Final Project Report on

# Machine Learning Techniques Based Pneumonia Classification from Chest X-ray

Pattern Recognition- ECSE 6610

Mohaiminul Al Nahian

RIN: 662026703

Date: December 10, 2021

# Table of Contents

# 1. Introduction:

Pneumonia is an infection of one or both of the lungs. It is caused by bacteria, viruses, or fungi. Human lungs are made up of small sacs called alveoli, which fill with air when a healthy person breathes. When an individual's lungs are affected by certain bacteria, virus or fungi, the alveoli are filled with pus and fluid, which makes breathing painful and limits oxygen intake. This situation is called Pneumonia. The general symptoms are cough, fever or sweating, shortness of breath, severe chest pain, loss of appetite, nausea and vomiting.

Sometimes pneumonia can be difficult to diagnose because the symptoms are so variable and are often very similar to those seen in a cold or influenza. The most common diagnosis techniques are blood test, chest x-ray, pulse oximetry and lung CT-scan. In case of chest x-ray method, Radiologists use chest x-ray images to look for distinct white spots in the lung area to identify pneumonia infection along with patient's medical history and symptoms. A healthy lung is normally filled with air, which allows x-ray to pass through it easily. When a person has pneumonia, the air sacs are filled up with fluid, which attenuates the x-ray from passing through, causing white spots in x-ray films.

The goal of this project has been to identify pneumonia from chest x-ray images using different machine learning and pattern recognition-based techniques. The idea has been to distinguish between a normal person's lung and a pneumonia affected person's lung by taking different features of lung image and classifying it. It should be mentioned that bacterial pneumonia and viral pneumonia have different levels and types of symptoms on a patient's body, but to differentiate between different pneumonia types solely from chest x-ray without knowing any other symptoms is a more demanding work and has been out of scope for this project.

To classify normal lungs and pneumonia affected lungs, here two approaches have been adopted. Firstly, taking some features from the x-ray images and then fitting a classifier model to differentiate between two classes. Secondly, using the image data itself to train a neural network model for classification of the two classes. The performance metrics chosen is accuracy, precision and recall of test data. Both of the methods show promising result and we get to the conclusion that with proper and careful preprocessing and selection of image features either manually or by a neural network, the task of detecting pneumonia from chest x-ray can be made automatic with human level performance.

The rest of the report is organized as follows- first a literature review has been given about related works in this field. Then a description of the dataset used for this project has been given. After that, the proposed Feature Based Method and Neural Network based method has been discussed in detail. Then the next section shows the results of different experimentations. Then Observations on the results and Remarks for improvement has been shared. Finally, a conclusion has been drawn.

## 2. Related Works:

The use of machine learning techniques to classify chest x-ray images is not a very recent domain of research. Oliveria et al. [1] proposed a machine learning-based network that classifies pediatric CXR images as pneumonia or normal images based on wavelet transform coefficients and KNN classifier. Sousa et al. [2] proposed a pneumonia detection algorithm and tried five machine learning classifiers (KNN, naive Bayes, multi-layer perceptron, decision tree and SVM) combined with dimensionality-reduction techniques such as principle component analysis (PCA). Yao et al. [3] proposed a machine learning-based automated system that identifies five diseases, including pneumonia. Their extracted feature vector contained 25 texture features, such as the mean, variance, energy, and correlation and employed SVM classifier. Depeursinge et al. [4] similarly compared the performances of five machine learning classifiers (naïve Bayes, KNN, J48 decision trees, multi-layer perceptrons, and SVM) in pneumonia detection. They extracted a total of 39 texture-based attributes and optimized the parameters of each classifier by gridsearching. The performance was evaluated by McNemar's statistical tests and the accuracy measure. They concluded that SVM achieved the best values of each metric, with a correct prediction rate of 88.3%.

However, most of these works mentioned are a few years old. Recently, to the author of this report's surprise, we see that most people are using sophisticated deep learning models to predict chest x-ray images these days. Especially, with the outbreak of covid-19 pandemic, there has been hundreds of papers published in the last 2 years for covid-19 detection using chest x-ray images based on deep learning techniques. Covid-19 chest x-ray images contain similar white/ opaque patterns that we see in pneumonia images. Albahli et al. [5] used DenseNet121, InceptionResNetV2 and ResNet152V2 to detect 14 chest diseases. Wang et al. [6] classified pathologies from CXRs using various pre-trained models (AlexNet , GoogleNet , VGG16, and

ResNet). Rahaman et el. [7] examined 15 different pre-trained CNN models to find the most suitable one for covid-19 detection. Alam et. El [8] proposed a feature fusion based deep learning technique where they fused the HOG feature of x-ray images with CNN features and then classified covid-19 using a deep learning architecture. Many authors around the world have taken similar approaches to solving the problem, either employing deep learning directly on the image or taking some features from an intermediate layer of the CNN and combine it with hand crafted features to employ a classification model afterwards.

## 3. Proposed Dataset:

Chest X-ray Images Dataset (Pneumonia) published by Kaggle.com (https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia) has been used for this project. This dataset contains 5956 x-ray images divided into two classes; PNEUMONIA and NORMAL lungs. The dataset contains 5216 training samples of which 1341 are NORMAL and 3875 images having PNEUMONIA. The testing set has 624 testing samples. It is split into 234 NORMAL Images and 390 PNEUMONIA images. The training dataset has almost thrice as many pneumonia samples than normal samples, so it is an imbalanced dataset. The images had different size range such as 1840x1480, 1296x1296, 1300x976 etc.

## 4. Proposed Method:

In this project, the classification of chest x-ray has been addressed by two approaches. First, feature extraction-based classification and second, Neural Network classifier.

### i. Feature Extraction-Based Classification:

At first, all the training and testing images have been resized to fixed size (300, 300). Then from the X-ray images some features have been extracted. A small description of the extracted features are given below:

***Color Histogram:***

For digital images, a color histogram represents the number of pixels that have colors in each of a fixed list of color ranges, that span the image’s color space, the set of all possible colors. As we know that the diseased x-ray has more brighter pixels than normal x-rays, the color histogram feature has been chosen. Normalized RGB color histograms with 8 bins have been extracted for all the images of the dataset.

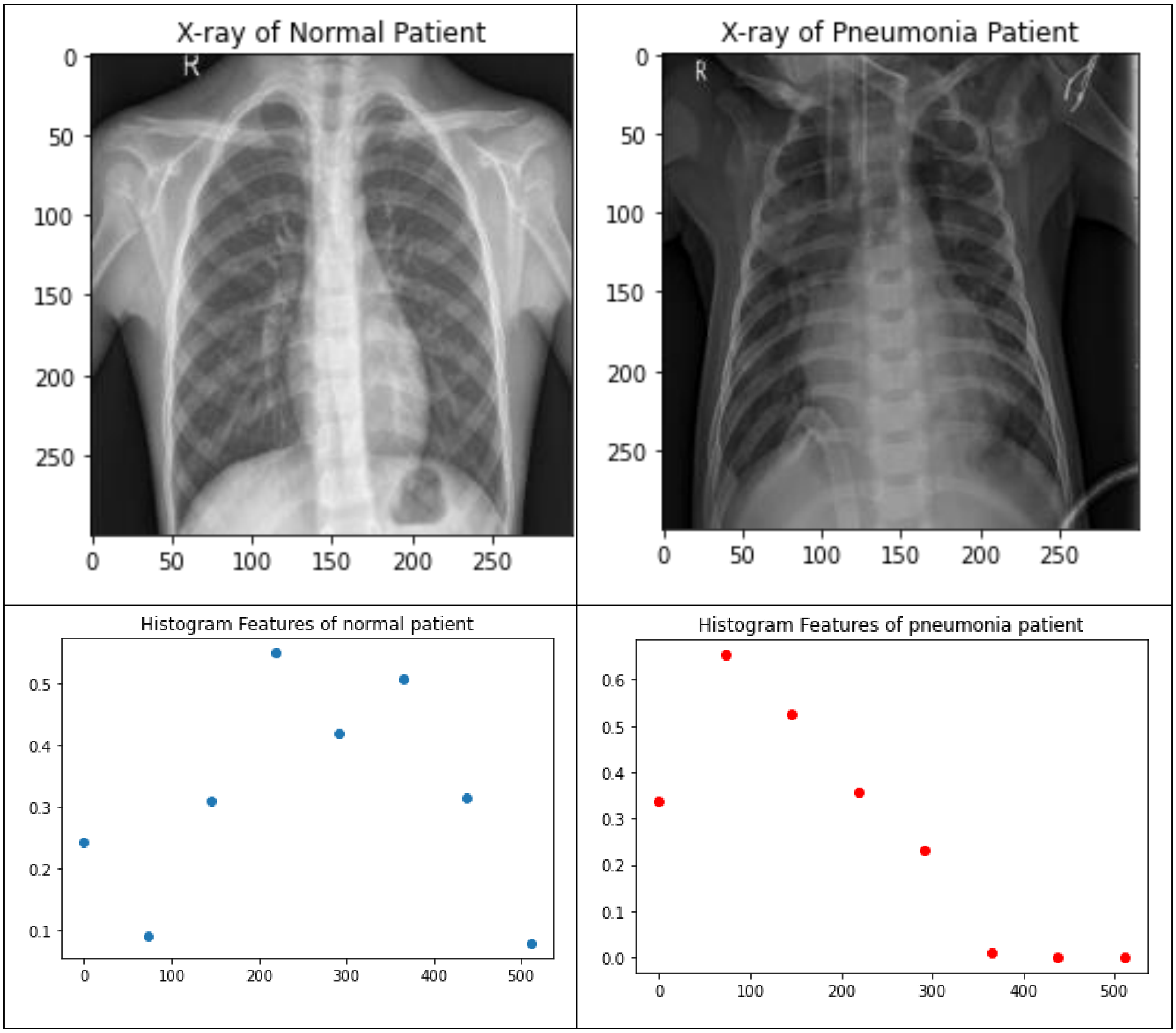

Figure: Example of Chest X-ray Image from 2 classes with their Color Histogram

### *Haralic Feature:*

Haralick texture features are used to describe the “texture” of an image, i.e, to quantify and represent the appearance or consistency of an image. Haralick features are derived from the Gray Level Co-occurrence Matrix (GLCM). This matrix records how many times two gray-level pixels adjacent to each other appear in an image. 13 values can be extracted from the GLCM to quantify texture.

### *Hu Moments:*

Hu moment feature has 7 values that come from an image. These values represent the invariants with respect to translation, scale, and rotation. The first one is analogous to the moment of inertia around the image's centroid, where the pixels' intensities are analogous to physical density. The Hu moments for both class images have been calculated and added in the feature vector set.

### *HOG Feature:*

The HOG or Histogram of Oriented Gradient descriptor focuses on the structure or the shape of an object. In the case of edge features, we only identify if the pixel is an edge or not, whereas, HOG is able to provide the edge direction as well. This is done by extracting the gradient and orientation of the edges. For calculating HOG features, the images have been resized to (128,128). Then the features have been calculated using 8 directions, (32,32) pixel per cell and (2,2) cells per block.

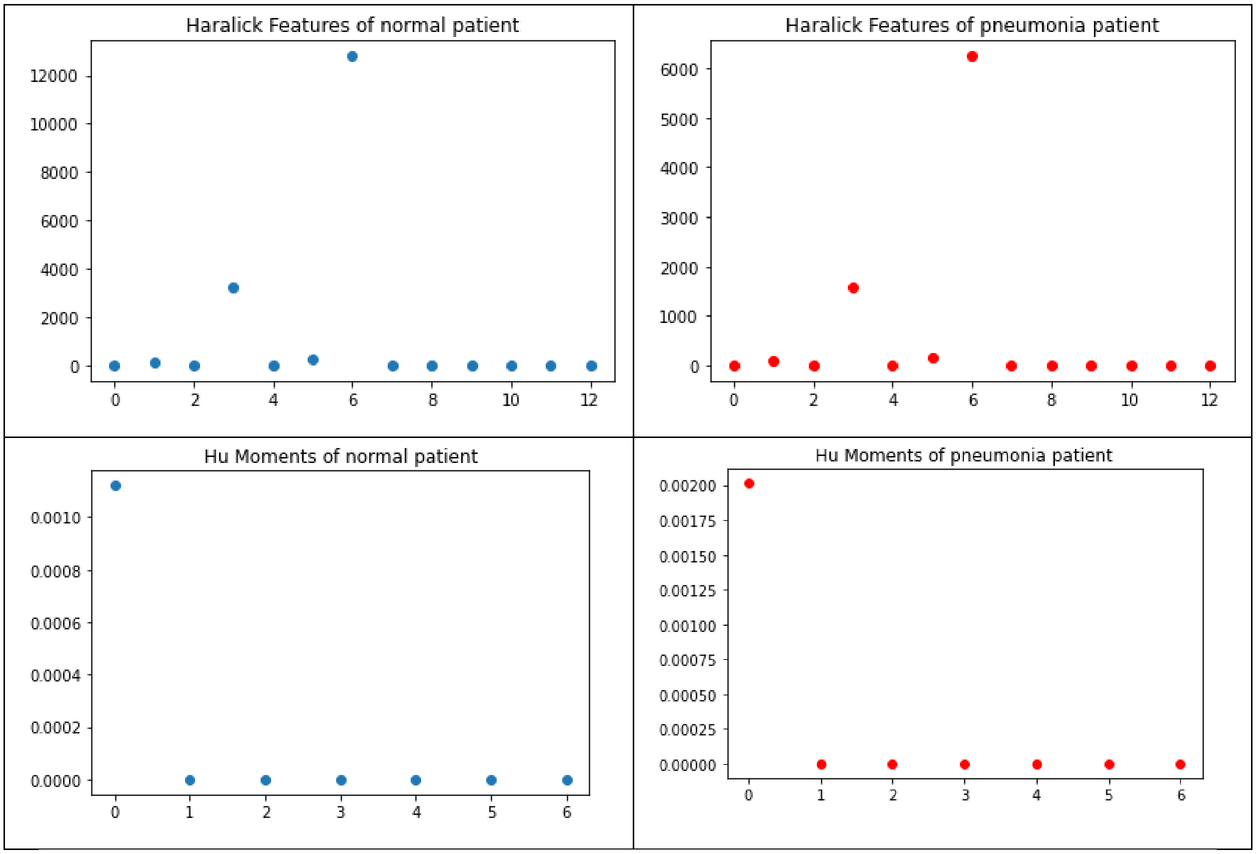

Figure: Typical Haralick Feature and Hu Moment Values of Two Images from Two Classes

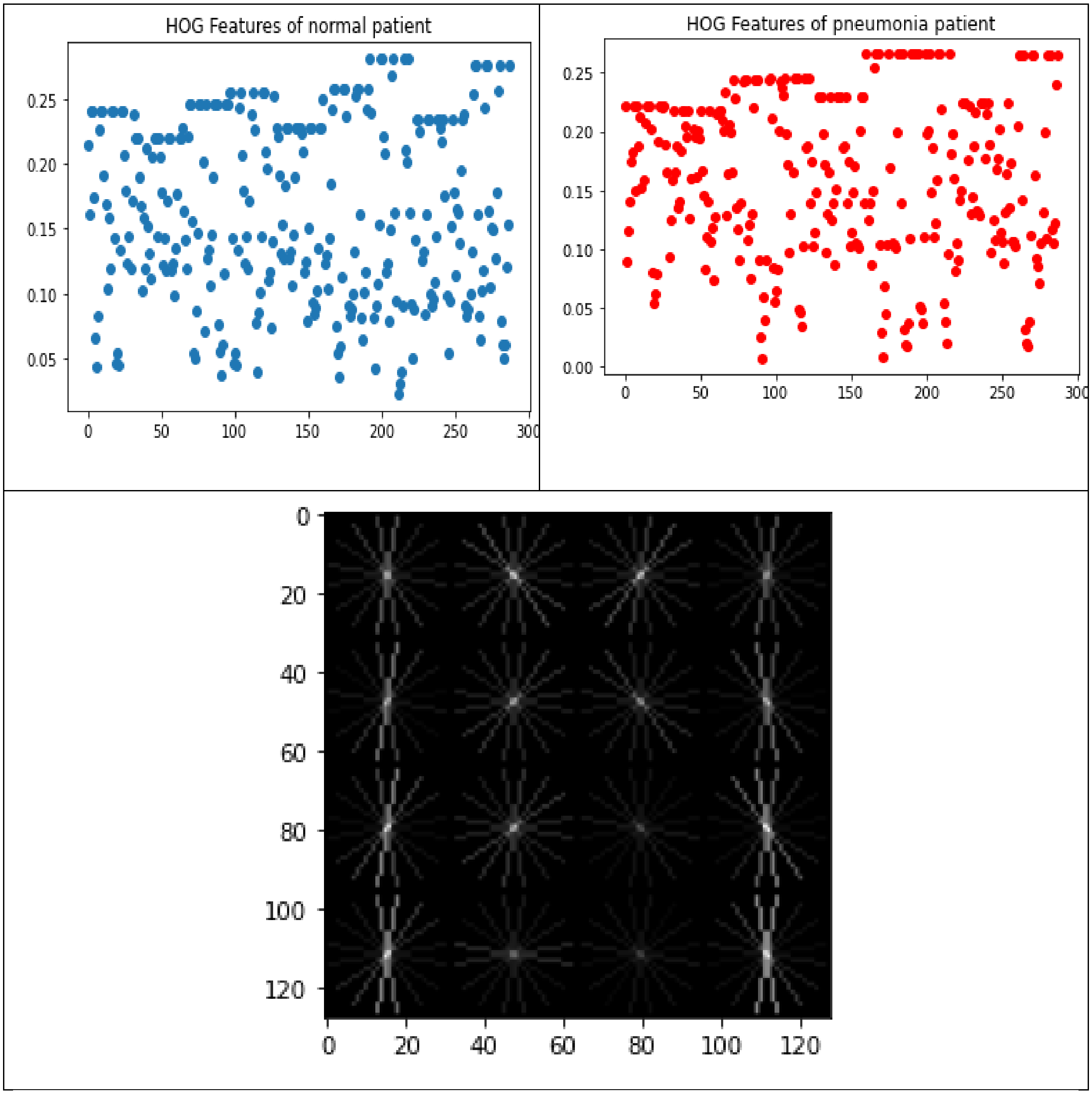

Figure: HOG features of a normal and a diseased lung. Also, a visual representation of a typical HOG image

Calculating all the features, the values of individual features have been normalized between 0 and 1. The concatenated feature vector has a feature length of 820 for each image.

## Classification Based on the Selected Features:

After extracting the normalized features we fit the training data into different machine learning classifiers. The classifiers used in this project are Logistic Regression with Newton-cg solver, KNN with 3 neighbors, Naïve Bayes algorithm with Complement Naïve Bayes (Because the dataset is unbalanced) and Support Vector Machine (SVM). The test data has been classified using the fitted models to get the output predictions. In addition to that, heuristically, the output prediction labels of three best resulting methods of the four (LR, K-NN, SVM) have been averaged to get a new prediction which also show interesting result. The 10-fold cross validation on the training data showed promising result. The picture below shows the cross-validation result on training data. The actual test results are shown on the result section.

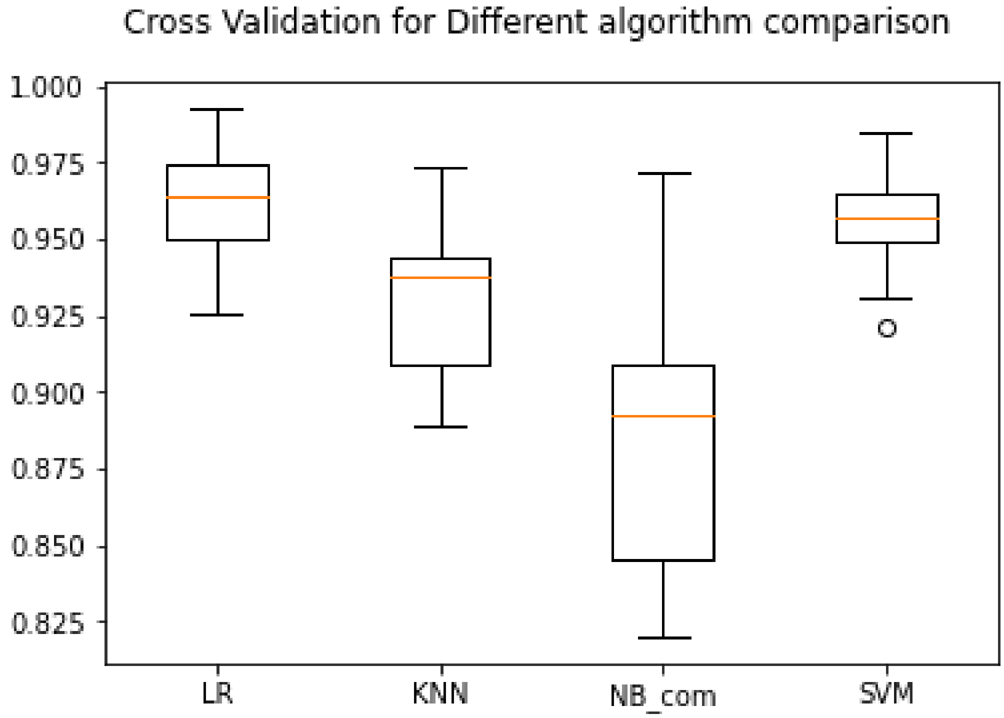

Figure: 10-Fold Cross Validation Accuracy for Logistic Regression, K-NN, Naïve Bayes and SVM

### ii. Deep- Learning based classifier:

The problem with hand-crafted feature selection is that it is very difficult to get the exact features that will be good for fitting a machine learning model. For this reason, I wanted to train a deep neural network to see if the test results could be improved. An improved result on deep learning would mean that with careful selection of hand-crafted features, the result of the machine learning techniques mentioned previously can be improved further.

For deep learning based classification, transfer learning technique has been adopted. The reason is, it is very hard to train a neural network from scratch, especially when the dataset is very small and also imbalanced. Furthermore, the training has been done using the whole image, rather than a segmented image for only the region of interest. As a result, I took a pre-trained ResNet-18 model trained on ImageNet dataset. The model has been finetuned with a learning rate of 0.005. The whole training data runs for 20 epochs and the training and testing loss and accuracy has been recorded. In my personal computer, with core i7 processor, 16GB RAM and nvidia Geforce GTX 1050Ti GPU, the training takes 24 minutes and 1 second to complete. After training completion, the best model among the 20 epochs is selected which is used to classify the test data. The result of classification is shown on the result section of this report.

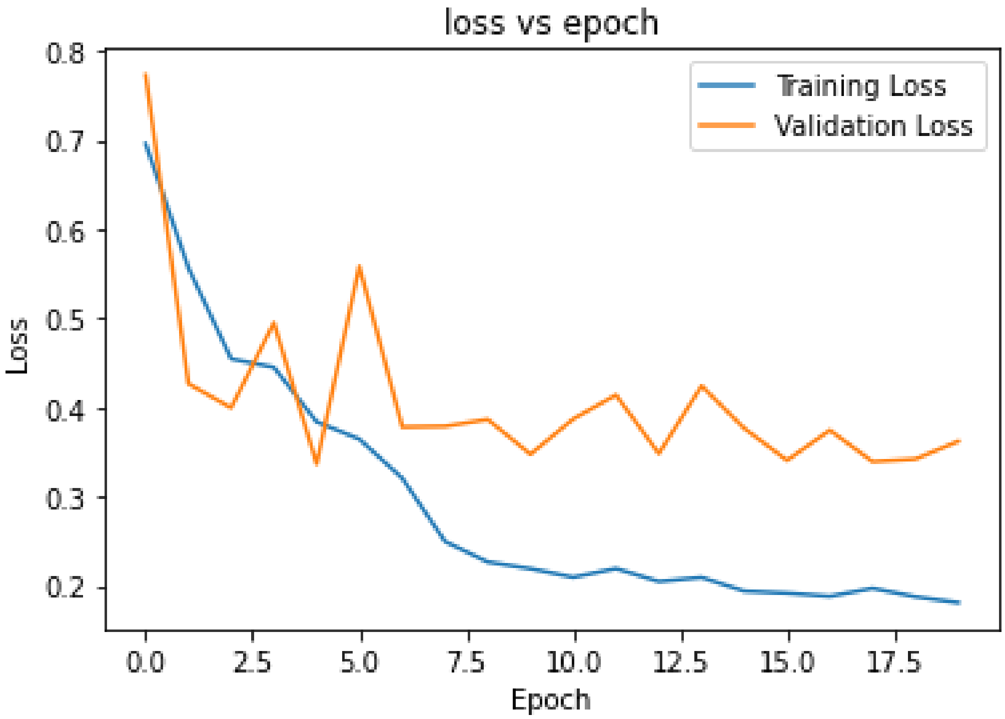

Figure: Loss vs Epoch graph for Neural Network

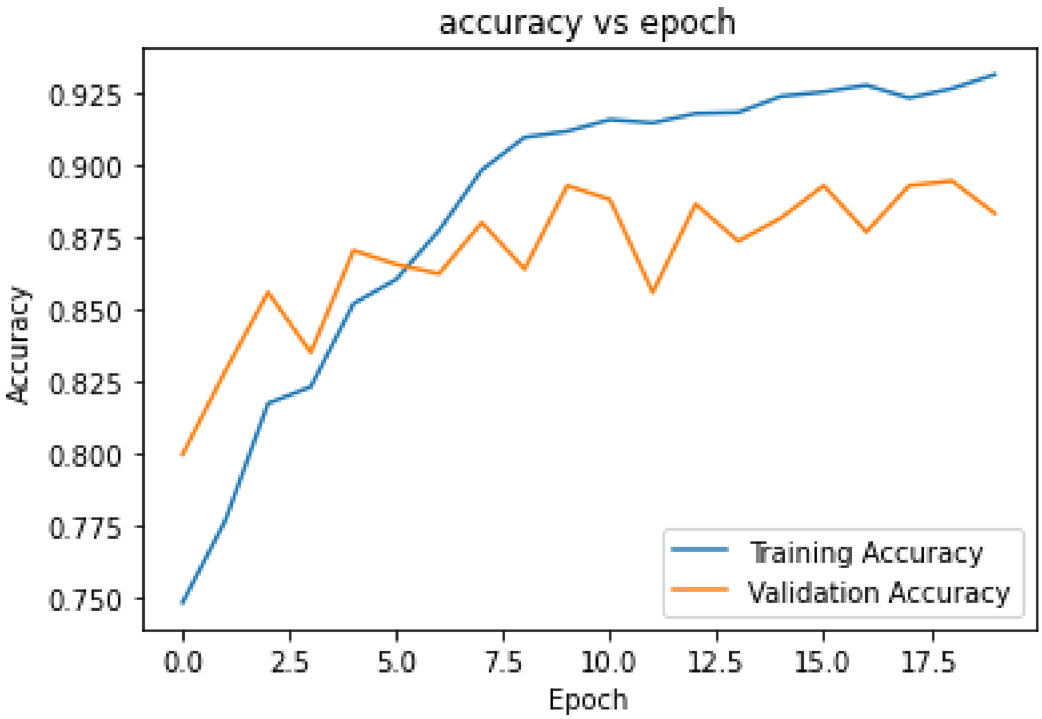

Figure: Accuracy vs Epoch Graph for Neural Network

## 5. Experimental Results:

The metrics which have been used to evaluate a methods performance for this project are Accuracy, Precision and Recall. Accuracy is the number of datapoints from the test set that has been predicted correctly over the total number of test data points. Accuracy is a good measure of performance when the test set has equal number of datapoints from both classes. But if the the number of datapoints from one class is significantly higher, then accuracy could be misleading. For example, if a test set has 90 class1 image and 10 class2 image and the model predicts every point as class1, then the accuracy would be 90% but the model has failed. Similarly, in this case the test set has 234 normal images and 390 pneumonia images. So, in addition to accuracy, the precision and recall for each class has been calculated.

Precision is defined by the following expression

$$Precision(Class_i) = \frac{\#\ of\ Class_i\ point\ Classified\ as\ Class_i}{(\#\ of\ Class_i\ point\ classified\ as\ class_i) + (\#\ of\ class_j\ point\ classified\ as\ class_i)}$$

Recall is defined by the following expression

$$Recall(Class_i) = \frac{\# \, of \, Class_i \, points \, Classified \, as \, Class_i}{(\# \, of \, Class_i \, point \, classified \, as \, class_i \,)+(\# \, of \, class_i \, points \, classified \, as \, class_j)}$$

Using the Metrics described, we get the following results on Test Data for the methods described previously:

| Method Name | Accuracy | Precision for Normal Class | Recall for Normal Class | Precision for Pneumonia Class | Recall for Pneumonia Class |
|---|---|---|---|---|---|
| Logistic Regression | 0.78 | 0.94 | 0.43 | 0.74 | 0.98 |
| K-NN | 0.81 | 0.92 | 0.54 | 0.78 | 0.97 |
| Naïve Bayes | 0.72 | 0.72 | 0.41 | 0.72 | 0.9 |
| SVM | 0.74 | **0.95** | 0.33 | 0.71 | **0.99** |
| Ensemble Average of Prediction (LR+KNN+SVM) | **0.82** | 0.91 | **0.63** | **0.81** | 0.96 |

It can be seen that the ensemble average has better accuracy, Normal class Recall and Pneumonia class Precision. It is also noted that all the four models struggle to get a good Recall value on Normal Class, where K-NN has the highest recall value of 0.54. It means that the models are predicting many of the normal images as having Pneumonia. One reason could be that the training data had almost three times as many Pneumonia images than Normal images, so the classifiers seem to have become biased towards predicting more images as having Pneumonia.

Now, for the Neural Network, similar metrics have been used namely Accuracy, Precision and Recall. But as the test data has been used to validate the neural network model and the best model has been chosen based on lowest validation error, which we get from test set prediction error. So, another new metric has been introduced to theoretically give an upper bound of the out-of-sample error for datapoints that the network has never seen. This upper error bound is called Hoeffding bound, defined as:

$$E_{out} \leq E_{Test} + \sqrt{\frac{1}{2N} \ln \frac{2}{\delta}}$$

Here, N is the number of Test Set Samples which in this case is 624. $\delta$ is the tolerance value. $\delta = 0.01$ means that we have a 99% confidence that the above equation will be true.

Adding the estimate of out of sample error with the other metrics, the performance of Neural Network is shown below:

| **Finetuned ResNet-18** | |
|---|---|
| **Metric Name** | **Value** |
| Accuracy | **0.89** |
| Precision for Normal Class | **0.96** |
| Recall for Normal Class | **0.75** |
| Precision for Pneumonia Class | **0.87** |
| Recall for Pneumonia Class | 0.98 |
| Test Error, $E_{Test}$ | 0.11 |
| Upper Bound of Out of Sample Error, $E_{out}$ (for $\delta = 0.05$) | <0.11+0.053=0.164 |

## 6. Observations and Remarks for Improvement:

Here, it is seen that the Neural network performed better than the other methods in all metrics except Recall for Pneumonia class. However, the performance of feature based method is almost similar to Neural Network in all cases, except Recall for Normal Class. So, we may assume that, a careful selection of new features may improve the performance of the feature based methods significantly.

Also, using $\delta = 0.05$, we can theoretically say that we have a 95% confidence that the out of sample error will be below 0.164 for any new test dataset for this neural network.

Here it can also be mentioned that the features have been taken from whole image to classify diseases which are associated only with the lung part. But if a segmentation of lungs could be performed to extract only the lung before getting features or training the neural network, we may assume that the performance could improve significantly. Because, outside the lung area, the image has no important information, rather both normal and Pneumonia images have similar pattern outside the lungs. So, they just add noise and degrades the performance of the machine learning algorithm.

## 7. Conclusion:

In this project, we have exploited the power of four important image features to distinguish between normal and pneumonia affected x-ray images. We have fitted our training image features to train four machine learning models named Logistic Regression, K-NN, Naïve Bayes and SVM. We have predicted the test image features with our fitted models and got reasonable accurate results. We have also proposed a heuristic ensemble approach to show that the results from individual best models can be averaged to get a more generalized result. We have also used a pre-trained neural network architecture and finetuned with the training images to predict the outputs of testing images. For the neural network, we have shown a theoretical calculation to predict the generalized performance of this network on unknown images with 95% confidence. Finally, we have remarked that with better feature selection and some pre-processing, we can improve our results to matchup the performance of the state-of-the art methods performing similar tasks of image classification.

## 8. References:

1. L. L. G. Oliveira, S. A. Silva, L. H. V. Ribeiro, de R. M. Oliveira, C. J. Coelho, and A. L S. Andrade, "Computer-aided diagnosis in chest radiography for detection of childhood pneumonia." International journal of medical informatics, 77, no. 8 (2008): 555-564.
2. R. T. Sousa, O. Marques, F. A. A. M. N. Soares, I. I. G. Sene, L. L. G. De Oliveira, and E. S. Spoto, “Comparative performance analysis of machine learning classifiers in detection of childhood pneumonia using chest radiographs,” Procedia Comput. Sci., vol. 18, pp. 2579–2582, 2013, doi: 10.1016/j.procs.2013.05.444.

3. J. Yao, A. Dwyer, R. Summers, D. M.-A. radiology, and undefined 2011, "Computer-aided diagnosis of pulmonary infections using texture analysis and support vector machine classification," Elsevier.
4. A. Depeursinge et al., "Comparative performance analysis of state-of-the-art classification algorithms applied to lung tissue categorization," J. Digit. Imaging, vol. 23, no. 1, pp. 18–30, Feb. 2010, doi: 10.1007/s10278-008-9158-4.
5. Albahli S, Rauf HT, Algosaibi A, Balas VE. AI-driven deep CNN approach for multi-label pathology classification using chest X-Rays. PeerJ Comput Sci. 2021 Apr 20;7:e495. doi: 10.7717/peerj-cs.495. PMID: 33977135; PMCID: PMC8064140.
6. 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," Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017, vol. 2017-Janua, pp. 3462–3471, May 2017, doi: 10.1109/CVPR.2017.369.
7. Rahaman MM, Li C, Yao Y, Kulwa F, Rahman MA, Wang Q, Qi S, Kong F, Zhu X, Zhao X. Identification of COVID-19 samples from chest X-Ray images using deep learning: A comparison of transfer learning approaches. J Xray Sci Technol. 2020;28(5):821-839. doi: 10.3233/XST-200715. PMID: 32773400; PMCID: PMC7592691.
8. Alam, N.-A.-; Ahsan, M.; Based, M.A.; Haider, J.; Kowalski, M. COVID-19 Detection from Chest X-ray Images Using Feature Fusion and Deep Learning. *Sensors* **2021**, *21*, 1480. https://doi.org/10.3390/s21041480

## 9. APPENDIX: CODES

The codes used for this project can be accessed through the following link. The codes are written in python with jupyter notebook:

https://drive.google.com/drive/folders/1UuW6DwOdQykocSO2Mc1b4eo0eLDcJMar?usp=sharing