AFRICAN CENTER OF EXCELLENCE IN DATA SCIENCE

Detection of pneumonia by chest x-ray using machine learning techniques.

Prepared by: Samuel Nahimana

Reference Number: 220001174

A Dissertation Submitted in Partial Fulfillment of the requirements for the degree of

MASTER OF SCIENCE IN DATA SCIENCE (BIOSTATISTICS)

In the college of Business and Economics

University of Rwanda

Supervisor: Dr. Melanie Fernandez Pradier.

March 2022

Declaration

I declare that this dissertation entitled Detection of pneumonia by chest x-ray using machine learning techniques is the result of my work and has not been submitted for any other degree at the University of Rwanda or any other institution.

Samuel Nahimana

Date: March 25, 2022

Signature 

Approval sheet

This dissertation entitled Detection of pneumonia by chest x-ray using machine learning techniques written and submitted by Samuel Nahimana in partial fulfillment of the requirements for the degree of Master of Science in Data Science majoring in Biostatistics is hereby accepted and approved. The rate of plagiarism tested using Turnitin is 13 % which is less than 20% accepted by the African Centre of Excellence in Data Science (ACE-DS).

Academic Supervisor: Dr. Melanie Fernandez Pradier

Head of Training: Dr. Kabano H. Ignace

Dedication

I dedicate my thesis to my friends and my family. I would like to extend my special thanks to my parent whose encouragement was outstanding and strengthened me to carry on. Thanks to the friends who have been supportive in this period of school. My fellow in church and the entire GBU family for their prayer and advice. Thanks to my classmates whose ideas, break time, and challenging each other has made me strong to finish my thesis.

Acknowledgment

Throughout the writing of this dissertation, I have received a great deal of support and assistance.I would first like to thank my supervisor, Professor Pavlos Protopapas, whose expertise was invaluable to be able to use convolutional neural networks and transfer learning. Dr. Weiwei Pan was encouraging and added value to me in all areas of writing a thesis, her correction, guidance, and the weekly meeting made this thesis possible. Dr. Melanie Fernandez Pradier was always available for any questions, and her guidance, her correction, and the recommendation from first to the end improved my level of analysis and academic writing. Your insightful feedback pushed me to sharpen my thinking and brought my work to a higher level.

I also give special thanks to the University of Rwanda especially the African Centre of Excellence in Data Science through the head of teaching who had organized and established collaboration with Harvard University. It was the basis of getting the opportunity to be supervised by Professors from a reputable university globally. I cannot finish without thanking my classmates and teammates with whom we shared skills to support each other, and encouragement.

In addition, I would like to thank my parents, for their wise counsel and sympathetic ear. You are always there for me. Finally, I could not have completed this dissertation without the support of my friends, and my colleagues, who provided stimulating discussions as well as happy distractions to rest my mind outside of my research.

Abstract

Pneumonia causes death every year worldwide, especially under five-year-old children 15 percent of death are accounted for it. Chest x-rays are primarily used for the diagnosis of pneumonia disease. However, there is a low number of trained radiologists. It is a challenging task to examine chest x-rays when there is a high number of patients, particularly in the sub-Saharan region. There is a need to improve the diagnosis accuracy and reduce radiologists' caseload. In this work, an efficient and generalizable model for the detection of pneumonia trained on chest x-ray images was developed. It could support radiologists in their decision-making process. A modern approach based on convolutional neural network (CNN) and DenseNet201 pre-trained models were used to solve this problem. Both methods are supervised learning approaches in which the network predicts the result based on the quality of the dataset used. Data augmentation techniques were used to augment the training dataset in a more balanced way. DenseNet201 model outperformed the CNN model. Finally, the model is evaluated with F1-score as it helps to bring balance in case there is a class imbalance. The final DenseNet201 model achieved an F1-score of 95.59 on the unseen data from the Guangzhou Women and Children's Medical Center pneumonia dataset and was able to generalize better to different data with an F1-score equal to 94.29 on patients with age less than 30 in the National Institute of Health (NIH) chest x-ray image dataset. The developed model is generalizable to new data, especially at young ages. Therefore, it can be used for a pneumonia diagnosis and can reduce the caseload for the radiologists.

Keywords: pneumonia; chest x-ray images; convolution neural network (CNN); DenseNet201; computer-aided diagnostics

Contents

Tables

Figures

Chapter 1: Introduction

Pneumonia is an acute inflammation of the lower respiratory tract and limits patients to breathing sufficient oxygen to reach the bloodstream. In addition to that, it is caused by viruses, bacteria, or fungi. [1] It is transmitted through airborne droplets from cough or sneezing, through blood especially during or after birth. The symptoms of pneumonia are cough or difficult breathing, fever, sweating and shaking chills, fatigue, chest pain, Nausea, vomiting or diarrhea, and confusion. [2] [3] And the most susceptible are children with not a well-developed immune system, malnourished, who come from poor families, hungry, and who live in the least developed or overpopulated areas [4] , for adults smokers are most susceptible. [5]

Pneumonia is the leading cause of death worldwide, especially in children under 5 years old who account for 15 percent. [4] Developing countries have much more cases for instance in Rwanda 19 percent of death in children are due to pneumonia. [6] [7] WHO has also taken this issue seriously and aims to reduce pneumonia mortality in 2025, especially in children under 5, to less than 3 per 1,000 births. [4]

Traditionally pneumonia is diagnosed with different exams such as medical history, physical exams, chest x-ray to examine the presence of lung inflammation, blood tests, pulse oximetry, blood oxygen tests, sputum tests, polymerase chain reaction (PCR) test, computed chest tomography (CT) scan to examine how much of lungs affected. [8] Chest radiology x-ray tests are considered the gold standard for diagnosing pneumonia. [9]

A chest radiology exam requires a radiologist expert to read and interpret the image. Nevertheless, developing countries like Rwanda and others do not have enough human resources in the health sector, especially experienced radiologist experts to interpret chest x-ray images. For instance, in Rwanda ratio of health professionals compared to all populations in 2011 was 0.72 per 1000. It is lower than the ratio recommended by WHO 2.3 per 1000. [10] There is also a low number of radiologists experts in Africa for example in 2015 was reported that in Rwanda there were only 11 radiologists trained abroad. [11]

Pneumonia is a social problem, with a high death rate, and a high caseload for radiologists, and the traditional way of treatment has not solved the problem as desired. There is a need for technology to improve traditional methods and to support the low number of radiologists. Machine

learning models can support and improves treatment. [12] We used a Convolutional neural network and deep learning models to diagnose pneumonia through chest x-ray images. Machine learning techniques are very useful nowadays in medical imaging especially convolutional neural networks and deep learning models to solve several problems in healthcare.

Convolutional neural networks are flexible models to extract features from images. [13] Pranav Rajpurkar et al reported that they developed an algorithm that exceeds practicing radiologists. [14] Stephen et al also used Convolutional Neural Network (CNN) to identify the presence of pneumonia in x-ray images. [15] And others used deep learning pre-trained models to detect pneumonia from chest x-ray images. Previous researchers have tried to solve the problem using different methods. [16] [17] But most of them focused on part of the training model and have not reported on model generalizability to different data. Assessing how the model predicts unseen data is critical especially in healthcare applications as the model intends to be used in human life.

The contribution of our research is to use a convolutional neural network and pre-trained model to detect pneumonia through x-ray images as inputs that can generalize well to new data. Different methods were applied to reduce overfitting. Images from another dataset different from training data were used to validate the generalizability of the model to new data, for it is important to learn the stability and generalizability of the model so that the success and limitation of the model may be known. We used interpretability approaches such as saliency map to assess model robustness, by inspecting what the model relies on to classify.

The output of the research will be an automated computer-assisted system that will help to detect pneumonia from x-ray images and will help to speed up the medical diagnosis of pneumonia diseases, reduce diagnosis time and increase the efficiency and reliability of pneumonia diagnosis. The result or product of the research can be used in hospitals as most hospitals have radiology machines. For example, in Rwanda, all district and referral hospitals have radiology x-ray machines in a total of 60 various hospitals. [11] Therefore, the research product will be an added value and will transform medicine from the traditional way of pneumonia diagnosis to the modern way of technology which is an automated computer-assisted system in line with Rwanda's national target for 2019-2020 in healthcare was to empower healthcare workers to incorporate technology to diagnose and treat diseases. [18]

The pre-trained DenseNet201 model outperformed the convolutional neural network (CNN) model. With the DenseNet201 model we achieved an F1-score of 95.59 to classify normal and pneumonia patients, the false-negative was low, with only 11 people out of 390 in the test set, it was also able to generalize better to a new data with an F1-score of 86.28 and on patients with age less than 30 with F1-score 94.29. Age is an important factor influencing model performance. Therefore, the model is reliable to deploy to new data, especially for younger ages. By changing the threshold to the higher boundary and lower boundary at a range with no false positive and false negative, the model could reduce the caseload of the radiologist by 55 percent.

CNN model achieved an F1-score of 94.74 to classify normal and pneumonia patients, the falsenegative was low, only 12 people out of 390 in the test set, it was also able to generalize better to a new data with an F1-score of 86.28 and on patients with age less than 30 with F1-score 91.75. Age is an important factor influencing model performance. Therefore, the model is reliable to deploy to new data, especially for younger ages. By changing the threshold to the higher boundary and lower boundary at a range with no false positive and false negative, the model could reduce the caseload of the radiologist by 45 percent.

1.1. Research objectives

Our main objective is to detect pneumonia with chest x-rays images using machine learning techniques. These are specific objectives to achieve the main one. Firstly, to train a model to identify chest x-ray images from patients with pneumonia against healthy ones. This means that we want a trained model that can detect whether the images are pneumonia or not. Secondly, to compare a convolutional neural network (CNN) with a pre-trained model and also to learn boundary probabilities where the model makes the right prediction with no false predictions. So that the radiologists may intervene if the model is not sure but if the model is sure there would be no need for intervention. It will help to reduce the caseload of radiologists. Thirdly to transfer a trained model to different data to learn the generalizability of the model. The model able to do that is more reliable and more likely to be deployed in real life.

These are the research questions that will help us to achieve our objectives when answered. Firstly, does the trained model detect pneumonia efficiently? secondly, does the CNN model have better accuracy than pre-trained models, and at which range of prediction probabilities does the model

make the right prediction? Thirdly does the model generalize better to new data? when does the model generalize better?

The rest of this work is organized as follows. Chapter 1 deals with the background of the research. Chapter 2 deals with some related work. In Chapter 3, we describe our proposed methods. Chapter 4 presents some results obtained using our proposed models and interpreting the results. Chapter 5 includes a discussion of the findings. The conclusion and recommendation are given in the last section.

1.2. Definition of terms

Learning Rate: It is a tuning parameter in an optimization algorithm that determines the step size at each iteration while moving toward a minimum loss function.

Epoch: This is a hyper-parameter that determines how many times the learning algorithm will work throughout the training set.

Batch size: Batch size is a hyper parameter that determines the number of samples to process before updating the internal parameters of the model.

Chapter 2 Background:

2.1. Background of convolution neural network

2.1.1. Neural Network

Neural networks (NNs) are a collection of connected units called nodes and layers to process data and produce results in the desired form. The concept itself is inspired by the human brain as neurons of the human brain also process the amount of information and understand the information. [19] Neural networks have an incredible capability to extract meaning from complex or unspecific data, NNs can be used to extract patterns and detect trends too complex to be examined by humans or other computing techniques. Neural networks are composed of three parts which are input, hidden, and output layers.

Input units are raw information fed into the neural network and are mostly called the input layer. The second is the hidden layer which applies weights to the inputs and performs nonlinear transformations. The third unit is output which is the output according to the input and hidden layer. Convolution neural networks are extensions of neural networks.

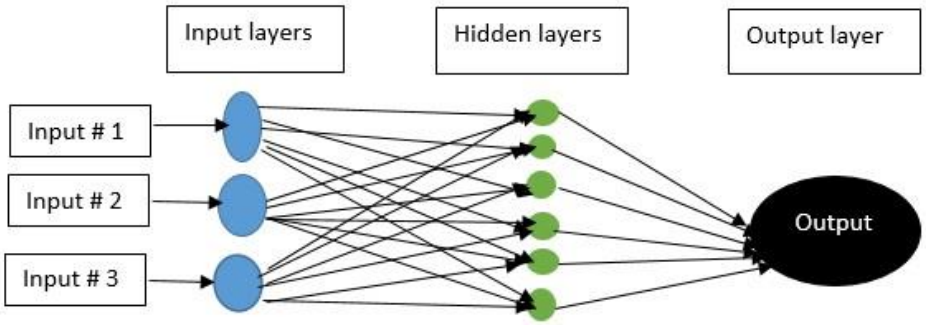


Figure 1: Neural network

2.1.2. Convolutional neural network

A convolutional neural network (CNN) is a sub-class of neural networks built of a convolutional layer, pooling layer, softmax layer, and final output. CNN has two main parts. The first part is Feature Extraction which helps to extract, separate and identify various features of the image. It consists of convolutional and pooling layers. The convolutional layer each holds a collection of filters, and kernels, which are rectangles of numbers. The pooling layer examines the elements that form an input block, calculates only one of those values, and stores that single value as output instead of all of the input elements.

The most common use of the pooling layer is to reduce the size of its input. Batch normalization was used as regularization on each batch of data, and resulted in an average of 0 and a standard deviation of 1 to reduce weights from growing too large. The dropout rate helps to reduce overfitting as dropping out some nodes according to the dropout rate from the neural network. The second part is the fully connected layer that uses the extracted features to predict the image class using the activation function such as softmax, sigmoid, relu, and hyperbolic tangent (tanh). [20]

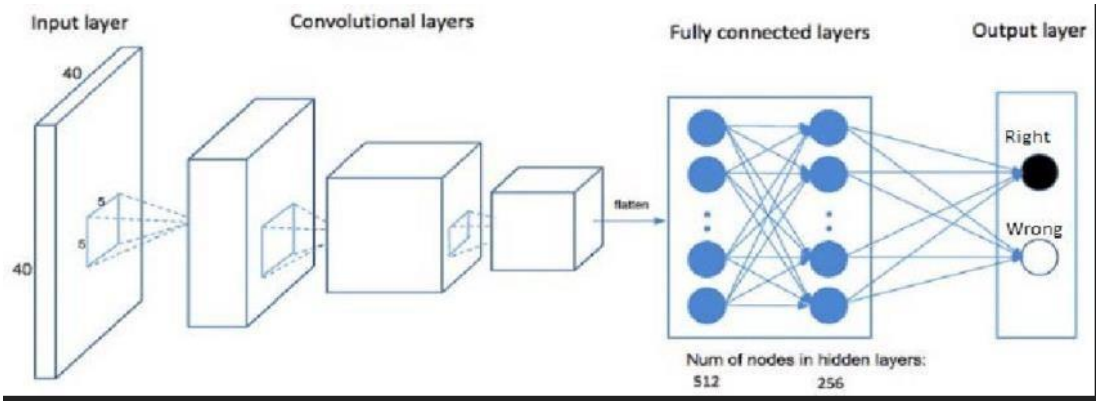


Figure 2: Convolutional Neural network structures

2.2. Transfer learning background

Transfer learning is machine learning that focuses on storing knowledge gained in solving a problem and using that knowledge to different but related problems. [21] For instance, knowledge of classifying cars could apply to classify motorcycles, and in daily life learning to play guitar can facilitate learning to play piano, and learning mathematics can facilitate learning statistics. The use of transfer learning is motivated by the fact that people can use knowledge learned in previous problems or situations to solve new problems better and faster. [22] Transfer learning improves the prediction especially when the dataset is small. [23] For such cases using a pre-trained model trained in big datasets could improve classification performance.

These are three major structures of transfer learning. The first type is ConvNet as a fixed feature extractor where we use ConvNet pre-trained model, only remove the last fully connected and train only the last fully connected layer for classification and prediction. The second way is to train only the last few convolutional neural networks and fully connected layers. The third type is to predict or classify using a pre-trained model without any training. In our case, the first type was used.

2.2.1. DensNet201

A dense convolutional neural network (DenseNet) is a pre-trained model with the structure of connecting each layer to every other layer in a feed-forward way which means that feature maps of all previous layers are inputs to all following layers. DenseNet has the following advantages compared to other models, they reduce the vanishing gradient problem, Enhances the spread of features, encourage the reuse of features, and reduce considerably the number of parameters. [24] We wanted to compare CNN built from scratch and pre-trained to learn the best model in terms of performance metrics and generalizability. DenseNet201 was chosen for it was outperforming other models in the previous work [16], And its strengths in reducing vanishing gradients problems.

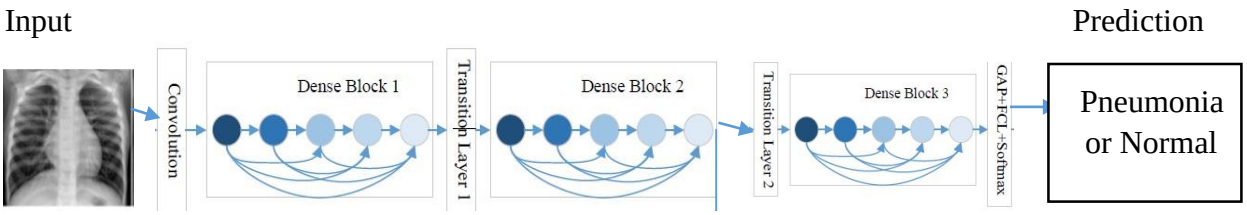


Figure 3: DenseNet201

DenseNet201 is composed of the input layers, convolutions, three dense blocks, and transitional layers that contain batch normalization, with one by one global average pooling (GAP), one by one convolution layer, and with stride made of two. [25]

2.3 Performance Metrics

In classification, we need a classifier to measure the success of the model. There are several classifier estimates those are accuracy, F1-score, recall, precision, and confusion matrix. Accuracy is the percentage of the total number of correctly classified compared to the total number of subjects. Though the accuracy is not always the most suitable metric especially when there are imbalanced data, sometimes is biased toward the majority class. Therefore, other metrics will help to understand better the model performance as a compliment.

Accuracy = (**total number of correctly classified**)/(**total number of subjects**).

Precision is the percentage of true positive compared to all predicted positive and also called positive predictive value, Precision= $TP/(TP + FP)$.

The recall is the rate of true positive compared to all actual true positive and also called sensitivity.

Recall= $TP/(TP + FP)$. F1-score is the harmonic mean of precision and recall, it brings balance between recall and precision. F1-score= $(precision * recall)/(precision + recall) * 2$

A confusion matrix is a tabular representation of actual versus predicted values that contain true positive, true negative, false positive, and false negative. True positives (TP) are defined as the number of positive cases that were predicted positively. True negatives (TN) is the number of negative cases that were predicted negative. False positives (FP) are the number of negative cases that were predicted positively. False negatives (FN) are the number of positive cases that were predicted negative. And threshold as decision boundary probability to classify the groups.

Table 1: Confusion Matrix

Status	Predicted Negative	Predicted Positive
Actual Negative	True Negative	False Positive
Actual Positive	False-negative	True Positive

2.4. Saliency maps

The saliency map: is a CNN visualization technique that helps us to recognize the area CNN is looking at, at the time of classification and highlight that area where the pixels are most contributing to predicting a particular class. A saliency map is the derivative of the class score in respect of the input image. Saliency maps help to diagnose the failure of the model. (Selvaraju et al., 2020) There are different techniques of visualization among them are Gradient-based backpropagation, deconvolutional networks, guided backpropagation algorithm, class activation maps, Grad-CAM, and guided Grad-CAM. We used Grad-CAM to visualize the pixels where a model is focusing to predict pneumonia or normal. Grad-CAM is Gradient-weighted Class Activation Mapping to visualize saliency map. The formula of Grad-CAM.

$$\alpha_k^c = \frac{1}{Z}\sum_{ij} \frac{\sigma y^c}{\sigma A_{ij}^k}$$

$\sum_{ij}$ Summing of all elements of the k activation map. α_k^c is the importance of feature map k for class c. Z is some normalization. A_{ij}^k is the i,j element of the k activation map of the last activation layer. σy^c is the logit for class c.

$$LGrad - CAM = ReLU(\sum k\ \alpha_k^c\ A^k)$$

We computed the gradient of the logits y^c of category c concerning the activation maps of the last convolutional layer, and average the gradients in each feature map to give us a score of importance. After feature maps are combined and applied ReLU activation to its total.

Chapter 3: Related work

Previous related works have tried to solve this problem by using different methods and techniques. Traditional pneumonia is diagnosed by radiologists examining the chest x-ray images and computed tomography in case there is a complication. [9]

Stephen et al used a convolutional neural network (CNN) to identify the presence of pneumonia in x-ray images and reported a training accuracy of 0.94814. But test metrics were not reported to assess the overfitting of a model and its stability. [10] Hashmi and others used different pre-trained deep learning models such as ResNet18, DenseNet121, InceptionV3, Xception, and MobileNetV2 to detect pneumonia from x-ray images, but they didn't report on sensitivity analysis, model stability, and model generalizability to new data. [16] Elshennawy and Ibrahim developed the model with ResNet152V2 and MobileNetV2, a Convolutional Neural Network (CNN), and a Long Short-Term Memory (LSTM) and reported ResNet152v2 was outperforming others.

However, metrics on the test set were not reported to validate the model generalizability to different images not used in training. [17] Hussain and others used the MobileNet deep learning model reported that is good for constrained resources devices and achieved a higher training accuracy of 99.1%. Yet test metrics were not reported. [26] Training accuracy does not show any information about model generalizability to new data. It is possible to attain higher training accuracy while models make worse predictions to other data. El Asnaoui and others used Deep Convolutional Neural Network (DCNN) architectures for binary classification of pneumonia images and different pre-trained model such as VGG16, VGG19, DenseNet201, Inception_ResNet_V2, Inception_V3, Resnet50, MobileNet_V2 and Xception, and Resnet50 had higher accuracy compared to others with training accuracy of 96.61 percent and F1-score equal to 96.67. They reported training accuracy only without comparing it with test metrics to learn if the model overfitted. Therefore, they missed part of model generalizability to new data. [13]

There are other similar related works where CNN or pre-trained model was used to classify chest x-ray images into more than two groups for instance to classify them into normal, pneumonia, or covid-19. And to localize the areas of infection in the thorax by using a transfer learning model, and an average accuracy of 0.97 was achieved [27]. To classify Chest x-ray images of patients into 4 classes normal, pneumonia, tuberculosis, and Covid-19 at an accuracy of 98.9 percent by using the DenseNet-161 pre-trained model and to classify the severity of Covid-19. [28]

And to classify x-ray images into three classes viral, bacterial pneumonia, or Normal, and visualize the areas of infection by using a deep learning approach. [29]

Convolution neural network has also been used in other medical image disease detection and classification. [30] Because of its capability to extract and select information such as edges, colors, and textures in an image. [31] It has also been used in breast cancer detection. [32] [33] dental image diagnostics. [34] Brain tumor classification [35], and many others. The pre-trained model has performed with higher accuracy than a convolutional neural network built from scratch because of its strong feature extraction, especially for a model built on a small dataset.

Chapter 4: Data Exploration

4.1. Data Sources

We used public datasets hosted by Kaggle online data science platform. [36] Dataset has 5,863 chest x-ray images with a position view of anterior-posterior. [37] In a dataset, there are 2 categories pneumonia and normal. Data were collected from Guangzhou Women and Children's Medical Center. Methods of retrospective cohorts of pediatric patients one to five years old were used to collect images. The data were pre-cleaned before, by removing images with low quality, and the diagnosis of images was rated by two expert physicians and checked by the third expert as an evaluator. This assures the quality of data used to train a model.

We used also the second dataset that serves as validation to validate the model performance. Dataset is called NIH Chest x-ray and is comprised of 112,120 x-ray images with disease labels from 30,805 unique patients. [38, 39] Not all images in the dataset were used, only samples of images were used. The process and basis of selection are explained in detail in the sections ahead.

4.2. Data Exploration

4.2.1. Training Dataset

Table 2: Dataset folder

Dataset	Train set	Validation set	Test set	Total
Number of images	5216 (89.07 %)	16 (0.27 %)	624 (10.65 %)	5856

The data were divided into training with 89.07 percent of total images, test with 10.65 percent of total images, and validation set with 0.27 percent of total images. Train and validation datasets were combined to make a train set. For model validation, a different dataset was used there was no need to duplicate the process.

4.2.2. Class representation

Table 3: Class representation

Category	Normal	Pneumonia	The ratio of Normal to Pneumonia
Images	1341	3875	1 to 3

There is a significant difference in the number of pneumonia images from normal images with the ratio of normal to pneumonia images equal to 1 to 3. Therefore, there is a class imbalance. Class imbalance sometimes brings some complications to the model to generalize on the new data. When there is a class imbalance model makes a better prediction for the majority class but fails for the

minority class, as it has only a few instances of minority classes. This was taken into consideration by applying methods to reduce the effect of class imbalance.

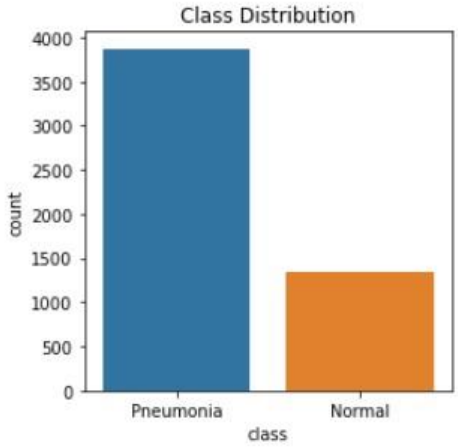


Figure 4: Class Distribution

4.2.3. Image representation

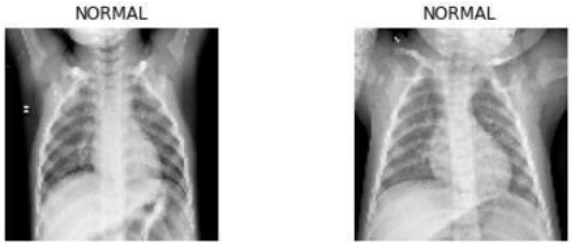


Figure 5: Normal images

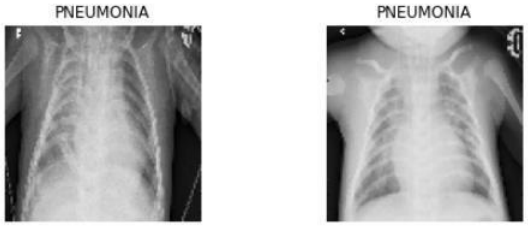


Figure 6: Pneumonia Images

The image of pneumonia shows a much lighter area than normal images which might be fluid in the lungs as a chest x-ray test helps to show fluid in the chest as a presence of pneumonia. The figure above is a sample representation of images, used to build a pneumonia detection model, by inputting normal and pneumonia images so that the model learns to classify both images.

4.2.4. Average Image

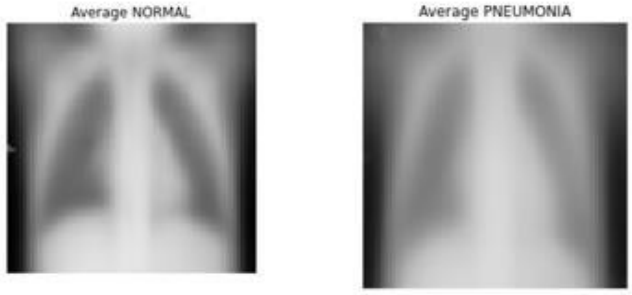


Figure 7: Average image

The average of images was calculated from all images. Pixel values of each image were converted into a matrix so that operations of images may be possible. Average values of each pixel were calculated in all observations. Visually looking at the two pictures there is a slight difference between pneumonia and normal. The pneumonia patient image has a larger lighter area than normal.

4.2.5. Contrast Between Average Images

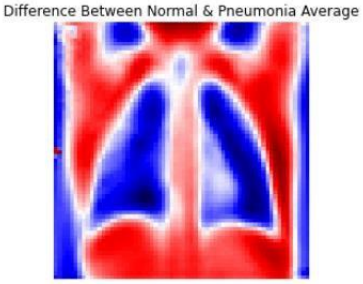


Figure 8: Contrast between average Images

By using an average of images, we computed the contrast between pneumonia and normal image to show the difference between the two averages. We subtracted the average of pneumonia from an average of normal images to get their contrast. The lighter space in the image shows the difference. This means that pneumonia images and normal images differ in terms of image structure.

4.2.6. Variability

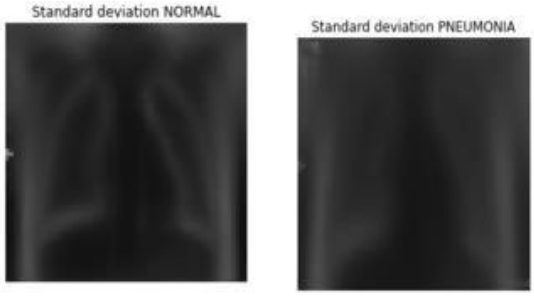


Figure 9: Standard deviation of images

We used standard deviation to show which image is most valuable. It was calculated as the mean deviation of each pixel from the mean. Visually an image of a pneumonia patient, there is more variability different from normal images where a clearer image of the thorax is slightly visual. There is a clear difference between normal and pneumonia x-ray images whether by image representation, average, and standard deviation of image. Therefore, we can build a model to differentiate between normal and pneumonia x-ray images. Because of class imbalance we used methods of weighting class to increase the weight of the minority class.

4.2.7. Validation Dataset

Table 4: Diseases Distribution

Diseases	Frequency
Infiltration	19894
Effusion	13317
Atelectasis	11559
Nodule	6331
Mass	5782
Pneumothorax	5302
Consolidation	4667
Pleural Thickening	3385
Cardiomegaly	2776
Emphysema	2516
Edema	2303
Fibrosis	1686
Pneumonia	1431
Hernia	227
No finding	60361

The national institute of health (NIH) dataset was composed of images with 14 different thorax diseases and another class of health images. It was not well cleaned as the hospital data are. Using

the NIH dataset, we can know how the model can predict the different datasets. The table below shows the distribution of disease in an NIH dataset.

The dataset includes the number of thorax diseases such as Infiltration with 19894 images, Effusion with 13317 images, Atelectasis with 11559 images, Nodule with 6331 images, Mass with 5782 images, Pneumothorax with 5302 images, Consolidation with 4667 images, Pleural Thickening with 3385 images, Cardiomegaly with 2776 images, Emphysema with 2516 images, Edema with 2303 images, Fibrosis with 1686, Pneumonia with 1431 images, Hernia with 227 images, No finding or normal images with 60361 images, some images was having more than one diseases and 60361 was only images without any presence of any 14 diseases.

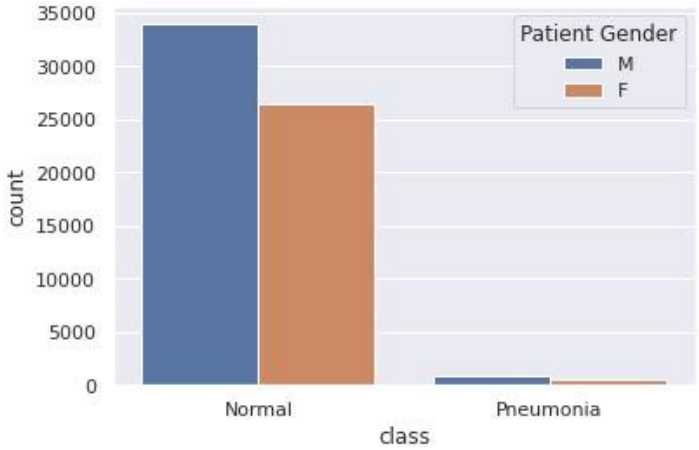


Figure 10: Class distribution by gender

Our interest is pneumonia and normal images, therefore we selected only images with pneumonia diseased and normal categories, this figure shows the big difference between normal and pneumonia images, therefore there is a class imbalance in both gender.

4.2.8. Position of images

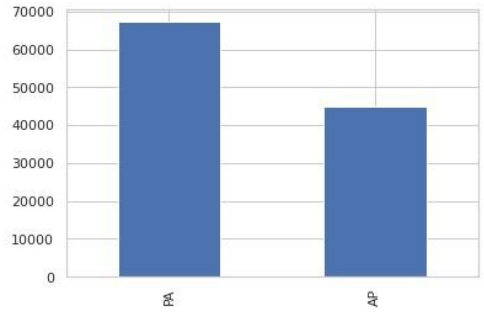


Figure 11: Position view of images

The NIH dataset has images with a two-position view anterior-posterior (AP) and posterioranterior (PA). The view position can affect any shape of images, so in the analysis, we have separated

images by view position to understand any effect on classification. As in another dataset of training, we only have one position view AP.

4.2.9. Age distribution of patients

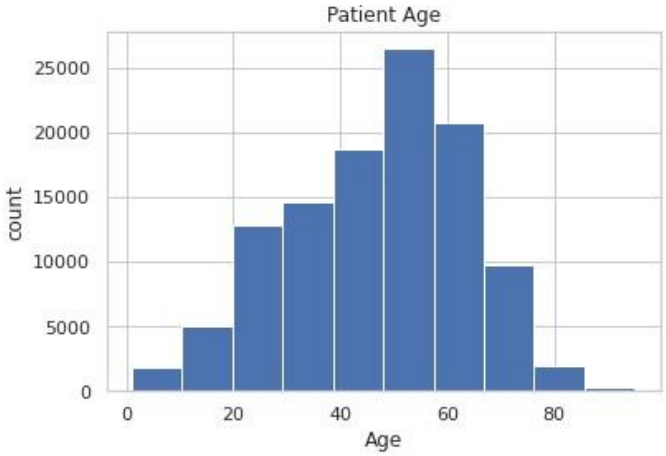


Figure 12: Age Distribution of patients

The NIH dataset has images of patients with ages from 0 to 100 years old which is different from the training dataset which has only images of 0 to 5 years old. Therefore, using the NIH dataset to validate the model will help us to know how a model will predict the patients of different ages regardless that it was trained on younger patients' images.

4.3. Data Preprocessing

4.3.1. Training dataset

We had images in three folders train, test, and validation which we added to the training set, for we used another dataset for validation. Then we imported data from directories of kaggle and resized the images to have the same size of 224 by 224. We created a list with images and classes, after that we separated the images column from the classes column. To make any operation possible like data augmentation, the list of images and list of classes were converted to numpy, this preprocessing was also applied to the test set with normalization to have the same format.

4.3.2. Data augmentation

We augmented data to change the image view so that we might have different variability so that the model may learn all possible varieties of images. First, we rescale images by 1/255, we used a shear range equal to 0.2, zoom range equal to 0.2, horizontal flip we made it false, brightness range of [0.2, 1.0], and validation split of 0.2.

4.3.3. Validation dataset (NIH data)

We performed different transformations on this dataset to have the same format as the data used to train a model. First, we have selected two classes from 15 classes, normal and pneumonia images. We removed people with an age older than 100 years old. We deleted all other variables and remained with 5 columns which are image path and classes with image index, gender, age, and position view. Because of class imbalance, there was a large number of normal images than pneumonia images, therefore we sampled some number of normal images equal to 1,207 which is 2 percent of all images, and pneumonia images equal to 1430.

When we used some images to predict. Later we found that the position view of images was affecting the prediction. Then we selected normal images with a position view of the Posterior Anterior, for pneumonia with a position view of the Anterior-posterior. After that, we remained with the dataset of 793 normal images, and 800 pneumonia images. Then we read the images from a directory, resized images to 224 by 224, transformed images into numpy array and rescaled the images by 1. /255.

Chapter 4: Methodology

4.1. Study Design

We inputted chest x-ray images transformed to the same scales and augmented them to have image varieties. Then we trained the model to classify diseased and healthy x-ray images. To validate model generalizability, we inputted x-ray images from another dataset to test its generalizability to new data.

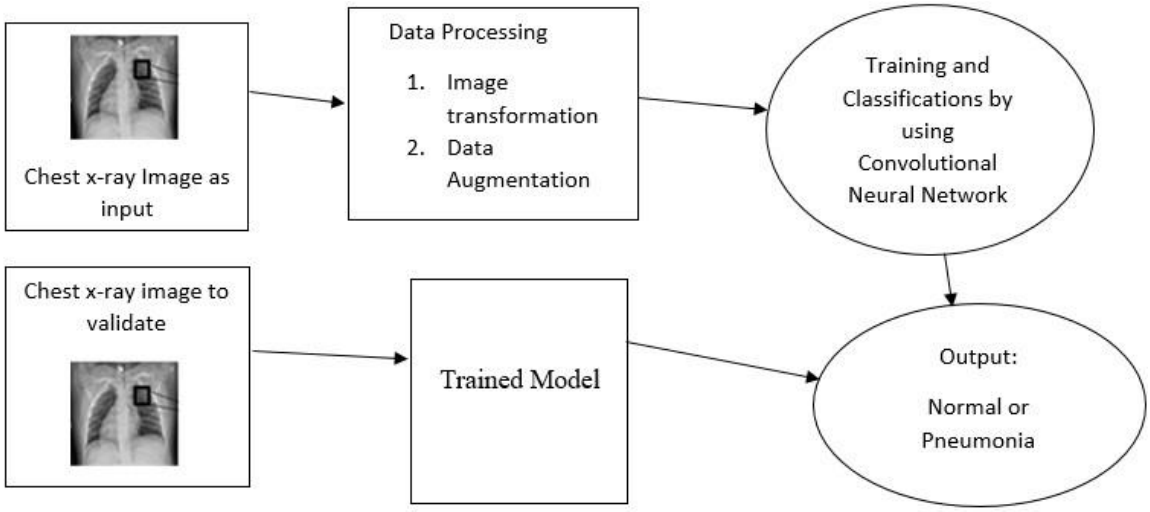


Figure 13: Conceptual framework

4.2. Methods

We used a convolution neural network and DenseNet201 as defined in Chapter 1. CNN structure was composed of 5 blocks, each block has two convolution layers, we applied Batch Normalization and dropout rate different at each layer, max-pooling of 2 by 2, a stride of 3 by 3, filters from 16 to 512 same filter by block, and used sigmoid activation function to classify into two classes. We optimized the model with Adam optimizer and used different learning rates between 0.000007 to 0.01 but the best is 0.0005. We used binary cross-entropy as a loss function. We used accuracy as a metric. We fitted the model with 55 epochs and 256 batch sizes.

CNN requires a big dataset for better accuracy. Because our dataset is not large, we also used a pre-trained model, which initially was trained on a big dataset to increase the model performance. Therefore, we used DenseNet201 pre-trained before, and we trained only the fully connected layer, so we used DenseNet201 architecture until its last convolution layer. Then we added 3 dense layers with nodes from 256, 128, and 64. And we used the dropout rate to reduce overfitting with a rate of 0.2 and 0.3 respectively, and also we added the last layer with 1 dense node for binary

classification with a sigmoid activation function. We optimized the model with Adam optimizer, and we used different learning rates between 0.00001 to 0.0012 but the best is 0.0009. We used binary cross-entropy as a loss function. We used accuracy as a metric. We fitted the model with 55 epochs and 256 batch sizes.

The training dataset and validation dataset both have a class imbalance. Class imbalance affects overall model prediction probabilities to be biased toward the majority class. To avoid this case, we have used class weight balance methods to avoid any model bias toward the majority class. Class weight is calculated as the total number of samples in the dataset divided by the total number of samples in respective classes multiplied by the number of classes.

$$w_j = \frac{samples}{(classes * samples_j)}$$

w_j is the weight of class, samples are the total number of samples, classes are the number of classes, and samples j are the total number of samples for each class.

The experiments were evaluated by accuracy to understand the number of cases rightly predicted. But accuracy is not a good metric in case there is class imbalance. F1-score was used as it helps to bring the balance between two classes and to reduce the effect of class imbalance, recall as sensitivity rate, and precision as positive predictive value. Confusion matrix to define the number of true positive, true negative, false positive, and false negative. Also, the threshold was used as a decision probability to split the classes. Different thresholds were used to learn the range of probabilities the model makes the right prediction with no false positive and false negative.

To learn the interpretability and stability of the model Grad-CAM visualization was used to highlight the pixels' model used to classify a specific area of its location. Visualization of the model region of interest helps to interpret the success or failure of the model depending on the area of focus. Therefore, saliency maps help to diagnose the failure of the model [40].

Chapter 5: Experiments and results

This section contains all the experiments and results performed. We will use different metrics to define success such as accuracy, recall, precision, F1-score, and confusion matrix as defined in chapter 2, but our emphasis is F1-score and confusion matrix as both account for imbalanced classes. The performance metrics for train, testing, and validation on the different datasets are presented in tables, figures, and saliency maps to visualize the region of interest.

5.1. Performance Metrics

Table 5: Performance Metrics

Model	**Dataset**	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-score**	**Train accuracy (%)**
DenseNet201	**Children data**	94.39	94.04	97.17	95.58	98.49
CNN		93.26	92.64	96.92	94.73	97.0
DenseNet201	**NIH prediction**					
	< 20 Years of Age	87.93	85.05	98.66	91.35	
	< 30 Years of Age	93.13	90.47	98.44	94.29	
	all ages	84.16	76.9	98.25	86.26	
	Gender					
	Male	85.57	78.79	98.92	87.71	
	Female	85.42	78.52	97.33	86.92	
CNN	**NIH prediction**					
	< 20 Years of Age	86.61	82.95	97.33	89.57	
	< 30 Years of Age	89.75	86.3	97.93	91.75	
	all ages	84.78	77.62	97.125	86.28	
	Gender					
	Male	88.91	83.95	97.4	90.18	
	Female	82.11	74.31	96.74	84.06	
DenseNe201	**Another position view of images AP on normal and PA on pneumonia**	33.24	44.18	50	46.91	
CNN		34.034	43.87	34.15	38.39	

The model was trained on 89 percent of the data to learn to detect pneumonia with chest x-ray images, and 11 percent was used to test the generalizability of the model to the unseen data. When the model performs better on the training set but differently with lower performance metrics on the test set the model overfits. To reduce model overfitting different methods were used to avoid it. Different class weight was applied by each class, with the minority higher weight which is the normal class equal to 1.94, and with the majority class low weight which is equal to 0.67 to increase the weight of the small class and to penalize the weight of the higher class. Other methods such as data augmentation, learning rate, dropout rate, batch normalization, and regularizers were used to reduce model overfitting.

The DenseNet201 model accuracy on the test set is 94.39, precision is 94.04 percent, recall is 97.17, F1-score is 95.58, while training accuracy was 98.49 percent, CNN model accuracy is 93.26 percent, precision is 92.64 percent, recall is 96.92 percent, F1-score is 94.73, while training accuracy was 97.0 percent.

One of the objectives is to test the generalizability of our model to the different data, we used images from the NIH dataset to make the prediction. The DenseNet201 model prediction on the patient with less than 20 years old accuracy is 86.61 percent, precision is 85.05 percent, recall is 98.66 percent, and F1-score is 91.35. On patients less than 30 years of age accuracy is 93.13 percent, precision is 90.47 percent, recall is 98.44 percent, and F1-score is 94.29. For all patients accuracy is 84.16 percent, precision is 76.9 percent, and recall is 98.25 percent.

CNN model prediction on patients with age less than 20 years' accuracy is 87.2 percent, precision is 82.95 percent, recall is 97.33 percent and F-1 score is 89.57 percent. In patients with age, less than 30 years' accuracy is 89.75 percent, precision is 86.3 percent, recall is 97.93 percent and F-1 score is 91.75 percent. On patients of all age groups, accuracy is 84.78 percent, precision is 77.62 percent, recall is 97.125 percent and F-1 score is 86.28 percent.

The DenseNet201 model prediction on male patient accuracy is 85.57 percent, precision is 78.79 percent, recall is 98.92 percent, and F1-score is 87.71. On female patients accuracy is 85.42 percent, precision is 78.52 percent, recall is 97.33 percent, and F1-score is 86.92. CNN model prediction on male patient accuracy is 88.91 percent, precision is 83.95 percent, recall is 97.4 percent, and F1-score is 90.18. On female patients accuracy is 82.11 percent, precision is 74.31 percent, recall is 96.74 percent, and F1-score is 84.06.

The model performance on a set of images selected with anterior-posterior (AP) position view of normal images, and posterior-anterior (PA) position view of pneumonia images had prediction with an accuracy of 33.24 percent with the DenseNet201 model and 34.03 percent with the CNN model. It is not a better prediction, for that reason, another position view was used PA on normal and AP on pneumonia which had a better prediction.

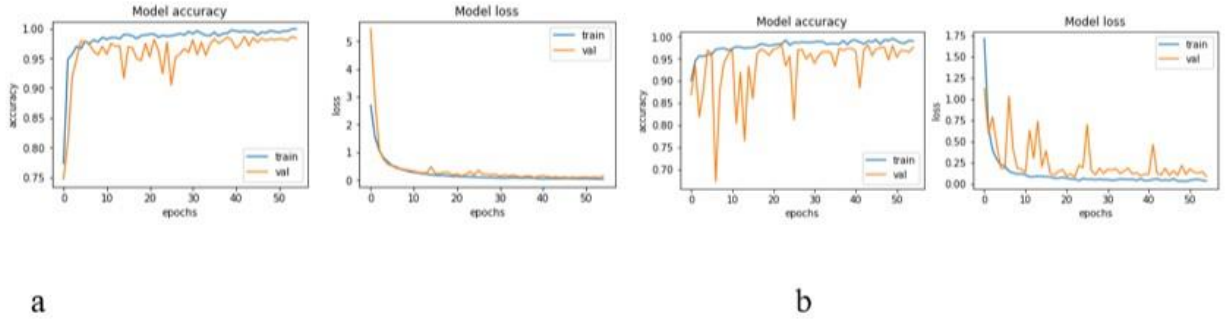


Figure 14: a: Trace plot of DenseNet201 b: Trace plot of CNN

The model to make the right prediction must converge. We know it by observing the trace plot of model loss, our model trace plot converged when the loss reaches its minimum state and does not decrease any more. both CNN and DenseNet201 model as shown in figure 14 has reached their minimum loss, therefore have converged.

5.2. Confusion matrix by model and dataset

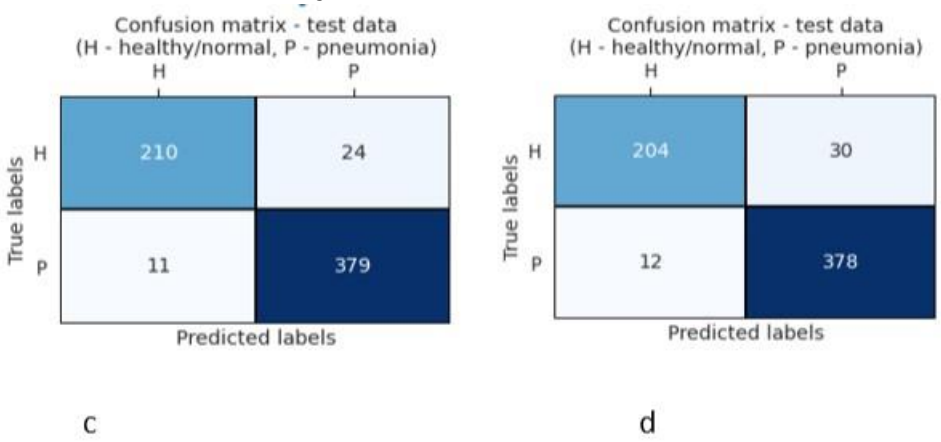


Figure 15: c: Children data set DenseNet201 d: Children data set CNN

Figure 15 shows the confusion matrix as the representation of true negative, false positive, false negative, and true positive. DenseNet201 model prediction on the test set out of 234 normal images 210 were correctly predicted, 24 were incorrectly predicted to be pneumonia, out of 390 pneumonia images 379 were correctly predicted but 11 were incorrectly predicted to be normal.

CNN model prediction on the test set out of 234 normal images 204 were correctly predicted, 30 were incorrectly predicted to be pneumonia, out of 390 pneumonia images 378 were correctly predicted but 12 were incorrectly predicted to be normal.

5.2.1. Confusion matrix of DenseNet201 model prediction with NIH data

5.2.1.1. Stratified by age

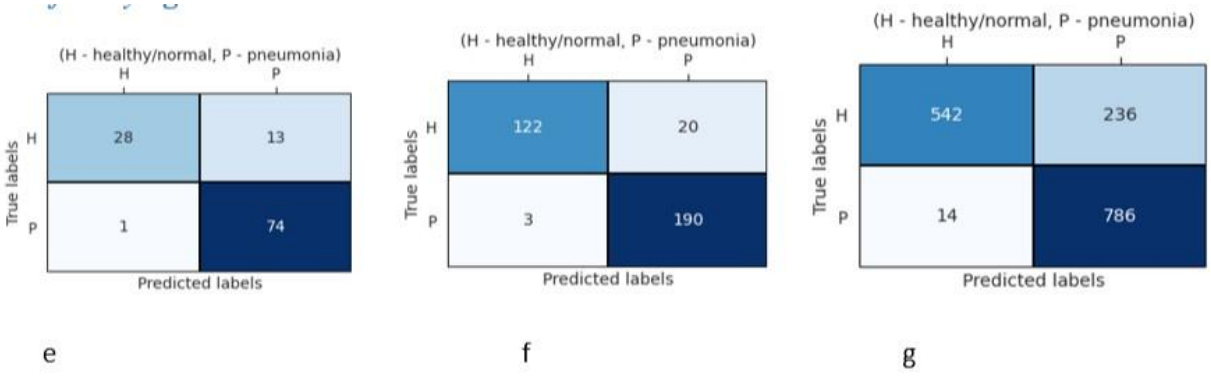


Figure 16: e: Less than 20 years old f: less than 30 years old g: all age group

5.2.1.2.Stratified by gender

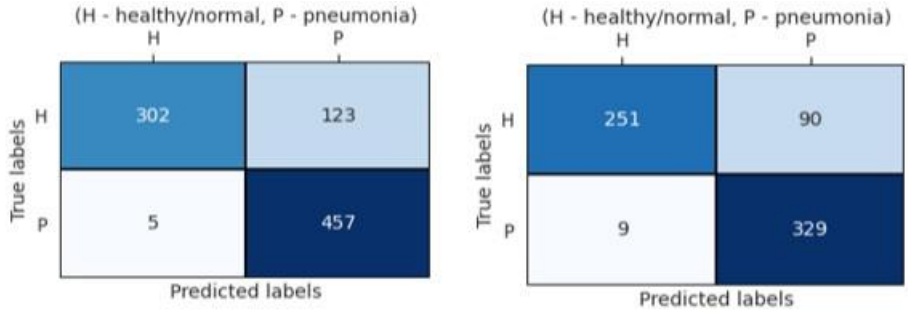


Figure 17: h: Male i: Female

These are the confusion matrices of model prediction on the NIH dataset, we predicted by age groups according to figure 16 to learn how models behave to different age groups and by gender in figure 17. DenseNet201 model prediction on the patient with less than 20 years of age out of 41 of normal images 28 were correctly predicted, 13 were incorrectly predicted to be pneumonia, out of 75 of pneumonia images 74 were correctly predicted but 1 was incorrectly predicted to be normal. On the patient with less than 30 years of age out of 142 normal images, 122 were correctly predicted, 20 were incorrectly predicted to be pneumonia, and out of 193 pneumonia images 190 were correctly predicted but 3 were incorrectly predicted to be normal. On all patients out of 778 normal images, 542 were correctly predicted, 236 were incorrectly predicted to be pneumonia, and

out of 800 pneumonia images 786 were correctly predicted but 14 were incorrectly predicted to be normal.

With male patients Out of 425 normal images, 302 were correctly predicted, 123 were incorrectly predicted to be pneumonia, and out of 462 pneumonia images 457 were correctly predicted but 5 were incorrectly predicted to be normal. In female patients Out of 341 normal images, 251 were correctly predicted, 90 were incorrectly predicted to be pneumonia, and out of 338 pneumonia images 329 were correctly predicted but 9 were incorrectly predicted to be normal.

5.2.2. Confusion matrix of CNN model prediction with NIH data

5.2.2.1. Stratified by age

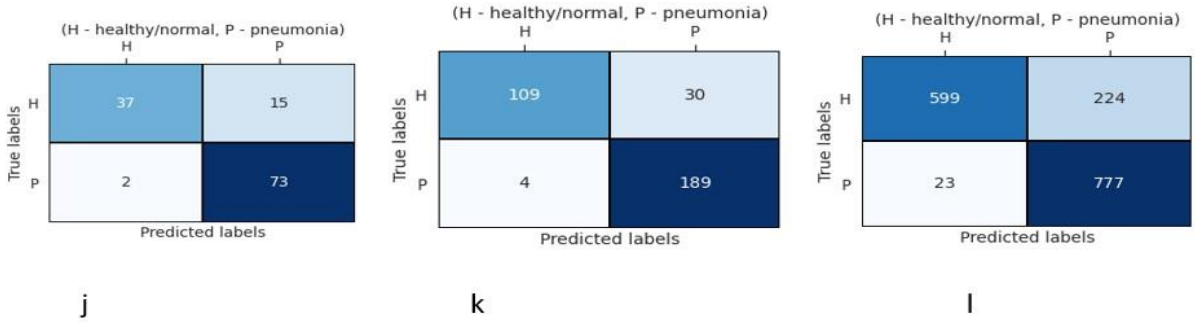


Figure 18: j: Less than 20 years old k: less than 30 years old l: all ages

5.2.2.2. Stratified by gender

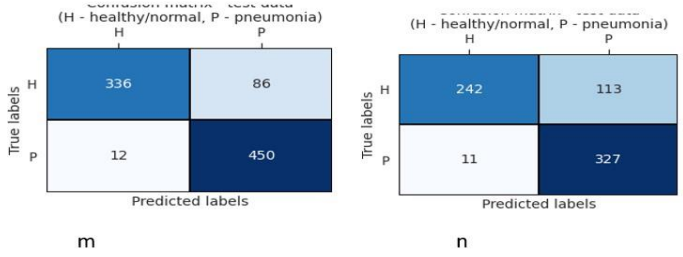


Figure 19: m: Male n: Female

Convolutional neural network (CNN) model prediction on the patient by age according to figure 18, patient with age less than 20 years of age out of 52 of normal images 37 was correctly predicted, 15 were incorrectly predicted to be pneumonia, out of 75 of pneumonia images 73 were correctly predicted but 2 were incorrectly predicted to be normal. CNN model prediction on the patient with less than 30 years of age, out of 139 normal images 109 were correctly predicted, 30 were incorrectly predicted to be pneumonia, out of 193 pneumonia images 189 were correctly predicted but 4 were incorrectly predicted to be normal. CNN model prediction on the patient of all age

groups out of 823 normal images 559 was correctly predicted, 224 were incorrectly predicted to be pneumonia, out of 800 pneumonia images 777 were correctly predicted but 23 were incorrectly predicted to be normal.

Figure 19 shows the confusion matrix of the CNN model by gender. Confusion matrix of CNN model of male patients out of 422 normal images, 336 were correctly predicted, 86 were incorrectly predicted to be pneumonia, out of 462 pneumonia images 450 were correctly predicted but 12 were incorrectly predicted to be normal. With female patients out of 355 normal images, 242 were correctly predicted, 113 were incorrectly predicted to be pneumonia, and out of 338 pneumonia images 327 were correctly predicted but 11 were incorrectly predicted to be normal.

5.3. The threshold where the model is super confident

5.3.1. DenseNet201 model

Figure 20 shows the Confusion Matrix of the DenseNet201 model by threshold. We have tried different thresholds the lower and higher between 0.0005 and 0.99999985 where the model makes the right prediction and there is no false negative and false positive in that range of probabilities. With the DenseNet201 model, we have found that when the model predicts with a probability above 99.999985 percent and probability below 0.0005 percent, the model prediction fall in that probability range there is 100 percent confidence that there is no false positive and false negative. Therefore, by summing 136 true negatives to 209 true positives equal to 345 correctly predicted out of 624 images, the model without any intervention of radiologists would reduce 55 percent of the caseload of radiologists and only intervenes for 45 percent of the patient. This is the confusion matrix of a threshold above 99. 999985 percent and achieved an F1-score of 69.78, the threshold above 0.0005 had an F1-score equal to 88.84.

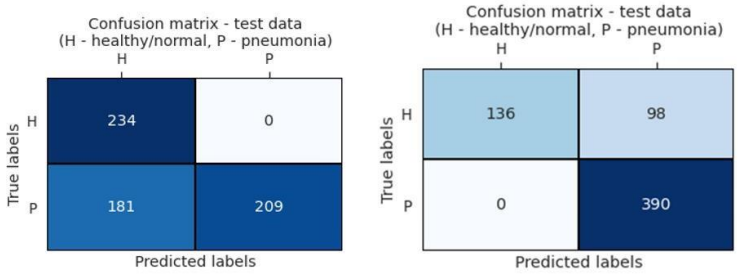
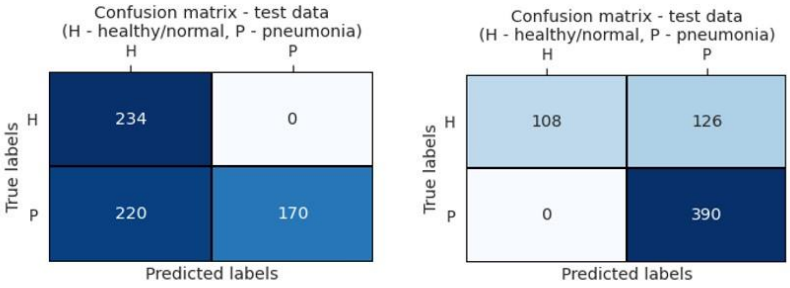


Figure 20: o: Threshold greater 0.99999985 p: threshold greater than 0.000005

5.3.2. CNN model



q r

Figure 21: Figure 21: q: threshold greater than 0.999999 r: threshold greater than 0.01

Figure 21 shows the Confusion Matrix of the CNN model by threshold. With lower and higher thresholds at threshold greater than 0.999999 F1-score equal to 64.74 and threshold greater than 0.01 F1-score equal to 86.09. By using the CNN model the prediction probabilities in the range between 0.001 to 0.999999 need the radiologist intervention for in that range there is a false positive and false negative. If the model prediction probability does not fall in that range, the model is highly confident and is correct for the prediction. The model can accurately or efficiently diagnose pneumonia without error. Therefore, by summing 108 true negatives and 170 true positives equal to 278 correctly predicted out of 624 images, the model will reduce 44.55 percent of the caseload to the radiologist only will intervene for 55.55 percent of patients.

5.4. Saliency maps

5.4.1. The saliency of DenseNet201 in Children Dataset test set

5.4.1.1. Correctly predicted

Figure 22: Pneumonia and predicted pneumonia

Figure 23: Actual normal but predicted normal

We used Grad-CAM to visualize saliency maps. They are presented by model and dataset. The probability of predicting pneumonia increases when the model is focusing on the central middle of the chest. The model predicted those images to be pneumonia because it has been highlighted in the middle of the chest, and predicted normal images because there is no highlight in the middle of the chest.

5.4.1.2. Incorrectly predicted

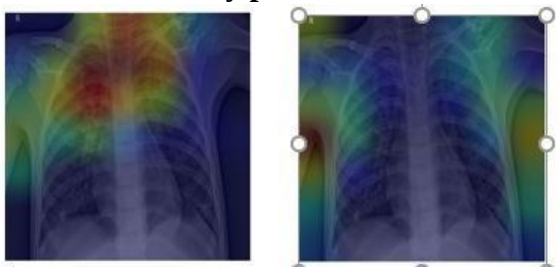


Figure 24: Actual pneumonia but predicted Normal

The model lost focus to the wrong pixels which are not the middle of the chest, therefore it was predicted incorrectly.

5.4.2. Saliency map of DenseNet201 model prediction with NIH data

5.4.2.1. Correctly predicted

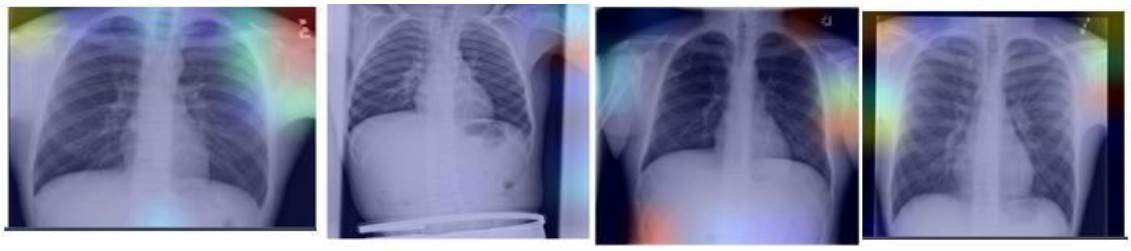


Figure 25: Normal and predicted Normal

Figure 26: pneumonia and predicted Pneumonia

The probability of predicting pneumonia increase when the model is focusing on the central middle of the chest. The model predicted those images to be normal because there is no highlight in the middle of the chest.

5.4.2.2. Incorrectly predicted

Figure 27: Normal but predicted Pneumonia

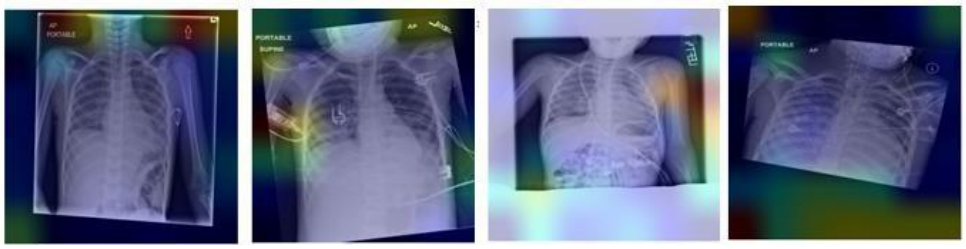


Figure 28: Pneumonia and predicted normal

The saliency maps fail to mark the correct pixels, they are not incorrectly classified in figure 27, the normal images are incorrectly classified as sick because there is a center-right marker, and in figure 28, the model failed to extract the pixels showing pneumonia presence.

5.4.3. The saliency map of the CNN model in the Children Dataset test set

5.4.3.1. Correctly predicted

Figure 29: pneumonia and predicted pneumonia

Figure 30: Actual normal predicted normal

The saliency map shows the presence of pneumonia by highlighting the middle of the chest. When the saliency maps highlighted other regions instead of the middle of the chest there is no presence of pneumonia.

5.4.3.2. Incorrectly predicted

Figure 31: Actual normal but predicted pneumonia

5.4.4. Saliency map of CNN Model Prediction with NIH data

5.4.4.1. Correctly predicted

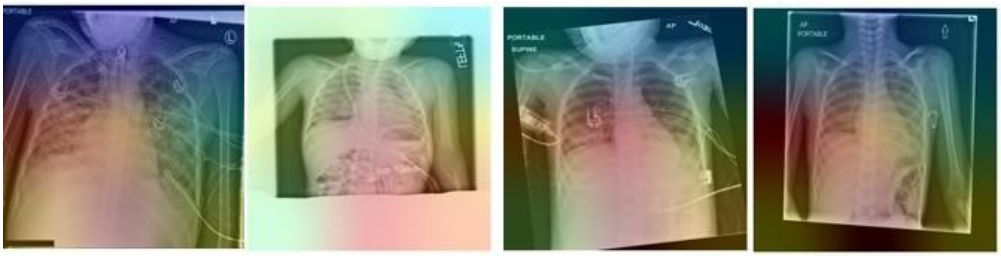


Figure 32: Pneumonia and predicted pneumonia

Figure 33: Actual normal and predicted normal

5.4.4.2. Incorrectly predicted

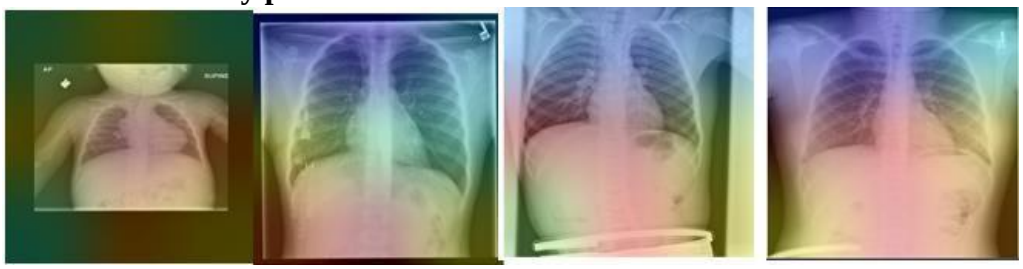


Figure 34: Actual normal but predicted pneumonia

The saliency maps help us show visually whether the pixel model based on to predict the presence of the disease is correct. As explained in the figures above when the saliency maps in the middle right of the chest are marked, there is pneumonia. But if there is no marking and is all over the picture there is no presence of pneumonia.

5.5. Sensitivity Analysis

Sensitivity analysis is the method of analyzing the variability of the model by modifying different parameters to assess the robustness of the model. The sensitivity analysis was examined by modifying the learning rate to optimize the model and we trained the model at different times to assess the change in model performance. Each change was calculated and all the results for each learning rate, mean and standard deviation, and confidence interval were recorded to understand the stability and variability of the model. Table 6 bellows contain all the results.

Table 6: DenseNet201 model with different learning rates on test Metrics.

learning rate	Accuracy	Recall	Precision	F1-score	Training Accuracy
0.00001	91.51	99.74	88.21	93.62	99.9
0.00009	94.55	98.21	93.41	95.75	99.7
0.0001	94.55	98.97	92.79	95.78	99.7
0.0002	94.55	99.23	92.58	95.79	99.2
0.0005	94.07	99.23	91.92	95.44	99.0
0.0009	94.39	97.17	94.04	95.59	98.5
0.001	92.63	98.97	90.19	94.38	98.3
0.0012	94.87	94.62	97.11	95.84	97.6
Average	93.89	98.27	92.53	95.27	99.0
Standard deviation	1.19	1.67	2.64	0.82	0.8
margin of error	0.99	1.40	2.20	0.69	0.7
lower bound	92.90	96.87	90.33	94.59	98.3
upper bound	94.88	99.67	94.73	95.96	99.7

We trained the DenseNet201 model with different learning rates from 0.00001 to 0.0012 the model changes by learning rate at a low level as the standard deviation shows. On average accuracy is equal to 93.89 percent and varies by 1.19 standard deviation and the 95 percent confidence interval falls between 92.90 and 94.88 percent. On average recall is equal to 98.27 percent, varies by 1.67 percent of standard deviation, and 95 percent confidence interval falls between 96.87 and 99.67 percent. On average precision equal to 92.53, varies by 2.64 percent of standard deviation, and the 95 percent confidence interval falls between 90.33 and 94.73. On average F1-score equal to 95.27, varies by 0.82 standard deviations, and the 95 percent confidence interval falls between 94.59 and

95.96. On average training accuracy is equal to 99.0 percent, varies by 0.8 standard deviations, and 95 percent confidence interval falls between 98.3 and 99.7 percent.

Table 7: CNN model with different learning rate

learning rate	Accuracy	Recall	Precision	F1-score	Training Accuracy
0.00007	92.15	92.05	95.23	93.61	94.2
0.00009	91.99	97.44	90.48	93.83	96.5
0.0001	89.42	98.97	86.16	92.12	95.9
0.0003	93.59	93.85	95.81	94.82	97.3
0.0005	93.27	96.92	92.64	94.74	97.0
0.0007	91.83	97.43	90.26	93.71	96.3
0.0009	90.06	96.92	88.32	92.42	96.0
0.0012	90.71	98.46	88.07	92.98	97.3
0.01	90.38	88.92	96.67	92.63	94.8
average	91.49	95.66	91.51	93.43	96.1
Standard deviation	1.44	3.36	3.77	0.97	1.1
margin of error	1.10	2.58	2.89	0.74	0.8
lower boundaries	90.38	93.08	88.62	92.69	95.3
Upper boundaries	92.59	98.24	94.41	94.17	97.0

We trained a convolutional neural network (CNN) model with different learning rates from 0.000007 to 0.01 the model changes by learning rate at a low level as the standard deviation shows. The results in table 7 show that on average accuracy is equal to 91.49 percent and varies by 1.44 standard deviation and the 95 percent confidence interval falls between 90.38 and 92.59 percent. On average recall is equal to 95.66 percent, varies by 3.36 percent of standard deviation, and 95 percent confidence interval falls between 93.08 and 98.24 percent. On average precision is equal to 91.51 and varies by 3.77 percent of standard deviation and the 95 percent confidence interval falls between 88.62 and 94.41, on average F1-score is equal to 93.43, which varies by 0.97 of standard deviation, and the 95 percent confidence interval falls between 92.69 and 94.17. On average training accuracy is equal to 96.1 percent, varies by 1.1 standard deviations, and 95 percent confidence interval falls between 95.3 and 97.0 percent.

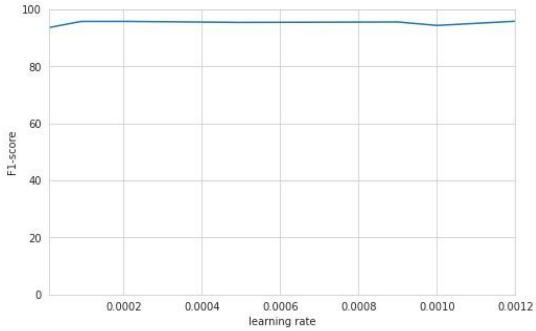


Figure 35: DenseNet201 learning rate by F1-score

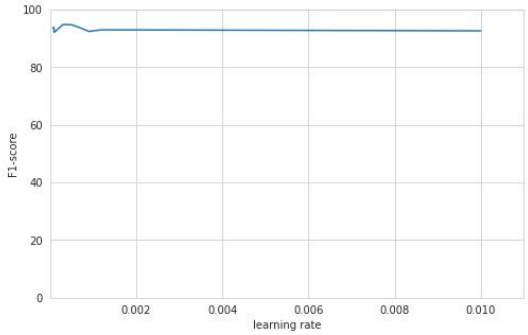


Figure 36: CNN learning rate by F1-score

Visually, as the two figures (figure 35 and figure 36) above show, changes in learning rate do not change much F1-score metric on both models, figures show stability by changing the learning rate parameter. Then DenseNet201 and CNN have stability because the parameter change does not change much in the performance measurements, especially in the F1-score.

Chapter 6: Discussion

The pneumonia detection model was developed by using chest x-ray images. First of all, we started with the following research questions: does the trained model detect pneumonia efficiently? Secondly, does the convolutional neural network (CNN) model have better accuracy than pretrained models? At which range of predictions probabilities does the model make the right prediction? Thirdly does the model generalize better to new data? when does the model generalize better?

CNN and DenseNet201 models were used to answer those questions. The DensetNet201 model was able to predict accurately on test set images with F1-score equal to 95.59 and with 95 percent confidence that by changing model parameter F1-score falls between 94.59 and 95.96. This means that model can detect approximately 96 out of 100 patients. Therefore, the model will efficiently make the right classification only if it fails on 4 patients, the number of false-positive was 24 which is higher than false-negative equal to 11. Based on the fact that the number of false-negative is lower, the model was able to recognize much better those who have pneumonia which is our main goal.

CNN model was almost equal to DensetNet201 on test set images with F1-score equal to 94.74, with 95 percent confident that by changing model parameter F1-score falls between 92.69 and 94.17 and was able to detect pneumonia approximately 95 out of 100 patients model make efficiently right predictions, and fail on 5 patients, the number of false-positive is 12, the number of false-negative equal to 30. Both DenseNet201 and CNN detect pneumonia at a higher score and do not vary a lot with changes in model parameters. Therefore, the models are efficient and stable to detect pneumonia.

The second research question is, does our model make better predictions? And at which range of probabilities, does the model make the right prediction and have no false predictions? As long as the model could be deployed in the health system it is important to know the probabilities where the model makes the right predictions without the radiologists' intervention. With DenseNet201 when we have above 99.999985 percent probabilities all cases model predicted pneumonia with no false positive. Below 0.0005 percent probability, all cases model predicted normal with no false negative. The only radiologist would intervene if model predictions fall between 0.0005 and

99.999985 percent. Therefore, 55 percent of the caseload of patients to the radiologists would be reduced.

With the CNN model when the probability falls Between 0.01 and 0.999999, the model could make the right prediction mixed with a false positive and false negative. So this is where the intervention of radiologists comes in as the model predictions are not always accurate in that range. But below 0.01 all instances are normal and above 0.999999 all instances are pneumonia. Therefore, the model will reduce 44 percent of the caseload to the radiologist only will intervene for 56 percent of patients. DenseNet201 can reduce 55 percent and CNN reduces 45 percent of the caseload. DenseNet201 outperforms CNN by almost 10 percent above the number of caseload reduction.

The third research question is, does the model generalize better to new data? when does the model generalize better? we used another dataset to test the model generalizability to new data and we obtained good accuracy in almost all groups of patients, the patients with age less than 20 years DenseNet201 model is F1-score equal to 91.56, CNN model is 89.57, on patients of age less than 30 DenseNet201 model is accurate approximately with F1-score equal to 94.29, CNN model is 91.75, on all patients, DenseNet201 model was able to classify images accurately with F1-score equal to 86.28, CNN model with F1-score equal to 86.29.

DenseNet201 model predictions on male patients F1-score equal to 87.71 while on female patients equal to 86.92, and CNN model predictions on male patients F1-score equal to 90.18 while on female patients equal to 84.06. Therefore, DenseNet201 can generalize better to new data as it can predict well to the validation dataset and makes better predictions with patients below 30 years old age while it was trained on images of patients of age less than 6 years old.

Whether DenseNet201 and CNN model, their model predictions on validation datasets vary by age, and the CNN model varies by gender. The shape, quality of image, and age of patients can influence the prediction, the bad quality of images which is not clear confuses the model and does not make the right predictions, the model predicted better the younger ages than older patients. DenseNet201and CNN models predict higher F1-score, are stable, has better generalizability to the new data especially younger age, with good shape and quality images, but DenseNet201 outperforms the CNN model. The DenseNet201 made a better prediction to the test data and

validation dataset, therefore the model can generalize better to a different dataset, is more reliable, and the model likely deploys in real-life, especially to a younger age.

Chapter 7: Conclusion and Recommendations.

7.1. Conclusion

Pneumonia is a killing disease but if patients receive better treatment is recoverable. Therefore, our research is a benefit to the method of diagnosis already existing that could shape the methodology of pneumonia disease diagnosis. To bring faster, more effective solutions for pneumonia disease. To reduce caseload for radiologists 44 percent of reduction with the CNN model, and 55 percent of reduction with the DenseNet201 model mostly in developing countries where the number of pneumonia patients is high and death is still high. The model developed is reliable and trustworthy to provide solutions in pneumonia detection.

7.2. Recommendations.

We recommend health institutions integrate this system of computer-aided diagnosis in the hospitals to help and support radiologists for better decision-making on pneumonia treatment. We recommend health hospitals record quality and clear x-ray images electronically to facilitate the development of more quality models to detect pneumonia. Future research would address the Covid-19 detection with chest x-ray using machine learning techniques, to detect cancer, pneumonia, and tuberculosis with chest x-ray images using machine learning techniques.

References:

[1] M. D. Daniel M. Musher, M.D., and Anna R. Thorner, "Community-Acquired Pneumonia _ Enhanced Reader.pdf," *The New England Journal of Medicine*. 2014.

[2] A. C. Noordam, A. B. Sharkey, P. Hinssen, G. Dinant, and J. W. L. Cals, "Association between caregivers' knowledge and care seeking behaviour for children with symptoms of pneumonia in six sub-Saharan African Countries," *BMC Health Services Research*, vol. 17, no. 1, pp. 7–9, 2017, doi: 10.1186/s12913-017-2060-3.

[3] World health organization, "Pneumonia causes and symptoms," *WHO publications*, 2019, [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/pneumonia

[4] S. Qazi *et al.*, "Ending preventable child deaths from pneumonia and diarrhoea by 2025. Development of the integrated Global Action Plan for the Prevention and Control of Pneumonia and Diarrhoea," *Archives of Disease in Childhood*, vol. 100, pp. S23–S28, 2015, doi: 10.1136/archdischild-2013-305429.

[5] J. H. Reynolds, G. McDonald, H. Alton, and S. B. Gordon, "Pneumonia in the immunocompetent patient," *British Journal of Radiology*, vol. 83, no. 996, pp. 998–1009, 2010, doi: 10.1259/bjr/31200593.

[6] N. Gupta *et al.*, "Causes of death and predictors of childhood mortality in Rwanda: A matched case-control study using verbal social autopsy," *BMC Public Health*, vol. 18, no. 1, pp. 1–9, 2018, doi: 10.1186/s12889-018-6282-z.

[7] National Institute of Statistics of Rwanda, *Rwanda Demographic and Health Survey 2014-15 - Final Report*. 2015. doi: March, 2016.

[8] D. R. Murdoch *et al.*, "Breathing New Life into Pneumonia Diagnostics," *JOURNAL OF CLINICAL MICROBIOLOGY*, vol. 47, no. 11, pp. 3405–3408, 2009, doi: 10.1128/JCM.01685-09.

[9] S. Andronikou, P. Goussard, and E. Sorantin, "Computed tomography in children with community-acquired pneumonia," *Pediatric Radiology*, vol. 47, no. 11, pp. 1431–1440, 2017, doi: 10.1007/s00247-017-3891-0.

[10] C. Cancedda *et al.*, "Health professional training and capacity strengthening through international academic partnerships: The first five years of the human resources for health program in rwanda," *International Journal of Health Policy and Management*, vol. 7, no. 11, pp. 1024–1039, 2018, doi: 10.15171/ijhpm.2018.61.

[11] DA. Rosman1*, J. J. Nshizirungu2, E. U. , Emmanuel Rudakemwa2 , Crispin Moshi3 , Jean de Dieu Tuyisenge3, and Louise Kalisa, "Imaging In The Land Of 1000 Hills," pp. 1–7, 2015.

[12] S. L. Goldenberg, G. Nir, and S. E. Salcudean, "A new era: artificial intelligence and machine learning in prostate cancer," *Nature Reviews Urology*, vol. 16, no. 7, pp. 391–403, 2019, doi: 10.1038/s41585-019-0193-3.

[13] K. El Asnaoui, Y. Chawki, and A. Idri, "Automated Methods for Detection and Classification Pneumonia based on X-Ray Images Using Deep Learning," *arXiv*, 2020.

[14] P. Rajpurkar *et al.*, "CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning," *arXiv*, pp. 3–9, 2017.

[15] O. Stephen, M. Sain, U. J. Maduh, and D. U. Jeong, "An Efficient Deep Learning Approach to Pneumonia Classification in Healthcare," *Journal of Healthcare Engineering*, vol. 2019, no. 2, pp. 7–21, 2019, doi: 10.1155/2019/4180949.

[16] M. F. Hashmi, S. Katiyar, A. G. Keskar, N. D. Bokde, and Z. W. Geem, "Efficient pneumonia detection in chest xray images using deep transfer learning," *Diagnostics*, vol. 10, no. 6, pp. 1–23, 2020, doi: 10.3390/diagnostics10060417.

[17] N. M. Elshennawy and D. M. Ibrahim, "Deep-Pneumonia Framework Using Deep Learning Models Based on Chest X-Ray Images," *Diagnostics*, vol. 10, no. 9, pp. 1–16, 2020, doi: 10.3390/diagnostics10090649.
[18] J. Hanson, a Global, and R. Cdc, "Republic of Rwanda," *... .Wikischolars.Columbia.Edu*, no. June, pp. 1–34, 2010.
[19] J. Bell, "Chapter 5 - Artificial Neural Networks," *Machine Learning: Hands-On for Developers and Technical Professionals*, pp. 91–116, 2014.
[20] Andrew Glassner, "Deep Learning, Vol. 1: From Basics to Practice; Vol. 2: From Basics to Practice," vol. 2, p. 914, 2018.
[21] J. Lu, V. Behbood, P. Hao, H. Zuo, S. Xue, and G. Zhang, "Transfer learning using computational intelligence: A survey," *Knowledge-Based Systems*, vol. 80, pp. 14–23, May 2015, doi: 10.1016/J.KNOSYS.2015.01.010.
[22] S. Panigrahi, A. Nanda, and T. Swarnkar, "A Survey on Transfer Learning," *Smart Innovation, Systems and Technologies*, vol. 194, pp. 781–789, 2021, doi: 10.1007/978-981-15-5971-6_83.
[23] W. Zhao, "Research on the deep learning of the small sample data based on transfer learning," *AIP Conference Proceedings*, vol. 1864, no. August 2017, 2017, doi: 10.1063/1.4992835.
[24] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, vol. 2017-Janua, pp. 2261–2269, 2017, doi: 10.1109/CVPR.2017.243.
[25] S. H. Wang and Y. D. Zhang, "DenseNet-201-Based Deep Neural Network with Composite Learning Factor and Precomputation for Multiple Sclerosis Classification," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 16, no. 2s, 2020, doi: 10.1145/3341095.
[26] A. Hussain, A. Khan, H. Yar, and N. Fayyaz, "Efficient Deep learning Approach for Classification of Pneumonia using Resources Constraint Devices in Healthcare," 2019.
[27] L. Brunese, F. Mercaldo, A. Reginelli, and A. Santone, "Explainable Deep Learning for Pulmonary Disease and Coronavirus COVID-19 Detection from X-rays," *Computer Methods and Programs in Biomedicine*, vol. 196, p. 105608, 2020, doi: 10.1016/j.cmpb.2020.105608.
[28] A. Shelke *et al.*, "Chest X-ray classification using Deep learning for automated COVID-19 screening," *medRxiv*, no. December 2019, 2020, doi: 10.1101/2020.06.21.20136598.
[29] S. Rajaraman, S. Candemir, I. Kim, G. Thoma, and S. Antani, "Visualization and interpretation of convolutional neural network predictions in detecting pneumonia in pediatric chest radiographs," *Applied Sciences (Switzerland)*, vol. 8, no. 10, 2018, doi: 10.3390/app8101715.
[30] K. Suzuki, "Overview of deep learning in medical imaging," *Radiological Physics and Technology*, vol. 10, no. 3, pp. 257–273, 2017, doi: 10.1007/s12194-017-0406-5.
[31] F. P. An and Z. W. Liu, "Medical Image Segmentation Algorithm Based on Feedback Mechanism CNN," *Contrast Media and Molecular Imaging*, vol. 2019, 2019, doi: 10.1155/2019/6134942.
[32] M. Sadeghi *et al.*, "Feedback-based Self-improving CNN Algorithm for Breast Cancer Lymph Node Metastasis Detection in Real Clinical Environment," *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, pp. 7212–7215, 2019, doi: 10.1109/EMBC.2019.8857432.

[33] N. Wahab, A. Khan, and Y. S. Lee, "Transfer learning based deep CNN for segmentation and detection of mitoses in breast cancer histopathological images," *Microscopy*, vol. 68, no. 3, pp. 216–233, 2019, doi: 10.1093/jmicro/dfz002.
[34] S. C. K. Chu *et al.*, "Orienting Health Systems for Maternal Health – the Sri Lankan Experience," *Development*, vol. 48, no. December, pp. 399–406, 2012, doi: 10.1057/palgrave.jors.2601024.
[35] H. A. Khan, W. Jue, M. Mushtaq, and M. U. Mushtaq, "Brain tumor classification in MRI image using convolutional neural network," *Mathematical Biosciences and Engineering*, vol. 17, no. 5, pp. 6203–6216, 2020, doi: 10.3934/MBE.2020328.
[36] Paul Mooney, "Chest X-Ray Images (Pneumonia)," *Kaggle*, Jun. 01, 2018. https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia (accessed Aug. 01, 2021).
[37] D. Kermany, K. Zhang, and M. Goldbaum, "Labeled Optical Coherence Tomography (OCT) and Chest X-ray Images for Classification," vol. Version 2, 2018, doi: 10.17632/rscbjbr9sj.2.
[38] National Institutes of Health, "NIH Chest X-rays," *Kaggle*. https://www.kaggle.com/nih-chest-xrays/data
[39] R. (NIH/CC/DRD) [E] Summers, "CXR8," *National Institutes of Health - Clinical Center*, 2017. https://nihcc.app.box.com/v/ChestXray-NIHCC/folder/36938765345
[40] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," *International Journal of Computer Vision*, vol. 128, no. 2, pp. 336–359, 2020, doi: 10.1007/s11263-019-01228-7.

Detection of Pneumonia by chest x-ray images using machine learning technique

ORIGINALITY REPORT

SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS
13%	10%	9%	5%

PRIMARY SOURCES

1	www.mdpi.com Internet Source	2%
2	doctorpenguin.com Internet Source	1%
3	Submitted to The University of Law Ltd Student Paper	1%
4	Submitted to University of Hull Student Paper	1%
5	arxiv.org Internet Source	<1%
6	"Artificial Neural Networks and Machine Learning – ICANN 2017", Springer Science and Business Media LLC, 2017 Publication	<1%

7	Submitted to Bournemouth University Student Paper	<1%
8	scholar.uwindsor.ca Internet Source	<1%
9	"Artificial Intelligence and Blockchain for Future Cybersecurity Applications", Springer Science and Business Media LLC, 2021 Publication	<1%
10	en.wikipedia.org Internet Source	<1%
11	Asmaa Abbas, Mohammed M. Abdelsamea, Mohamed Medhat Gaber. " : Self Supervised Super Sample Decomposition for Transfer learning with application to COVID-19 detection ", Cold Spring Harbor Laboratory, 2020 Publication	<1%
12	Submitted to University of Rwanda Student Paper	<1%
13	Submitted to Jabatan Pendidikan Politeknik Dan Kolej Komuniti Student Paper	<1%

14	arno.uvt.nl Internet Source	<1%
15	www.kaggle.com Internet Source	<1%
16	Adhiyaman Manickam, Jianmin Jiang, Yu Zhou,Abhinav Sagar, Rajkumar Soundrapandiyan, R. Dinesh Jackson Samuel. "Automated pneumonia detection on chest X-ray images: A deep learning approach with different optimizers and transfer learning architectures", Measurement, 2021 Publication	<1%
17	Submitted to Nottingham Trent University Student Paper	<1%
18	d-nb.info Internet Source	<1%
19	"Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries", Springer Science and Business Media LLC, 2018 Publication	<1%

20	www.freepatentsonline.com Internet Source	<1%
21	"Intelligent Information and Database Systems", Springer Science and Business Media LLC, 2020 Publication	<1%
22	"International Conference on Innovative Computing and Communications", Springer Science and Business Media LLC, 2022 Publication	<1%
23	Jihao Shi, Yuanjiang Chang, Changhang Xu,Faisal Khan, Guoming Chen, Chuangkun Li. "Real-time leak detection using an infrared camera and Faster R-CNN technique", Computers & Chemical Engineering, 2020 Publication	<1%
24	Submitted to The University of Buckingham Student Paper	<1%
25	digitalcommons.njit.edu Internet Source	<1%

26	www.vdu.lt Internet Source	<1%
27	Submitted to University of Hertfordshire Student Paper	<1%
28	Yuvraj Sinha Chowdhury, Rupshali Dasgupta, Sarita Nanda. "Analysis of Various Optimizer on CNN model in the Application of Pneumonia Detection", 2021 3rd International Conference on Signal Processing and Communication (ICPSC), 2021 Publication	<1%
29	scholarcommons.sc.edu Internet Source	<1%
30	"Proceedings of the International Conference on Artificial Intelligence and Computer Vision (AICV2020)", Springer Science and Business Media LLC, 2020 Publication	<1%
31	Mohammad-H. Tayarani N.. "Applications ofartificial intelligence in battling against covid-19: A literature review", Chaos, Solitons &	<1%

Fractals, 2021
Publication

32	Joaquim de Moura, Lucia Ramos, Placido L. Vidal, Milena Cruz, Laura Abelairas, Eva Castro, Jorge Novo, Marcos Ortega. "Deep convolutional approaches for the analysis of Covid-19 using chest X-ray images from portable devices", IEEE Access, 2020 Publication	<1%
33	algotrading101.com Internet Source	<1%
34	www.ajronline.org Internet Source	<1%
35	www.hindawi.com Internet Source	<1%
36	"Wireless Algorithms, Systems, and Applications", Springer Science and Business Media LLC, 2019 Publication	<1%
37	Heather J. Zar, Thomas W. Ferkol. "The global burden of respiratory disease-Impact on child health", Pediatric Pulmonology, 2014 Publication	<1%

38	diagnosticpathology.biomedcentral.com Internet Source	<1%
39	www.nature.com Internet Source	<1%
40	www.researchsquare.com Internet Source	<1%
41	John H. Reynolds, Arpan K. Banerjee. "Imaging pneumonia in immunocompetent and immunocompromised individuals", Current Opinion in Pulmonary Medicine, 2012 Publication	<1%
42	Kui Jiang, Zhongyuan Wang, Peng Yi, Junjun Jiang, Jing Xiao, Yuan Yao. "Deep Distillation Recursive Network for Remote Sensing Imagery Super-Resolution", Remote Sensing, 2018 Publication	<1%
43	Mohamed Abdel-Zaher, Mustafa Hisham, Retaj Yousri, M. Saeed Darweesh. "LightWeight Convolutional Neural Network For Fire Detection", 2021 International Conference on Electronic Engineering (ICEEM), 2021 Publication	<1%

44 Muhammad Owais, Hyo Sik Yoon, TahirMahmood, Adnan Haider, Haseeb Sultan, <1%

Kang Ryoung Park. "Light-weighted ensemble network with multilevel activation visualization for robust diagnosis of COVID19 pneumonia from large-scale chest radiographic database", Applied Soft Computing, 2021

Publication

---

45 Olaide N. Oyelade, Absalom E. Ezugwu, <1%

Haruna Chiroma. "CovFrameNet: An enhanced deep learning framework for COVID-19 detection", IEEE Access, 2021

Publication

---

46 Shailender Kumar, Pankaj Singh, Mohit <1%

Ranjan. "A review on deep learning based pneumonia detection systems", 2021
International Conference on Artificial Intelligence and Smart Systems (ICAIS), 2021

Publication

---

47 DalSpace.library.dal.ca <1%

Internet Source

<1%

48 Liangrui Pan, boya ji, Xiaoqi wang, shaoliang peng. "MFDNN: Multi-channel feature deep neural network algorithm to identify Covid19 chest X-ray images", Cold Spring Harbor Laboratory, 2021
Publication

<1%

49 www.jmir.org
Internet Source

<1%

50 Submitted to Liverpool John Moores University
Student Paper

<1%

51 Md Manjurul Ahsan, Md Tanvir Ahad, Farzana Akter Soma, Shuva Paul et al. "Detecting SARS-CoV-2 From Chest X-Ray Using Artificial Intelligence", IEEE Access, 2021
Publication

<1%

52 Noorul Wahab, Asifullah Khan, Yeon Soo Lee. "Transfer learning based deep CNN for segmentation and detection of mitoses in breast cancer histopathological images", Microscopy, 2019
Publication

<1%

53 ijisrt.com
Internet Source

<1%

54	jips.jatsxml.org Internet Source	<1%
55	link.springer.com Internet Source	<1%
56	publikacio.uni-eszterhazy.hu Internet Source	<1%
57	"Medical Image Computing and Computer-Assisted Intervention – MICCAI 2017", Springer Nature, 2017 Publication	<1%
58	Abdullahi Umar Ibrahim, Mehmet Ozsoz,Sertan Serte, Fadi Al-Turjman, Salahudeen Habeeb Kolapo. " Convolutional neural network for diagnosis of viral pneumonia and -19 alike diseases ", Expert Systems, 2021 Publication	<1%
59	Mohamed Ramzy Ibrahim, Karma M. Fathalla, Sherin M. Youssef. "HyCAD-OCT: A Hybrid Computer-Aided Diagnosis of Retinopathy by Optical Coherence Tomography Integrating Machine Learning and Feature Maps Localization", Applied Sciences, 2020 Publication	<1%

Okeke Stephen, Mangal Sain, Uchenna Joseph Maduh, Do-Un Jeong. "An Efficient Deep Learning Approach to Pneumonia Classification in Healthcare", Journal of Healthcare Engineering, 2019 <1%
Publication

docs.google.com <1%
61 Internet Source

etd.lib.nsysu.edu.tw <1%
62 Internet Source

insightsimaging.springeropen.com <1%
63 Internet Source

64 Ali Mohammad Alqudah, Shoroq Qazan,Ihssan S. Masad. "Artificial Intelligence Framework for Efficient Detection and Classification of Pneumonia Using Chest Radiography Images", Research Square, 2020 <1%
Publication

65 Ankita Shelke, Madhura Inamdar, Vruddhi Shah, Amanshu Tiwari, Aafiya Hussain, Talha Chafekar, Ninad Mehendale. "Chest X-ray classification using Deep learning for automated COVID-19 screening", Cold Spring Harbor Laboratory, 2020 <1%
Publication

66 Shuaijing Xu, Hao Wu, Rongfang Bie. "CXNet-m1: Anomaly Detection on Chest X-Rays with Image-Based Deep Learning", IEEE Access, 2018 <1%
Publication

67 Juan David Gutiérrez. "Machine Learning to Forecast Medical Attentions of Pneumonia Cases in Colombian Cities: An implementation with Air Quality, Meteorological and Admission Data", Research Square, 2020 <1%
Publication

68	Mahmoud Said Elsayed, Nhien-An Le-Khac, Soumyabrata Dev, Anca Delia Jurcut. "DDoSNet: A Deep-Learning Model for Detecting Network Attacks", 2020 IEEE 21st International Symposium on "A World of Wireless, Mobile and Multimedia Networks" (WoWMoM), 2020 Publication	<1%
69	Nishit Pareshbhai Nakrani, Jay Malnika,Satyam Bajaj, Harihar Prajapati, Vivaksha Jariwala. "Chapter 9 Pneumonia Identification Using Chest X-Ray Images with Deep Learning", Springer Science and Business Media LLC, 2020 Publication	<1%
70	Internet Source	<1% hdl.handle.net

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		