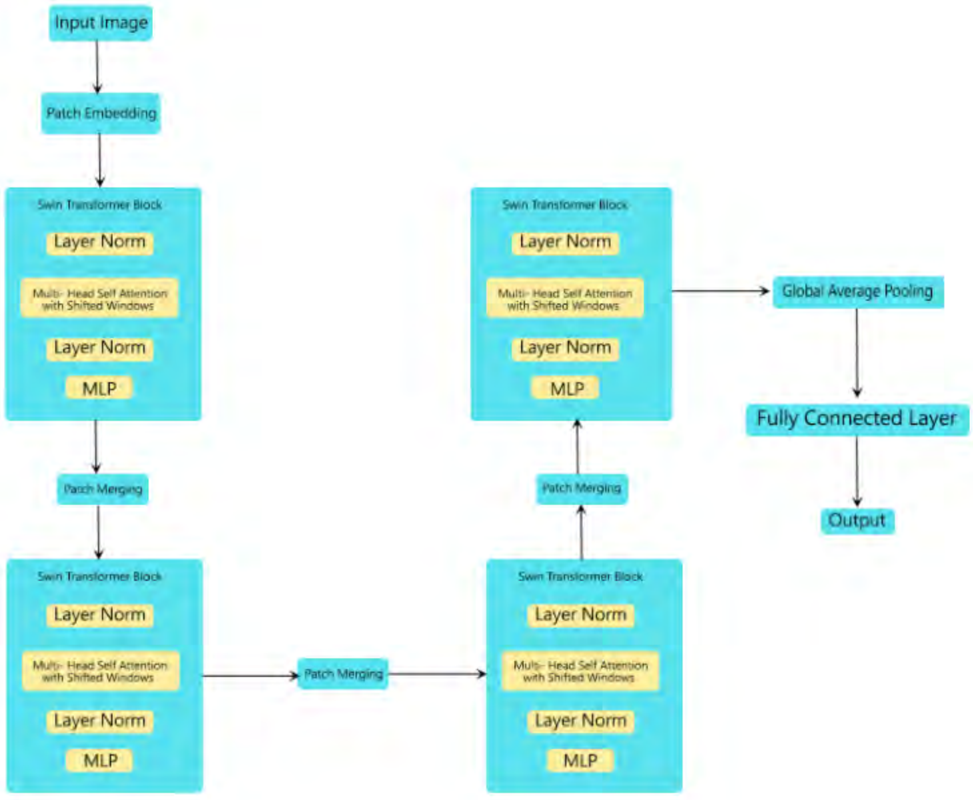


Figure 3.12: Workflow of Swin Transformer [22]

lung structure. This is vital for medical imaging tasks where subtle changes in small regions can indicate disease, but understanding the overall context is equally important. The shifted window attention mechanism enables the model to capture cross-window dependencies, allowing it to process both local and global contexts, which is especially beneficial in medical image analysis where abnormalities may span across multiple regions of the X-ray. Pretraining on large datasets allows Swin Transformer to transfer its general knowledge from tasks like image classification to more specialized tasks, such as chest disease detection. Fine-tuning on the X-ray dataset helps the model adapt to specific medical features, such as lung opacity or abnormal tissue structures. Its high testing accuracy reflects the model's robustness to variations in X-ray images, including differences in angles, lighting, and patient demographics. Data augmentation during training further enhanced the model's robustness, making it highly reliable for clinical applications.

3.7 Volo (Vision Outlooker) for Chest Disease Classification

3.7.1 Introduction to Volo

Volo, or Vision Outlooker, is a vision model designed to bridge the gap between traditional convolutional neural networks (CNNs) and modern transformer-based

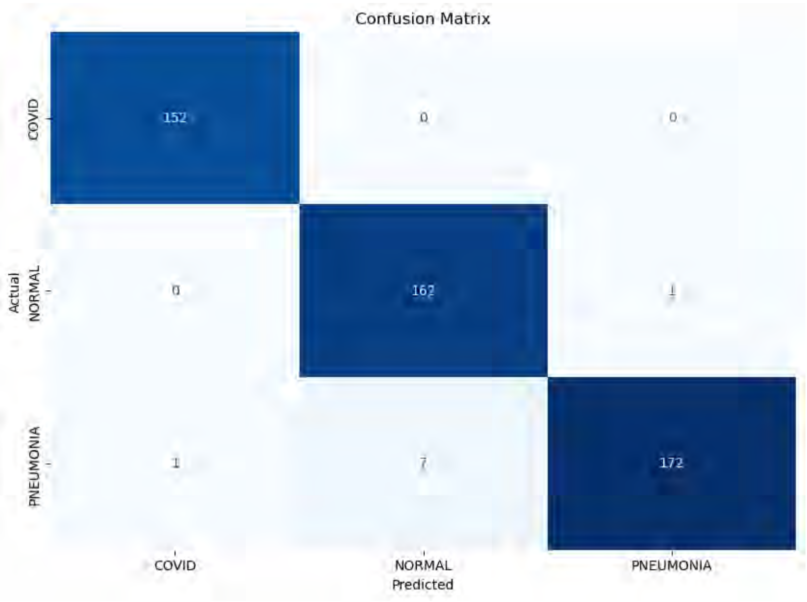


Figure 3.13: Confusion Matrix of Swin Transformer

architectures. It introduces a novel outlook attention mechanism that effectively captures both local and global features. By combining CNNs' proficiency in low-level feature extraction with transformers' ability to capture global context, Volo provides a comprehensive approach to image analysis.
In this study, Volo demonstrated impressive performance in chest disease classification using X-ray images, achieving a training accuracy of 98.43% and a testing accuracy of 98.0%. This makes Volo one of the top-performing models, showcasing its capability to handle the complex challenges of medical image classification with precision.

3.7.2 Volo Architecture

Volo introduces an innovative architecture that combines convolutional operations and transformer-like attention mechanisms to address key challenges in vision tasks, such as capturing long-range dependencies while maintaining computational efficiency.

Outlook Attention Mechanism

- **Outlook Attention:** Volo's core innovation is the outlook attention mechanism, which extracts both local and global features from input images. This mechanism enables the model to focus on critical areas of the image, akin to traditional attention mechanisms in transformers, but with improved computational efficiency.

- Spatial aggregation is performed over local patches, which are then combined into global feature representations. This design allows the model to capture detailed, fine-grained features, such as small lesions in chest X-rays, while also comprehending the broader image structure.

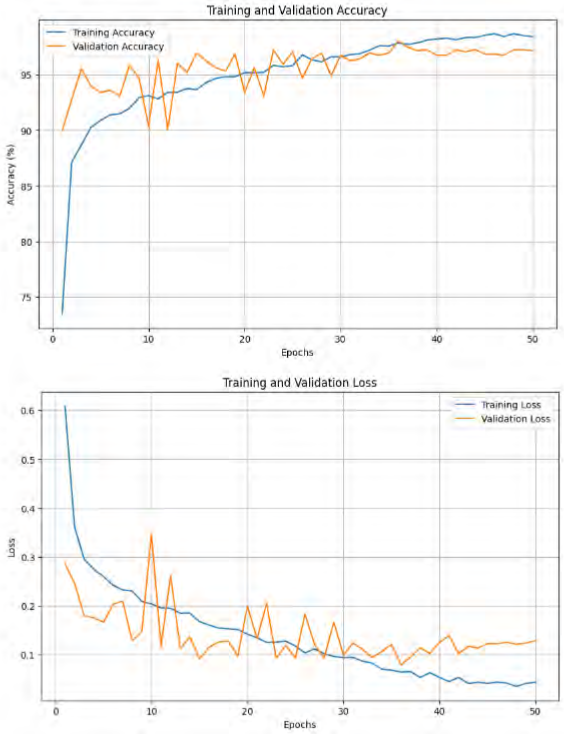


Figure 3.14: Accuracy and Loss graph for VOLO-D1

Patch Embedding

- Similar to ViT, Volo divides the input image into patches, treating each as a token. These patches are then flattened and linearly embedded into vectors, which are processed by the outlook attention layers. This approach enables efficient handling of high-resolution medical images like chest X-rays.
- For example, a 256x256 pixel X-ray image can be divided into 16x16 patches, with each patch embedded into a 768-dimensional vector.

Multi-stage Processing

- Volo's architecture is divided into multiple stages, with each stage processing the input image at different levels of abstraction. In the initial stages, the model focuses on low-level features like edges and textures, while later stages capture higher-level semantic information, such as lung structure or disease indicators.
- Stage 1 processes patches through outlook attention layers, capturing local details and relationships between neighboring patches.
- In stages 2 and beyond, multi-head self-attention and feedforward layers refine the feature maps and learn complex patterns, such as lung opacity or abnormal tissue structures in X-rays.

Transformer-like Encoder

- In addition to the outlook attention mechanism, Volo retains traditional transformer encoder components, including multi-head self-attention and feedforward layers. These enable the model to capture long-range dependencies between different regions of the image, essential in medical imaging where abnormalities can span multiple areas.

- **Layer Normalization and Residual Connections:** These mechanisms are employed after each attention and feedforward layer to ensure stable training and smooth gradient flow.

Classification Head

- After processing through multiple outlook and self-attention layers, the final output is passed through a classification head consisting of fully connected layers. A softmax layer produces a probability distribution over possible classes, such as healthy, pneumonia, or tuberculosis.

- In this study, the classification head was fine-tuned to distinguish between various chest diseases using X-ray images, leveraging both the global context and local features extracted by previous layers.

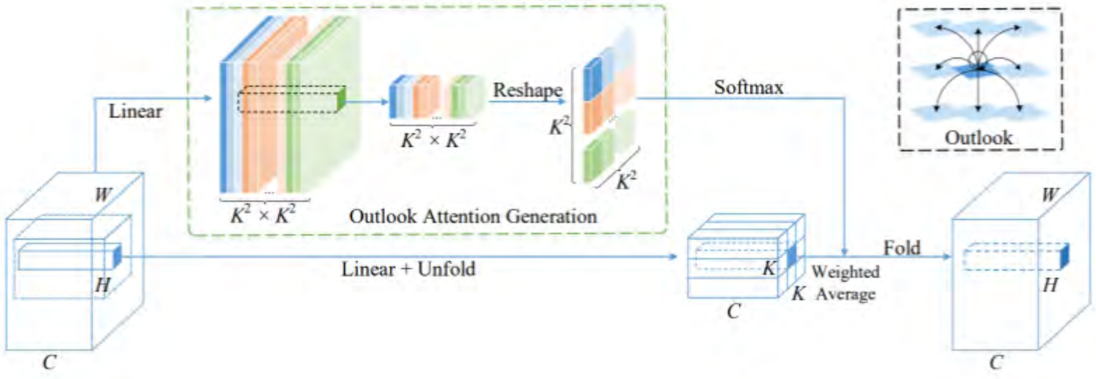


Figure 3.15: VOLO-D1 Architecture

3.7.3 Training Process of Volo on Chest X-ray Dataset

Dataset Preparation

- **Resizing and Patch Division:** Chest X-ray images were resized to 256x256 pixels and divided into 16x16 patches. These patches were then embedded into high-dimensional vectors for processing by the outlook attention layers.

- **Data Augmentation:** Techniques such as rotation, brightness adjustments, and random cropping were applied during training to enhance the model's generalization. These augmentations helped improve robustness against real-world variations in medical imaging data.

Training Parameters

- **Optimizer:** The AdamW optimizer was used for parameter updates, providing a balance between speed and stability, especially in transformer-based architectures like Volo.
- **Learning Rate:** A learning rate of 1e-4 was employed, along with a cosine annealing schedule to gradually reduce the learning rate during training.
- **Batch Size:** A batch size of 32 was selected to optimize GPU usage while preventing memory overload.

Training Epochs

- **Number of Epochs:** Volo was trained for 50 epochs, achieving a training accuracy of 98.43% and a testing accuracy of 98.0%. This extensive training period allowed the model to converge to high accuracy while avoiding overfitting.
- **Early Stopping:** Early stopping was implemented to halt training when no improvement in validation accuracy was observed over several epochs.

Training Efficiency

- Volo's outlook attention mechanism significantly improved training efficiency by reducing the complexity typically associated with self-attention mechanisms. This allowed the model to achieve high accuracy without the heavy computational demands of models like ViT.

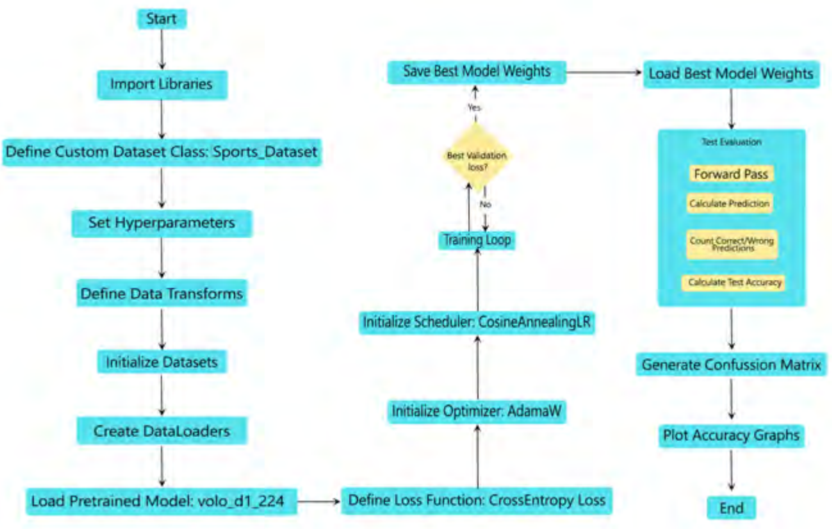


Figure 3.16: Workflow for VOLO-D1 [30]

3.7.4 Performance of Volo on Chest Disease Classification

Training Accuracy

Volo achieved a training accuracy of 98.43%, demonstrating its ability to effectively learn patterns from the training dataset. The model's capability to capture both local features, such as small lesions, and global structures, like lung opacity, contributed to this high accuracy.

Testing Accuracy

Volo achieved a testing accuracy of 98.0%, demonstrating strong generalization to unseen X-ray images. The small gap between training and testing accuracies suggests the model avoided overfitting, allowing it to accurately predict outcomes on new images, even those with subtle disease patterns.

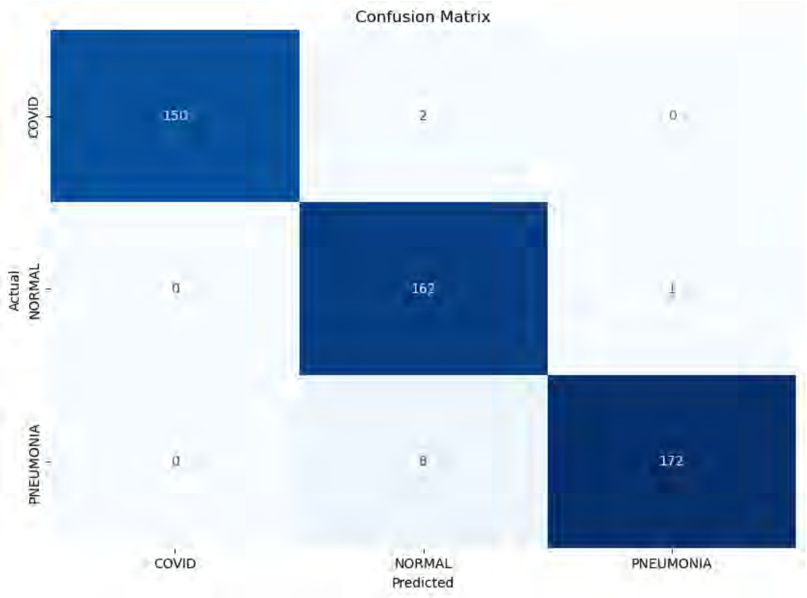


Figure 3.17: Confusion Matrix of VOLO-D1

3.7.5 Strengths of Volo

Efficient Attention Mechanism

Volo's outlook attention mechanism balances attention across local and global regions of the image, allowing the model to capture important features while maintaining computational efficiency. This makes it suitable for high-resolution X-ray images without excessive resource consumption.

Multi-scale Feature Extraction

Volo's multi-stage architecture enables the extraction of multi-scale features, a critical requirement for medical image analysis. It captures small abnormalities while understanding the larger structure of the lungs, making it highly effective for chest disease classification tasks.

Pretraining and Fine-tuning

Pretraining on large datasets like ImageNet provides Volo with a foundation of general image features, which are fine-tuned for chest X-ray patterns. This process enhances the model's ability to identify disease markers in medical images.

Attention Mechanism for Explainability

Volo's outlook attention mechanism offers a degree of explainability by highlighting the regions of the image the model focuses on during prediction. This is valuable in medical applications where understanding the model's reasoning can provide additional insights for clinicians.

3.7.6 Limitations of Volo

Computational Complexity

Although Volo is more efficient than many transformer-based models, it still demands significant computational resources compared to simpler CNN architectures. The inclusion of both outlook attention and transformer-like layers increases overall complexity, making deployment in resource-limited environments challenging.

Data Dependency

Volo, like most transformer-based models, requires access to a large dataset to reach optimal performance. While fine-tuning on chest X-ray data enhanced its accuracy, the model's full potential may only be realized with even larger datasets.

Window Size and Attention Sensitivity

Volo's performance can be highly sensitive to the size of the image patches and attention mechanism configuration. Tuning these hyperparameters requires careful experimentation, and incorrect settings could lead to suboptimal performance.

3.8 FocalNet (Focal Modulated Network) for Chest Disease Classification

3.8.1 Introduction to FocalNet

Focal Modulated Network (FocalNet) is a transformer-based architecture designed to improve feature extraction at different scales without compromising computational efficiency. FocalNet introduces a lightweight focal modulation mechanism that enables the model to focus on both local and global features simultaneously. This is especially beneficial for tasks like chest disease classification using X-ray images, where fine-grained details and global context are equally important for accurate diagnosis. In this study, FocalNet achieved 99.43% training accuracy and 97.60% testing accuracy, making it one of the top-performing models with strong learning and generalization capabilities.

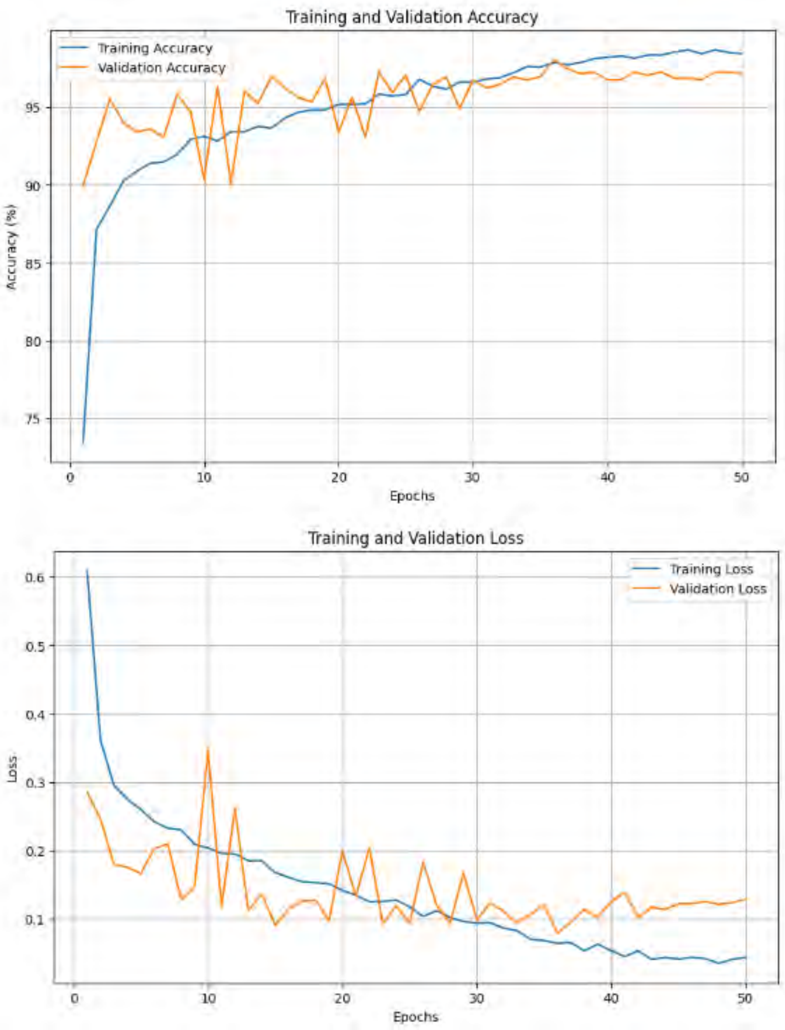


Figure 3.18: Accuracy and Loss graph for FocalNet

3.8.2 FocalNet Architecture

FocalNet's architecture builds on transformer-like models while introducing focal modulation layers that allow for efficient multi-scale feature extraction. This design enables the model to capture important details in medical images while keeping computational costs manageable.

Focal Modulation Mechanism

- **Focal Modulation:** The key innovation in FocalNet is its focal modulation mechanism, which models both local and global representations in a depth-wise manner. This operates similarly to transformers but is more computationally efficient.

- **Local and Global Focus:** In early layers, FocalNet extracts local textures (e.g., edges) from X-ray images, which are essential for detecting small lesions or nodules. As the network deepens, FocalNet captures global information (e.g., overall lung structure or opacity), providing a broader context for diagnosis.

- **Multi-scale Representation:** Focal modulation layers operate across multiple scales, allowing the model to capture fine-grained details and global patterns simultaneously.

Patch-based Input

- Like ViT and Swin Transformer, FocalNet divides the input image into patches, treating each patch as a token that passes through focal modulation layers. This approach is efficient for high-resolution images like chest X-rays.
- For example, a 256x256 pixel X-ray image can be divided into 16x16 patches, each embedded into high-dimensional vectors, which are processed by focal modulation layers.

Hierarchical Design

FocalNet employs a hierarchical design similar to CNNs, progressively reducing the spatial resolution of feature maps while increasing the number of channels. This design is crucial for capturing multi-scale features, which are important for accurate chest disease classification.

- **Stage 1 (Local Feature Extraction):** FocalNet starts by focusing on extracting local features through convolutional layers. These features are passed to focal modulation layers to aggregate local information.
- **Stage 2 (Global Feature Integration):** In later stages, the model integrates global information from different regions of the image, capturing high-level features like lung structure or diffuse opacity in chest X-rays.

Transformer-like Encoder

- FocalNet retains key elements of transformers, such as multi-head self-attention (in a modified form) and feedforward layers. These components help capture long-range dependencies between different parts of the image, which is essential for tasks like chest disease classification.

Classification Head

- The final output from the focal modulation layers is passed through a classification head, consisting of fully connected layers followed by a softmax function. This outputs probabilities for each class (e.g., healthy, pneumonia, tuberculosis). The classification head is fine-tuned specifically for chest disease diagnosis.

Pretraining and Fine-tuning

- FocalNet benefits from pretraining on large image datasets like ImageNet to learn general visual features. It is then fine-tuned on the chest X-ray dataset to adapt to specific medical patterns, such as abnormal lung textures, masses, or pleural effusion.

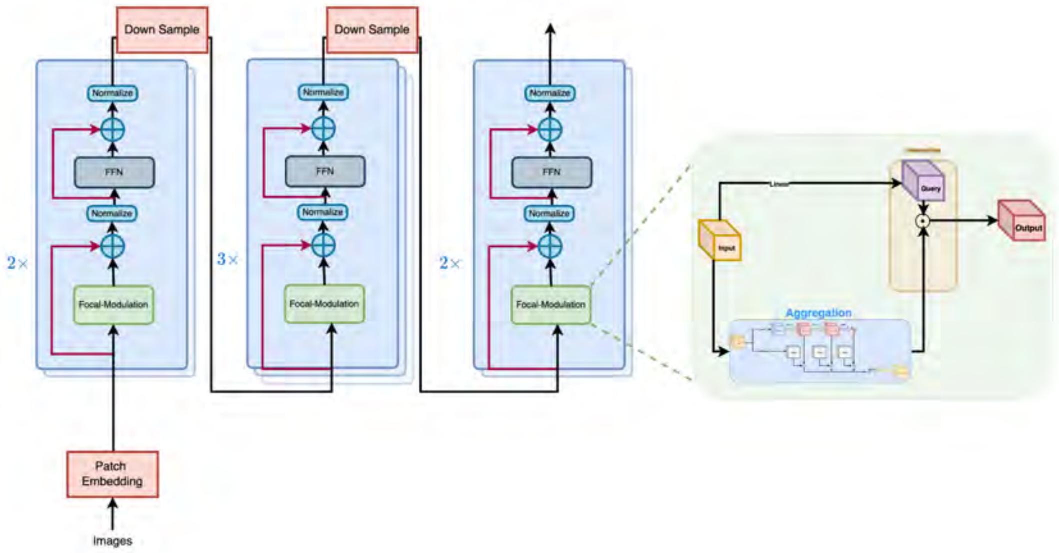


Figure 3.19: FocalNet Architecture

3.8.3 Training Process of FocalNet on Chest X-ray Dataset

Dataset Preparation

- **Image Resizing:** Chest X-ray images were resized to 256x256 pixels and divided into patches that served as input to the focal modulation layers.

- **Data Augmentation:** Techniques like random rotations, brightness adjustments, and flipping were applied during training to improve the model's generalization ability, which is critical for medical imaging tasks where variations in patient positioning or image quality can affect performance.

Training Parameters

- **Optimizer:** AdamW was used for stable and efficient training, well-suited for transformer-based models due to its ability to handle sparse gradients and weight decay effectively.

- **Learning Rate:** The learning rate was set to 1e-4, with a cosine annealing schedule to gradually reduce the learning rate, ensuring smooth convergence and avoiding overshooting.

- **Batch Size:** A batch size of 32 was used to balance memory efficiency with training speed.

Training Epochs

- FocalNet was trained for 50 epochs, ensuring the model converged and reached peak performance. Validation loss was monitored, and early stopping was implemented to avoid overfitting.

Training Efficiency

- Despite its complex architecture, FocalNet's focal modulation mechanism made it more efficient than traditional transformers by avoiding the quadratic complexity of global self-attention, allowing it to achieve strong performance without excessive computational demands.

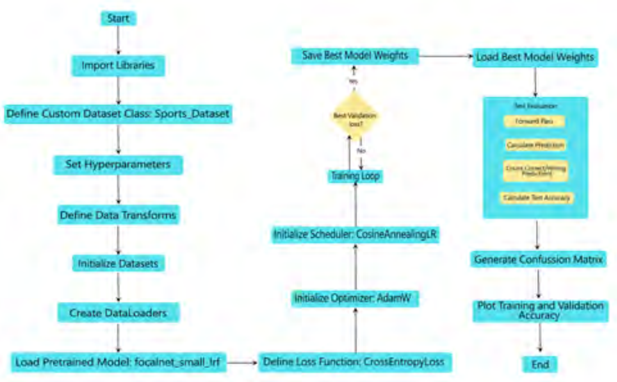

Start
Import Libraries
Define Custom Dataset Class: Sports_Dataset
Set Hyperparameters
Define Data Transforms
Initialize Datasets
Create DataLoaders
Load Pretrained Model: focalnet_small_lrf
Define Loss Function: CrossEntropyLoss
Initialize Optimizer: AdamW
Initialize Scheduler: CosineAnnealingLR
Training Loop
Best Validation loss?
Yes
No
Save Best Model Weights
Load Best Model Weights
Test Evaluation
Forward Pass
Calculate Prediction
Count Correct/Wrong Predictions
Calculate Test Accuracy
Generate Confussion Matrix
Plot Training and Validation Accuracy
End

Figure 3.20: Workflow for FocalNet [34]

3.8.4 Performance of FocalNet on Chest Disease Classification

Training Accuracy

FocalNet achieved a remarkable 99.43% training accuracy, indicating that the model effectively learned the patterns in the chest X-ray dataset. This high accuracy demonstrates FocalNet's ability to capture both local and global features, such as small lesions, lung opacity, and tissue structures.

Testing Accuracy

FocalNet achieved a 97.60% testing accuracy, reflecting strong generalization. The minimal gap between training and testing accuracy suggests that FocalNet did not overfit and can accurately predict outcomes on unseen X-ray images. This is crucial for medical imaging tasks where reliable generalization is essential for real-world deployment.

3.8.5 Strengths of FocalNet

Efficient Multi-scale Feature Extraction

The focal modulation mechanism in FocalNet allows for efficient multi-scale feature extraction, enabling the model to focus on fine-grained details (e.g., small lesions) and broader patterns (e.g., lung opacity) without the computational costs of traditional self-attention mechanisms.

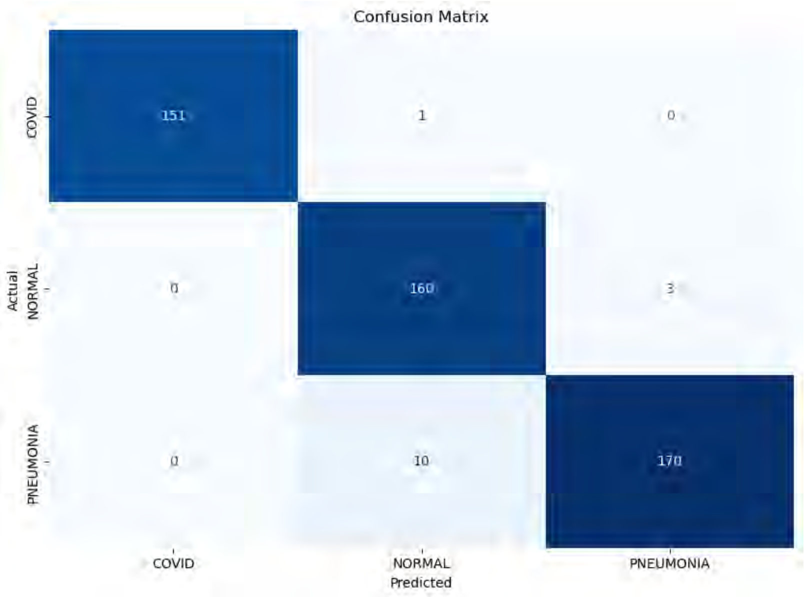


Figure 3.21: Confusion Matrix of FocalNet

Strong Generalization

FocalNet's high testing accuracy demonstrates its ability to generalize well to unseen data, making it ideal for medical imaging tasks where generalization across different patient populations and imaging conditions is critical for clinical use.

Pretraining and Fine-tuning

Like other transformer-based models, FocalNet benefits from pretraining on large datasets to learn general image features. Fine-tuning on the chest X-ray dataset helps the model specialize in patterns relevant to chest disease classification, boosting its performance on medical imagery.

Efficient Use of Attention

FocalNet's attention mechanism is more efficient than traditional self-attention, avoiding the quadratic complexity of transformers. This allows the model to handle high-resolution images like chest X-rays without the computational demands of models like ViT or Swin Transformer.

Robustness to Data Variations

FocalNet's performance on the chest X-ray dataset shows its robustness to data variations such as different image resolutions, patient positioning, and imaging conditions. Data augmentation techniques during training improved the model's resilience to these variations.

3.8.6 Limitations of FocalNet

Complexity

While FocalNet is more efficient than traditional transformers, it remains more complex than CNN-based models. The focal modulation layers, though efficient, add computational overhead compared to simpler architectures like custom CNNs, limiting FocalNet's use in settings with limited computational resources.

Data Dependency

Like most transformer-based models, FocalNet requires large amounts of data to perform optimally. While fine-tuning mitigated this limitation, the model's full potential can only be realized with access to large datasets, which can be a challenge in medical environments with limited data.

3.9 Vitamin (CNN-based Architecture with Attention Mechanisms) for Chest Disease Classification

3.9.1 Introduction to Vitamin

Vitamin is a modified version of CNN with attention mechanisms, replacing the block convolution (blockcv) module with a Vecoder attention-based backbone to focus on important parts of an image. While recent Vision Transformers use attention mechanisms exclusively, Vitamin combines CNN for local feature extraction with attention layers to capture more meaningful relationships within the image. This model was designed to leverage both approaches for tasks like chest disease classification using X-ray images. However, in this study, Vitamin underperformed compared to other models, achieving 86.20% training accuracy and 82.73% testing accuracy.

3.9.2 Vitamin Architecture

The network architecture is similar to traditional CNN-based approaches but includes additional layers to incorporate attention mechanisms, allowing the model to focus on regions likely to indicate abnormalities while ensuring computational efficiency from CNNs.

Convolutional Layers

- **Local Feature Extraction:** Like standard CNNs, Vitamin begins by extracting local features through convolutional layers. These layers detect low-level features such as edges, textures, and shapes, crucial for identifying abnormalities in X-ray images like lesions, nodules, or lung opacity.

- **ReLU Activation:** After each convolution, a ReLU (Rectified Linear Unit) activation function is applied to introduce non-linearity, enabling the model to learn more complex patterns in the data.

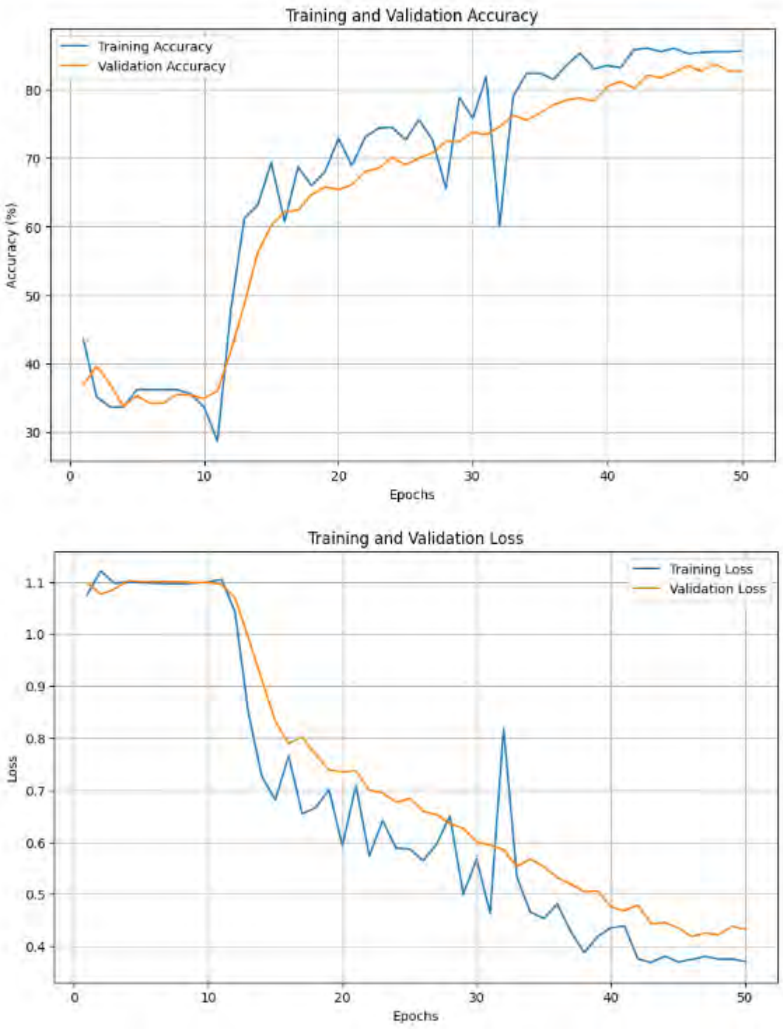


Figure 3.22: Accuracy and Loss graph for ViTamin

Attention Mechanisms

- Vitamin employs attention layers between convolutional layers, allowing the model to focus on regions of the X-ray image likely to indicate disease, similar to how people inspect X-rays—focusing on suspicious areas while ignoring irrelevant ones.

- **Self-attention:** While transformers use self-attention, Vitamin simplifies this mechanism. Each pixel in the feature map evaluates neighboring pixels to determine which are relevant, enabling the model to focus on regions exhibiting disease-like patterns.

- **Channel Attention:** Vitamin also integrates channel attention, which prioritizes the most informative feature channels, ensuring the model does not waste capacity on irrelevant or redundant features.

Pooling Layers

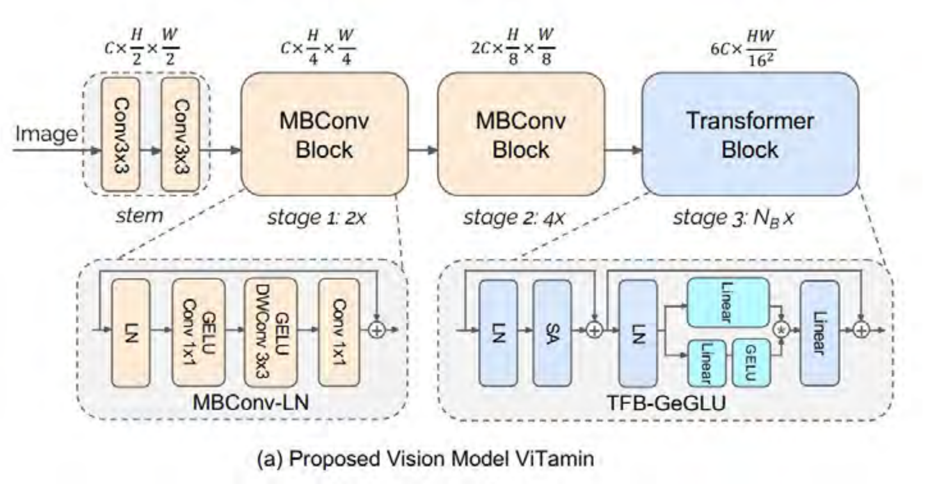
- **Max Pooling:** After each convolutional layer, max pooling reduces the size of feature maps, downsampling the image while retaining important features. Max pooling helps the model become translation invariant, enabling it to recognize features (like a tumor) regardless of location in the image.

Fully Connected Layers

After the convolution and attention layers, the features are flattened into a vector and passed through fully connected (dense) layers to make a prediction. Dropout layers are used between fully connected layers to prevent overfitting. Dropout randomly deactivates a portion of neurons during training, forcing the model to learn more robust, generalized features.

Classification Head

The final layer of Vitamin is a softmax layer that outputs probabilities for each class. In this study, the classes likely corresponded to chest diseases like pneumonia, tuberculosis, and healthy conditions, with the model predicting the condition based on the X-ray.



(a) Proposed Vision Model ViTamin

Figure 3.23: ViTamin Architecture

3.9.3 Training Process of Vitamin on Chest X-ray Dataset

Dataset Preparation

- **Image Resizing:** Chest X-rays were resized to 256x256 pixels to standardize inputs across all models, including Vitamin, and to fit within memory constraints during training.

- **Data Augmentation:** Various augmentation techniques, such as rotation, horizontal flipping, and brightness adjustments, were applied to make the model more robust to real-world variations in medical images.

Training Parameters

- **Optimizer:** The Adam optimizer, well-suited for deep learning models, was used, helping Vitamin converge faster compared to simple stochastic gradient descent (SGD).
- **Learning Rate:** A learning rate of 1e-4 was chosen for balancing training speed and stability, with a scheduler to reduce the learning rate if validation accuracy plateaued.
- **Batch Size:** A batch size of 32 was used, optimizing GPU resources without overwhelming memory limits.

Training Epochs

Vitamin was trained for 50 epochs, but it failed to achieve the high performance seen in models like ViT or Swin Transformer.

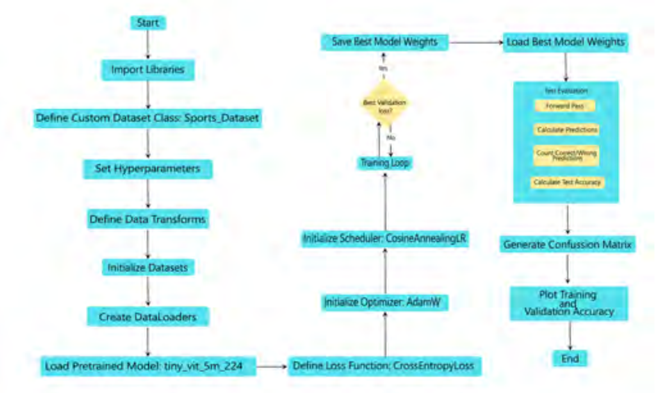

Start
Import Libraries
Define Custom Dataset Class: Sports_Dataset
Set Hyperparameters
Define Data Transforms
Initialize Datasets
Create DataLoaders
Load Pretrained Model: tiny_vit_5m_224
Define Loss Function: CrossEntropyLoss
Initialize Optimizer: AdamW
Initialize Scheduler: CosineAnnealingLR
Training Loop
No
Best Validation loss?
Yes
Save Best Model Weights
Load Best Model Weights
Test Evaluation
Forward Pass
Calculate Predictions
Count Correct/Wrong Predictions
Calculate Test Accuracy
Generate Confussion Matrix
Plot Training and Validation Accuracy
End

Figure 3.24: Workflow for ViTamin

Challenges in Training

Despite extensive training, Vitamin struggled to reach the desired accuracy levels. The model had difficulty distinguishing between disease types, likely contributing to its lower accuracy compared to other models. This could be due to the attention mechanism not being strong enough to compensate for the loss of global information when using convolutional layers.

3.9.4 Performance of Vitamin on Chest Disease Classification

Training Accuracy

Vitamin achieved 86.20% training accuracy after 50 epochs. While this is acceptable for many tasks, it is lower compared to other models tested in this study. This lower accuracy may stem from Vitamin's limited ability to capture both local and global features in X-ray images, which is critical for distinguishing between different chest diseases.

Testing Accuracy

The model's testing accuracy was 82.73%, reflecting its generalization ability. However, this result indicates that the model struggled to generalize to unseen X-ray images, likely due to overfitting and a failure to learn the broader patterns necessary to classify chest diseases correctly across diverse images.

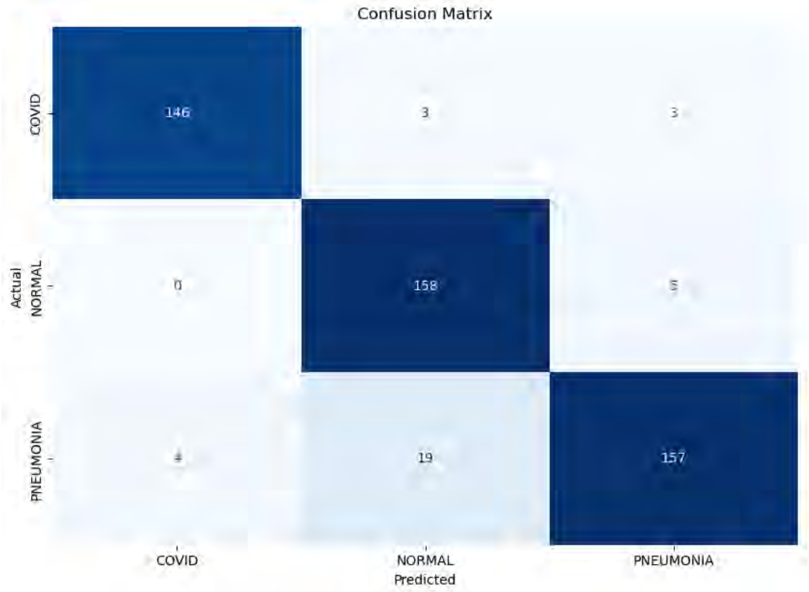


Figure 3.25: Confusion Matrix of ViTamin

3.9.5 Strengths of Vitamin

Integration of Attention Mechanisms

Vitamin incorporated attention mechanisms within a CNN framework for the first time, allowing it to localize critical regions in X-ray images. This helped the model focus on relevant areas, such as potential disease zones, while filtering out irrelevant sections like the background.

Efficient Local Feature Extraction

The CNN-based architecture enabled Vitamin to quickly capture essential local features, important for detecting small abnormalities in X-rays. However, this local focus was insufficient for capturing global features, limiting the model's overall performance.

Computational Efficiency

Vitamin required significantly less computational power and memory compared to transformer models like ViT and Swin Transformer. This made it more suitable for environments with limited resources, such as smaller clinics or hospitals without high-end GPUs.

Flexibility and Simplicity

The simplicity of Vitamin's CNN architecture made it easy to train and fine-tune. Its flexible design allowed for experimentation with various attention mechanisms and regularization techniques, making it adaptable for a wide range of image classification tasks.

3.10 Result Analysis

The Custom CNN model exhibited performance similar to transformer-based models, despite its simpler architecture. Its results were comparable to models like ViT, Swin Transformer, and FocalNet, showing that the CNN architecture was effective at distinguishing between chest diseases in this dataset. The model's strong performance with fewer epochs also suggested that the Custom CNN was computationally efficient.

Model	COVID Precision	COVID Recall	COVID F1-Score	NORMAL Precision	NORMAL Recall	NORMAL F1-Score	PNEUMONIA Precision	PNEUMONIA Recall	PNEUMONIA F1-Score
ViT	0.98	0.98	0.98	0.95	0.95	0.95	0.96	0.97	0.96
Swin	0.99	0.98	0.99	0.95	0.96	0.96	0.97	0.97	0.97
Vitamin	0.93	0.93	0.93	0.78	0.90	0.84	0.88	0.76	0.81
Volo	0.99	0.99	0.99	0.99	0.96	0.97	0.97	0.99	0.98
FocalNet	0.99	0.99	0.99	0.94	0.99	0.96	0.99	0.96	0.97
CNN	0.97	0.97	0.97	0.95	0.95	0.95	0.96	0.97	0.96

Table 3.6: Performance comparison of different models on chest disease classification

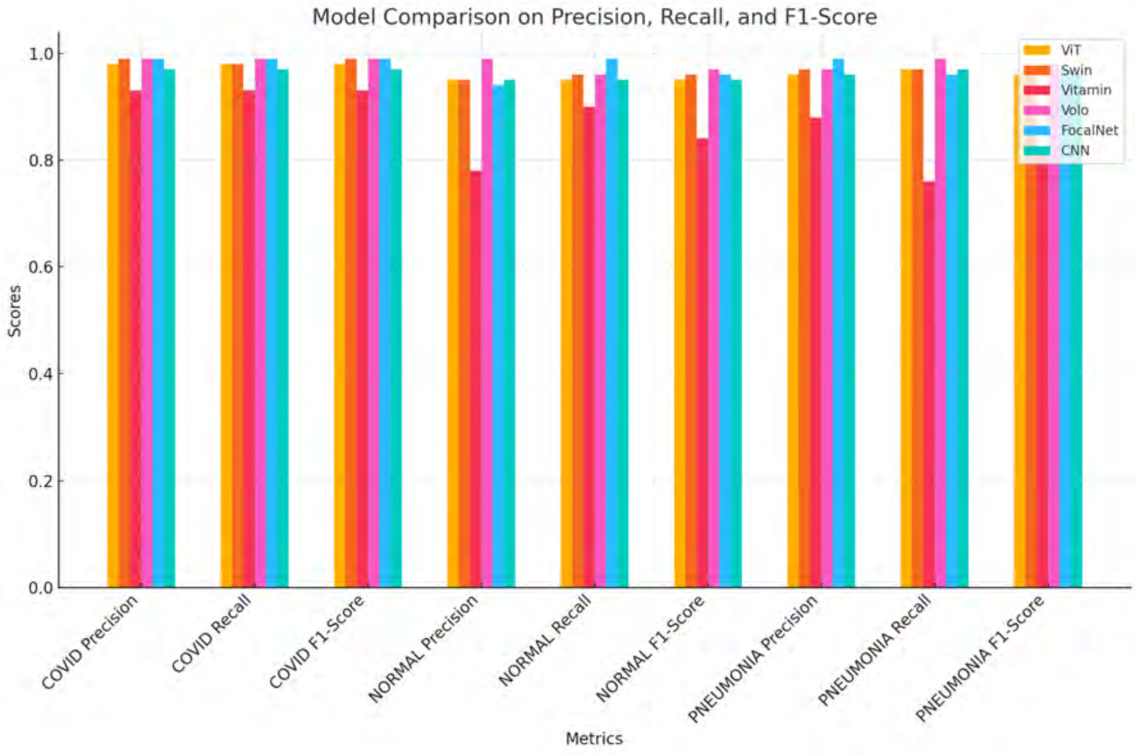


Figure 3.26: Model Comparison

3.10.1 Summary Insights

Top Performers

Volo and FocalNet achieved near-optimal scores in precision, recall, and F1-score, making them highly effective for classifying chest diseases. These models excelled by capturing both local features (e.g., small lesions) and global features (e.g., overall lung structure).

CNN vs Transformers

The Custom CNN performed well compared to transformer-based models, even though it required fewer epochs. However, pre-trained transformers like Swin Transformer and ViT excelled when fine-tuned on larger datasets, especially in recognizing more complex image patterns and fine-tuning tasks.

Lower Performer

Vitamin struggled, particularly in classifying *NORMAL* and *PNEUMONIA* cases. Its lower precision and recall suggest that it was less effective in identifying subtle features in the X-rays, which are critical for differentiating between these diseases.

Pretraining

Models such as ViT, Swin Transformer, and Volo benefited significantly from pretraining on large image datasets. This gave them a strong foundation for transfer learning on the chest X-ray dataset. Fine-tuning allowed these models to adapt to the specific patterns of chest diseases, improving their classification accuracy.

3.10.2 Insight Conclusion

Transformer-based models like Volo, Swin Transformer, and FocalNet demonstrated exceptional performance in classifying chest diseases due to their advanced attention mechanisms and multi-scale feature extraction capabilities. The Custom CNN also performed very well, proving that CNNs remain competitive for medical image classification tasks. On the other hand, Vitamin underperformed, highlighting the importance of model choice when addressing medical imaging challenges.

Chapter 4

Conclusion

Detecting lung diseases at the right time can save someone's life. In Low and Middle-Income Countries, it is extremely difficult to detect these pulmonary diseases due to the socio-economic condition of those countries and the limited number of healthcare professionals. To solve this problem and to achieve satisfactory efficiency, our automated lung disease detection can play a vital role. Our proposed method uses a customized CNN model to differentiate between pneumonia, COVID-19, and normal CXRs. With the help of our model, we can create an effective and error-free network. Disease detection will become effortless and less time-consuming. This model can be implemented in all kind of pulmonary disease detection along with their variants as well stages. For this model to detect more and more pulmonary diseases, the model just have to be trained on the diseases' CXR images. Therefore, it would be possible for this model to detect any disease in near future. Lastly, bypassing our data through our customized CNN model we think we have developed an effective model that is capable of detecting different lung diseases with the highest accuracy and makes health care better for developing countries so that they can easily detect diseases even without the help of health professionals and prevent any untimely death.

Bibliography

[1] D. SS, "Early diagnosis of spinal tuberculosis by mri," *J Bone Joint Surg*, vol. 76, pp. 863–869, 1994.

[2] T. T. Zacharia, J. Shah, D. Patkar, H. Kale, and V. Sindhwani, "Mri in ankle tuberculosis: Review of 14 cases," *Australasian radiology*, vol. 47, no. 1, pp. 11–16, 2003.

[3] R. Kaila, A. M. Malhi, B. Mahmood, and A. Saifuddin, "The incidence of multiple level noncontiguous vertebral tuberculosis detected using whole spine mri," *Clinical Spine Surgery*, vol. 20, no. 1, pp. 78–81, 2007.

[4] E. Busi Rizzi, V. Schinina, M. Cristofaro, *et al.*, "Detection of pulmonary tuberculosis: Comparing mr imaging with hrct," *BMC infectious Diseases*, vol. 11, no. 1, pp. 1–7, 2011.

[5] C. Liu, Y. Cao, M. Alcantara, *et al.*, "Tx-cnn: Detecting tuberculosis in chest x-ray images using convolutional neural network," in *2017 IEEE international conference on image processing (ICIP)*, IEEE, 2017, pp. 2314–2318.

[6] U. Lopes and J. F. Valiati, "Pre-trained convolutional neural networks as feature extractors for tuberculosis detection," *Computers in biology and medicine*, vol. 89, pp. 135–143, 2017.

[7] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2097–2106.

[8] A. Afzali, F. B. Mofrad, and M. Pouladian, "Feature selection for contour-based tuberculosis detection from chest x-ray images," in *2019 26th National and 4th International Iranian Conference on Biomedical Engineering (ICBME)*, 2019, pp. 194–198. DOI: 10.1109/ICBME49163.2019.9030395.

[9] F. Pasa, V. Golkov, F. Pfeiffer, D. Cremers, and D. Pfeiffer, "Efficient deep network architectures for fast chest x-ray tuberculosis screening and visualization," *Scientific reports*, vol. 9, no. 1, p. 6268, 2019.

[10] J. Zeng, Z. Liu, G. Shen, *et al.*, "Mri evaluation of pulmonary lesions and lung tissue changes induced by tuberculosis," *International Journal of Infectious Diseases*, vol. 82, pp. 138–146, 2019.

[11] K. Munadi, K. Muchtar, N. Maulina, and B. Pradhan, "Image enhancement for tuberculosis detection using deep learning," *IEEE Access*, vol. 8, pp. 217 897–217 907, 2020. DOI: 10.1109/ACCESS.2020.3041867.

[12] T. Ozturk, M. Talo, E. A. Yildirim, U. B. Baloglu, O. Yildirim, and U. R. Acharya, "Automated detection of covid-19 cases using deep neural networks with x-ray images," *Computers in biology and medicine*, vol. 121, p. 103 792, 2020.

[13] T. Rahman, A. Khandakar, M. A. Kadir, *et al.*, "Reliable tuberculosis detection using chest x-ray with deep learning, segmentation and visualization," *IEEE Access*, vol. 8, pp. 191 586–191 601, 2020.

[14] S. Rajaraman and S. K. Antani, "Modality-specific deep learning model ensembles toward improving tb detection in chest radiographs," *IEEE Access*, vol. 8, pp. 27 318–27 326, 2020.

[15] D. Verma, C. Bose, N. Tufchi, K. Pant, V. Tripathi, and A. Thapliyal, "An efficient framework for identification of tuberculosis and pneumonia in chest x-ray images using neural network," *Procedia Computer Science*, vol. 171, pp. 217–224, 2020.

[16] D. Buonsenso, D. Pata, E. Visconti, *et al.*, "Chest ct scan for the diagnosis of pediatric pulmonary tb: Radiological findings and its diagnostic significance," *Frontiers in Pediatrics*, vol. 9, p. 583 197, 2021.

[17] M. K. Puttagunta and S. Ravi, "Detection of tuberculosis based on deep learning based methods," in *Journal of Physics: Conference Series*, IOP Publishing, vol. 1767, 2021, p. 012 004.

[18] S. Rajaraman, L. R. Folio, J. Dimperio, P. O. Alderson, and S. K. Antani, "Improved semantic segmentation of tuberculosis—consistent findings in chest x-rays using augmented training of modality-specific u-net models with weak localizations," *Diagnostics*, vol. 11, no. 4, p. 616, 2021.

[19] S. Tripathi, S. Shetty, S. Jain, and V. Sharma, "Lung disease detection using deep learning," *Int. J. Innov. Technol. Explor. Eng*, vol. 10, no. 8, 2021.

[20] S. S. Alahmari, B. Altazi, J. Hwang, S. Hawkins, and T. Salem, "A comprehensive review of deep learning-based methods for covid-19 detection using chest x-ray images," *Ieee Access*, 2022.

[21] M. Bhandari, T. B. Shahi, B. Siku, and A. Neupane, "Explanatory classification of cxr images into covid-19, pneumonia and tuberculosis using deep learning and xai," *Computers in Biology and Medicine*, vol. 150, p. 106 156, 2022.

[22] Y. Chen, J. Feng, J. Liu, B. Pang, D. Cao, and C. Li, "Detection and classification of lung cancer cells using SWIN Transformer," *Journal of Cancer Therapy*, vol. 13, no. 07, pp. 464–475, Jan. 2022. DOI: 10.4236/jct.2022.137041. [Online]. Available: https://doi.org/10.4236/jct.2022.137041.

[23] I. Haq, T. Mazhar, Q. Nasir, *et al.*, "Machine vision approach for diagnosing tuberculosis (tb) based on computerized tomography (ct) scan images," *Symmetry*, vol. 14, no. 10, p. 1997, 2022.

[24] M. E. Khokhar, N. H. Khan, S. Bilal, and S. Ameer, "The diagnostic accuracy of magnetic resonance imaging (mri) for detection of spinal tuberculosis (tb)," *Journal of Islamabad Medical & Dental College*, vol. 11, no. 2, pp. 110–113, 2022.

[25] S. Kumar, S. Shastri, S. Mahajan, *et al.*, “LiteCovidNet: A lightweight deep neural network model for detection of COVID-19 using X-ray images,” *International Journal of Imaging Systems and Technology*, vol. 32, no. 5, pp. 1464–1480, Jun. 2022. DOI: 10.1002/ima.22770. [Online]. Available: https://doi.org/10.1002/ima.22770.

[26] K. Santosh, S. Allu, S. Rajaraman, and S. Antani, “Advances in deep learning for tuberculosis screening using chest x-rays: The last 5 years review,” *Journal of Medical Systems*, vol. 46, no. 11, p. 82, 2022.

[27] S. Shastri, I. Kansal, S. Kumar, K. Singh, R. Popli, and V. Mansotra, “CheXImageNet: a novel architecture for accurate classification of Covid-19 with chest x-ray digital images using deep convolutional neural networks,” *Health and Technology*, vol. 12, no. 1, pp. 193–204, Jan. 2022. DOI: 10.1007/s12553-021-00630-x. [Online]. Available: https://doi.org/10.1007/s12553-021-00630-x.

[28] S. Urooj, S. Suchitra, L. Krishnasamy, N. Sharma, and N. Pathak, “Stochastic learning-based artificial neural network model for an automatic tuberculosis detection system using chest x-ray images,” *IEEE Access*, vol. 10, pp. 103 632–103 643, 2022. DOI: 10.1109/ACCESS.2022.3208882.

[29] L. Venkataramana, D. V. V. Prasad, S. Saraswathi, C. Mithumary, R. Karthikeyan, and N. Monika, “Classification of covid-19 from tuberculosis and pneumonia using deep learning techniques,” *Medical & Biological Engineering & Computing*, vol. 60, no. 9, pp. 2681–2691, 2022.

[30] L. Yuan, Q. Hou, Z. Jiang, J. Feng, and S. Yan, “VOLO: Vision Outlooker for Visual Recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–13, Jan. 2022. DOI: 10.1109/tpami.2022.3206108. [Online]. Available: https://doi.org/10.1109/tpami.2022.3206108.

[31] A. Y. Yusuf and Y. Hagos, “Spinal tuberculosis: Mri findings in a case series of 35 patients,” *GSC Advanced Research and Reviews*, vol. 13, no. 2, pp. 098–102, 2022.

[32] G. M. M. Alshmrani, Q. Ni, R. Jiang, H. Pervaiz, and N. M. Elshennawy, “A deep learning architecture for multi-class lung diseases classification using chest x-ray (cxr) images,” *Alexandria Engineering Journal*, vol. 64, pp. 923–935, 2023.

[33] A. L. Association, *Our impact*, Sep. 2023. [Online]. Available: https://www.lung.org/about-us/our-impact.

[34] T. Gulsoy and E. B. Kablan, “FocalNeXt: A ConvNeXt augmented FocalNet architecture for lung cancer classification from CT-scan images,” *Expert Systems with Applications*, p. 125 553, Oct. 2024. DOI: 10.1016/j.eswa.2024.125553. [Online]. Available: https://doi.org/10.1016/j.eswa.2024.125553.

[35] J. Ko, S. Park, and H. G. Woo, “Optimization of vision transformer-based detection of lung diseases from chest X-ray images,” *BMC Medical Informatics and Decision Making*, vol. 24, no. 1, Jul. 2024. DOI: 10.1186/s12911-024-02591-3. [Online]. Available: https://doi.org/10.1186/s12911-024-02591-3.