AI-Driven Pneumonia Diagnosis: Detection and Cause Classification

KVNS Vishnu Vardhan, Chaitanya Bharathi Institute of Technology (A), Hyderabad

Shlok Agarwal, Chaitanya Bharathi Institute of Technology(A), Hyderabad

1 Introduction

1.1 Background

Pneumonia is a severe lung infection that has a big hand in global mortality and morbidity. It can be caused by bacterial, viral or fungal infection, while bacterial and viral causes are the most common types. Chest X-ray images are the primary diagnostic process of pneumonia, but radiologically identifying the infection requires an expert. We propose an automated deep-learning based architecture that can assist radiologists to diagnose and identify the cause of the infection with good accuracy.

1.2 Problem Statement

Traditional diagnosis is not effective in differentiating between viral pneumonia and bacterial pneumonia just using radiological data. This makes it challenging to deal with anti-biotic misuse and misdiagnosis of cause for the infection. An AI based model can assist in medical decision-making process to make the process less tedious.

1.3 Objective

The objective of this study is to develop a deep learning model for classifying chest X-rays into normal, bacterial pneumonia, and viral pneumonia. It aims to enhance diagnostic accuracy, assist radiologists, and reduce antibiotic misuse.

1.4 Scope of study

This study utilizes a publicly available pneumonia dataset consisting of labeled chest X-ray images. A Convolutional Neural Network (CNN) is implemented to classify images into three categories: normal, bacterial pneumonia, and viral pneumonia. The model is trained and evaluated using various deep learning techniques with performance metrics such as accuracy, precision, recall, and F1-score.

2 Literature Study

2.1 Existing Methods for Pneumonia Diagnosis

Traditionally, pneumonia diagnosis involves chest X-ray imaging, blood tests, and sputum analysis. While effective, these methods are time-consuming and require specialized

medical equipment, which may not always be available in resource-limited settings. This delay in diagnosis can impact treatment effectiveness, especially in critical cases.

2.2 Gaps in Existing Research Using Deep Neural Networks

Deep Neural Networks can be a promising tool for such automation tasks, but existing research has several limitations:

- Focus on Early Detection: Most studies focus on early detection but do not differentiate between bacterial and viral pneumonia, this can help in treating critical cases, preventing drug resistance, and assessing contagiousness.
- Larger networks: Most studies have used model like Resnet-50, DenseNet and U-Net like architectures which are very computationally extensive for applications where low-resource environments are present.

With our trainable parameters of only about 10 million, our ensemble model is significantly lighter than existing approaches, making it easier to train and deploy.

3 Methodology

3.1 Dataset Description

The dataset used in this study was obtained from Guangzhou Women and Children's Medical Centre, as cited in [1]. It consists of JPEG images with a total size of approximately 2GB. The training set contains 5,232 images, including 1,349 images of normal lungs, 2,538 images of bacterial pneumonia, and 1,345 images of viral pneumonia.

3.2 Preprocessing

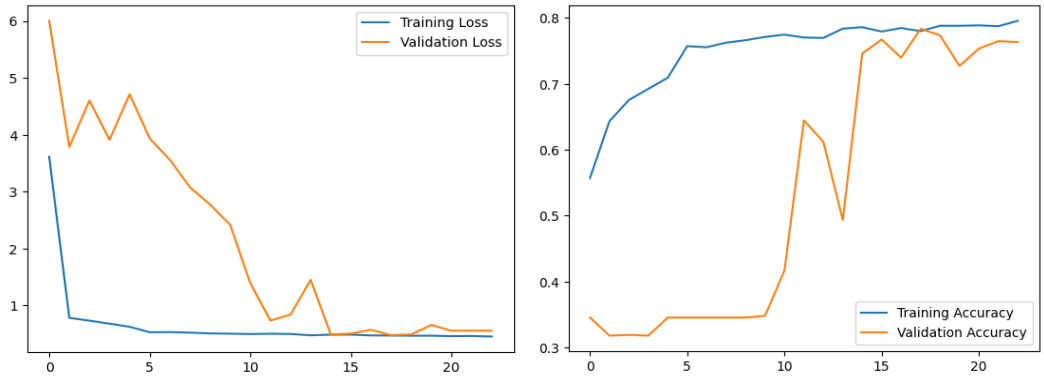
Preprocessing was performed using OpenCV, NumPy, and Pandas to standardize the dataset. All images were resized to 244×244 pixels, and grayscale images were converted to 3-channel images by stacking them along three channels. To handle class imbalance, we tested two approaches. The first method involved augmenting the minority classes (normal and viral pneumonia) and undersampling the bacterial pneumonia class to balance the dataset. Since the imbalance was significant, pure upsampling was avoided to prevent overfitting. In the second approach, we only undersampled the bacterial pneumonia class, as the other two classes had nearly equal samples. After testing both methods, we observed similar results, indicating that either approach could be used without significantly affecting model performance. Finally, the training data was split into 80% training and 20% validation, while the test set remained unchanged as it was already predefined.

3.3 Model Developments and Experiments

First model: Custom CNN Architecture

We started off with a custom network where the weights of the convolution layers are initialised using the ones of VGG16. The model consisted of five convolutional blocks, each followed by batch normalization and max pooling layers. The convolutional layers used filter sizes ranging from 32 to 256, allowing the network to progressively learn

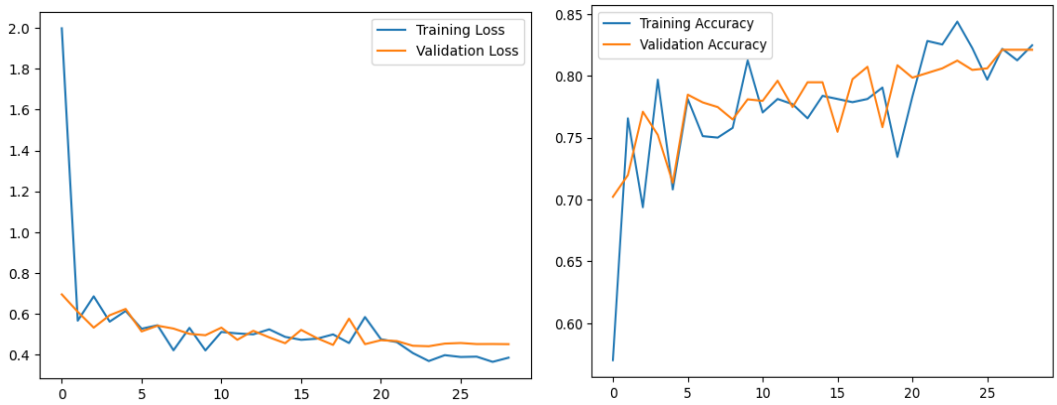
hierarchical features from chest X-ray images. After each convolution we applied batch normalisation to make the training more stable, while max pooling helped in reducing spatial dimensions and retaining essential features. To prevent overfitting, we added additional dropout layers. The dense layers had ReLU activation function, followed by softmax in the output layer. During training, the model was compiled along with the RMSprop optimiser and categorical cross entropy loss. Additionally, to tackle overfitting , we implemented Early Stopping and Learning Rate optimisers like ReduceLROnPlateau. This model served as the baseline for our subsequent prototypes. The statistical metrics for its performance are as follows:




```
The test loss is 0.4740603566169739,the test accuracy is 0.8002002835273743
```

Augmenting all the classes

As the initial model was mildly overfitting and the accuracy was not satisfactory, we went back to our preprocessing and augmented the images of all classes, to prevent overfitting . This gives the model more perspectives of an image, to train better. This reduced the mild overfitting, but there was no significant increase in accuracy. The approach also resulted in a much smoother convergence , but the metrics were still not satisfactory. The performance scores are as follows:

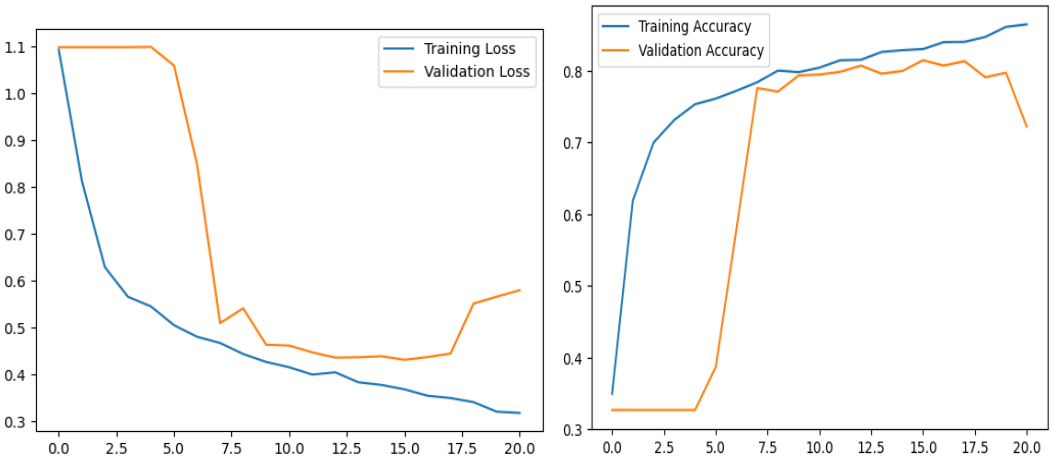



```
The test loss is 0.40523025393486023,the test accuracy is 0.813470184803009
```

Since the graph looked very pleasing, we tried the same experiment by neglecting the early stopping and increased the number of training epochs and unfreezing the vgg16 layers, but it caused the model to overfit. Hence we discarded the approach.

Second model : VGG16 with Depth wise convolution

This time we changed our approach and tested with Separable Convolution layers with batch normalisation. Initially, all VGG16 layers were frozen, except for block5_conv3, which was set to be trainable to allow learning of high-level abstract features. In this network, we start with 2 Conv2D layers then multiple SeparableConv2D layers to reduce computational complexity. max pooling was used to downsample feature while after each layer we also applied batch normalisation. For training, we used the Adam optimizer with a learning rate of 0.0001 and categorical cross-entropy as the loss function, while keeping all the other hyperparameters almost the same as before. Though Depthwise convolution is a computationally efficient approach, the parameter space added up to almost 400 MB which makes it difficult in implement in realtime. This approach combines transfer learning, depthwise separable convolutions, and fine-tuning . The accuracy metrics are as follows:

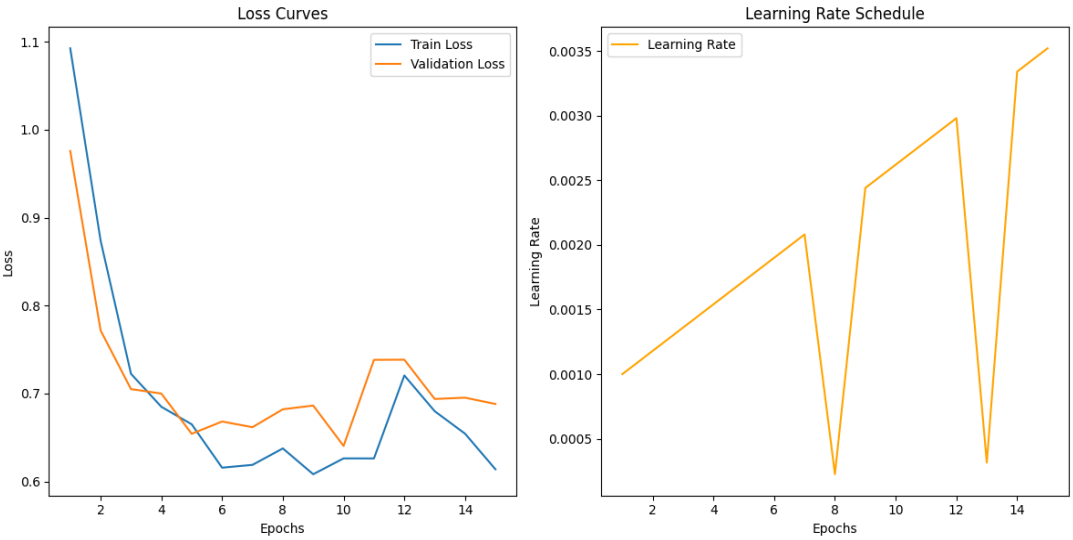


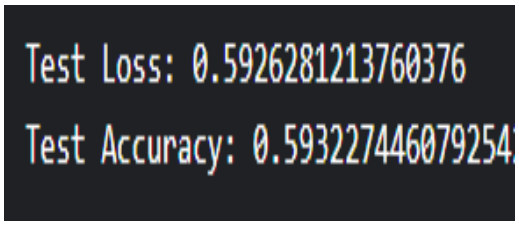

```
The test loss is 0.5210524797439575,the test accuracy is 0.756384551525116
```

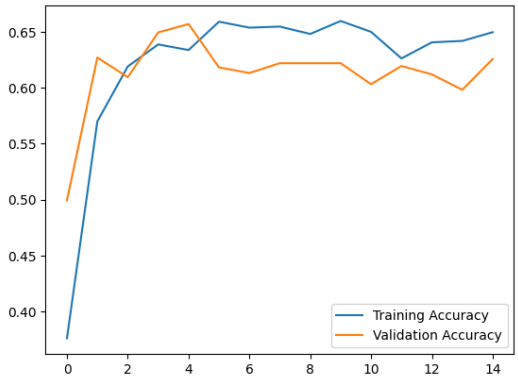
The above graph shows that the convergence is not very smooth and also the testing accuracy is not satisfactory.

Third model : Inception-Based CNN with Auxiliary Classifiers

We developed a custom CNN architecture inspired by the Inception model, which parallelly processes inputs with various kernel sizes to capture more complex features. We start with convolutional and max-pooling layers for initial feature extraction, followed by stacked Inception modules that apply 1×1, 3×3, and 5×5 convolutions along with max pooling. To address risk of vanishing gradients, we used two auxiliary classifiers to assist in training. Additionally, we also introduced One Cycle scheduler to adjust learning rate dynamically. This architecture effectively balances depth, computational efficiency, and robust feature extraction, making it a powerful alternative to standard CNNs . The accuracy metrics and the learning rate scheduler metrics are as follows:



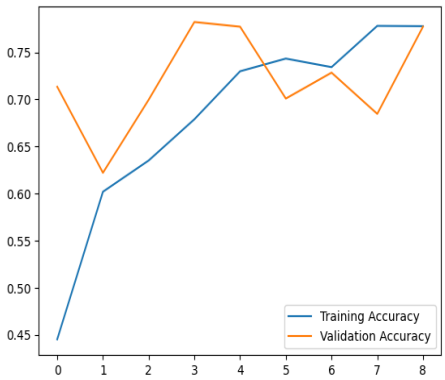




The model metrics showed that the testing accuracy was not upto the mark and the model had already started to overfit around 15 epochs.

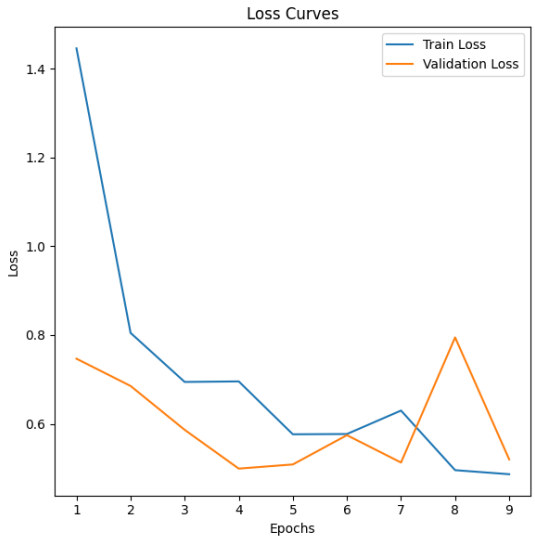
Ensemble method: Ensemble of VGG19 and inception

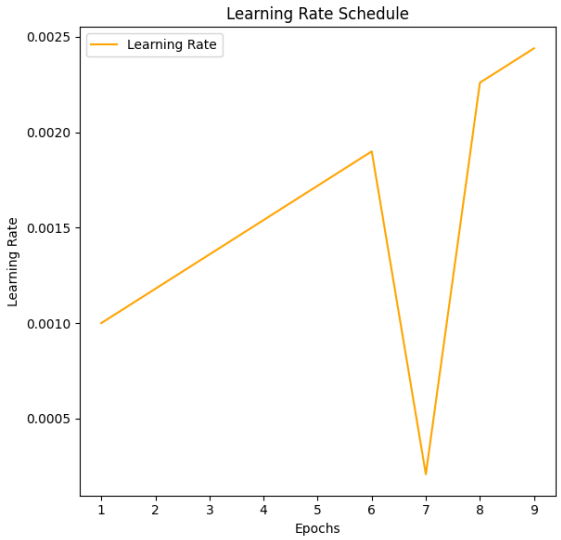
We used a hybrid architecture combining VGG19 and an Inception-based network, where VGG19 was used for feature extraction and the extracted features were then passed through Inception modules, which process the input at multiple scales using 1×1, 3×3, and 5×5 convolutions alongside max pooling. Along with the fully connected layers (512 and 256 neurons) followed by dropout (0.3 and 0.2) to prevent overfitting, and a softmax layer for classification into three categories. or optimization, we used RMSprop with a learning rate of 0.000001, and a one-cycle learning rate scheduler to dynamically adjust learning rate. While VGG19 alone and Inception based architectures alone produced unsatisfactory results, their ensemble produced rather better outputs than a few other architectures.



	precision	recall	f1-score	support
0	0.94	0.96	0.95	1349
1	0.66	0.83	0.73	1300
2	0.76	0.56	0.65	1345
accuracy			0.78	3994
macro avg	0.79	0.78	0.78	3994
weighted avg	0.79	0.78	0.78	3994

Accuracy: 0.7822





The model achieved an accuracy of 78.22%, with the highest performance in detecting normal cases (precision: 0.94, recall: 0.96, F1-score: 0.95). Bacterial pneumonia had a recall of 0.83, indicating good sensitivity, but a lower precision of 0.66, suggesting some misclassifications. Viral pneumonia showed weaker performance (precision: 0.76, recall: 0.56, F1-score: 0.65), indicating difficulty in distinguishing it from other classes. The macro and weighted averages (0.78 F1-score) confirm balanced overall performance, but class-wise disparities suggest room for improvement.

Final model: Two-Phase fine-tuning of the ensemble model

we implemented a two-phase training strategy using VGG19 as a feature extractor combined with Inception modules for multi-scale feature extraction. The intuition is that Phase 1 takes larger steps toward convergence by training only the Inception layers, leveraging frozen VGG19 features. In Phase 2, once nearing convergence, the top VGG19 layers are unfrozen, and SGD with momentum is used for smaller, precise steps, ensuring refined learning without overfitting.

Phase 1: Feature Extraction

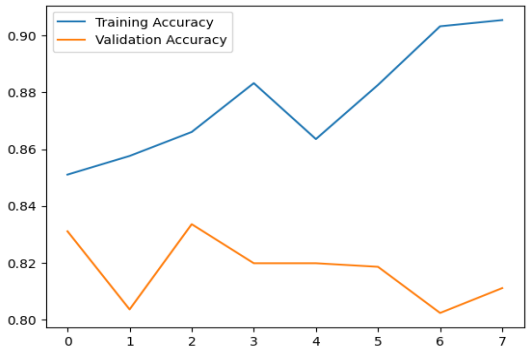
First we freeze all layers of VGG19 and only allow the inception layers for training, this will continue for 100 epochs with an RMSprop optimiser. Early stopping and a one-cycle learning rate scheduler were applied to optimize convergence. Once we are about to approach the convergence, the 1st phase of training is stopped.

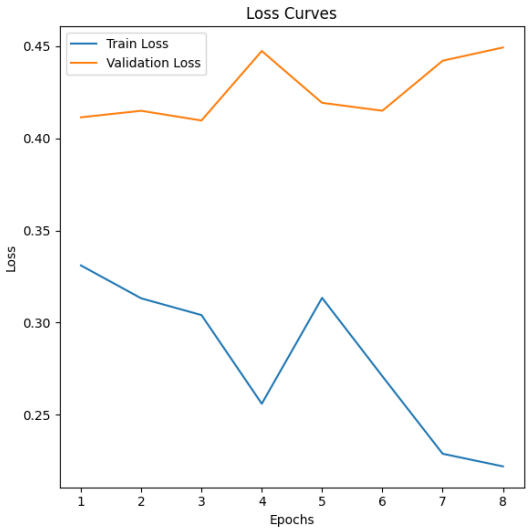
Phase 2: Fine-Tuning

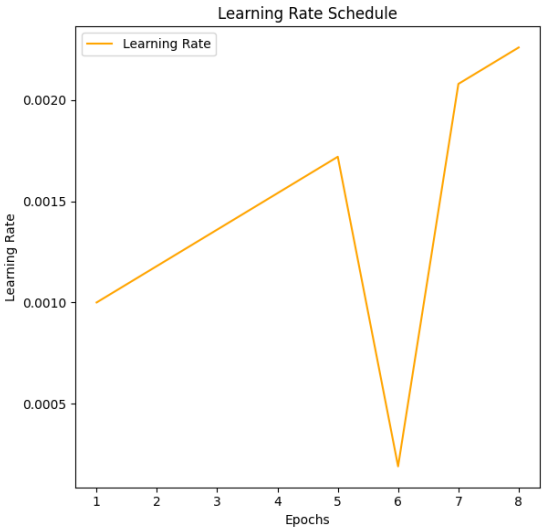
After the feature extraction, the top layers of VGG19 are unfrozen, this time we switch to SGD optimiser. The optimizer is switched to SGD with momentum and Nesterov acceleration, which takes smaller, more precise steps towards convergence. This phase runs for 15 epochs, ensuring fine-tuning without overfitting. The combination of large-step learning in Phase 1 and small-step refinement in Phase 2 helps the model achieve optimal performance. The performance statistics after the 2nd phase are as follows:

	precision	recall	f1-score	support
0	0.99	1.00	1.00	1349
1	0.85	0.80	0.82	1300
2	0.82	0.86	0.84	1345
accuracy			0.89	3994
macro avg	0.89	0.89	0.89	3994
weighted avg	0.89	0.89	0.89	3994

Accuracy: 0.897







4 Results

The ensemble model achieved an overall accuracy of 89.71%, showing a significant improvement over previous iterations. Normal cases were classified with near-perfect precision (0.99) and recall (1.00), leading to an F1-score of 1.00. Bacterial pneumonia had an F1-score of 0.82, indicating good classification but some overlap with viral pneumonia. Viral pneumonia had a higher recall (0.86) than bacterial pneumonia (0.80), meaning the model was better at detecting viral cases but had slightly lower precision. The macro and weighted F1-scores (0.89) confirm balanced overall performance. For binary classification (grouping bacterial and viral pneumonia as "abnormal" and normal lungs as "normal"), the model performed exceptionally well, achieving a precision of 0.92, recall of 0.96, and an F1-score of 0.94, making it highly effective in distinguishing between healthy and pneumonia-affected lungs.

Metric	Value
Precision	0.92
Recall	0.96
F1-score	0.94

Table 1: (Normal vs. Pneumonia)

Class	Precision	Recall	F1-score
Normal (0)	0.99	1.00	1.00
Bacterial (1)	0.85	0.80	0.82
Viral (2)	0.82	0.86	0.84
Overall Accuracy		0.8971	
Macro Average	0.89	0.89	0.89
Weighted Average	0.89	0.89	0.89

Table 2: Three-Class Classification Metrics

As part of future scope of this project, we believe that exploring the use of transformer models like vision transformers could potentially improve classification performance by capturing long-range dependencies and richer spatial features, but it with a certain level of compromise with the computational efficiency.

5 References

1. Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images for Classification
2. A Deep Learning Approach Considering Image Background for Pneumonia Identification Using Explainable AI (XAI), IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS, VOL. 21, NO. 4, JULY/AUGUST 2024
3. M3 Lung-Sys: A Deep Learning System for Multi-Class Lung Pneumonia Screening from CT Imaging, IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS, VOL. 24, NO. 12, DECEMBER 2020
4. ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health
5. National Institutes of Health Chest X-Ray Dataset
6. Deep Supervised Domain Adaptation for Pneumonia Diagnosis From Chest X-Ray Images, IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS, VOL. 26, NO. 3, MARCH 2022