A Comparative Analysis of Interpretable Classification Models for Pneumonia Chest X-ray Dataset

Amitash Nanda
Electrical and Computer Engineering
University of California, San Diego
A59011408

Vaibhav Bishi
Electrical and Computer Engineering
University of California, San Diego
A59009958

Abstract

Image-based Deep Learning methods have been very effective for various medical diagnostic tasks, surpassing medical professionals' performance. However, the black-box nature of most of the Computer Vision models has restricted clinical use. Also, medical datasets have heterogenous quality due to different artifacts and biases. It is therefore essential to develop a method that will detect and classify potential diseases based on images, induce explainability while identifying the most important features, and exploit this to improve the model performance and reliability. In this study we implemented model interpretability for selected image classification models on NIH Chest X-Ray dataset by calculating and visualizing the best SHAP values.

1 Introduction

Deep neural networks have achieved near-human accuracy levels in various classification and prediction tasks, including image, text, speech, and video data. However, the networks are mostly treated as black-box functions, mapping a given input to the classification output. Deep-Learning models involve a critical use case of medical diagnosis, planning, and control which requires a considerable level of trust associated with the system output. In short, the neural network model should provide human-understandable justifications for its output, leading to insights into the inner workings. So, we call such models deep interpretable networks. Model Interpretation is a very active area among researchers in academia and industry. High model interpretability may help people break several bottlenecks of deep learning. For example, learning from a few annotations, learning via human-computer communications at the semantic level, and semantically debugging network representations.

Deep Learning methods have been very effective for various medical diagnostic tasks and have sometimes outperformed human experts. However, the black-box nature of the algorithms has restricted their clinical use. Recent explainability studies aim to show the features that influence a model's decision at best. AI-based medical devices are becoming more common in imaging fields like radiology and histology, interpretability of the underlying predictive models is crucial to expanding their use in clinical practice.

Pneumonia accounts for over 15% of all deaths of children under five years old internationally. In 2015, 920,000 children under the age of 5 died from the disease. More than 1 million adults are hospitalized with Pneumonia, and around 50,000 die from the disease yearly in the USA alone (CDC, 2017). The diagnosis of Pneumonia on CXR is complicated because of several other lung conditions such as uid overload(pulmonary edema), bleeding, volume loss(atelectasis or collapse), lung cancer, or post-radiation or surgical changes. However, detecting Pneumonia in chest X-rays is a challenging task that relies on the availability of expert radiologists. While many deep learning techniques may

provide state-of-the-art predictive performance, interpretable deep learning models are necessary for regulatory approval of their ability to explain their predictions. Further, it can reveal biases and failure modes, as seen in the case of (Oakden-Rayner, L. (2017)) [1].

In this project, initially, we performed multiclass classification on pulmonary diseases based on the NIH chest X-rays sample dataset. Then, we used chest radiographs, a special 2D high-resolution greyscale medical image to detect Pneumonia. We performed a comparative analysis of detecting Pneumonia using different image classification models like (custom CNN, VGG-16, and ResNet-50). Further, we used model interpretability methods like SHAP value analysis to justify our classification model.

2 Related Work

Model interpretation has become very active among researchers in academia and industry. (Silva, W. et al., 2020) uses Interpretability for content-based medical image retrieval [3]. (Shaban et al. 2021) has given a survey about explainability and Interpretability in health and medical intelligence and decision making [2]. (Okeke Stephen et al., 2019) provides an efficient deep learning approach to pneumonia classification in healthcare [4].

3 Method

To determine the explainability of an image classification model, we decided to try out a few different models initially and then focus our time on tweaking the one that performed best doing hyperparameter tuning. Further, we implemented the SHAP value analysis using the best model accuracy received.

3.1 Architecture

In this project, we implement various Convolutional Neural Network (CNN) - based classification models to detect pulmonary diseases such as pneumonia from chest X-ray images and use SHAP analysis to make our models' choices explainable.

3.1.1 ResNet-18

A residual network or ResNet is an artificial neural network that helps build a deeper neural network by utilizing skip connections or shortcuts to jump over some layers. ResNet18 is a 72-layer architecture with 18 deep layers. This architecture aims to enable many convolutional layers to function efficiently. Firstly, we perform a multi-class classification of various diseases related to lungs based on the NIH Chest X-ray Dataset Sample. We use the ResNet18 model initialized with pre-trained weights for the classification task.

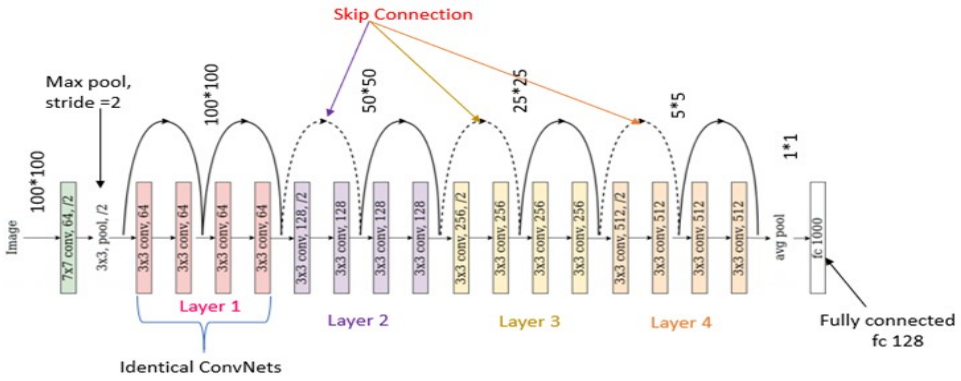

Figure 1: ResNet18 Original Architecture

3.1.2 Custom CNN

CNN is a class of artificial neural networks that analyzes visual data. It has multiple layers with three main categories: an input layer, an output layer, and a hidden layer that includes multiple convolutional layers, pooling layers, fully connected layers, and normalization layers. We decided to use CNNs in our project for all the classification and detection tasks because they perform well in visual data as they can learn more discriminative features in an image. We then shift our focus to the pneumonia detection problem. We implement three different CNN-based models for the specific task of pneumonia classification from the RSNA Pneumonia Detection Challenge Dataset. We use a custom CNN model, VGG16, and ResNet50 and compare their performance. The CNN model we have implemented is a 10-layer model built from scratch. It consists of 6 convolutional layers with ReLU activation, 3 max-pooling layers, several dropout layers, a global max-pooling layer, and a Fully Connected Network (FCN) with SoftMax activation at the end. The stride is taken to be one with zero padding. It has a total of 338,723 trainable parameters.

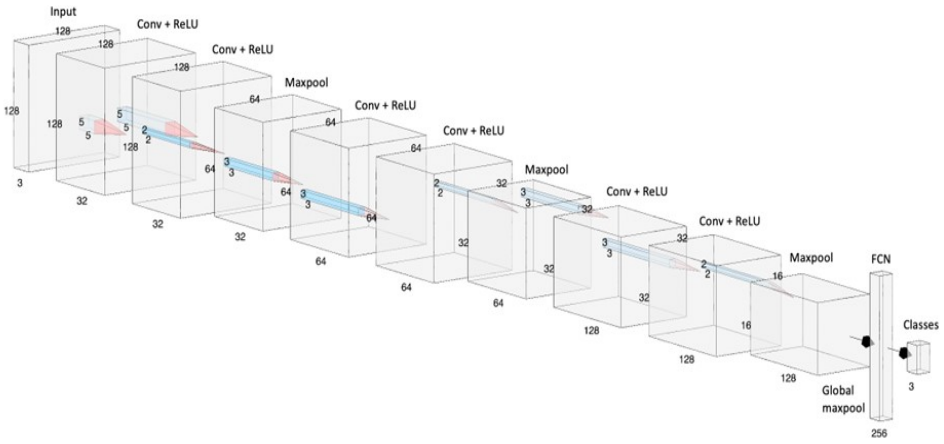

Figure 2: Custom CNN Architecture

3.1.3 VGG-16

VGG-16 is a convolutional neural network that is 16 layers deep. The unique thing about VGG16 is that instead of having a large number of hyper-parameter, it focuses on having convolution layers of a 3x3 filter with stride 1. It always uses the same padding and max pool layers of a 2x2 filter of stride 2. It follows this arrangement of convolution and max pool layers consistently throughout the whole architecture. We load a pre-trained version of the VGG16 network in the pneumonia classification task from the ImageNet database. The pre-trained network can classify images into 1000 object categories, such as keyboard, mouse, pencil, and many animals. As a result, the network has learned rich feature representations for a wide range of images. The network has an image input size of 224-by-224.

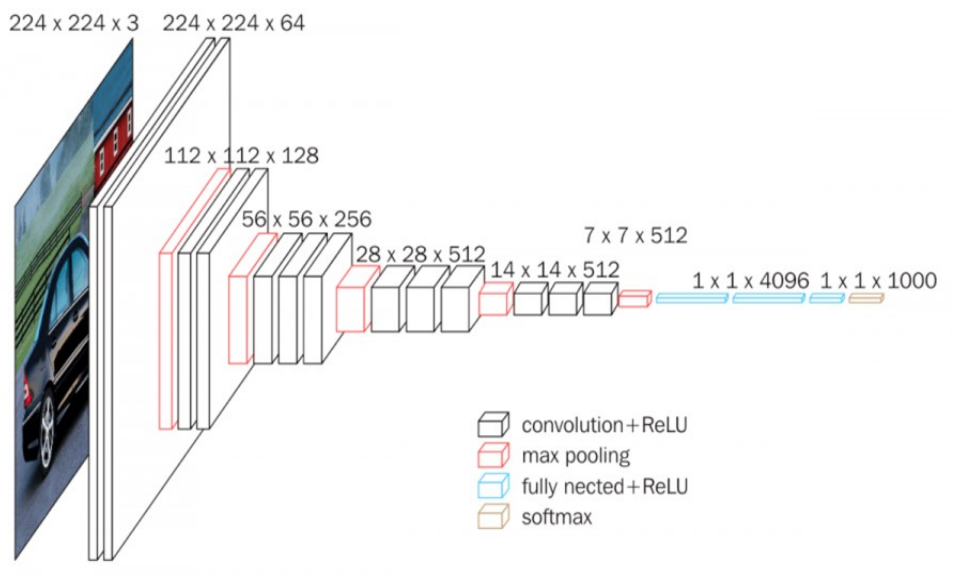

Figure 3: VGG16 Original Architecture

3.1.4 ResNet-50

ResNet, short for Residual Networks, is a classic neural network used as a backbone for many computer vision tasks. The ResNet-50 model has 5 stages, each with a convolution and Identity block. Each convolution block has 3 convolution layers, and each identity block has 3 convolution layers. The ResNet-50 has over 23 million trainable parameters. We load a pre-trained version of the ResNet-50 network in the pneumonia classification task from the ImageNet database

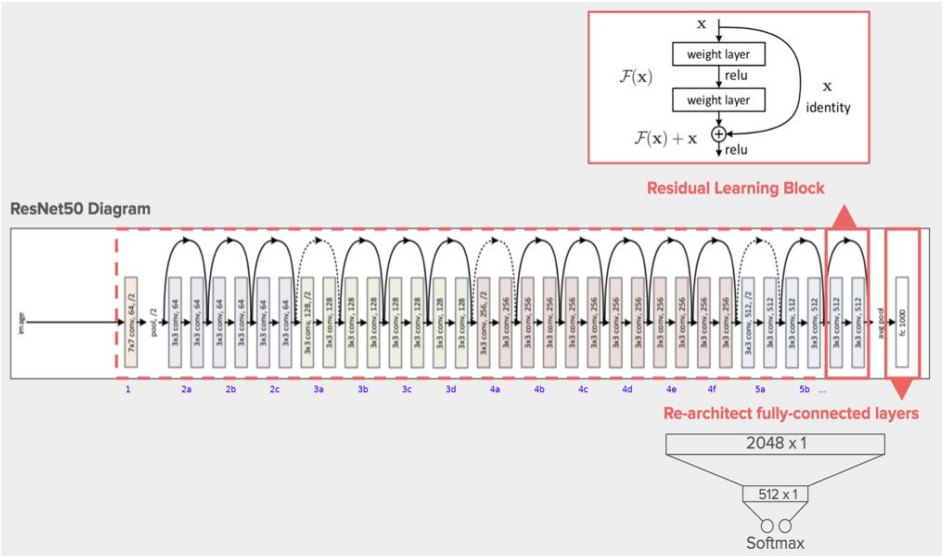

Figure 4: ResNet-50 Original Architecture

3.2 Interpretability using SHAP

We make our models interpretable using visual explanations based on the SHAP (SHapley Additive exPlanations) library. SHAP analysis is one of the most popular ways of explaining the model and understanding how the features of your data are related to the outputs. It's a method derived from coalitional game theory to provide a way to distribute the payout across the features fairly. In the context of our project, SHAP values help us visualize why our model made a particular decision based on pixel values. This essentially tells us what the local or global features are which contribute to pneumonia presence in the chest X-ray images.

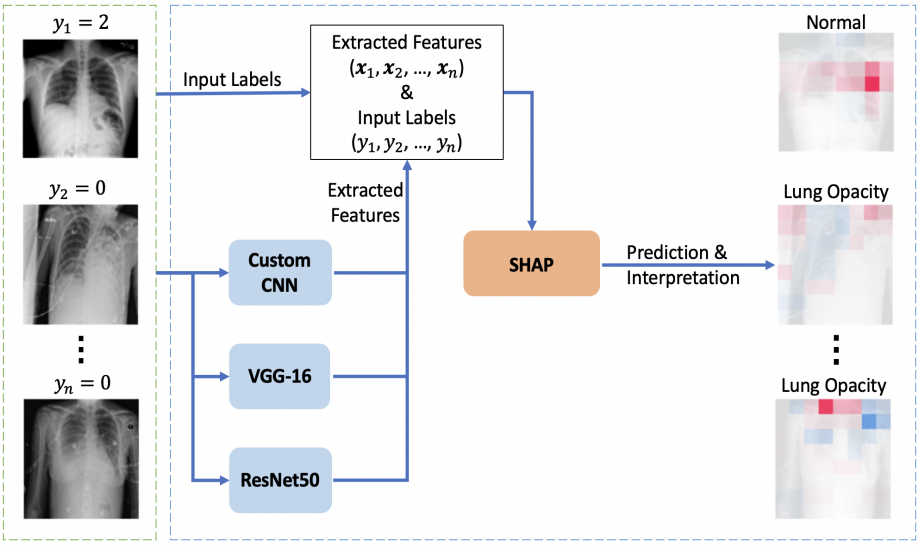

Figure 5: ResNet-50 Original Architecture

4 Experiments

4.1 Datasets

Our project involves use of the following datasets for experimentation:

4.1.1 NIH Chest X-ray Dataset Sample

This dataset is a sample of the famous NIH ChestX-ray14 dataset, which comprises 112,120 frontal-view X-ray images of 30,805 unique patients with text-mined fourteen common disease labels. The sample dataset takes 5% of the entire dataset consisting of 5,606 images with resolution 1024X1024. There are a total of 15 classes (14 for diseases and another class for "No findings") for the images as follows:

- Hernia - 13 images
- Pneumonia - 62 images
- Fibrosis - 84 images
- Edema - 118 images

- Emphysema - 127 images
- Cardiomegaly - 141 images
- Pleural_Thickening - 176 images
- Consolidation - 226 images
- Pneumothorax - 271 images
- Mass - 284 images
- Nodule - 313 images
- Atelectasis - 508 images
- Effusion - 644 images
- Infiltration - 967 images
- No Finding - 3044 images

4.1.2 RSNA Pneumonia Detection Challenge Dataset

This dataset consists of 30,227 Chest X-ray images meant for pneumonia detection. The images are classified into 3 labels based on lung opacity, 'Lung Opacity', 'Normal', 'Not Normal/No Lung Opacity'.

Lungs are full of air and therefore appear black in the X-ray image. If there is any inflammation in lungs, there would be some opacity in the lungs, and it can be indicative of possible pneumonia affliction in the patient. The class "Not Normal / No Lung Opacity" refers to the case where pneumonia was not present, but there was some other form of abnormality on the image that may mimic the appearance of pneumonia. The images are given in the DICOM format while the corresponding labels are given in a CSV file. Another CSV file contains detailed information about the positive and negative classes in the training set. The data contains the following attributes:

- patientId_- A patientId. Each patientId corresponds to a unique image.
- x_ - the upper-left x coordinate of the bounding box.
- y_ - the upper-left y coordinate of the bounding box.
- width_ - the width of the bounding box.
- height_ - the height of the bounding box.
- Target_ - the binary Target, indicating whether this sample has evidence of pneumonia.

4.2 Results

We implement two key ideas in the project. The results for each of the experiments are discussed below.

4.2.1 Multi-class classification of pulmonary diseases based on NIH Chest X-ray Dataset Sample

Firstly, we augment the dataset using random rotation and resize all images to 224X224 dimensions. We split the dataset into 5000 training, 303 validation, and 303 testing images. We implement the ResNet18 model (initialized with pre-trained weights) using the following attributes:

Optimizer: Adam, Learning rate: 0.001, Epochs: 30, Scheduler with factor: 0.1, Loss function: Weighted loss

The accuracy of the model in terms of classification for various diseases on the training and testing datasets is illustrated below.

Train Dataset Accuracy Report

	Labels	Acc
0	Cardiomegaly	67.38
1	Emphysema	72.80
2	Effusion	68.46
3	Hernia	86.48
4	Nodule	64.84
5	Pneumothorax	69.70
6	Atelectasis	65.84
7	Pleural_Thickening	69.02
8	Mass	66.82
9	Edema	72.82
10	Consolidation	69.88
11	Infiltration	63.30
12	Fibrosis	65.32
13	Pneumonia	76.56

Test Dataset Accuracy Report

	Labels	Acc
0	Cardiomegaly	67.656766
1	Emphysema	70.627063
2	Effusion	61.716172
3	Hernia	86.468647
4	Nodule	63.366337
5	Pneumothorax	66.006601
6	Atelectasis	66.996700
7	Pleural_Thickening	64.686469
8	Mass	63.696370
9	Edema	74.587459
10	Consolidation	63.696370
11	Infiltration	56.765677
12	Fibrosis	65.016502
13	Pneumonia	74.257426

Figure 6: Train and Test Accuracy Report

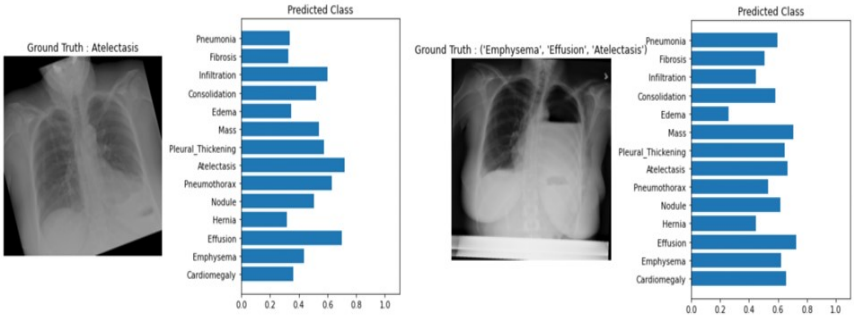

Figure 7: Model prediction on test sample images

4.2.2 Pneumonia detection using Chest X-ray images based on the RSNA Pneumonia Detection Challenge Dataset and interpretability using SHAP

Firstly, we check the target distribution and see that there are 31% of the patients with pneumonia while the remaining are disease-free. We also see from the class distribution that 39.1% of the patients are diagnosed with no lung opacity, 29.3% are normal and the rest are with lung opacity in Figure 8.

Pneumonia Classification:

The 3 classes are converted into one-hot encoding using the labelbinarizer function from Scikit-learn such that:

- [1 0 0] corresponds to lung opacity (0th class)
- [0 1 0] corresponds to no lung opacity / not normal (1st class)
- [0 0 1] corresponds to normal (2nd class)

We split the dataset into training, validation and testing using the train_test_split function from Scikit-learn. Now, we implement 3 different deep learning models for the task of pneumonia classification and observe their performance.

Custom CNN model:

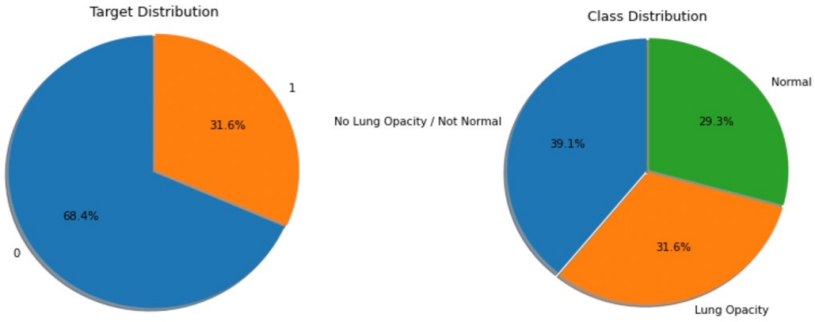

Figure 8: Data Distribution

We implement a custom-made basic architecture CNN model using the following attributes after some experimentation and hyperparameter tuning:

Optimizer: SGD with Nesterov momentum at 0.9, Learning rate: 0.001 Epochs: 100, Batch size: 64, Loss function: Categorical cross-entropy

We obtain an accuracy of 58.81% on the training set, 53.16% on the validation set and 51.52% on the test set. The training and validation accuracies and losses are plotted over the epochs.

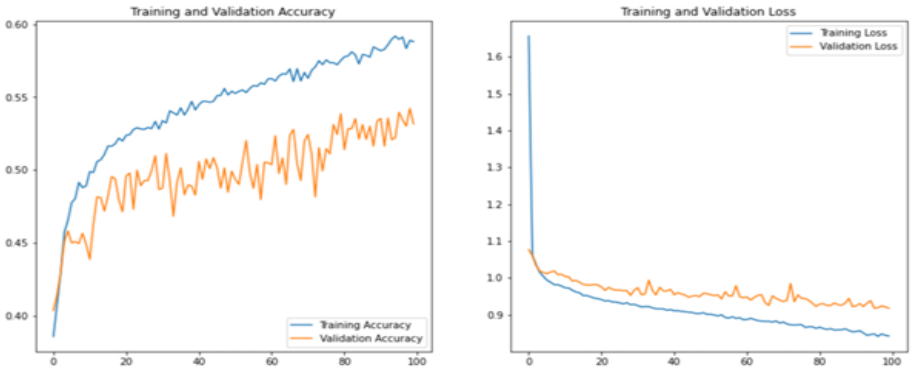

Figure 9: Accuracy and Loss Curve

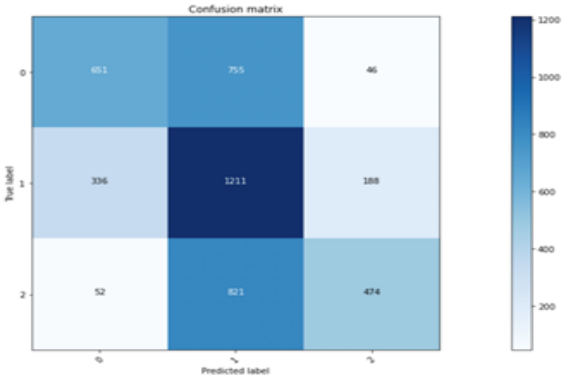

Figure 10: Confusion Matrix

VGG16 model: We load the VGG16 model and add 2 more hidden layers along with one Soft-Max layer as an output layer. We implement this model using the following attributes after some experimentation and hyperparameter tuning:

Optimizer: SGD with Nesterov momentum at 0.9 Learning rate: 0.001 Epochs: 50 Batch size: 64 Loss function: Binary cross-entropy

We obtain an accuracy of 88.72% on the training set, 79.41% on the validation set and 78.14% on the test set. The training and validation accuracies and losses are plotted over the epochs.

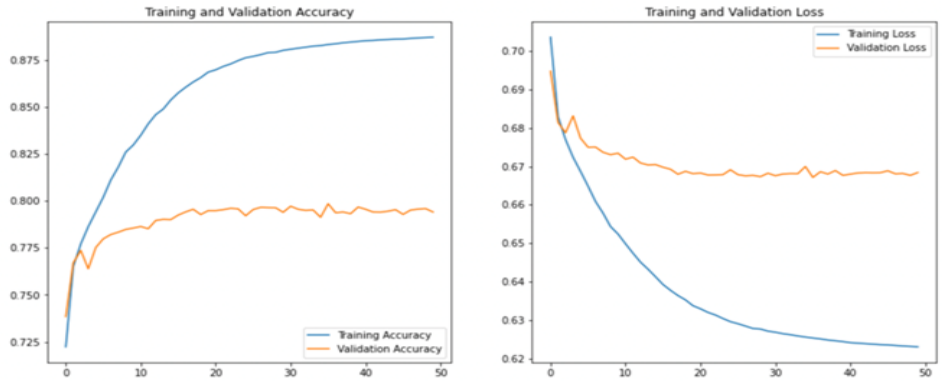

Figure 11: Accuracy and Loss Curve

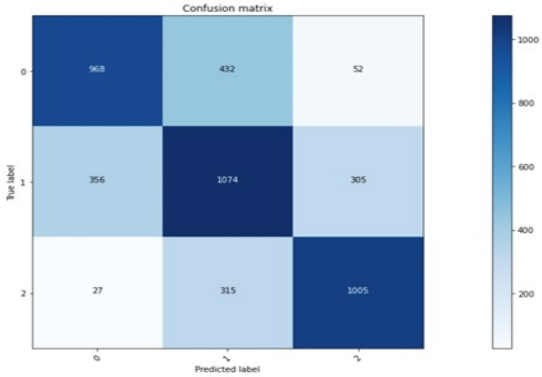

Figure 12: Confusion Matrix

ResNet50 model:

We load the ResNet50 model and add 2 more hidden layers along with one SoftMax layer as an output layer. We implement this model using the following attributes after some experimentation and hyperparameter tuning:

Optimizer: SGD with Nesterov momentum at 0.9 Learning rate: 0.001 Epochs: 50 Batch size: 64 Loss function: Binary cross-entropy We obtain an accuracy of 97.23% on the training set, 79.87% on the validation set and 79.52% on the test set. The training and validation accuracies and losses are plotted over the epochs.

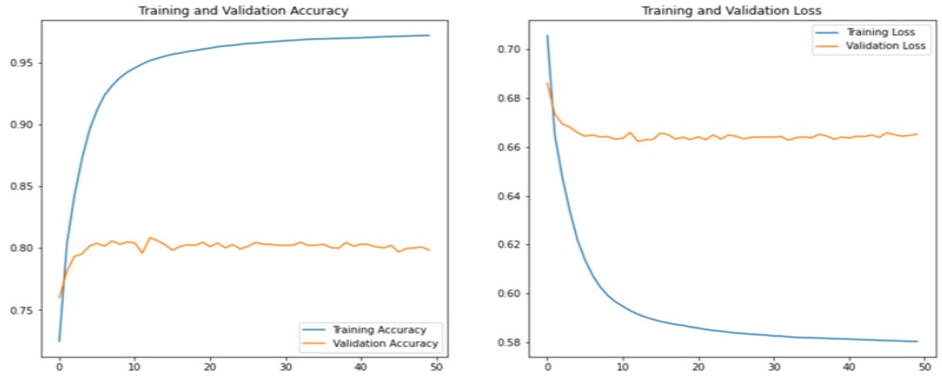

Figure 13: Accuracy and Loss Curve

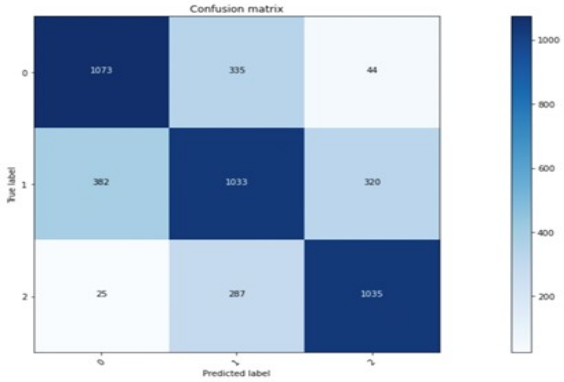

Figure 14: Confusion Matrix

Model Interpretability:

We implement model interpretability for all our models on some test images by calculating and visualizing the SHAP values using the explainer function from the SHAP library. The top 3 predictions for any input image are given from left to right along with the masks generated by SHAP on top the image.

Image classification tasks can be explained by the scores on each pixel on a predicted image, which indicates how much it contributes to classifying that image into a particular class. The red pixels represent positive SHAP values that contributed to classifying that image as that particular class, while the blue pixels represent negative SHAP values that contributed to not classifying that image as that particular class. Thus, we can now see why our model is classifying a chest X-ray image as any specific class, i.e., by emphasizing on certain regions in and around the lungs depending on the presence or absence of any lung opacity.

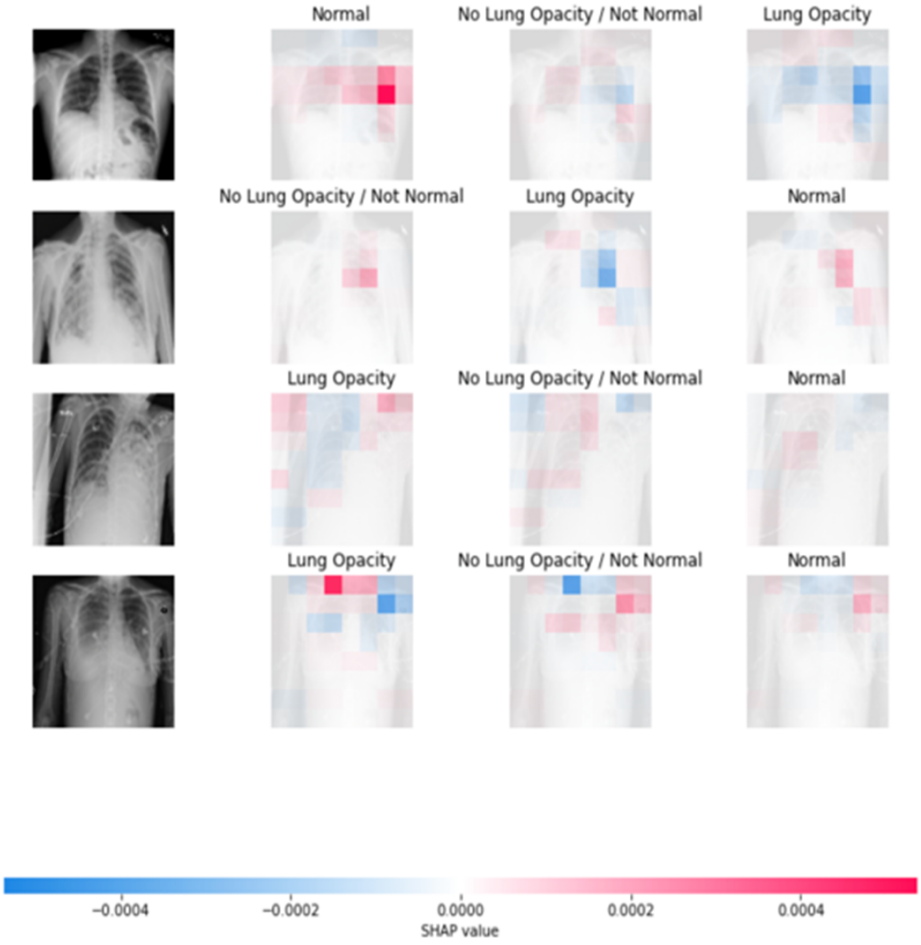

Figure 15: SHAP for Custom CNN

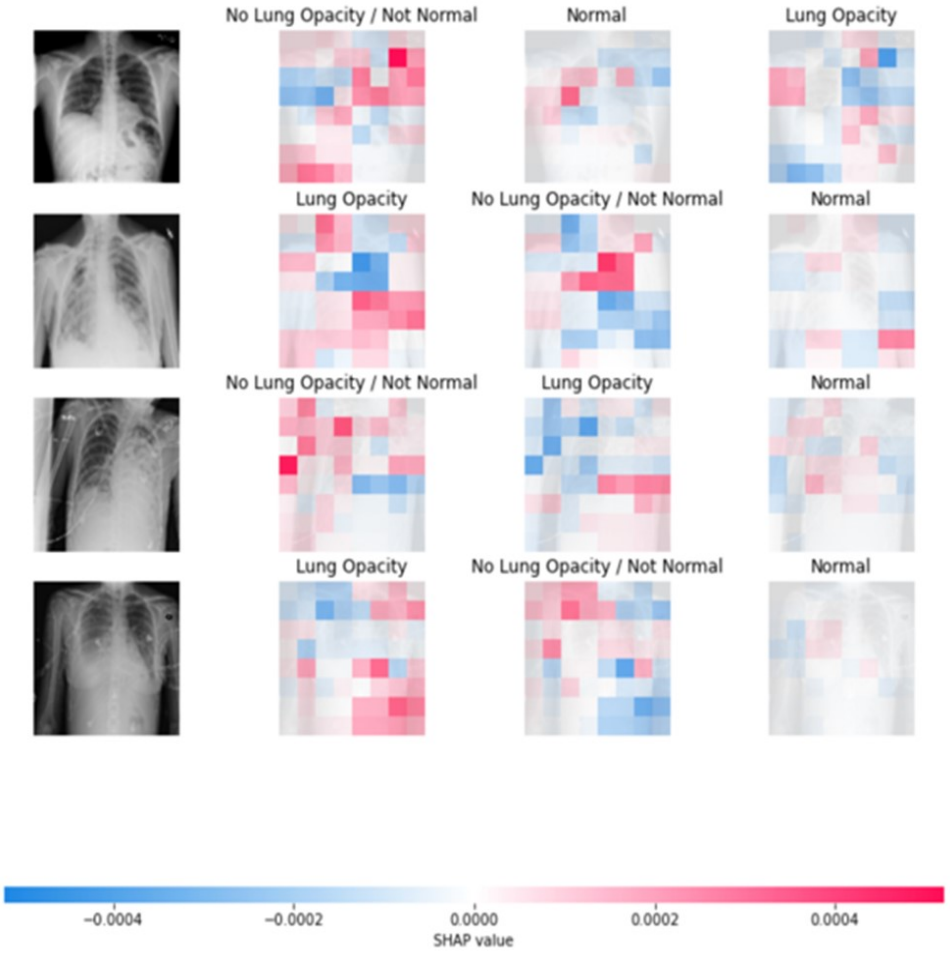

Figure 16: SHAP for VGG16

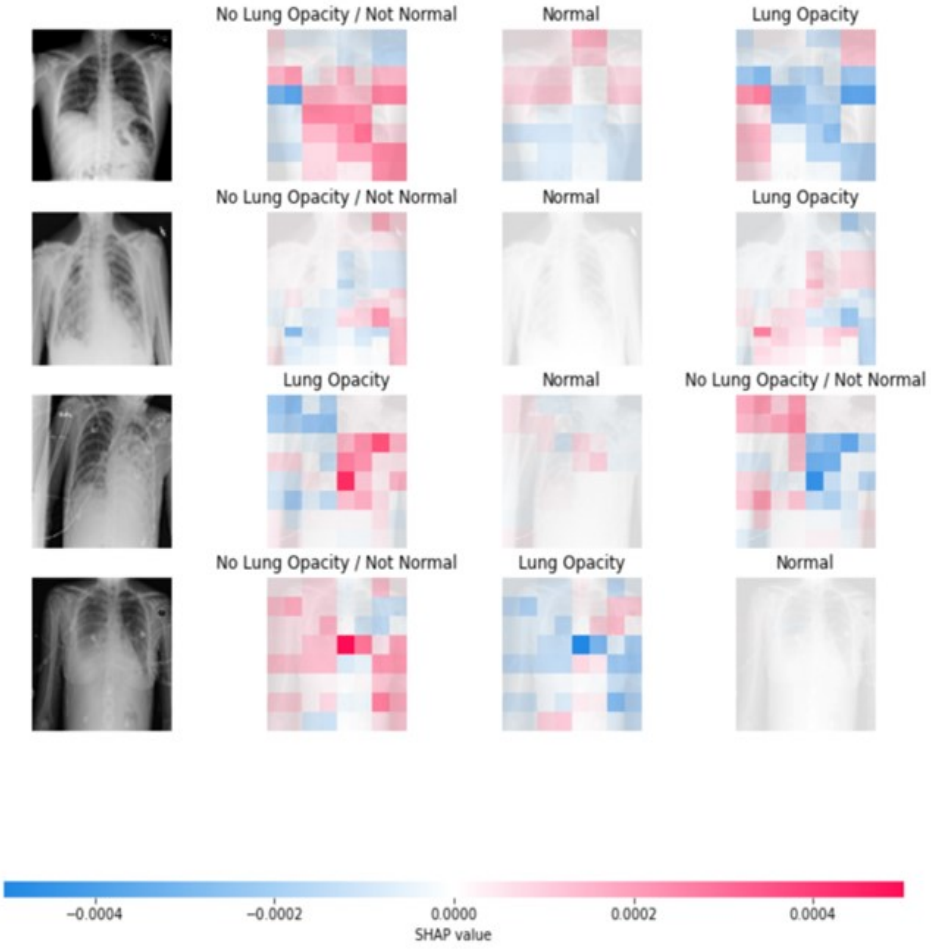

Figure 17: SHAP for ResNet50

5 Supplementary Material

Code and Video is enclosed

References

[1] L Oakden-Rayner. Exploring the chestxray14 dataset: problems. `https://lukeoakdenrayner.wordpress.com/2017/12/18/the-chestxray14-dataset-problems/`, December 2017.

[2] Arash Shaban-Nejad, Martin Michalowski, and David L Buckeridge. Explainability and interpretability: keys to deep medicine. In *Explainable AI in Healthcare and Medicine*, pages 1–10. Springer, 2021.

[3] Wilson Silva, Alexander Poellinger, Jaime S Cardoso, and Mauricio Reyes. Interpretability-guided content-based medical image retrieval. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 305–314. Springer, 2020.

[4] Okeke Stephen, Mangal Sain, Uchenna Joseph Maduh, and Do-Un Jeong. An efficient deep learning approach to pneumonia classification in healthcare. *Journal of healthcare engineering*, 2019, 2019.