AIML

Capstone Project

Group 2 - CV 1 - Pneumonia Detection

Group Members:

Malcolm Nicholas Monserrate

Minatchi S

Navin Kumar Patro

Shiva Chari

Soumya Ranjan Behera

Sunidhi Dixit

Contents

[1. Summary of problem statement, data and findings 3](#_Toc199685099)

[ Project Objective 3](#_Toc199685100)

[ Problem Statement 3](#_Toc199685101)

[ Dataset 4](#_Toc199685102)

[ Key Findings 5](#_Toc199685103)

[2. Summary of the Approach to EDA and Pre-processing 6](#_Toc199685104)

[ Exploratory Data Analysis (EDA): 6](#_Toc199685105)

[ Pre-processing Steps 8](#_Toc199685106)

[3. Deciding Models and Model Building 10](#_Toc199685107)

[ Basic CNN model: 10](#_Toc199685108)

[ Deeper CNN model: 12](#_Toc199685109)

[ MobileNetV2 (Pre-trained Model): 14](#_Toc199685110)

[ LeNet-5 Model: 15](#_Toc199685111)

[4. How to improve your model performance? 17](#_Toc199685112)

[ Data Preprocessing 17](#_Toc199685113)

[ Feature Selection / Engineering 17](#_Toc199685114)

[ Model Improvements 17](#_Toc199685115)

[ Training Enhancements 17](#_Toc199685116)

[ Evaluation Metrics 17](#_Toc199685117)

# Summary of problem statement, data and findings

## Project Objective

The primary objective of the Pneumonia Detection project is to develop an efficient and reliable system that utilizes deep learning algorithms to automatically detect and classify pneumonia from chest X-ray images. This system aims to:

1. Achieve a high level of accuracy in distinguishing between pneumonia-infected and healthy lung images, thereby supporting timely and accurate diagnosis for improved patient outcomes.
2. Ensure the scalability of the system to handle large volumes of medical imaging data, making it suitable for integration into clinical workflows.
3. Contribute to the broader field of medical diagnostics by showcasing the practical applicability of machine learning models in real-world healthcare settings, ultimately aiding in early intervention and reducing morbidity and mortality associated with pneumonia.

By fulfilling these objectives, the project seeks to provide a valuable tool for healthcare professionals, enabling prompt and accurate pneumonia detection and facilitating better patient care.

## Problem Statement

Domain: Healthcare

Pneumonia is a serious respiratory condition that requires rapid and accurate diagnosis. Manual analysis of chest X-ray images can be time-consuming and error-prone, especially in resource-limited settings. Our objective is to build a deep learning-based system that can **classify chest X-ray images into three categories**:

* **Normal**
* **Lung Opacity (Pneumonia detected)**
* **No Lung Opacity / Not Normal**

**A screenshot of a computer

AI-generated content may be incorrect.**

**A screenshot of a computer error

AI-generated content may be incorrect.**

## Dataset

The dataset used is derived from the **RSNA Pneumonia Detection Challenge**, which includes:

* Chest X-ray DICOM images
* Bounding box annotations for pneumonia regions
* Label metadata indicating the class (Normal, Lung Opacity, or No Lung Opacity)

A screenshot of a computer

AI-generated content may be incorrect.

## Key Findings

* Majority of lung opacity cases have clearly distinguishable patterns in the radiographs.
* Normal and “No Lung Opacity / Not Normal” classes are harder to distinguish without localization.
* There is a class imbalance with fewer "Normal" samples compared to "Lung Opacity".
* **Statistical Summary**

A screenshot of a computer

AI-generated content may be incorrect.

**patientId**

There are 26,684 unique patient IDs among 30,227 entries, indicating some patients have multiple associated records (possibly multiple bounding boxes or findings per patient).

The most frequent patientId appears 4 times.

**Bounding Box Coordinates (x, y, width, height)**

Only 9,555 rows have non-null values for x, y, width, and height, suggesting bounding boxes are present only for a subset of the data (likely the positive cases where pneumonia is detected).

The statistics for these columns:

x: Mean ≈ 394, Std ≈ 205, Range: 2 to 835

y: Mean ≈ 367, Std ≈ 149, Range: 2 to 881

width: Mean ≈ 218, Std ≈ 59, Range: 40 to 528

height: Mean ≈ 329, Std ≈ 158, Range: 45 to 942

This indicates bounding boxes vary significantly in size and position, reflecting the variable presentation of pneumonia in chest X-rays

**Target**

All 30,227 rows have a Target value (no missing data).

Target is likely binary (0 or 1), with a mean of 0.316. This suggests that about 31.6% of the cases are positive for pneumonia, and the rest are negative.

The minimum is 0, and the maximum is 1, confirming binary labeling.

**class**

There are 3 unique classes, with the most common being "No Lung Opacity / Not Normal" (11,821 occurrences).

The class distribution is imbalanced, with a significant portion of the data labeled as "No Lung Opacity / Not Normal".

# Summary of the Approach to EDA and Pre-processing

## Exploratory Data Analysis (EDA):

* **Image sample visualization:** Random chest X-rays across each class were visualized.

A collage of x-ray images of a chest

AI-generated content may be incorrect.

* **Class distribution:** Following plot confirmed **imbalance**, with “Normal” being the most frequent class.

A graph with numbers and a bar

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect..

* **Bounding box analysis:** For Lung Opacity images, multiple bounding boxes are often present, indicating severity.

A collage of x-ray images of a chest

AI-generated content may be incorrect.

* **Size distribution of bounding boxes:** Heatmaps showed that pneumonia regions are generally centralized in the lungs.

## Pre-processing Steps

* **DICOM conversion:** DICOM images were converted to pixel arrays using pydicom.

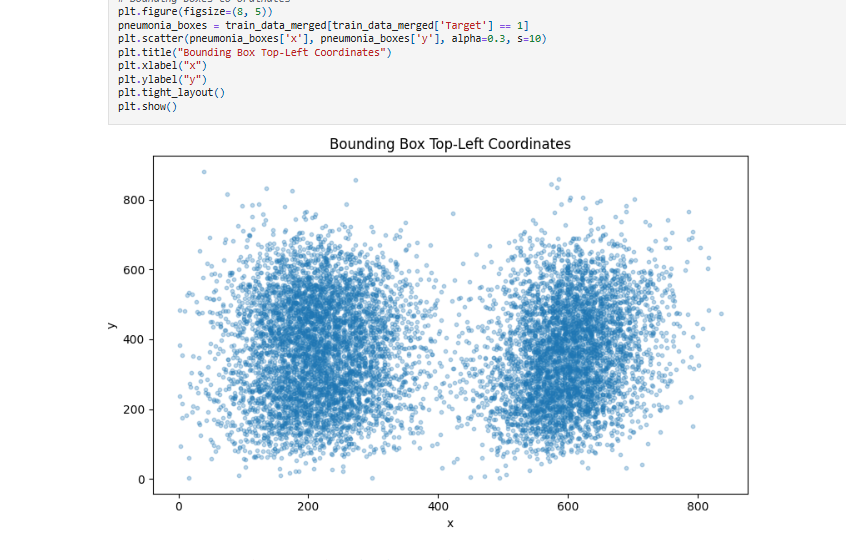
A screenshot of a computer code

AI-generated content may be incorrect.

* **Resizing:** Images resized to 224x224 for uniform input to the CNN model.
* **Normalization:** Applied per-channel mean and standard deviation normalization.
* **Augmentation:** To combat class imbalance and improve generalization, applied transformations like:
  + Random horizontal flips
  + Rotation ±10°
  + Zoom and cropping

**Meaningful Features Identified:**

* Pneumonia presence is visually detectable in dense and irregular lung patterns.
* In “Lung Opacity” images, lesion density and location in lungs (mostly lower lobes) are common.



* The pneumonia annotations are most densely concentrated in two symmetric regions, corresponding to the central areas of the left and right lungs. Darker red areas indicate higher annotation frequency.

A screenshot of a computer screen

AI-generated content may be incorrect.

# Deciding Models and Model Building

We developed four models with distinct CNN architectures to evaluate the feasibility of multiclass classification, including one model utilizing MobileNet with transfer learning.

## Basic CNN model:

* **Architecture:** A simple Convolutional Neural Network with fewer layers.

A screenshot of a computer

AI-generated content may be incorrect.

* **Purpose:** Acts as a baseline to measure the performance of more advanced models.
* **Performance:**
  + **Accuracy:** ~67.02%

A screenshot of a computer

AI-generated content may be incorrect.

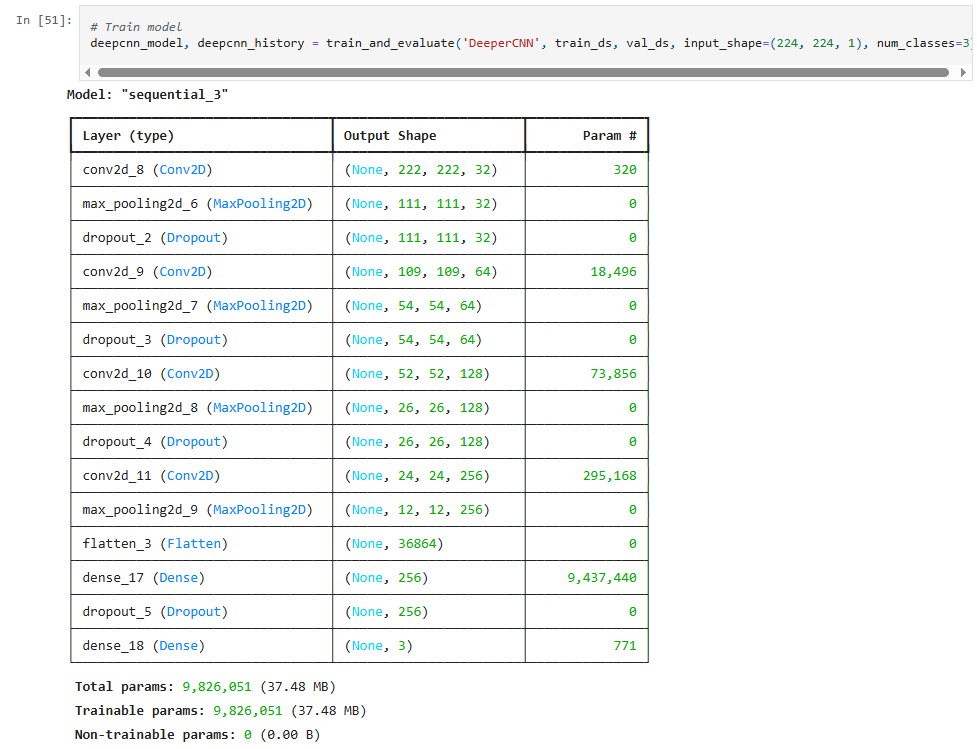
* + **Observations:** Balanced performance; particularly strong in identifying the **Normal** class (F1 = 0.7396).
  + **Limitation:** Struggles slightly with “No Lung Opacity / Not Normal” (F1 = 0.6041).

A screenshot of a computer

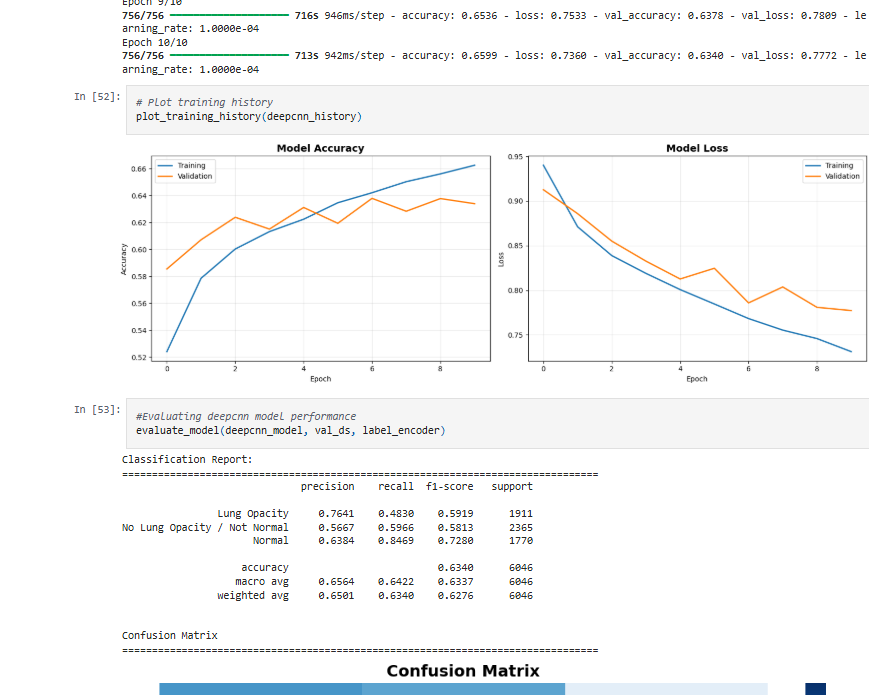
AI-generated content may be incorrect.

## Deeper CNN model:

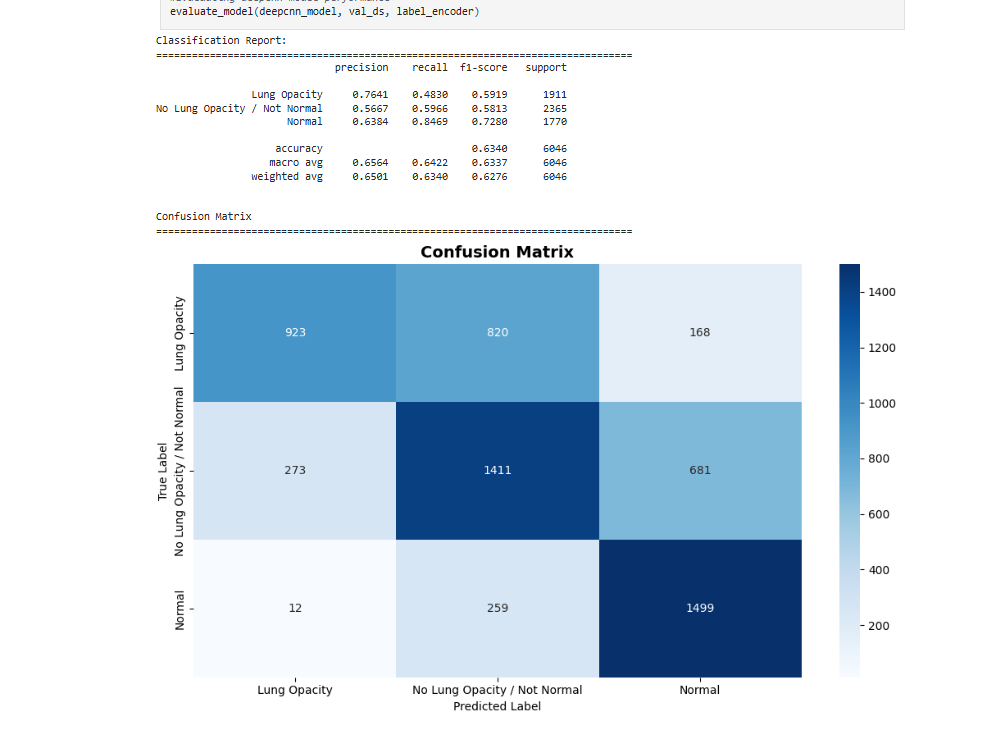
* **Architecture:** More convolutional layers and parameters to learn deeper features.



* **Performance:**
  + **Accuracy:** ~63.40%



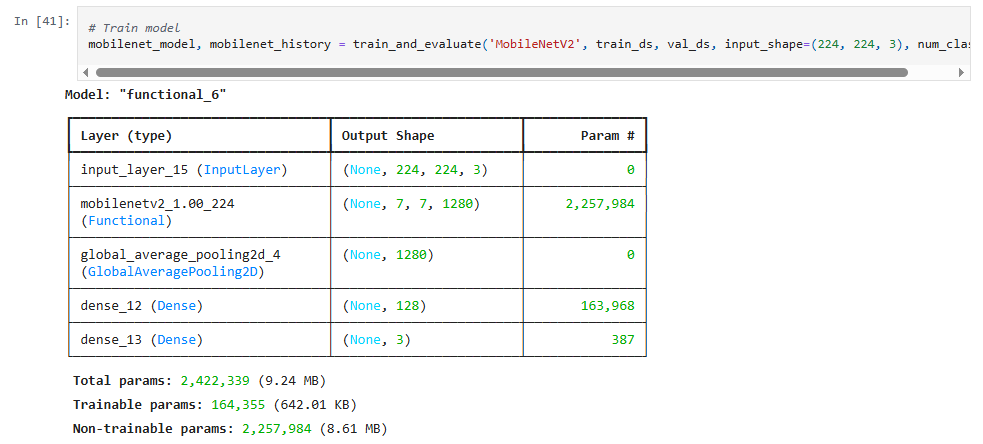
* + **Observations:**
    - Best at detecting **No Lung Opacity / Not Normal** (Recall = 0.5966).
    - Slightly underperforms in overall accuracy compared to BasicCNN and MobileNetV2.



* + **Limitation:** Sacrifices precision on Lung Opacity to boost recall.

## MobileNetV2 (Pre-trained Model):

* **Architecture:** Lightweight, pre-trained on ImageNet.



* **Purpose:** Uses transfer learning to achieve better generalization with less training data.
* **Performance:**
  + **Accuracy:** ~68.97% **(best overall)**

A screenshot of a computer

AI-generated content may be incorrect.

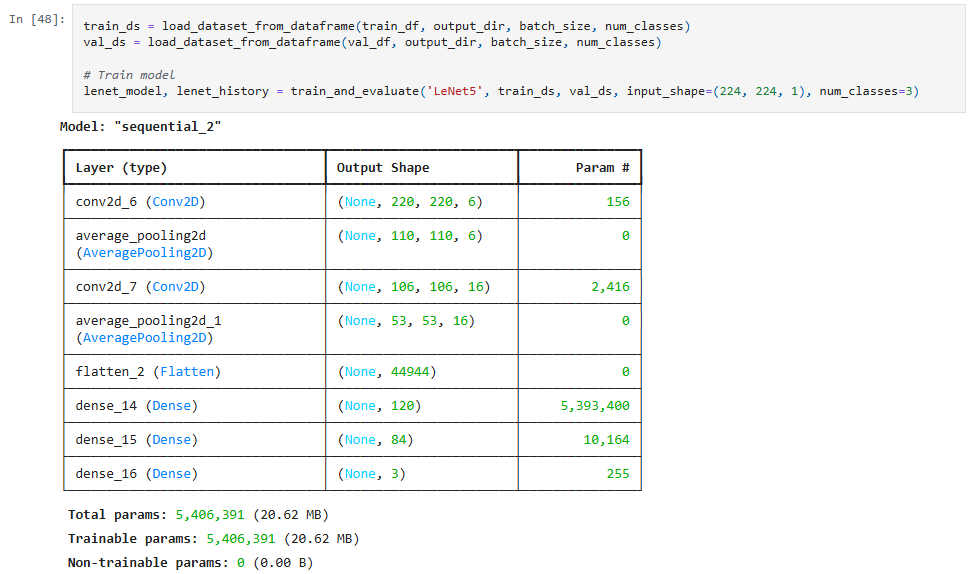
* + **Observations:**
    - Training accuracy improved steadily, reaching about 0.73, while validation accuracy peaked around 0.69 by the final epoch.
    - Training loss decreased consistently, while validation loss plateaued after an initial drop, indicating no major overfitting.
    - Highest overall F1-score (0.6895), particularly strong in classifying **Normal** cases (F1 = 0.7791).
    - Transfer learning helps in extracting better features.

A blue squares with black text

AI-generated content may be incorrect.

## LeNet-5 Model:

* **Architecture:** Classic CNN with limited depth.



* **Performance:**
  + **Accuracy:** ~62.39%

A screenshot of a computer

AI-generated content may be incorrect.

* + **Observations:** Performs relatively well for **Normal** (F1 = 0.6906), but weak for “No Lung Opacity / Not Normal” (F1 = 0.5711).

A screenshot of a computer

AI-generated content may be incorrect.

* + **Limitation:** Older architecture not well suited for complex classification tasks like chest X-ray interpretation.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision (Macro Avg)** | **Recall (Macro Avg)** | **F1-Score (Macro Avg)** |
| **Basic CNN** | 67.02% | 67.61% | 67.80% | 67.64% |
| **MobileNet** | **68.97%** | **69.95%** | **69.54%** | **69.62%** |
| **LeNet-5** | 62.39% | 63.80% | 62.74% | 63.09% |
| **DeepCNN** | 63.40% | 65.64% | 64.42% | 63.37% |

|  |  |
| --- | --- |
| **Model** | **Observation** |
| **Basic CNN** | Reliable performance in predicting **Lung Opacity** cases (1265/1760). |
| **MobileNet** | **Lung Opacity** class identification (1215/1767) and overall better balance across all classes. Less confusion across all axes. |
| **LeNet-5** | Lower performance on **No Lung Opacity**, indicating confusion with both other classes. |
| **DeepCNN** | Better at detecting **No Lung Opacity / Not Normal**, slightly weaker on **Lung Opacity** compared to MobileNet. |

# How to improve your model performance?

## Data Preprocessing

* Apply **data augmentation** (flips, rotations, brightness, etc.) to improve generalization.
* Ensure **pixel normalization** (0–1 range).
* Check **class imbalance**: Consider oversampling or using **Focal Loss** for better minority class performance.

## Feature Selection / Engineering

* Not directly applicable for raw images, but:
  + Try **autoencoders or PCA** for dimensionality reduction.
  + Combine CNN features with **metadata** (e.g., age, gender) if available.

## Model Improvements

* **Fine-Tune MobileNetV2 Further**
  + Unfreeze deeper layers and retrain using a **lower learning rate**.
* **Use Ensemble Models**
  + Combine predictions from **BasicCNN**, **DeeperCNN**, and **MobileNetV2** using majority voting or weighted averaging.
* **Redesign the Classifier Head**
  + Add **dense layers**, **dropout**, or **attention layers** after the base MobileNetV2 model.

## Training Enhancements

* Use learning rate schedulers like:
  + **ReduceLROnPlateau**
  + **Cosine Annealing**
* **Early stopping** to prevent overfitting.
* Perform **grid/random search** for hyperparameter tuning:
  + Learning rate, dropout, filters, batch size, etc.

## Evaluation Metrics

* Already evaluated using **Precision, Recall, F1-score**.
* Consider **Confusion Matrix** visualization to diagnose class-level misclassifications.
* Use **ROC-AUC** for further insights, especially if model is used in clinical decision-making.