Pneumonia Detection

Mini Project Report

Prof. Predrag Radivojac, Northeastern University, Spring 2020
Basar S. Chowdhury chowdhury.b@husky.neu.edu, Mayur Kurup kurup.m@husky.neu.edu

Abstract

In this project we aim to accurately diagnose pneumonia using convolutional neural networks with hope to augment the diagnosis by specialists. Chest radiographs are one of the most commonly available screening data. The dataset we are working with are images in DICOM (Digital Imaging and Communications in Medicine) format taken from NIH Clinical Center[1]*. Our primary motivation behind doing the project is getting familiar with medical image classification using neural networks. Diagnosis is a very sensitive and specialized field in healthcare, we think machine learning can have a huge impact on it. One of the most challenging problems in the medical field is proper and early diagnosis, as in the case of COVID-19, or any other life threatening disease. We hope to apply the theoretical concepts taught in the class with some new techniques that will help us understand the subject in practical terms also.*

Keywords*: Machine Learning, Pneumonia, Convolutional Neural Network, Support Vector Machines, Logistic Regression. Diagnosis, Image classification, Chest radiographs*

Objectives and significance

Pneumonia is a form of acute respiratory infection that affects the lungs and fills the alveoli with pus and fluid which makes it difficult for the person to breath and exchange the oxygen in lungs. According to WHO estimates pneumonia accounts for 15% of the deaths in children under 5 years, killing nearly 808,694 in 2017[4]. Cause of pneumonia could be from viruses, fungi or bacterial infection. In the USA pneumonia is one of the top 10 causes of deaths.

Major problem is diagnosis at the earlier stage before it becomes chronic and life threatening. Proper diagnosis requires review of the chest radiograph by highly trained specialists and past clinical records. What makes it difficult to diagnose the pneumonia in chest radiograph (CXR) is the presence of other fluids or any other underlying medical conditions. Usually the CXR of an infected person becomes opaque but there might be other causes of the opacity of the CXR. Even the positing of the person and depth of the inspiration can add noise to the CXR images adding more complexity to the diagnosis by the specialist. Even there is disagreement between the specialists which calls for a need to consult experts for their consensus on the classification.

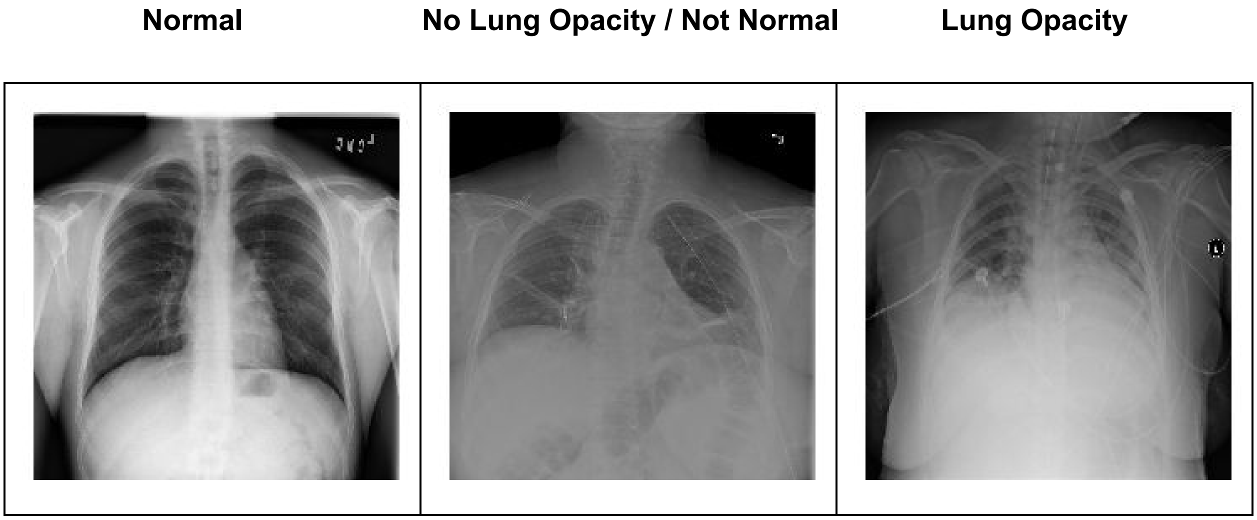
Using machine learning we can reduce the stress on the specialist and can contribute to the more accurate diagnosis of the disease at the earlier stage. In the eve of the current pandemic COVID-19, patients are coming to the testing centers in overwhelming numbers and it is difficult for doctors to manually go through every patient's X-ray images. Doctors may even make mistakes in such times which can cause a patient his life. Thus we are looking to create a solution that will enable medical professionals to use our algorithm which will detect Pneumonia in a patient just by analyzing his chest X-ray (CXR). This will save a lot of time for patients and doctors.

We will be using the Convolutional Neural Network (CNN) , our main model, Support Vector Machine and Logistic Regression. The main algorithm is based on the architectural variant of CNN called ResNet (residual network). Finally SVM and Logistic regression are for experimental purposes just to gain any hidden correlation or any rich insight between the image metadata and pneumonia.

Background

We are provided with the patient's information and their corresponding CXR images. Images provided here are not the full resolution (2000 x 2000) used by the specialist for diagnosis but compressed resolution of 1024x1024 which works well with the algorithms but not with radiologists.

In order to clearly understand our goal we need to know the difference between healthy lungs, lungs with other ailments and lungs with pneumonia infection.

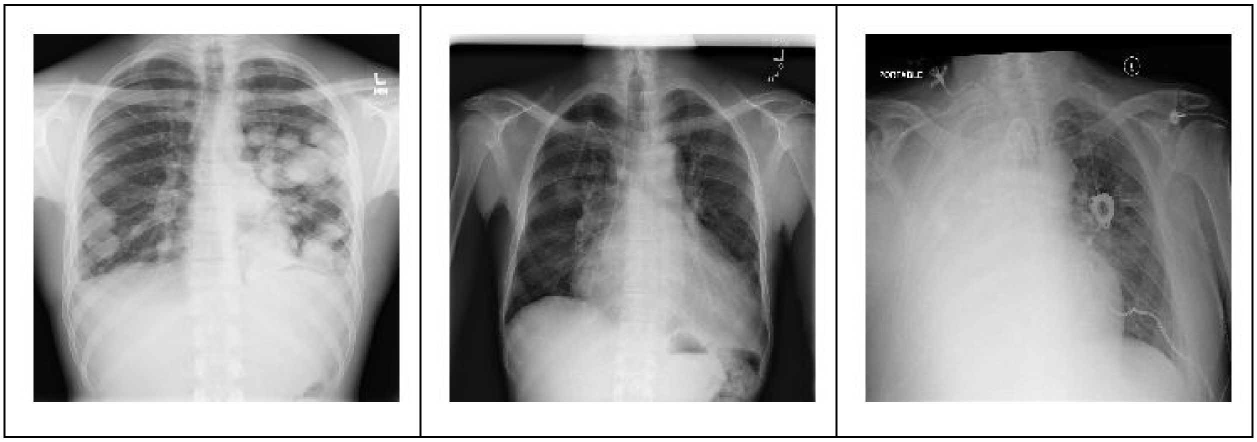


Examples of 3 different classes of patient CXR

Above we can see the CXR of the patients with three classes normal, no lung opacity / not normal and lung opacity . As the X-Ray permeates through the body to the other detector, the amount of intensity it receives depends on the density of the tissue it goes through.
Black patches correspond to the air, Grey corresponds to the fluid or some dense tissue and the white patches bones. Roughly more dense it is the more bright/white the patch will appear.

In the first image we can see it is more opaque and dark as it is filled with air. While the second image is overall hazy but it is not due to pneumonia. Our target is to identify the third type of images with opacity in the lower portions of the lungs. If we compare the second image with the third one we can see the opacity is more distributed, which is sometimes called ground glass opacity.

Consider other types of opacities with modules and nodes, which might be due to some other kind of infection but not pneumonia in the right one. On the right side the opacity is due to the enlarged heart rather than pneumonia. Our main challenge would be to distinguish between opacity due to non-pneumonia causes and opacity due to actual pneumonia.



Examples of opacity due to non-pneumonia causes
1. Due to another infection. 2. Due to an enlarged heart. 3. Missing lungs (surgically removed)

One other consideration would be in cases where the lungs might have been surgically removed altogether. So the question is what if the one of the lungs is removed. Is it opaque ?
Well it can be opaque or not. There might be many reasons which can explain the opacity in the half side of the image. It may be due to removed lungs, pneumonia or outside of the lungs filled with some fluid because of some other condition.

So how do we deal with that ? If the lungs are surgically removed then it should not be a problem, because the specialist would be aware of it beforehand. With the proper history and other information specialists can figure that out since this algorithm is supposed to augment the specialist. If the algorithm is used it most probably that the image provided meets the prior conditions of the pneumonia. So with this assumption it is safe to mark this opacity as pneumonia.

Previous work

Medical image processing has been extensively studied using Convolutional Neural Network (CNN) and Deep neural network (DNN). Recently CNN-motivated deep learning algorithms are the primary choice for image classification techniques. These network architectures are based on the trial and error principles. To design these models specialists depend on large no. of design decisions and intuitions guiding the manual search process.

CNNs outperform the DNNs because of the similarity of the classification based on the visual filter which is similar to the humans. CNNs are extremely good at detecting 2D and 3D shapes and feature extraction[5]. With the large availability of the datasets and resources CNN algorithms are outperforming the human specialist in speed and accuracy in diagnosis.

Dataset and Approach

Dataset

The original dataset is from the National Institute of Health - Clinical Center (NIH) with the size of 45.7GB[2]. But we are using a subset of the data which are approximately 4 GB due to resource constraints. All the image files are compressed in order to be used in the algorithm. There are two important files ‘stage_2_train_labels.csv’ and ‘stage_2_detailed_class_info.csv’. First file represent the table which contains patient ids with corresponding boxes and the second file contains the patient id with corresponding classification class i.e. Normal, No Lung Opacity / Not Normal, Lung Opacity.
All the image files are present in the ‘stage_2_train_images’ and ‘stage_2_test_images’. All the final test images are in the ‘stage_2_test_images’ folder. To save on resources since we don’t have powerful machines we split the actual train test in train/test with 80/20 ratio.

Methods for classification

Convolutional Neural Networks (CNN) are classes of deep neural networks with convolutional layers instead of hidden layers. These convolutional layers act as a pattern recognition layer and perform very well on the image classification. These layers are defined by convolutional kernels with *width and height*. They act as some kind of filter for detecting edges in the images. Mathematically these layers are sliding dot products.

Below is the illustration of the convolution step followed by pooling. First convolutional layer is feeded with tensor (28, 28, 1) as input and then it generates a tensor output (24, 24)x32(filters) i.e. a 3D tensor with 32 outputs of 24×24 pixel result of computing the 32 filters on the input. Next there is a pooling operation with a window of size (2x2) matrix which results in a final 3D tensor of size 12x12x(32 filters) . Pooling operation can be described as a compressing operation which reduces the size of the dimension of the data. It can be done using two variants called max-pooling or average pooling. Max-pooling chooses the max value from the pool window and similarly Average-pooling takes the average of the pool window.

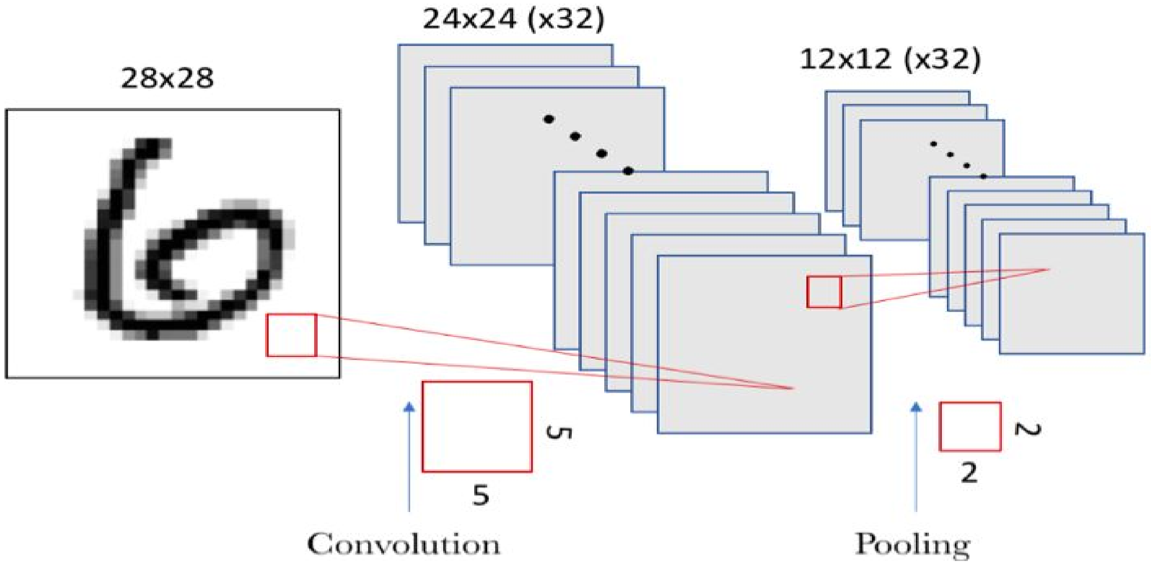
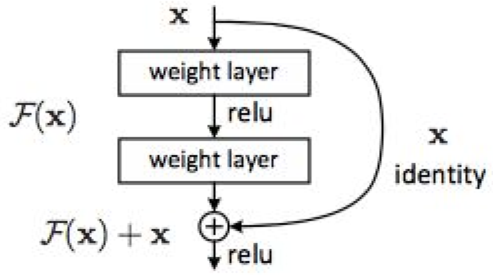


Fig: Example of generic convolution filter and pooling operation on image

There are various types of architecture for implementing CNN like LeNet, AlexNet, VGGNet, GoogLeNet, ResNet and ZFNet. We are using ResNet architecture in our project. The main idea behind the model is the shortcut connections. ResNet stands for residual network. Main difference between the traditional model and ResNet is instead of learning mapping *h(x)* we learn mapping ***F(x) = h(x) - x***. Where *h(x)* is the true model and *F(x)* is the residue. But our aim is to learn the true model *h(x)*, hence we can rearrange the equation as ***h(x) = F(x) + x*** . This will lead to an opportunity to skip intermediate subnets by making ***F(x) = 0 => h(x) = x***. Which implies output of a particular network is just the output of the last subnetwork making it more efficient on deep neural networks.


x
weight layer
relu
$\mathcal{F}(\mathbf{x})$
weight layer
x
identity
$\mathcal{F}(\mathbf{x}) + \mathbf{x}$
relu

The motivation behind ResNet is no matter how deep a network is it should perform equivalently to a shallow network. The argument for this is if a neural network can learn any complex function then it should also be able to learn the identity function f(x) = x i.e. input = output skipping few learning processes on some layers. But the problem with this is diminishing gradient descent and curse of dimensionality. So instead of learning a true mapping we try to learn the residue.

Support vector machine

SVM is a class of supervised learning algorithms used for classification and regression. They belong to a family of generalized linear models for classification. They are also called maximum margin classifiers which reduce empirical error and maximize the geometric boundary. SVM can be used to classify the non-linear boundaries in the metadata. Using the SVM on actual images for edge detection is out of scope for this project as it will take a lot of effort and time. So we thought to experiment with meta information in the images itself in hope to final any hyperplane which might classify the data.

Logistic regression

It is a probabilistic model of classification that in the original form uses a logit activation function. Basically it tries to predict the parameters of Bernoulli distribution using linear regression. This linear relationship can be represented as

$$\ell = \log_b \ (p / 1-p) = \beta_0+\beta_1 x1+\beta_2 x2$$

After exploring the data we are expecting to find a linear boundary between the image metadata and pneumonia patients. Intuitively any correlation would be superficial but we hope to get some interesting insights.

We will be referring to our assignments and other notes from the class to perform the SVM and Logistic regression classification.

Evaluation strategy

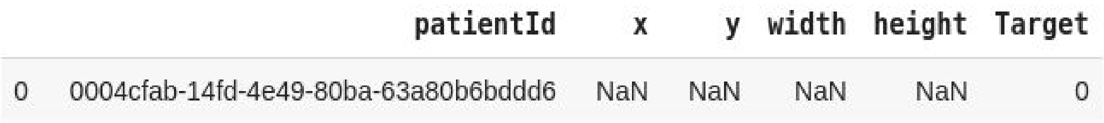
1. Convolution networks don't require much pre processing on the dataset. We already have a curated set of images. What was important to the model was image augmentation. In reality there might be many cases with misaligned images like horizontally flipped, resized images or boxed adding the rectangles within it. For this we used keras for generating sequential augmented images in specified batch sizes. It would have been great to train on other dataset and test them but couldn’t do it due to resource constraints. Another consideration is shuffling of the data before each epoch. Since we are using only one epoch this doesn't matter for this model.
2. For the support vector machine and Logistic regression we performed 10 fold stratified validation. As we were only experimenting with the parsed metadata dataframes we were able to do the 10 fold validation is reasonable time. Finally we tried to average the accuracies and plot roc curve and calculated their respective AUC values. Other import consideration might be random boosting and ensemble but this is just an experiment for casual insight so we left that.

Results

Exploratory Analysis

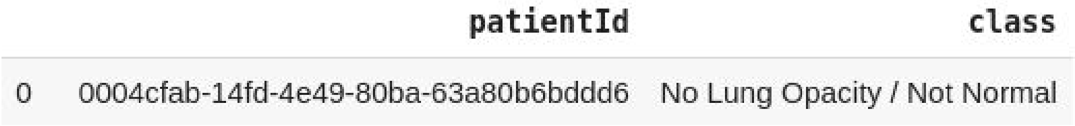
Starting with the types of the file contents and their data structures we figured that we have to map the tables and images into one dataframe. We tried to look into the different classes of images and understand their distribution.

Structure of the ‘`stage_2_train_labels.csv`’ file with shape (30227, 6)



	patientId	x	y	width	height	Target
0	0004cfab-14fd-4e49-80ba-63a80b6bddd6	NaN	NaN	NaN	NaN	0

Structure of the ‘stage_2_detailed_class_info.csv‘ file with shape (30227, 2):



	patientId	class
0	0004cfab-14fd-4e49-80ba-63a80b6bddd6	No Lung Opacity / Not Normal

Meta-data of the image file consists of various parameters which might be very useful while using SVM and Logistic regression. Some of the parameter are ‘Body Part Examined’, ‘Modality’, ‘View Position’, ‘Patient's Sex’ and ‘Patient's Age’.

Next is the distribution of the types of the classes in the dataset. It is important that we should have balances of each type of example so that our classifier can be accurately trained.

In the following image we can see that we have an almost balanced dataset for each class type which saved us a lot of juggling and processing. Here target 1 are all positive cases of pneumonia.

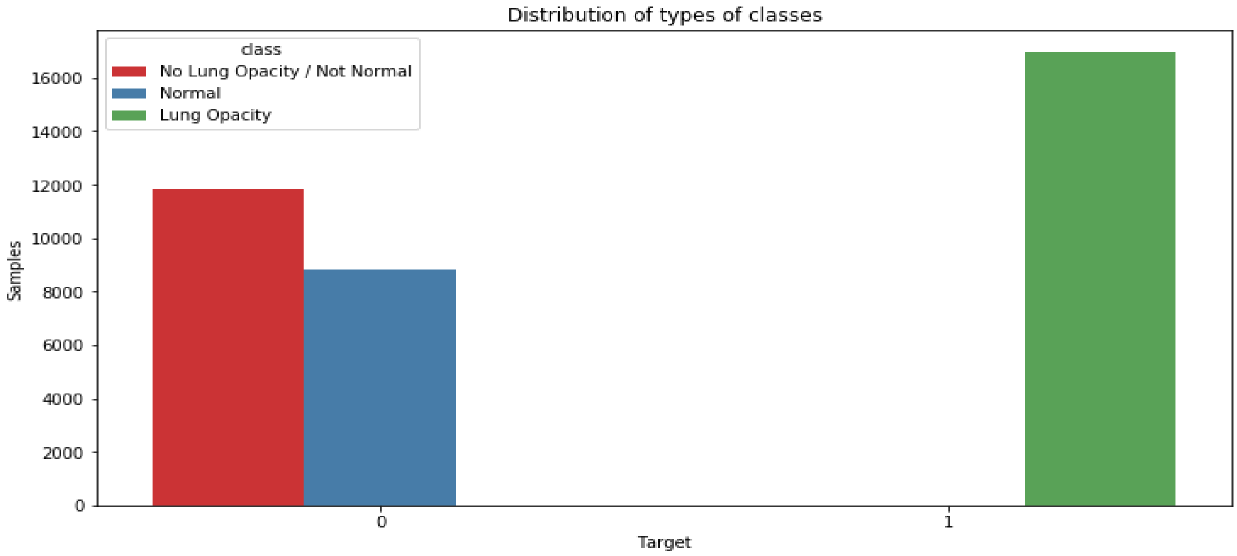


Fig: Distribution of the 3 different types of classes

Finding the lung opacity statistics (distribution and centers) :

Just to understand the range of the opaque boxes in the images, we grabbed the distribution of the centers of the boxes. X,Y are coordinates which refers to the top left corner of the boxes.

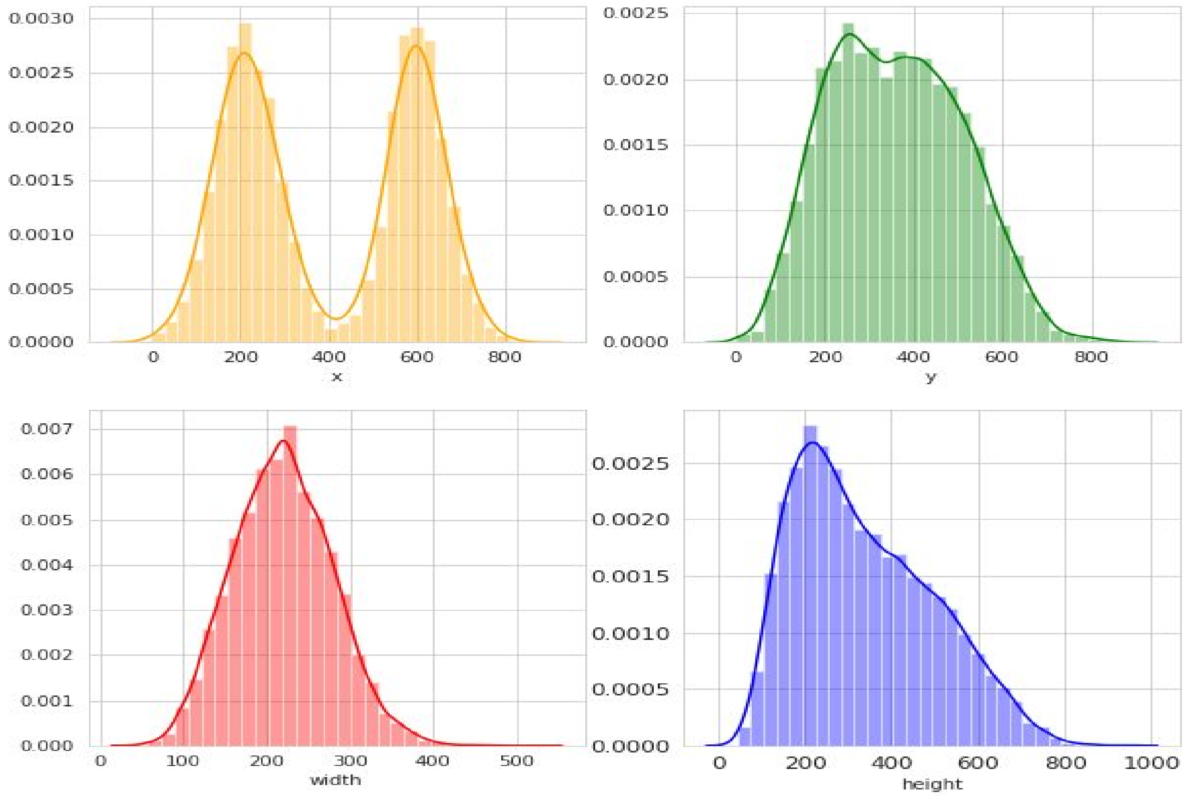


Fig: Distribution of the X,Y cord of the boxes in the infected person

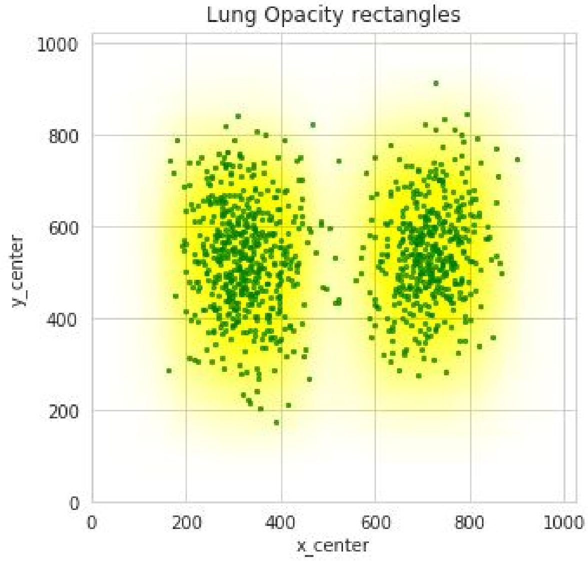


Fig: Coordinates of the centers of the rectangles (box)

Colormap Images:
Viewing different colormap images for a patient can help in understanding the depth of the opacity and accurately localize the boxes. Following is the patient no. 45 with pneumonia infection.

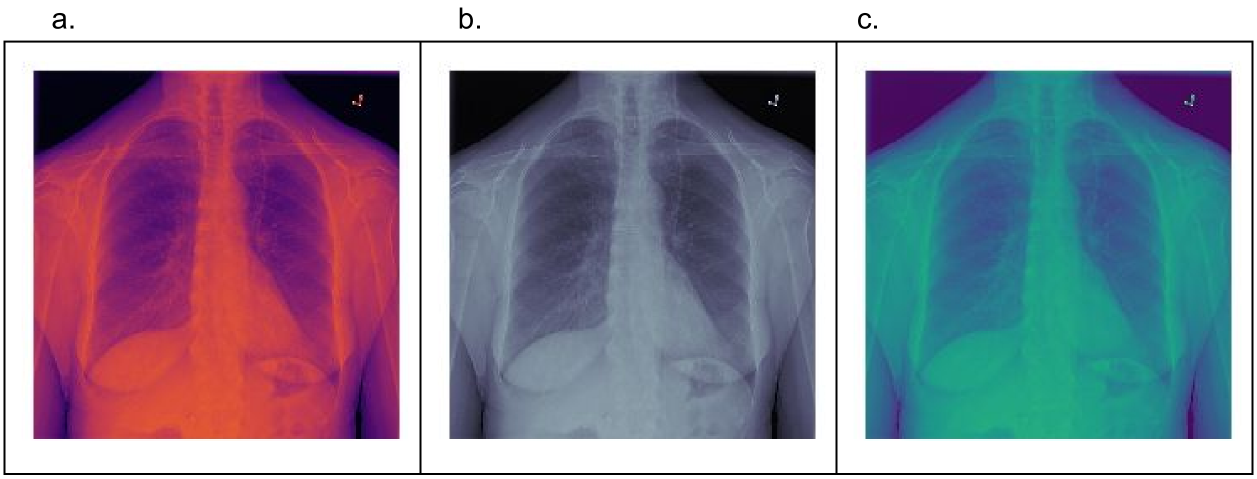


Fig: a Inferno, b bones and c viridis

Brief summary of the convolutional neural network model
We used only one epoch due to resource and time constraints. It takes nearly 7hrs to complete the training with one epoch. We are pretty sure that with the addition of more epochs we can achieve more better accuracy.

You can see the actual network summary and diagram in the project notebook. Below is the concise summary of our model.

Total params	3,771,169
Trainable params	3,767,073
Non-trainable params	4,096
Accuracy	0.96008945

Experiments for data insights and augmentation

Support Vector Machine with image metadata:
Just as an experiment we wanted to train an SVM with rbf kernel and see what are the outputs. Since the metadata contains some important information like persons age, sex, viewing angle,

modality and body parts which might be useful features. Following is the result of the 10 fold validations.

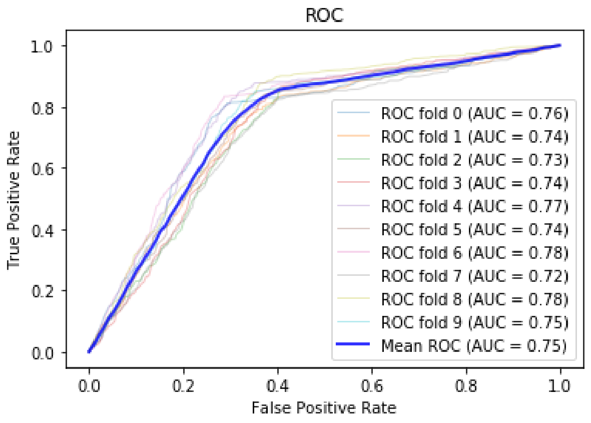
AUC = 0.7603543383903387
AUC = 0.7352565814095717
AUC = 0.7318865756588255
AUC = 0.739970282106051
AUC = 0.7748791587996797
AUC = 0.7414709017814711
AUC = 0.7813776824463083
AUC = 0.7150638596138742
AUC = 0.7753649039251859
AUC = 0.751559674605929
Mean AUC = 0.750701584668296
Accuracy = 0.7326245847176079
	

Table: AUC , ROC and accuracies for 10 fold validation with Support vector machine

Logistic Regression on image metadata

We also experimented with the logistic regression for classification. It actually doesn't make any sense here also because metadata don't have any indicator directory related to pneumonia otherwise specialists would have been using them.

To our surprise the accuracy was around 70%, which is way more than expected.

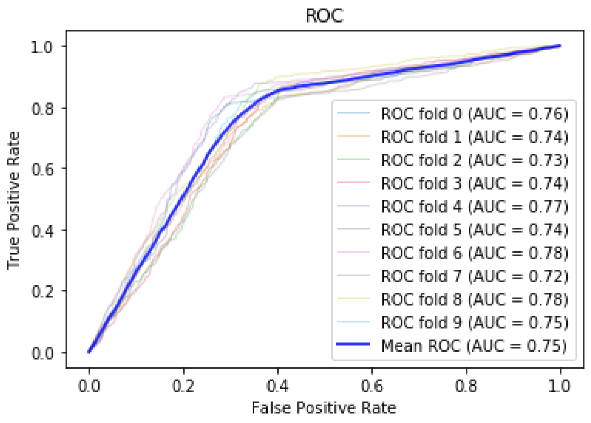
AUC = 0.7607636797016915
AUC = 0.7187631217423848
AUC = 0.7404661983916168
AUC = 0.7381497829777912
AUC = 0.7860680478312805
AUC = 0.7341965968948769
AUC = 0.7650417503314481
AUC = 0.7164905409231219
AUC = 0.7878211816654186
AUC = 0.7484251995470237
Mean AUC = 0.7495639166290029
Accuracy = 0.7327574750830564
	

Table: AUC , ROC and accuracies for 10 fold validation with Logistic regression

Conclusion

Referring to the above accuracies, our primary model based on CNN performed best on the dataset. The model is able to detect pneumonia in patient X-rays with a respectable accuracy of

96.01%. This research can be considered extremely useful to the medical professionals. The research also adds to the fact that Convolutional Neural Networks, despite their computational heaviness, are applicable to a wide range of problem statements, especially those where accuracy is a prime factor. So we are pretty confident that CNN based models can augment the specialists ability to diagnose pneumonia and help fast track the early detection.
The SVM and Logistic model although performed with a low accuracy, can still be used to filter out most of the patient X-rays. Only the 'YES' predicted X-rays can be further analysed by health professionals. Thus by using these algorithms we can save time of manually analysing all the patient X-rays. The meta data extracted from the DICOM images proved to be poor in regards to detecting Pneumonia. Given more rich data we expect the SVM to perform a lot better. Thus the fast computation of the RBF kernelized SVM can be used to an advantage to detect Pneumonia symptoms in X-rays.

Individual Task

Basar S Chowdhury

1. Performed some of exploratory analysis related to meda data extraction and image validations like missing values. Analysis of the different types of image classes and corner cases like missing lungs , heart conditions and other ailments.
2. Doing preliminary feasibility and resources needed in terms of hardware, literature and machine learning libraries.
3. Experiment with the convolutional neural network and different configurations, and data augmentations. Logistic regression and related accuracies AUC , ROC.
4. Final report compilation and code comments.

Mayur Kurup

1. Exploratory analysis with boxes and finding the central distribution of the rectangles.
2. I trained the final CNN model. Parsed the results for accuracies and other stats. This was the most time consuming part because of the long execution time of 7hrs.
3. Performed the Support Vector machine and metadata parsing for this.
4. Compiling the final tables and figures for the report.

References:

1. NIH News release: NIH Clinical Center provides one of the largest publicly available chest x-ray datasets to scientific community
2. Original source files and documents
3. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. IEEE CVPR 2017, http://openaccess.thecvf.com/content_cvpr_2017/papers/Wang_ChestX-ray8_Hospital-Scale_Chest_CVPR_2017_paper.pdf
4. Pneumonia facts WHO : https://www.who.int/news-room/fact-sheets/detail/pneumonia

5. Previous work : NCBI
6. ImageNet Classification with Deep Convolutional Neural Networks, Geoffrey E. Hinton, Ilya Sutskever, Alex Krizhevsky link
7. Keras data generator concepts : https://stanford.edu/~shervine/blog/keras-how-to-generate-data-on-the-fly
8. Learning Strict Identity Mappings in Deep Residual Networks, Xin Yu Zhiding Yu Srikumar Ramalingam, University of Utah, NVIDIA link
9. Scikit-learn : https://scikit-learn.org/stable/, Keras Documentation https://keras.io/ , Anaconda environment https://www.anaconda.com/ , Jupyter notebook https://jupyter.org/